

ADVANCED BAYESIAN MODELLING FOR THE ANALYSIS OF OUTBREAKS AND SHIFTING EPIDEMIOLOGICAL DYNAMICS

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY OF IMPERIAL COLLEGE LONDON
AND THE DIPLOMA OF IMPERIAL COLLEGE LONDON BY

ANDREA BRIZZI

IMPERIAL COLLEGE LONDON
DEPARTMENT OF MATHEMATICS

OCTOBER 2024

DECLARATION OF ORIGINALITY

I certify that this thesis and the research to which it refers are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline. Where work is derived from joint collaborations with the author, the author's original contribution will be clearly highlighted and the work of collaborators will be appropriately acknowledged.

- The Sections and Figures in Chapter 3 were copied, adapted and extended from the paper entitled “Refining Reproduction Number Estimates to Account for Unobserved Generations of Infection in Emerging Epidemics” published in *Clinical Infectious Diseases*.
- The Sections, Figures and Tables in Chapter 4 were copied, adapted and extended from the paper entitled “Spatial and temporal fluctuations in COVID-19 fatality rates in Brazilian hospitals” published in *Nature Medicine*.
- The Sections, Figures in Chapter 5 were copied, adapted and extended from the paper entitled “Longitudinal population-level HIV epidemiologic and genomic surveillance highlights growing gender disparity of HIV transmission in Uganda” published in *Nature Microbiology*. Sections 5.3.2 and 5.3.4 include further contributions which were not published with the paper.
- The Sections, Figures and Tables in Chapter 6 were copied, adapted and extended from the manuscript entitled “Age and gender profiles of HIV infection burden and viraemia: novel metrics for HIV epidemic control in African populations with high antiretroviral therapy coverage” submitted to *International Journal of Epidemiology* for review.

COPYRIGHT STATEMENT

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Licence](#) (CC BY-NC-ND 4.0).

Under this licence, you may copy and redistribute the material in any medium or format on the condition that; you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

ABSTRACT

The emergence, spread, and establishment of an infectious disease within a population brings about a plethora of challenges for public health organisations, whose aim is to reduce disease burden while having access to limited information. In this thesis, we develop statistical models and analyses to support public health response, addressing uncertainties that are inherent to epidemics. The work is divided into two parts, focusing on the last century's biggest pandemics.

In the first part, we focus on the emergence of novel pathogens and variants of concern, with applications to SARS-CoV-2. Firstly, we develop an adjustment to early reproduction number estimates, when generations of infections have not been reported. Our adjustment is shown to reduce early biases in simulation studies. Secondly, we develop a multi-strain Bayesian model to describe fluctuations in hospital fatality rates in Brazil following the emergence of the Gamma variant. By synthesising data from separate sources, we estimate the proportion of patients with either variant in hospitals across Brazil, and quantify the impact of healthcare pressures, variant, and location effects.

In the second part of the thesis, we describe changes in transmission dynamics and burden of HIV, using data from the Rakai Community Cohort Study. In the first project, we develop a phylogenetic pipeline to estimate HIV time since infection from viral sequences, and develop statistical models to refine estimates by incorporating testing histories and known transmission network. By dating transmissions, we are able to describe changes in transmission patterns, highlighting shifts in the age-profile of the sources. Finally, we provide detailed descriptions of shifts in the age and gender compositions of the burden of HIV and viraemia in Uganda. We obtain estimates at the age level by developing non-parametric models sharing information across age groups. We conclude by proposing novel metrics to inform prevention strategies.

ACKNOWLEDGEMENTS

I would like to express my gratitude to all those who have supported me throughout my PhD journey.

First and foremost I would like to thank my supervisor, Prof. Oliver Ratmann, for giving me the chance to work on fascinating problems, his caring guidance, his never-ending enthusiasm, his mentoring, and the excellent summer barbecues. Secondly, my gratitude is extended to my co-supervisors and collaborators. Notably, I would like to thank Prof. Axel Gandy for his insights and his infinite curiosity, Kate Grabowski PhD for sharing her invaluable experience in HIV research, Dr. Ilaria Dorigatti and Dr. Megan O'Driscoll for the collaboration, and Dr. Sikhulile Moyo for hosting me in Gaborone.

I would like to acknowledge the PANGAEA-HIV consortium, the Machine Learning for Global Health Network and the Rakai Health Sciences Program for their ambition to make a difference through innovative research. My heartfelt thanks is extended to the participants of the Rakai Community Cohort Study, whose commitment has allowed and keeps allowing researchers to understand and to fight the HIV pandemic.

My PhD journey would not have been as stimulating had I not been part of the EPSRC CDT in Statistics and Machine Learning. I would like to thank the CDT staff and coordinators for the excellent training, activities and support. A special thanks goes to the fellow students, who have been a constant source of inspiration, motivation and discussion.

I must also thank my friends: *grazie, merci, ευχαριστώ, gracias, obrigado, хвалю* and *thanks* to all of you.

My deepest gratitude goes to my family, for their unconditional love, support and encouragement. A huge thank to mamma Elena and papà Paolo for instilling in me the value of education and knowledge; to my brothers Stefano and Francesco for the affection and the laughs; to my grandparents Anna, Cesarina, Enzo and Luciano, for their wisdom; to my uncles Antonella and Antonello; and to my cousins Francesco and Maria Sole for the joy they bring.

Finally, Carmen. Thank you for your love, care, and for reminding me the value of the simple things.

CONTENTS

List of Figures	10
List of Tables	12
List of Acronyms	13
1 Introduction	15
1.1 The SARS-CoV-2/COVID-19 pandemic	16
1.2 The HIV/AIDS pandemic	18
1.3 The Rakai Community Cohort Study	20
1.4 Challenges	22
1.5 Thesis aims and Chapter descriptions	23
2 Technical Background	25
2.1 Basic epidemiological definitions	25
2.1.1 Population, incidence and prevalence	25
2.1.2 Individual level disease status pathogen	25
2.1.3 Incubation period, generation time and serial interval	26
2.1.4 Measures of transmissions speed	26
2.1.5 Transmission networks	27
2.2 Models of disease transmission	27
2.2.1 Compartmental models	27
2.2.2 Renewal models	29
2.2.3 EpiEstim	30
2.2.4 Epidemia	31
2.3 Statistical modelling	31
2.3.1 Bayesian inference	32
2.3.2 Posterior predictive distribution	32
2.3.3 Monte Carlo methods	33
2.3.4 Software for Bayesian computation	36
2.3.5 Convergence diagnostics and model fit	38
2.3.6 Non-parametric models	40
3 Refined R_0 estimates in emerging epidemics	43

3.1	Introduction	43
3.2	Contributions to the literature	44
3.3	Methods	46
3.3.1	Exponential Growth	46
3.3.2	EpiEstim with exponential back-imputation	46
3.3.3	Simulations	48
3.4	Findings	50
3.4.1	Method comparison	50
3.4.2	Impact of missed generations	51
3.4.3	Application to reported COVID-19 data	53
3.5	Sensitivity analyses	53
3.5.1	Sensitivity analysis on the generation interval distribution	53
3.5.2	Sensitivity analysis on the assumed reporting rate	54
3.5.3	Sensitivity analysis on exponential growth assumption	54
3.6	Discussion	58
4	Estimating Fluctuating Hospital Fatality Rates in LMIC settings	62
4.1	Introduction	62
4.2	Contributions to the literature	64
4.3	Methods	66
4.3.1	Data	66
4.3.2	Smoothed age-standardised fatality rates model	69
4.3.3	Bayesian multi-strain model	70
4.3.4	Key Epidemiological quantities uncovered	78
4.4	Findings	85
4.4.1	Fluctuations of in-hospital fatality rates around Gamma's emergence	85
4.4.2	In-hospital fatality rates are tracked by healthcare pressure indices	89
4.4.3	In-hospital fatality rates are mostly driven by regional and health-care inequities	90
4.4.4	Number of deaths attributable to resource limitations	92
4.5	Model validation and sensitivity analyses	93
4.6	Discussion	95
5	HIV-TSI: pipeline and applications	98
5.1	Introduction	98
5.2	Contributions to the literature	101
5.3	Methods	102
5.3.1	RCCS transmission cohort data 2010-2018	102
5.3.2	Identification of plausible source-recipient pairs	102

5.3.3	Phylogenetic pipeline for HIV time-since-infection	103
5.3.4	Estimation of infection data for source-recipient pairs	107
5.3.5	Modeling changes in transmission flow	115
5.3.6	Ethical consideration	116
5.4	Findings	118
5.5	Discussion	120
6	Profiles of HIV infection burden and viraemia	124
6.1	Introduction	124
6.2	Contributions to the literature	126
6.3	Methods	127
6.3.1	Population size and participation rates	127
6.3.2	Measurement of HIV serostatus and viraemia	127
6.3.3	Population-level <i>Human Immunodeficiency Virus</i> (HIV) prevalence estimates by age and gender	129
6.3.4	Age and gender profiles of <i>People Living With HIV</i> (PLHIV)	133
6.3.5	Prevalence of viral suppression among PLHIV by age and gender	133
6.3.6	Age and gender profile of people exhibiting viraemia	134
6.3.7	Age quantiles for burden profiles by gender	136
6.4	Findings	136
6.4.1	Age and gender burden profiles of HIV infection	136
6.4.2	“95-95-95” targets are achievable in Rakai	139
6.4.3	Age and gender profiles of the burden of viraemia	142
6.5	Discussion	143
7	Discussion	148
7.1	Conclusions	154
	Bibliography	156
	Sitography	202
	Appendices	206
A	Appendix for Fluctuating HFR in Brazil	206
B	Appendix for profiles of hiv infection burden and viraemia	223
C	Proofs	230
C.1	Proof of Algorithm 1 working.	230

LIST OF FIGURES

1.1	History of pandemics.	17
2.1	An example of a transmission network.	27
2.2	20 draws from a Gaussian process with squared exponential kernel	42
3.1	Visualisation of the exponential growth adjustment method	48
3.2	Distribution of mean R_0 estimates assuming a fixed reporting rate	49
3.3	Bias and coverage comparison by reporting rate.	50
3.4	Impact of missed generations on R_0 estimates.	51
3.5	Comparison of R_t estimates obtained with and without adjustment for COVID-19 data in 4 countries.	52
3.6	Sensitivity analysis on generation interval.	55
3.7	Sensitivity analysis on reporting rates.	56
3.8	Incidence curves simulated via a stochastic SEIR model with nonlinear incidence rates	57
3.9	Visualisation of the adjustment method applied to simulated data following non-exponential growth	58
3.10	Distribution of mean R_0 estimates obtained on simulated data following sub-exponential growth	59
4.1	Time trends in age-specific COVID-19 in-hospital fatality rates.	71
4.2	Spatiotemporal expansion of the SARS-CoV-2 Gamma variant across Brazil.	73
4.3	Estimated COVID-19 attributable hospital admissions by SARS-CoV-2 variant.	80
4.4	Estimated COVID-19 attributable deaths in hospitals by SARS-CoV-2 variant.	81
4.5	Analysis flowchart.	86
4.6	Spatio-temporal expansion of SARS-CoV-2 Gamma in Brazil and associated shocks in COVID-19 fatality rates in hospitals.	88
4.7	Time trends in age-standardised COVID-19 in-hospital fatality rates and pandemic healthcare pressure.	89
4.8	Estimated contribution of location effects, Gamma's infection severity, and pandemic healthcare pressure on COVID-19 in-hospital fatality rates.	91
4.9	Estimated effect of location on in-hospital fatality rates when using sequence data from Rede Genômica FioCruz.	93

5.1	Pipeline for estimating TSI from NGS data.	105
5.2	Schema illustrating the refinement of phylogenetic time since infection estimates.	113
5.3	Comparison of estimated infection dates in reconstructed source-recipient pairs.	114
5.4	Comparison of pipeline output for a set of 152 sequences appearing both in HIV-phyloTSI's training data and in the RCCS.	117
5.5	Phylogenetically reconstructed source-recipient pairs.	119
5.6	Time trends in age-specific sources of HIV infections in women and men.	120
6.1	Study flowchart (rounds 16 to 19).	128
6.2	Age- and gender-specific HIV prevalence estimates among study participants and first-time participants, 2013-2019.	130
6.3	Age- and gender-specific viral load suppression rate estimates among study participants and first-time participants with HIV, 2013-2019.	132
6.4	Age- and gender-specific viral load suppression rate estimates among census-eligible people with HIV.	135
6.5	Age and gender profiles of HIV burden in the Rakai Community Cohort Study, June 2018-May 2019 (round 19).	138
6.6	Trends in the proportions of PLHIV who have unsuppressed viraemia in the Rakai Community Cohort Study	141
6.7	Age and gender metrics for people exhibiting viraemia	144

LIST OF TABLES

1.1	Timeline of RCCS Survey Rounds	22
4.1	Temporal fluctuations in COVID-19 attributable in-hospital fatality rate, and avoidable COVID-19 attributable deaths in hospitals.	87
5.1	Description of HIV-phyloTSI predictors.	108
6.1	Characteristics of populations with and without HIV in the Rakai Community Cohort Study, June 2018-May 2019 (round 19)	137
6.2	People exhibiting viraemia in the Rakai Community Cohort Study, from July 2013-January 2015 to June 2018-May 2019	140

LIST OF ACRONYMS

AIDS	Acquired Immunodeficiency Syndrome
ART	AntiRetroviral Therapy
COVID-19	Coronavirus Disease 2019
ESA	Eastern and Southern Africa
Global Fund	Global Fund to Fight Aids
GP	Gaussian Process
HIC	High Income Countries
HIV	Human Immunodeficiency Virus
HMC	Hamiltonian Monte Carlo
LMIC	Low and Middle Income Countries
LOESS	Locally Estimated Scatterplot Smoothing
MAF	Minor Allele Frequency
MCMC	Markov Chain Monte Carlo
MSA	Multiple Sequence Alignment
MSM	Men who have Sex with Men
NGS	Next-Generation Sequencing
PANGEA	Phylogenetics And Networks for Generalised Epidemics in Africa
PEPFAR	US President's Emergency Plan for AIDS Relief
PHIA	Population-based HIV Impact Assessment
PLHIV	People Living With HIV
RCCS	Rakai Community Cohort Study
RHSP	Rakai Health Sciences Program
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2

TasP	Treatment as Prevention
TSI	Time Since Infection
UN	United Nations
UNAIDS	Joint United Nations Programme on HIV/AIDS
UTT	Universal Test and Treat
VoC	Variant of Concern
WHO	World Health Organization
SDG	Sustainable Development Goals
STD	Sexually Transmitted Diseases
VMMC	Voluntary Medical Male Circumcision

1

INTRODUCTION

THROUGHOUT human history, infectious diseases have posed one of the major threats to population survival. The Bubonic plague, now known as the Black Death, ravaged Europe between 1346 and 1353, claiming the lives of up to half of the European population [DeWitte, 2014]. The Columbian Exchange, following the discovery of America in 1492, led to a widespread intercontinental transfer of people, plants and products. This exchange facilitated the transfer of pathogens like smallpox and measles, which decimated native societies in the Americas [Jones, 2006, Koch et al., 2019]. These pandemics, surpassing the death toll of World War I [Britannica, 2024], left long-lasting impacts on the societies they affected [Diamond, 1999].

Despite advances in medicine and public health, infectious diseases continue to be a leading cause of death and poverty globally. One death in seven [Dattani and Roser, 2023] is attributable to existing infectious diseases, and the risk of global pandemics is on the rise due to various social, economic and ecological factors [Morse, 1995, Jones et al., 2008]. For instance, the growing interconnectedness of the world brought about by globalization enables pathogens to spread rapidly before they can be effectively identified and characterized.

To fight these threats, international cooperation is crucial. With 194 member states, the *World Health Organization (WHO)* is the leading international organization in the joint fight against infectious diseases, aiming to promote health and its democratisation [WHO, 1948]. For instance, International Health Regulations [WHO, 2005] compel members to meet standards for surveillance, reporting, and response to public health emergencies, and to provide support to other nations in need. Additionally, the WHO has contributed to the definition of the ambitious *Sustainable Development Goals (SDG)* from the *United Nations (UN)*, aiming to improve health and reduce worldwide poverty

by 2030 [United Nations, 2015] by ensuring healthy lives and promote well-being for all at all ages. The third of the 17 goals aims to end the epidemics of AIDS, tuberculosis, malaria, and neglected tropical diseases by 2030 [United Nations, 2016].

Ambitious vision needs strategy, strategy needs analysis, and analysis needs data. To achieve these goals and end some of the deadliest pandemics in history, effective interventions must be implemented based on the available evidence. This thesis aims to exemplify how statistical analysis of epidemiological data can be used to assess the effects of public health interventions, detect changes in transmission dynamics and inform public health policy updates. Specifically, I focus on the two most important pandemics of the last century: COVID-19 and AIDS.

1.1 The SARS-CoV-2/COVID-19 pandemic

In December 2019, the novel *Severe Acute Respiratory Syndrome Coronavirus 2* (SARS-CoV-2) was identified in China, following an outbreak of pneumonia cases associated with a seafood market in Hubei, China [Wu et al., 2020, Zhu Na et al., 2020]. Characterization of early cases revealed a respiratory disease similar to SARS, which had higher morbidity on men, people with diabetes and cardiovascular diseases. Symptoms included cough, fever and fatigue [Huang et al., 2020, Chen et al., 2020]. Subsequent studies revealed characteristic ideal for the virus' spread: efficiency in the person-to-person transmission [Li et al., 2020], possibility of pre-symptomatic or asymptomatic transmission [Rothe Camilla et al., 2020] and a relatively long incubation period [Wassie et al., 2020]. These factors led the WHO to declare the virus a Public Health Emergency of International Concern on the 30th of January 2020 [WHO, 2020a]. Less than two weeks later, on the 11th of February, the WHO officially named the associated disease *Coronavirus Disease 2019* (COVID-19) [WHO, 2020b]. Despite the introduction of travel restriction in China, SARS-CoV-2 spread globally, and by February 2020 more than 80,000 cases were reported across all inhabited continents [WHO, 2020b], which successively led the WHO to announce a global COVID-19 pandemic on the 11th of March 2020 [Cucinotta and Vanelli, 2020].

The unprecedented spread of transmission forced governments worldwide to adopt drastic non-pharmaceutical interventions in an attempt to reduce the burden on health-care systems. The response included lockdowns, quarantines, school closures, physical distancing, mask mandates and other measures [Romano et al., 2022, Perra, 2021], which were estimated to be effective in reducing the spread of the virus [Davies et al., 2020, Flaxman et al., 2020]. Simultaneously, massive investments were made in vaccine developments which could increase immunization against COVID-19 and relax restrictions. The scale and speed of the vaccination campaigns were remarkable: within a year

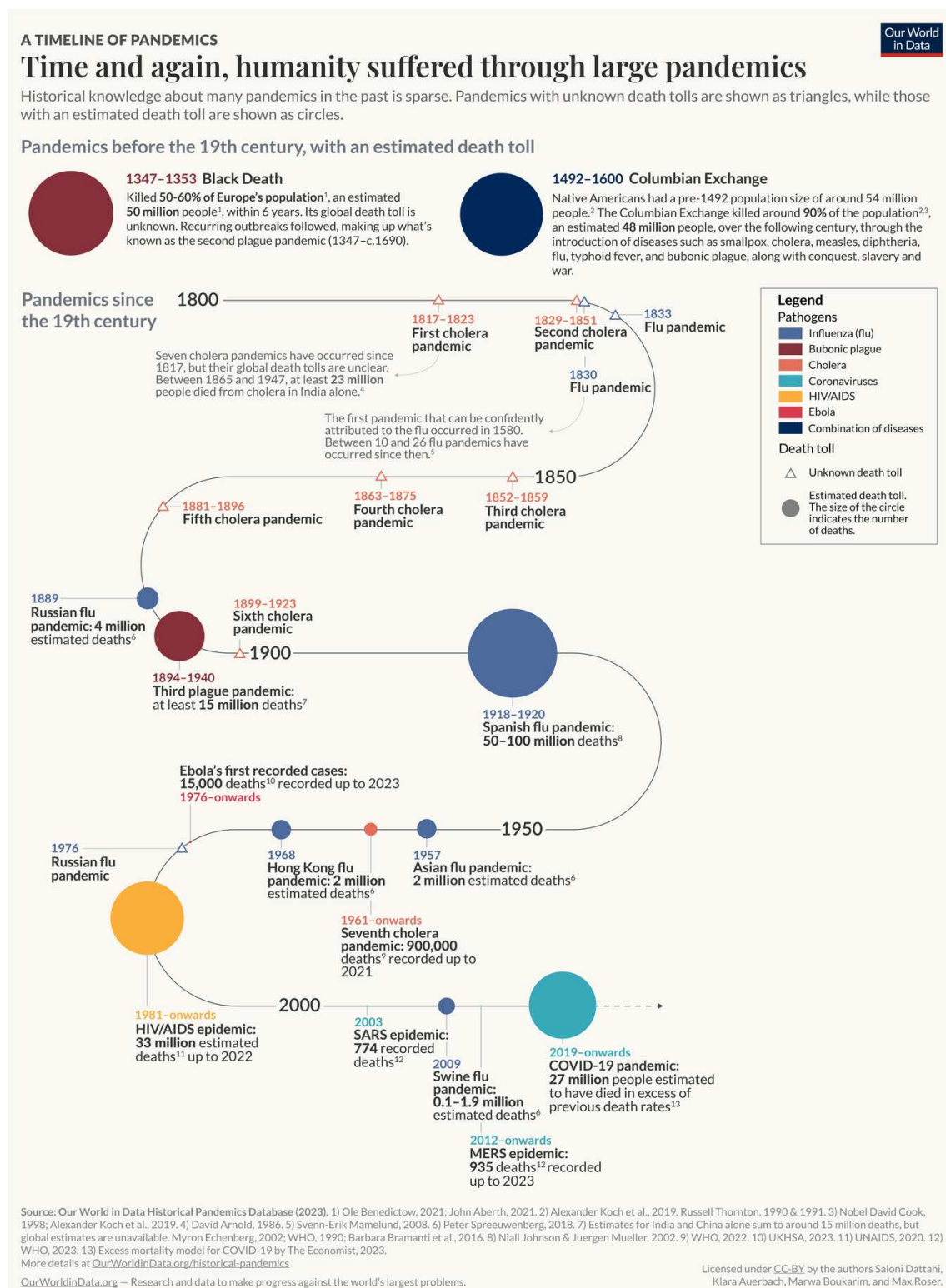


Figure 1.1: History of pandemics. Figure obtained from [Dattani and Roser \[2023\]](#)

of the first administration of a COVID-19 vaccine outside of clinical trials in December 2020 [NHS England, 2020], at least half of the world's population had received at least one dose [Mathieu et al., 2020]. Nonetheless, access to vaccines was not homogeneous and reflected structural and socio-economic disparities between and within countries. For example Gozzi et al. [2023] estimated that up to a half of COVID-19 deaths could have been averted in *Low and Middle Income Countries* (LMIC) countries had they had similar early access to vaccine as *High Income Countries* (HIC).

Although containment measures and vaccination programs were effective in slowing transmission, scientist consistently highlighted the possibility of the emergence of variants which could escape immunity by cumulating evolutionary advantageous mutations [Fontanet et al., 2021, Mascola et al., 2021]. Over time, a variety of different lineages emerged, whose increased transmissibility had the potential to destabilize already struggling healthcare systems. These public health concerns led the WHO to the label as *Variant of Concern* (VoC) any viral lineage satisfying one or more of the following criteria: increased transmissibility, increased virulence, change in clinical disease presentation, or decrease in effectiveness of public health and social measures or available diagnostics, vaccines, and therapeutics [WHO, 2024]. One of the first strains to be characterised as a VoC was Alpha (B.1.1.7). First detected in November 2020 in the United Kingdom, it was characterised by higher transmissibility compared to wildtype [Davies et al., 2021, Volz et al., 2021] and quickly became the dominant variant in the country, reaching a prevalence of 95% among COVID-19 cases in February 2021 [Public Health England, 2021]. Similarly, other variants of concern emerged around that time: Beta (B.3.5.1) and Omicron (B.1.1.529) were first detected in South Africa in May 2020 and November 2021 respectively, Gamma (P.1) in Brazil in December 2020, Delta (B.1.617.2) in India 2020 and Omicron in November 2021 [Duong, 2021, Fujino et al., 2021, Naveca et al., 2021, Faria et al., 2021, Carrazco-Montalvo et al., 2022]. All of these were characterised by selective advantages over their predecessors.

1.2 The HIV/AIDS pandemic

Acquired Immunodeficiency Syndrome (AIDS) made its first, sudden appearance in the 1980's. The first reported cases dated 1981, and concerned 19 white homosexual individuals exhibiting cancer and lung infections [Centers for Disease Control (CDC), 1981]. Additionally to *Men who have Sex with Men* (MSM), subsequent diagnoses involved people with haemophilia, people who inject heroin and Haitians [Centers for Disease Control (CDC), 1982, Kagaayi and Serwadda, 2016], all suffering similar symptoms and facing severe social stigma. Increasing concern and research efforts led to the independent discoveries of the causative virus, HIV, both in the USA [Gallo et al., 1984] and

in France [Barré-Sinoussi et al., 1983]. Associated with higher income countries, non traditional sexual practices and stigma, HIV was either deemed unlikely to be spreading in Africa, or worse denied altogether [Hitchcock, 1986].

The introduction of HIV in humans was later estimated to have occurred in the 1930's in Cameroon [Korber et al., 2000]. According to the bushmeat theory, human HIV infection was a result of zoonosis from chimpanzees due to hunting with poor hygienic conditions [Kurpiers et al., 2015]. The geographic reach of the virus expanded shortly after, following rivers that were major route of travel and commerce [Sharp and Hahn, 2011]. Although the exact extent of the spread remains unknown, it is believed that outbreaks of AIDS had already been happening in Central Africa between 1960 and 1980, for example through increases in cases of cryptococcal meningitis in Kinshasa [Molez, 1998]. By the 1990, there were an estimated 5.8 million cases of HIV in Africa, but no treatment yet available [Ortiz-Ospina and Roser, 2016, Our World in Data, 2016].

The timing and the extent of the response to AIDS differed widely across the African continent, but its impact was felt everywhere. Following reports of increasing incidence in Kampala in 1987 [Carswell and Lloyd, 1987], Uganda was the first country to mount a national response to HIV, effectively managing to slow prevalence through educational and destigmatizing campaigns [Kagaayi and Serwadda, 2016]. On the other extreme, the South African government, sometimes described as denialist, failed to address the rise from less than 1% to 30% in antenatal HIV prevalence from 1999 to 2004 [Simelela et al., 2015]. Overall, the impact of HIV in *Eastern and Southern Africa (ESA)* in the 1990s was devastating. More than half of adult mortality was attributable to HIV [Gregson et al., 1994], with widespread mother-to-child transmission and high rates of orphanhood as a consequence [Preble, 1990].

Between the late 1990s and the early 2000s, the growing international concerns motivated by the high mortality, together with the optimism of decreasing costs of HIV treatment, paved the way for the creation of international organizations planning the response to the pandemic. The foundations of the *Joint United Nations Programme on HIV/AIDS (UNAIDS)* in 1994, the *Global Fund to Fight Aids (Global Fund)* in 2002 [Swiss Agency for Development and Cooperation SDC., 2002], and of the *US President's Emergency Plan for AIDS Relief (PEPFAR)* in 2003 brought about substantial investments to fund research and scale up implementation of evidence-based interventions. Notably, some of the most effective interventions are *Voluntary Medical Male Circumcision (VMMC)*, which is estimated to reduce up to half the risk of transmission towards men [Auvert et al., 2005, Gray et al., 2007, Bailey et al., 2007] and *AntiRetroviral Therapy (ART)*, which suppresses viral replication in an infected host and is highly effective in reducing mortality [Mills et al., 2011, Herbst et al., 2009, Reniers et al., 2014] and prevent

onward transmission [Attia et al., 2009, Cohen et al., 2011]. The optimism sparked by these findings, together with the ambition of ending the epidemic by 2030 [United Nations, 2015] led to the definition of ambitious global targets. In 2015, the majority of world countries [UNAIDS, 2022] committed to the UNAIDS “90-90-90” targets, requiring that 90% of PLHIV are aware of their status, 90% of those aware are on ART, and 90% of those on ART achieve viral suppression by 2020 [UNAIDS, 2015]. The rationale was *Treatment as Prevention (TasP)*, i. e. the idea of guaranteeing easy access to treatment to minimize the number of new HIV cases. Despite the success in scaling-up ART administration, treatment uptake was heterogeneous across population subgroups. Notably, it was lower in younger individuals, men and in key populations [Frescura et al., 2022]. To address these disparities, in December 2020 UNAIDS introduced the analogous “95-95-95” targets, with the additional constraint that each target should be met by all population subgroups [UNAIDS, 2020].

Despite all the advances, the fight against HIV is still ongoing. In 2022, there were an estimated 630,000 (480,000-880,000) deaths worldwide and 1.3 million new infections globally, of which more than half occurred in ESA [UNAIDS, 2023]. Further, even though HIV incidence is declining, this reduction is geographically heterogeneous and the UNAIDS incidence targets for 2020 were missed [Mahy et al., 2021]. It is unclear whether the 2030 SDG goal of ending AIDS as a public health threat will be met, even if “95-95-95” targets are achieved by 2025. One possible reason for this is that different population subgroups contribute to HIV incidence with different strength. Understanding which population subgroups contribute most strongly allows to design targeted interventions and speed up HIV incidence reductions. In my thesis, a principal focus concerns using surveillance data to characterize which age-groups contribute most to ongoing transmission.

1.3 The Rakai Community Cohort Study

HICs and countries in Sub-Saharan Africa face different HIV epidemics, and as such require different approaches to control them. While in places such as the USA the epidemic is “concentrated”, meaning it is mostly confined to specific population subgroups, in ESA the epidemic is “generalized”, meaning HIV infection is firmly established in the general population [WHO, 2013]. As such, the role of African-based research in the local fight against HIV has been, and still is, crucial in characterising the problems and in evaluating the effectiveness of different interventions.

The *Rakai Health Sciences Program (RHSP)* is one of the oldest HIV research programs in ESA. Its establishment in 1987 was a response to the first identification of clinical AIDS in the Rakai district Uganda in 1982, and successively took up the role as a leading

“hypothesis generating and hypothesis testing” institution [Gray and Wawer, 2008] in the fight against the epidemic. This experimental role was perfectly exemplified by the Rakai STD Control for Prevention Study. Taking place between 1994 and 1998, it aimed to assess whether the cure of *Sexually Transmitted Diseases* (STD) could reduce the risk of HIV acquisition [Wawer et al., 1999]. Although the results were negative, the study was successively extended and renamed to *Rakai Community Cohort Study* (RCCS). Table 1.1 reports some of the key published findings, together with a timeline of the study rounds and of the interventions that were implemented.

The RCCS is an open, population-based, community HIV surveillance cohort of individuals aged 15 to 49 years old. The study includes communities in and around the Rakai District, in South Central Uganda, bordering Lake Victoria to the east and Tanzania to the south. Initially, it included 30 agrarian and semi-urban communities, but the number has since then increased to 40, notably with the addition of four fishing communities around Lake Victoria in 2011 [Chang et al., 2016]. In fact, 2011 coincided with the implementation of priority measures in these four hyperendemic communities, and the RCCS aimed to study the effect of *Universal Test and Treat* (UTT), i. e. treatment initiation immediately after the first positive test [Chang et al., 2017]. This measure was then extended to all communities in 2016.

The study is conducted in rounds, and each round is subdivided into three phases. Firstly, an informational community mobilisation event is carried out to identify eligible cohort participants. Secondly, a population census is executed. This involves a GPS mapping of all the households, followed by visits in which household heads report the number of people living in the household, as well as their gender and age. Individuals are then invited to participate in the study. Participation is only allowed for residents for at least 6 months, except for fishing communities where the residency requirement is 1 month plus a declared intention to stay. Finally, the RCCS round starts. Participant consent is obtained privately by a trained RCCS interviewer. Following consent, participants report either their birth date or current age themselves, and accompanying documentary evidence is requested. Then, the participant is interviewed on demographic characteristics, sexual behaviours, healthcare-seeking behaviour, and successively tested for HIV. In case of positive test, the participant is offered counselling and referral to care if needed. Starting from round 15, serum samples are collected for potential HIV-1 viral sequencing and viral load measurements.

Table 1.1: **Timeline of RCCS Survey Rounds.** RCCS survey round dates with contemporaneous interventions and related work.

Round	Start	End	Notable interventions	Related work
A1	Feb 1994	Dec 1995	Interventions to reduce STD transmission	Start of [Wawer et al., 1999]
A6	Aug 1995	Aug 1996		
B1	Jun 1996	May 1997		
B6	Jun 1997	Oct 1998		
W1	Apr 1999	Feb 2000		
W2	Feb 2000	Feb 2001	Measures for PMTCT	End of [Wawer et al., 1999] [Gray et al., 2001] Start of [Grabowski et al., 2017]
W3	Apr 2001	May 2002		
R9	Jul 2002	Aug 2003		
R10	Sep 2003	Nov 2004		
R11	Feb 2005	Jun 2006		
R12	Aug 2006	Jun 2008		
R13	Jun 2008	Jul 2009		
R14	Jan 2010	Jun 2011		
R15	Aug 2011	Jun 2013		
R16	Jul 2013	Jan 2015		
R17	Feb 2015	Sep 2016	ART if CD4 < 250 CT testing VMMC effect on incidence CHP (VMMC + ART)	Start of [Monod et al., 2023b] [†]
R18	Oct 2016	May 2018		
R19	Jun 2018	May 2019		
R10	Sep 2003	Nov 2004	ART if CD4 < 250	[Chang et al., 2016]; start of [Ratmann et al., 2020] End of [Ratmann et al., 2020]; start of Chapter 6 [†] End of [Grabowski et al., 2017] End of [Monod et al., 2023b] [†] End of Chapter 6 [†]
R11	Feb 2005	Jun 2006	CT testing VMMC effect on incidence	
R12	Aug 2006	Jun 2008	CHP (VMMC + ART)	
R13	Jun 2008	Jul 2009		
R14	Jan 2010	Jun 2011	ART if CD4 < 500	
R15	Aug 2011	Jun 2013	4 Lake Victoria fishing communities added to RCCS UTT in fishing communities and for key populations UNAIDS sets “90-90-90” targets.	
R16	Jul 2013	Jan 2015		
R17	Feb 2015	Sep 2016	UTT for all	
R18	Oct 2016	May 2018		
R19	Jun 2018	May 2019	Community outreach efforts (PrEP) [RHSP, 2018]	

* **Abbreviations:** Prevention of Mother to Child Transmission (PMTCT); Sexual Transmitted Diseases (STD); Combination HIV Prevention (CHP); Pro-Exposure Prophylaxis (PrEP).

[†] Project I contributed to.

1.4 Challenges

The statistical analysis of epidemiological data is crucial for public health planning. Many challenges lie in the interpretation of often limited, biased data to quickly assess the severity of a situation and the potential impact of different interventions. Some of the most relevant challenges tackled in this thesis are:

- **Missing data.** An epidemic is always going to be only partially observed: some cases are not going to be reported for a multitude of reasons, such as lack of an easily accessible global testing infrastructure. In both HIV and SARS-CoV-2 cases, individuals might also be unaware of their own infectious status due to absence of symptoms or long incubation periods.
- **Data aggregation and evidence synthesis.** Multiple data sources may be informative to answer a question of interest. These sources might measure the same quantity at different locations, time or granularity (such as cases), or they might measure different quantities that provide a wider context (vaccination, census, phylogenetics). This poses the problem of how to weight evidence coming from

different sources, especially when some sources of data may be conflicting [De Angelis et al., 2015].

- **Biases and confounding.** In order to generalize a statistical model inference, it is often crucial to understand how the data was collected in relation to the population of interest. In the case of HIV in ESA, testing rates are typically higher in women and older individuals. Not accounting for this might lead to underestimates of the contributions to the burden of HIV by younger individuals.
- **Limited data availability in LMIC.** This thesis focuses on the analysis of data from LMIC countries, where the impact of communicable diseases is most severe, but data are often less available or reliable. Analyses that are easily performed in HIC might be impossible in LMIC countries, and alternatives must be found.
- **Simplicity, transparency and interpretability.** When informing public health interventions, statistician should aim to “find models that are complex enough to reflect sufficient details of the system but simple enough not to get lost in the jungle of details” [Kretzschmar et al., 2022]. To guarantee that model outputs are used for the intended scope, it is essential that their assumptions, purposes and limits are clearly communicated [Grimm et al., 2020].
- **Moving targets.** Public health interest may change in light of new evidence, appearance of unexpected problems or changes in the epidemic itself as per implementation of interventions. For example, the emergence of new pathogen variants may prompt analysts to assess changes in transmission risks and severity of the disease.

1.5 Thesis aims and Chapter descriptions

Before the start of my PhD in October 2020, my research was supposed to be centered on the HIV epidemic in ESA. At the time however, most epidemiologist were focused on the ongoing COVID-19 epidemic, estimating its spread and impact in Europe and elsewhere in the world. Notably, Imperial College was at the forefront of the research, as exemplified by the publication of early, high impact papers [Flaxman et al., 2020, Ferguson et al., 2020, Verity et al., 2020, Nouvellet et al., 2021]. The opportunity of working on urgent and high-impact projects temporarily directed me towards COVID-19 projects, and it was only after 2021 that I started working on HIV. This shift is reflected in the structure of my thesis: Chapter 3 and Chapter 4 deal with COVID-19, while Chapter 5 and Chapter 6 deal with HIV. Despite this dichotomy, common motivations underlie the two sections. Each project involved the analysis of surveillance data to provide key

strategical information to policy makers at different stages of a pandemic. An almost chronological ordering emerged:

- In Chapter 3, we address biases in the estimation of the basic reproduction number in the very early stages of an epidemic, when the first infections are likely not to have been reported.
- In Chapter 4, we assess changes in the mortality of COVID-19 across Brazil, after the emergence of a variant of concern with increased immune escape. The project, aiming to quantify the effect of the novel variant on hospital mortality, is an example of how an epidemic can change due to evolutionary pressures on the pathogen, and how its impact can be heterogeneous across locations.
- In Chapter 5 we demonstrate how the dynamics of a long-established epidemic such as HIV can change through interventions, by describing changes in transmission flows in the RCCS. Here, we focus on the importance of dating HIV infections events to improve comparisons of transmission patterns over time. We develop a pipeline to estimate individual-level HIV time since infection, and a novel statistical model to estimate infection dates for individuals within a transmission network.
- In Chapter 6, we further describe temporal changes in the age and gender contributions to the burden of HIV and viraemia. Based on these analyses, we propose metrics to help prioritize prevention efforts aimed at ending HIV as a public health threat by 2030.

In summary, this thesis addresses statistical problems arising at different stages of an epidemic, ranging from early surveillance data biases to changes in transmission dynamics and ending with challenges related to bringing to control of the deadliest pandemics in history.

The previously mentioned Chapters are preceded by a technical background chapter (Chapter 2) and followed by a discussion (Chapter 7).

2

TECHNICAL BACKGROUND

2.1 Basic epidemiological definitions

One of the first aims of “quantitative” epidemiology is to measure the burden of a disease in a population, and to understand properties of transmission dynamics. In this section, we provide some of the basic definitions which are necessary to understand the rest of the thesis.

2.1.1 Population, incidence and prevalence

A *risk population*, or simply *population*, is the collection of individuals we are interested in studying. A *category*, or *population subgroup* of the population is a group whose constituents share some common characteristics such as sex, age, location or ethnicity. The *prevalence* of a disease is the proportion of individuals within the population who have been infected by the pathogen of interest. By definition, it is always a number between 0 and 1. *Incidence* is the rate of new infections over a specified period of time for the population at risk for infection. For HIV, it is typically reported in person-years. For example, an incidence rate of 0.01 infections per person-year means that 1% of the susceptible population is expected to be infected in a given year.

2.1.2 Individual level disease status pathogen

A *pathogen*, or *agent* is a microorganism such as a bacterium or a virus that may cause the disease of interest upon infection. An individual is *infected* if she was subject to invasion by the pathogen of interest. These are also denoted as *cases*. An infected individual is *infectious* if it has the potential to transmit the disease to another individual.

An individual is *susceptible* if she is not currently infected and could potentially develop the disease. A *transmission event* is the process by which an infectious individual transmits the disease to a susceptible individual. In this context, the person who transmitted the disease is called *infector* or *source* (of transmission). The person who got infected is called *infectee* or *recipient*. A pair of infector and infectee is also called a *transmission pair*.

2.1.3 Incubation period, generation time and serial interval

The *incubation period* denotes the time delay between infection and appearance of the first evidence, or symptom of the disease. The *generation interval* is the time delay between the infection of an infected person and the infection of her infector. In practice, infection dates are not observed, while it is easier to observe the onset of symptoms, which justifies the following definition. The *serial interval* is time delay between the development of symptoms in an infector and in her infectee. Note that while the generation interval must always be a positive number, the serial interval can be negative if the infectee develops symptoms before the infector.

2.1.4 Measures of transmissions speed

The *growth rate* r_t is defined as the per capita change in the number of new cases per unit of time, evaluated at time t . This measure is sometimes reported in an alternative form, the *doubling time*, which equals $\log(2)/r_t$, and represents the time it takes for the number of cases to double. Although the growth rate is a useful measure of the speed of transmission, other metrics because are often easier to interpret because they provide deeper insights into the potential for disease spread and the effectiveness of control measures. The *basic reproduction number* R_0 is the average number of secondary cases generated by a single infectious individual in a completely susceptible population. The *instantaneous reproduction number* R_t is the average number of secondary cases generated by a single infectious individual at time t . These measures are a function of both the growth rate and of the generation interval [Wallinga and Lipsitch, 2007], and relate to key epidemiological quantities such as the final size of an epidemic [Andreasen, 2011] and the critical herd immunity threshold [Thompson et al., 2020, Hethcote, 2000]. Some questions of interest can be answered by either considering the growth rate or the reproduction number. For instance, the number of cases at time t is growing if and only if $R_t > 1$ or equivalently $r_t > 0$.

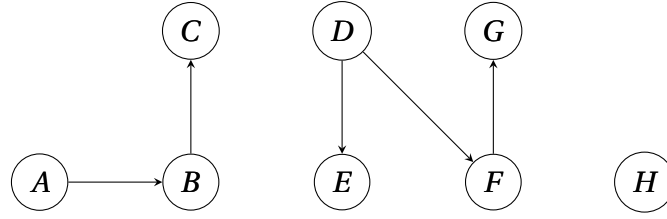


Figure 2.1: An example of a transmission network. Nodes represent individuals, and edges transmissions from a source to a recipient. Note no node has more than 1 incoming arrows, and some nodes may not be connected to others. Further, this transmission network can be broken into three smaller networks, one per connected component.

2.1.5 Transmission networks

A *transmission network* is a directed graph $\mathcal{N} = (V, E)$, where V is the set of nodes and E is the set of edges. Each edge $e \in E$ is a pair of nodes $e = (v_1, v_2)$, representing a transmission event from v_1 to v_2 . Given $\mathcal{N}$ we define the *source* of a node $v \in V$ as the set $\text{source}(v) = \{u \in V : (u, v) \in E\}$. Similarly, we define the *recipients* of a node $v \in V$ as the set of nodes $\text{recipients}(v) = \{u \in V : (v, u) \in E\}$. Note that two edges cannot be directed to the same node, as an individual can only be infected once. An example of a transmission network is shown in Fig. 2.1.

A network is said to be *connected* if for any pair of distinct nodes $v_1, v_2 \in V$, there exists a collection of edges in E connecting v_1 and v_2 . In this instances, a network can be also represented as a tree with a unique *root node*, i. e. a node v without any incoming edges, or $\text{source}(v) = \emptyset$. We can also define the *distance* between two nodes $v_1, v_2 \in V$ as the length of the shortest path connecting them, and the *depth* of a node v as the length of the longest path connecting the root node to another node in V .

2.2 Models of disease transmission

In practice, it is infeasible to observe all cases of a disease in a population and explicitly compute the key epidemiological quantities described previously. For instance, it is impossible to count how many individuals were infected by each case and explicitly compute the reproduction number. Instead, we rely on mathematical models which impose a certain structure of the disease transmission dynamics to infer quantities of interest. These include compartmental models and renewal models.

2.2.1 Compartmental models

Compartmental models are some of the simplest and most common models used in epidemiology to describe the dynamics of disease transmission. They were first introduced by Kermack et al. [1927] and have since been extended in a variety of ways. These

models involve a population of individuals who fall into different *compartments* as a function of their disease status. Each compartment is associated with different properties, and individuals can transition between compartments according to pre-specified dynamics. The dynamics are typically expressed as a system of ordinary differential equations, or as stochastic processes, such as Markov chains. The former case gives rise to deterministic dynamics, while the latter gives rise to stochastic dynamics.

SIR models

The most common compartmental model is the Susceptible-Infected-Recovered (SIR) model, which divides individuals into three compartments. At time t , we define:

$\mathcal{S}(t)$ the number of *susceptibles*. This counts individuals who are not infected and can be infected.

$\mathcal{I}(t)$ the number of *infectious*. This counts individuals who are infected and can transmit the disease.

$\mathcal{R}(t)$ the number of *recovered*. This counts individuals who have recovered from the disease and are assumed to be immune.

A fixed population of size N is assumed, and so the following relationship holds:

$$N = \mathcal{S}(t) + \mathcal{I}(t) + \mathcal{R}(t). \quad (2.1)$$

It is assumed that the disease is transmitted through direct contact between susceptibles and infectious individuals, and that the population is homogeneously mixing, meaning that each individual has the same probability of coming into contact with any other individual. The dynamics of a *deterministic* SIR model are described by the following system of ordinary differential equations:

$$\frac{d\mathcal{S}(t)}{dt} = -\beta \frac{\mathcal{S}(t)\mathcal{I}(t)}{N} \quad (2.2a)$$

$$\frac{d\mathcal{I}(t)}{dt} = \beta \frac{\mathcal{S}(t)\mathcal{I}(t)}{N} - \gamma \mathcal{I}(t) \quad (2.2b)$$

$$\frac{d\mathcal{R}(t)}{dt} = \gamma \mathcal{I}(t), \quad (2.2c)$$

where β is the transmission rate and γ is the recovery rate. Under this model, the computation of the reproduction numbers is trivial:

$$R_0 = \frac{\beta}{\gamma}. \quad (2.3a)$$

$$R_t = R_0 \frac{\mathcal{S}(t)}{N}. \quad (2.3b)$$

Alternatively, *stochastic* SIR dynamics can be imposed by replacing the ordinary differential equations with a continuous-time Markov chain [Allen, 2017].

SEIR models

For most diseases, an infected person does not immediately become infectious, but rather some time needs to pass before they can transmit the disease. The Susceptible-Exposed-Infectious-Recovered (SEIR) model is an extension of the SIR model which accounts for this latent period, and introduces an *exposed* compartment. Compared to the previously defined quantities, we here introduce $\mathcal{E}(t)$ which denotes the number of *exposed* individuals at time t . The dynamics of a *deterministic* SEIR model are described by the following system of ordinary differential equations:

$$\frac{dS(t)}{dt} = -\beta \frac{S(t)\mathcal{I}(t)}{N} \quad (2.4a)$$

$$\frac{d\mathcal{E}(t)}{dt} = \beta \frac{S(t)\mathcal{I}(t)}{N} - \sigma \mathcal{E}(t) \quad (2.4b)$$

$$\frac{d\mathcal{I}(t)}{dt} = \sigma \mathcal{E}(t) - \gamma \mathcal{I}(t) \quad (2.4c)$$

$$\frac{d\mathcal{R}(t)}{dt} = \gamma \mathcal{I}(t), \quad (2.4d)$$

where σ is the rate at which exposed individuals become infectious. As for the SIR model, the reproduction numbers can be computed as in Eq. (2.3), and stochastic dynamics can be imposed by replacing the ordinary differential equations with a continuous-time Markov chain.

2.2.2 Renewal models

Renewal models are an alternative class of models used to describe incidence dynamics, i. e. the evolution of the number of cases over time. While compartmental models simultaneously track how many individuals within a population are in each compartment, renewal models only focus on those individuals that were previously infected. This results in a simpler model structure, where, for example, there is no explicit need to specify population size.

Renewal models predict changes in incidence $I(t)$ at time t as a function of previous incidence and transmissibility $\beta(t, \tau)$ [Fraser, 2007] as follows:

$$\mathbb{E}[I(t) | (I(u))_{u < t}] = \int_0^\infty \lambda(t, \tau) I(t - \tau) d\tau. \quad (2.5)$$

Specifically, $\lambda(t, \tau)$ describes the rate of secondary cases generated at time t by an individual who was infected at time $t - \tau$. It is common to assume the decomposition

$\beta(t, \tau) = R(t)w(\tau)$, where $R(t)$ is the instantaneous reproduction number and $w(\tau)$ is the generation interval distribution [Fraser, 2007, Pakkanen et al., 2023], resulting in the following equation:

$$\mathbb{E}[I(t)|(I(u))_{u < t}] = R(t)\Lambda(t) \quad (2.6)$$

$$\text{with } \Lambda(t) = \int_0^\infty w(t)I(t-\tau)d\tau. \quad (2.7)$$

In practice, data is available at discrete time points [Cori et al., 2013], and so the integral in Eq. (2.5) is often replaced by a sum over a finite number of time points. We therefore obtain the following discrete-time renewal model:

$$\mathbb{E}[I_t|(I_u)_{u < t}] = R_t\Lambda_t = R_t \sum_{s=0}^{\infty} w_s I_{t-s}, \quad (2.8)$$

where the subscripts are indicative of discrete indexing and I_t denotes the number of cases at time t . Below are two examples of renewal-based models widely available through R packages.

2.2.3 EpiEstim

The EpiEstim package provides software to estimate the time-varying reproduction number following the methodology in Cori et al. [2013]. Without loss of generality, we assume that the time unit of interest are days, and we let $t = 0$ denote the first day for which incidence is reported. Here, the observed cases $(I_t)_{t \geq 0}$ are modelled with a Poisson likelihood with expectation given by Eq. (2.8), as follows:

$$\mathbb{P}(I_t|(I_s)_{0 \leq s < t}, w, R_t) = \frac{(R_t\Lambda_t)^{I_t} e^{-R_t\Lambda_t}}{I_t!}. \quad (2.9)$$

To avoid high variance in the estimates, the reproduction number is assumed to remain constant over sliding windows of time with length τ . Assuming a Gamma prior distribution on $R_{t,\tau}$, the posterior distribution is Gamma distributed with parameters:

$$R_{t,\tau} \sim \text{Gamma}(a_{\text{post}}, b_{\text{post}}) \quad (2.10a)$$

$$a_{\text{post}} = a_{\text{prior}} + \sum_{s=t-\tau+1}^t I_s \quad (2.10b)$$

$$b_{\text{post}} = \left(\sum_{s=t-\tau+1}^t \Lambda_s + b_{\text{prior}}^{-1} \right)^{-1}, \quad (2.10c)$$

where a and b refer to the shape and scale parameters, respectively. Note that the sliding windows are overlapping, and the estimation of the different $R_{t,\tau}$ is performed separately i. e. not with a unique, joint model.

One limitation of this model concerns the explicit reliance on a constant reporting process. The likelihood of cases in day t is fully determined by the number of previously reported cases. As such, changes in the reporting process, such as changes in ascertainment, may confound reproduction number estimates.

2.2.4 Epidemia

Epidemia is an alternative software based on the renewal equation that explicitly models the reporting process to avoid confounding [Scott et al., 2020, Bhatt et al., 2023]. This is achieved with two key differences. Firstly, the renewal equation is imposed on the latent incidence process $(i_t)_t$ which models the number of cases in the entire population. Secondly, reported cases $(Y_t)_t$ are modelled as a noisy observation of the latent incidence process, for example by accounting for ascertainment rates α_t or infection-to-outcome delays distributions π . For instance, we may model:

$$Y_t \sim \text{Negative Binomial} \left(\alpha_t \sum_{k < t} i_k \pi_k, \phi \right), \quad (2.11)$$

where ϕ is an overdispersion parameter. Compared to EpiEstim, the model is more flexible. It allows for both parametric and non-parametric inference of R_t through and for multiple observation processes, such as daily hospitalizations and daily deaths. As a drawback, Epidemia requires an explicit definition of the starting date of the epidemic as well as a prior on the number of cases present that day which typically leads to identifiability issues in R_0 estimates. The reason concerns transmissions in early stages. High transmissibility can be either explained by high R_0 or by many initial cases or importations. If there is no data on the number of early cases, the model will have no way to distinguish between the two sources of variation.

2.3 Statistical modelling

The main approach to statistical inference in this thesis is Bayesian. This decision can be justified by a number of reason: flexibility in the design of the model, ability to incorporate prior knowledge, and the necessity to accurately quantify uncertainty. These justification are all crucial in epidemiological applications [Dunson, 2001] where the data is often limited, and effects of interventions applied in different settings may be available.

2.3.1 Bayesian inference

Bayesian statistical inference aims to describe unknown quantities of interest through probability statements. It provides a framework for estimating unknown quantities of interest, by updating prior beliefs in light of observed data. This process occurs within the context of a family of statistical models, each of which represents a data-generating process deemed plausible. The models are distinguished by their *parameters*, denoted by θ , which govern the behaviour of the model. The *likelihood function* $p(y|\theta)$, measures the probability of observing the data y under a given parameter value. The core of Bayesian inference lies in updating our beliefs about the parameters θ after observing the data y . This is achieved through Bayes' theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \quad (2.12)$$

Here, $p(\theta)$ and $p(\theta|y)$ represent the *prior* and *posterior* distribution respectively, encoding our beliefs about the parameter before and after observing the data. The term $p(y) = \int p(y|\theta')p(\theta')d\theta'$ is denoted as *evidence*. As it is constant with respect to θ , we can rewrite Eq. (2.12) as:

$$p(\theta|y) \propto p(y|\theta)p(\theta). \quad (2.13)$$

The specification of the prior distribution is a key aspect of Bayesian inference, and the choice should be made to reflect our prior beliefs on the parameter of interest. *Informative* priors encode strong beliefs about the parameters. These are used when substantial prior information is available, such as previous studies or expert knowledge. *Non-informative* priors encode weak beliefs about the parameter, and are used when little prior information is available. *Weakly informative* priors encode beliefs that are intermediate between the two extremes. In practice, prior strength is sometimes controlled to address identifiability issues.

2.3.2 Posterior predictive distribution

After updating the probabilistic description of the parameter of interest θ , a Bayesian analyst may be interested in how this updated knowledge reflects on future data. This corresponds to evaluating the *posterior predictive distribution*, which is the distribution of future data $\tilde{y}$ given the observed data y . If the data $y, \tilde{y}$ is conditionally independent given the parameter θ , the posterior predictive distribution can be computed as follows:

$$p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y) d\theta \quad (2.14a)$$

$$= \int p(\tilde{y}|\theta, y) p(\theta|y) d\theta \quad (2.14b)$$

$$= \int p(\tilde{y}|\theta) p(\theta|y) d\theta. \quad (2.14c)$$

Hence, the posterior predictive distribution can be interpreted as a weighted average over the conditional predictions $p(\tilde{y}|\theta)$ over the posterior distribution $p(\theta|y)$.

2.3.3 Monte Carlo methods

In the previous sections, we introduced the posterior distribution $p(\theta|y)$ as the main object of interest in Bayesian inference. In most cases, this distribution is only known up to a normalizing constant as the evidence is unknown. Therefore, bespoke methods and techniques are required to either summarize this distribution or to sample from it.

Monte Carlo methods are a class of algorithms that rely on random sampling to obtain samples from a target distribution f of interest. These algorithms can be either *exact*, or *asymptotically exact*. The former category includes rejection sampling, which requires the knowledge of an envelope function that bounds the target distribution. The latter category is more general as it only requires knowledge of the target distribution up to a normalizing constant. Asymptotically exact methods include *Markov Chain Monte Carlo (MCMC)* methods, which involve constructing a Markov Chain whose stationary distribution is (or resembles) the target distribution.

A further distinction can be made between first and second-generation *MCMC* methods. The first generation methods, such as the Metropolis-Hastings algorithm and Gibbs sampling, are simple and flexible but can struggle with slow convergence, particularly in high-dimensional or complex parameter spaces. These methods rely on random walks or conditional sampling, which can lead to inefficient exploration of the posterior distribution. In contrast, second-generation MCMC methods, such as Hamiltonian Monte Carlo, leverage gradient information to guide the exploration of the parameter space, resulting in more efficient sampling and faster convergence.

Rejection sampling

Rejection sampling requires the knowledge of an envelope distribution g such that $f(\theta) \leq Mg(\theta)$ for some $M > 0$. The algorithm proceeds by sampling x from g and accepting it with probability $f(\theta)/(M \cdot g(\theta))$. The accepted sample will have the de-

sired distribution f . Additionally, if the samples from g are independent, the resulting samples from f will be independent too.

Gibbs Sampling

When rejection sampling is not feasible, it is often possible to construct a Markov Chain (that is a sequence of samples) whose stationary distribution is the target distribution. A basic example of a Markov Chain Monte Carlo approach is *Gibbs sampling*. This algorithm is especially useful when the joint distribution of the parameters of interest is complicated, but the marginal distribution for each parameter conditional on the others is simple.

Mathematically, for a p -dimensional parameter vector $\theta = (\theta_1, \dots, \theta_p)$, we define θ_{-i} to be the $p - 1$ -dimensional vector of all parameters except θ_i , and we define the *full conditional* distribution of θ_i as $p(\theta_i | \theta_{-i})$. Gibbs sampling is initialized with $\theta^{(0)} \in \text{supp}(f)$ and inductively obtains samples $\theta^{(n+1)}$ from $\theta^{(n)}$ by sampling coordinate-wise from the full conditional distributions:

$$\theta_i^{(n+1)} \sim f\left(\theta_i \mid \theta_1^{(n+1)}, \dots, \theta_{i-1}^{(n+1)}, \theta_{i+1}^{(n)}, \dots, \theta_p^{(n)}\right). \quad (2.15)$$

Under mild conditions, repeating this process for $n = 0$ to N will produce a chain $(\theta^{(n)})_{n=1}^N$ with stationary distribution matching the target distribution f [Robert and Casella, 2004].

Metropolis Hastings algorithm

In practice, we cannot always guarantee that the full conditional distributions are easy to sample from, and it is often necessary to resort to more sophisticated methods to explore the parameter space. The *Metropolis Hastings* algorithm represents one of the most general approaches to propose and accept new samples in such a way that the resulting chain has the desired stationary distribution. The algorithm involves sampling candidates θ' from a proposal distribution $q(\theta' | \theta)$, and accepting them with probability

$$\alpha(\theta, \theta') = \min\left(1, \frac{f(\theta')q(\theta | \theta')}{f(\theta)q(\theta' | \theta)}\right). \quad (2.16)$$

Intuitively, the proposal q represents an “easy-to-sample-from” distribution which is used to explore the support of the target distribution f , while the accept-reject mechanism discriminates against samples that are less likely to come from f as opposed to the current sample.

The advantages of the Metropolis Hastings algorithm and of Gibbs sampling are the simplicity and flexibility, and the usage is widespread for problems where θ has low dimensionality. However, the speed of convergence of the chain to the desired stationary regime may be compromised by multi-modal target distributions or by correlation structures among the parameters.

Hamiltonian Monte Carlo

A potential solution to the shortcomings of common implementations of the Metropolis Hastings algorithm is *Hamiltonian Monte Carlo (HMC)*. HMC utilizes gradient information of the target distribution to guide the exploration of the parameter space, and hence facilitate convergence to the stationary regime.

The technical implementation involves augmenting the parameter space with a momentum variable ρ of the same dimension of θ . This means we are now interested in obtaining samples from a joint distribution $f(\theta, \rho)$. As in the Metropolis Hastings algorithm, HMC involves proposing a candidate (θ', ρ') , and then accepting it with a given probability. In theory, the improvement lies in an acceptance probability equal 1, which is achieved by following specific dynamics in the augmented space. In practice, the dynamics are approximated, and therefore the computation of the acceptance-rejection mechanism (Eq. (2.16)) is still required. The function guaranteeing efficient exploration is built upon intuition from physics. This is denoted the *hamiltonian* function and represents the logarithm of the augmented target distribution:

$$H(\theta, \rho) = -\log f(\theta, \rho). \quad (2.17)$$

Enforcing independence between θ and ρ , we then obtain:

$$H(\theta, \rho) = -\log f(\theta) - \log f(\rho). \quad (2.18)$$

In practice, it is common to chose $\rho \sim \text{MVN}(0, \Sigma)$, where the covariance matrix Σ is generally tuned from a set of initial draws in the training phase, denoted as *warmup*. Given a previous sample $(\theta_{i-1}, \rho_{i-1})$, sampling from Eq. (2.18) involves:

- Sampling $\rho^* \sim p(\rho)$.

- Evolving the system with initial conditions (θ_{i-1}, ρ^*) according to the hamiltonian dynamics for a fixed amount of time ϵL :

$$\frac{d\theta}{dt} = \frac{\partial H}{\partial \rho} \quad (2.19a)$$

$$\frac{d\rho}{dt} = -\frac{\partial H}{\partial \theta}. \quad (2.19b)$$

In practice, the dynamics are approximated by the leapfrog algorithm. This algorithm involves taking L discrete steps of size ϵ , after which the candidate (θ', ρ') is proposed.

- Accepting the candidate with probability

$$\alpha(\theta, \rho, \theta', \rho') = \min(1, \exp(H(\theta, \rho) - H(\theta', \rho'))). \quad (2.20)$$

Again, note that if the dynamics in Eq. (2.19) are followed exactly, the acceptance probability is equal to 1. Once enough samples have been generated, it is possible to discard the momentum variables ρ yielding samples from the target distribution $f(\theta)$. More details on the implementation of HMC can be found in Betancourt [2018].

No-U-Turn Sampler (NUTS) The No-U-Turn Sampler (NUTS) is an extension of HMC that addresses the need for manual tuning of hyperparameters. NUTS automatically adjusts the step size and the number of leapfrog steps during the sampling process, eliminating the need for user intervention. This is achieved by recursively building a trajectory of proposals until the algorithm detects a “U-turn”, at which point the trajectory is terminated [Homan and Gelman, 2014a].

2.3.4 Software for Bayesian computation

The increase in popularity of Bayesian methods in the last 30 years would not have been possible without the development of software simplifying its usage. One of the main goals of such pieces software is to abstract away sampling algorithms from model specification. This way, an user can focus on modelling, by describing the likelihood and prior through a *modelling language*. Different modelling languages exist that cater to different needs, but they can be broadly grouped based on the underlying inference engine [Štrumbelj et al., 2024]. The first general purpose modelling languages emerged starting from the 1990s, and were mainly based on Gibbs samplers. With the popularisation of automatic differentiation through differentiable programming, HMC became a viable alternative to Gibbs sampling, leading to a second wave of modelling languages. Recently, with the advances in software acceleration and the need to work

with more complex models, a wave of general purpose probabilistic programming languages emerged, which provide a unified interface to different inference engines.

Gibbs samplers based software

Gibbs sampling represents a natural solution to sample from the posteriors of Bayesian networks. These models admit the following representation for the joint distribution of the parameters and data $\mathcal{V} = (y, \theta)$:

$$p(\mathcal{V}) = \prod_{v \in \mathcal{V}} p(v | \mathcal{P}(v)), \quad (2.21)$$

where $\mathcal{P}(v)$ denotes the parents of v in the graph, that is the smallest set of nodes in $\mathcal{V}$ such that $v \perp\!\!\!\perp \mathcal{P}(v) | \mathcal{P}(v)$.

The BUGS (Bayesian inference Using Gibbs Sampling) language [Gilks et al., 1994] was one of the first to provide a high-level interface to specify such models, requiring the user to define the full conditional distributions of the parameters in the right-hand side of Eq. (2.21). Improvements in portability and in sampling (e.g. through data augmentation techniques [Hobert, 2009]) were successively implemented in JAGS (Just another Gibbs Sampler) [Plummer, 2003].

Despite the success of these programs, Gibbs sampling faces some fundamental limitations. In particular, convergence to the posterior distribution may be slow if two or more parameters are highly correlated, slowing down the exploration of the parameter space. This issue is exacerbated in high-dimensional models, where the number of parameters is large.

HMC based software and Stan

HMC can be an attractive alternative to Gibbs sampling, as its performance scales better to high-dimensional models. Improvements in performance are achieved by using gradients of the posterior distribution to guide the exploration of the parameter space. These improvements hence comes at the cost of two additional requirements: the need to compute the gradient of the posterior distribution, and the need to specify hyper-parameters of the inference engine. The first problem was solved with the advent and popularisation of differentiable programming, which makes it possible to automate the computation of gradients of a wide range of functions. The second problem is typically addressed by the inclusion of warmup phases, and bespoke heuristics and tuning procedures.

The most popular HMC software is Stan [Carpenter et al., 2017], which implements samplers such as NUTS in C++, but provides interfaces to different programming languages,

such as R and Python. Most Bayesian models described in this thesis were implemented in Stan. Specifically, we used the `cmdstanr` [Gabry et al., 2023] and `RStan` [Stan Development Team, 2024] interfaces to Stan from within the R programming language [R Core Team, 2022].

General purpose probabilistic programming languages

General purpose probabilistic programming languages aim to provide a unified interface to different inference engines. In fact, there currently are no inference engines that are optimal for (or can handle) all types of models. For instance, `HMC` requires the computation of gradients, which are not available for models with discrete parameters. The aim of general purpose probabilistic languages is to provide a unified interface to different inference engines, while maintaining the same syntax for model specification. `Pyro` [Bingham et al., 2019] is one of the most popular general purpose probabilistic programming languages, which implements stochastic variational inference, and `HMC`.

We conclude this section by acknowledging that the development of automatic differentiation and hence of gradient-based methods has not only driven the development of modern Bayesian software but it has also made previously infeasible approaches computationally tractable. These tools have facilitated the development of new modelling techniques, including variational inference [Kucukelbir et al., 2017] and variational auto-encoders [Kingma and Welling, 2013], and plethora of other neural networks based approaches, such as convolutional neural networks [LeCun et al., 1989], generative adversarial networks [Goodfellow et al., 2014], or neural additive networks [Agarwal et al., 2020]. Despite their lack of interpretability, which is often crucial in epidemiological applications, these methods have been successfully applied to a wide range of problems, including diagnosis of `COVID-19` from chest X-rays [Rajpoot et al., 2024], or extensions of Cox survival models [Peroni et al., 2022]. Recent work also involves merging neural network approaches and (approximate) Bayesian inference, for example employing neural networks to simulate draws from Gaussian processes [Semenova et al., 2022], or learning amortized estimators for likelihood-to-evidence ratios to enable posterior inference in intractable likelihood settings [Hermans et al., 2019].

2.3.5 Convergence diagnostics and model fit

The main challenge in using `MCMC` methods is to ensure that the samples obtained are representative of the target distribution. In theory, the Markov Chains are guaranteed to converge to the stationary distribution. However, in practice, the convergence might be too slow or the samples might fail to reach some local modes of the target distribution. Multiple diagnostics tools exist to assess the convergence of `MCMC`:

- *Trace plots* are a simple visual tool to assess the convergence of the chain. They show the evolution of the parameters of interest (on the y -axis) over iterations (on the x -axis). A *well mixing* chain will show a random walk behaviour, with low autocorrelation between samples and with trajectories which do not indefinitely get stuck in local modes.
- *Gelman-Rubin Statistics*, or $\hat{R}$ is a diagnostic tools to assess whether different **MCMC** chains converged to the same distribution. More specifically, it compares the within-chain variance to the between chain variance for the draws of a parameter of interest [Gelman and Rubin, 1992, Vehtari et al., 2021]. Values of $\hat{R} \leq 1.01$ are generally considered satisfactory.
- The *effective sample size* quantifies the information in the samples obtained from the **MCMC** chain while accounting for the autocorrelation between samples. Intuitively, this measure can be interpreted as the number of independent samples that would provide the same amount of information as the **MCMC** draws for a given parameter. Different sample size requirements are needed to estimate different summary statistics for the posterior distribution: for instance, an accurate estimation of the posterior mean requires fewer samples than an accurate estimation of the posterior 95% credible interval.

A **MCMC** scheme may converge to a target distribution, but the target distribution may not reflect the desired properties of the model. For instance, the posterior distribution may not be able to explain the observed data, which may indicate a misspecification of the likelihood or prior. Several diagnostic tools exist to assess the quality of the model fit:

- *Leave-one-out cross-validation* is a diagnostic tool to assess the predictive performance of the model. It involves fitting the model to all but one data point, and then predicting the left-out data point [Vehtari et al., 2017]. The process is repeated for all data points, and the predictive distribution is compared to the observed data.
- *Prior predictive checks* involve generating data from the model by sampling parameters from the prior distributions. The rationale is to assess whether the prior assumptions lead to data with plausible properties. If not, the chosen priors may be inappropriate.
- *Posterior predictive checks* involve generating data from the model by sampling parameters from the posterior predictive distribution. The posterior distribution of a summary statistic is then computed and compared to the summary statistic

of the observed data. If the latter doesn't fall within the bulk of the posterior distribution, the model may be misspecified.

Convergence diagnostics and model checking are a fundamental component of the Bayesian workflow, and are crucial to ensure that the results obtained are reliable and interpretable [Gelman et al., 2020].

2.3.6 Non-parametric models

Non-parametric models are a class of models which do not assume a fixed number of parameters, but rather let the size of the data to dictate the complexity of the model. Typical example applications involve estimating the shape of a function over a continuous domain, or estimating the distribution of a random variable. In this thesis, two types of non-parametric models are used: LOESS and Gaussian Processes. For both applications, given a compact domain $D \subset \mathbb{R}^d$, the goal is to estimate a function $f : D \rightarrow \mathbb{R}$ from a set of observations $(x_i, y_i)_{i=1}^n$ such that $y_i = f(x_i) + \epsilon_i$.

Locally Estimated Scatterplot Smoothing

Locally Estimated Scatterplot Smoothing (LOESS) is a non-parametric regression method which involves fitting a polynomial surface to the data using local fitting. Local fitting means that for every point x , the fitted function $\hat{f}$ is obtained by fitting a polynomial curve to a subset of the data. Further, the points are weighted by their distance from x , through weighted least squares. The behaviour of the fit is controlled by three main parameters.

- The *span* $\alpha \in (0, 1)$, which corresponds to the proportion of points used for the fit at each point.
- The degree of the polynomial used for the fit.
- The distance metric used to weight the points.

In our applications, we used the `loess` function from the `stats` package in R [R Core Team, 2022], and controlled the behaviour of the fit through the `span` parameter. By default, the degree of the polynomial is set to 2, and the distance metric is the tricubic weighting function, given by:

$$w(d) = \left(1 - \left|\frac{d}{m}\right|^3\right)^3, \quad (2.22)$$

where m is the maximum distance between the point x and the points used for the fit. Note that once the fit is obtained, no uncertainty quantification is provided. This means

we cannot make probabilistic statements on say, how likely the true function is to be above a certain value at a given point.

Gaussian Processes

One of the most popular tools that allows to make probabilistic statements on the properties of the fitted function is *Gaussian Process (GP)* regression. A **GP** is a mathematical object which imposes a distribution over the space of functions supported in a domain D , such that for any finite set of points $x_1, \dots, x_n \in D$:

$$(f(x_1), \dots, f(x_n)) \sim \mathcal{N}(\mu(x_1, \dots, x_n), K(x_1, \dots, x_n)), \quad (2.23)$$

where μ is the *mean function* and K is the *covariance function*. To specify a **GP**, one needs to define a mean function $\mu: D \rightarrow \mathbb{R}^D$ and a kernel $k: D \times D \rightarrow \mathbb{R}^{d \times d}$, such that $(K(x_1, \dots, x_n))_{ij} = k(x_i, x_j)$. Through reparametrization, the mean function can often be assumed to be zero, and the focus lies on the choice of a kernel k from a parametric family. In the applications reported in this thesis, we used the *squared exponential* or *radial basis function* kernel, which is defined as:

$$k_{SE}(x, x') = \alpha^2 \exp\left(-\frac{(x - x')^2}{2\rho^2}\right), \quad (2.24)$$

where α^2 is the *variance* of the process and ρ is the *lengthscale*. The lengthscale determines the distance over which the GP output changes: a shorter lengthscale results in more rapid, short-range fluctuations, while a longer lengthscale yields smoother, long-range changes. The variance dictates the extent to which functions can deviate from their mean.

Fig. 2.2 illustrates the behaviour of a **GP** with a squared exponential kernel as the lengthscale and variance are varied. An interactive version of this plot can be found at https://abriz97.shinyapps.io/shiny_gp_viz/. Other popular examples include:

- The *Matérn* kernel, with parameter ν :

$$k_{Mat}(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|x - x'|}{\rho} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{|x - x'|}{\rho} \right), \quad (2.25)$$

where K_ν is the modified Bessel function of the second kind.

- The *periodic* kernel, with parameters α^2, ρ, λ :

$$k_{Per}(x, x') = \alpha^2 \exp\left(-\frac{2 \sin^2(\pi|x - x'|/\lambda)}{\rho^2}\right). \quad (2.26)$$

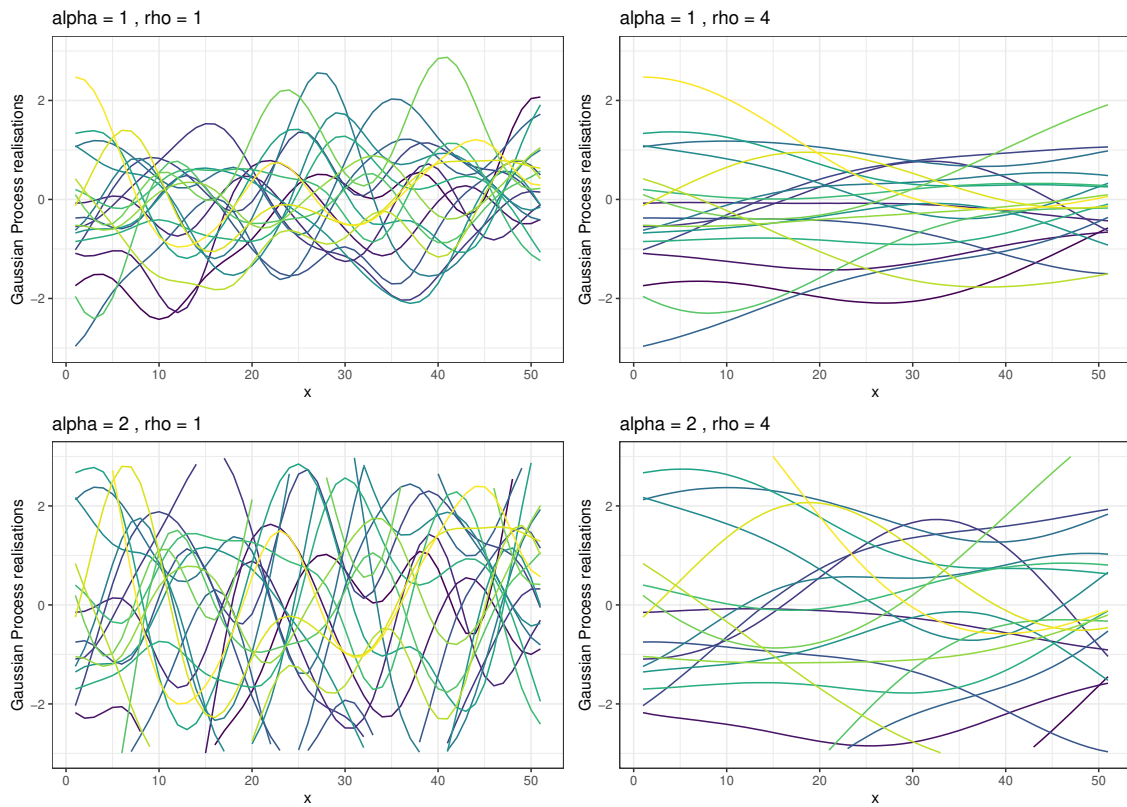


Figure 2.2: 20 draws from a Gaussian process with squared exponential kernel. The lengthscale and variance of the kernel are varied to illustrate the behaviour of the process. The lengthscale determines the distance over which the GP output changes, while the variance dictates the extent to which functions can deviate from their mean.

3

REFINING BASIC REPRODUCTION NUMBER ESTIMATES IN EMERGING EPIDEMICS

3.1 Introduction

OUTBREAKS of epidemics are generally viewed with apprehension by public health systems, as novel infectious agents can introduce considerable uncertainty and potentially severe effects on societies worldwide. A primary aim of epidemiologists during these phases is to utilize scarce data to address question on the type, scale and timing of the measures required to mitigate the epidemic [Polonsky et al., 2019]. For example, understanding the speed at which a disease is transmitted allows for calculating short-term demand of hospital resources. In this context, the basic and instantaneous reproduction numbers, R_0 and R_t , represent key epidemiological parameters (Section 2.1.4). R_t represents the average number of secondary infections caused by a single infectious individual at time t and R_0 represents the average number of secondary infections generated by a typical infection in a completely susceptible population.

In the last couple of decades, several methods have been developed to estimate R_t given case data and a generation interval distribution. For example, Wallinga and Teunis [2004] base their methodology on determining likelihoods of chains of infections, [Forsberg White and Pagano, 2008] use branching process theory, and Bettencourt and Ribeiro [2008] base their approach on SIR differential equations. In practice, one of the most popular methods is EpiEstim [Cori et al., 2013] due to its clear assumptions, an accessible software implementation, and its performance in detecting changes in transmissibility patterns [Gostic et al., 2020].

However, EpiEstim and other commonly used statistical methods, typically suffer from systematic overestimation of the reproduction number [Mercer et al., 2011], which need to be considered in the interpretation of the results. Indeed, although overestimates may be preferred to underestimates in the early-stage response, the expected decrease in bias over time [O’Driscoll et al., 2021] may be misattributed to initial interventions, hence overestimating their effect. For example, Xiao et al. [2021] attributed decreases in the instantaneous reproduction number of COVID-19 to public health interventions, even though the largest decreases were estimated to predate their implementation.

There are multiple reasons which may contribute to the overestimates. Some of these reasons are inherently hard to tackle. For example, estimates of transmissibility found in the literature are often biased toward higher values because outbreaks with subcritical transmission rates are inherently less likely to be noticed [Polonsky et al., 2019]. On the other hand, it is possible to control for some sources of bias by understanding how the reported case data departs from the underlying assumptions of the different methods. For example Thompson et al. [2019] highlight the importance of accounting for imported cases, which contribute to an increase in the force of transmission.

In this Chapter, we present the work featured in the article “Refining Reproduction Number Estimates to Account for Unobserved Generations of Infection in Emerging Epidemics” published in *Clinical Infectious Diseases*, and presented as a poster at the “8th International Conference on Infectious Disease Dynamics” in December 2021. We propose an adjustment to EpiEstim to account for missing generations of infections and use simulated data to show it reduces the bias in early reproduction estimates. The contributions of the authors are acknowledged thereafter: Ilaria Dorigatti and Megan O’Driscoll were my project supervisors and conceived the study. They and I, Andrea Brizzi, developed the methods, and wrote the paper. I implemented the methods and performed the analyses.

This Chapter is structured as follows: Section 3.2 summarises the main contributions of the paper. Section 3.3 introduces reproduction number estimators and describes the proposed adjustment. Section 3.4 summarises our results both on simulated data and on real count data from the COVID-19 pandemic. Section 3.5 assess the sensitivity of our adjustment to different assumptions on generation intervals, reporting rates, and model misspecification. Section 3.6 discusses the implications of our findings.

3.2 Contributions to the literature

In a previous simulation study, O’Driscoll et al. [2021] demonstrated that statistical methods widely adopted in outbreak analysis to quantify the reproduction number can

be biased, especially in the early stages of an outbreak. Indeed, empirically observed incidence data suffer from a variety of issues, including underreporting, and ignoring their effects can result in biased estimates which are prone to misinterpretation. Due to their simplistic assumptions, popular inference methods are especially vulnerable to these issues, which typically involve more sources of variance than are accounted for. For instance, EpiEstim assumes a Poisson distribution on the number of cases, which does not account for the typical over-dispersion observed in real data and yields overconfident estimates.

In this work, we consider the effects of under reporting in the first generations of infections, for instance due to the absence or suboptimal testing, reporting and surveillance, on the estimates of the basic (R_0) and effective (R_t) reproduction numbers obtained from widely adopted methods. Using simulated data, we show that reproduction number estimates are systematically overestimated and we propose a simple adjustment that overcomes the bias. Briefly, the adjustment involves fitting a linear exponential growth model to the first reported cases and then using this model to back-impute prior incidence. The rationale behind this approach is that it is unlikely that the first reported case corresponds to the actual first case of infection; therefore, more individuals than reported could have contributed to transmission.

We choose to apply our adjustment to EpiEstim, rather than other methods, for pragmatic reasons. Namely EpiEstim has a popular implementation through a R package and is widely used in practice. In this work, we compare the estimates produced by EpiEstim with and without our adjustment when applied to early SARS-CoV-2 incidence data reported in the USA, the UK, Italy and Sweden. Using one week worth of data, the adjusted EpiEstim method produces R_0 estimates that are consistent with results in the literature obtained from fitting transmission models to at least one month's worth of data, demonstrating that timely and reliable inference of the reproduction number is possible in the initial phases of the epidemic with our adjustment. Our analysis shows how a simple adjustment can reduce initial bias in reproduction number estimates in a newly emerging epidemic when the interpretation of reproduction number estimates is crucial for public health decision making.

The code necessary to reproduce our original analyses was made publicly available at https://github.com/abriz97/R0_simulations. Since then, I implemented the adjustment in the R package EpiEstim version v.2.2-5 alongside a vignette [Brizzi, 2024] to justify and describe its use.

3.3 Methods

In this section, we first give a brief overview of the inference methods considered in our work (Sections 2.2.3, 3.3.1 and 3.3.2), and then describe our simulation scenarios in Section 3.3.3.

3.3.1 Exponential Growth

One of the simplest methods to characterise the speed of an epidemic, as measured by the growth rate r , is to fit the log-transformed incidence data by linear regression:

$$\operatorname{argmin}_{c,r} \sum_t (\log(I_t) - (c + r t))^2. \quad (3.1)$$

When observations with 0 cases are present, linear regression requires the addition of a small constant to each incidence data point, so that the logarithm of each data point is well defined. If the generation interval distribution w is known, the reproduction number R can be expressed as the inverse of the moment generating function M_w of w evaluated at $-r$ [Wallinga and Lipsitch, 2007]:

$$R = (M_w(-r))^{-1} = \left(\sum_{s=1}^{\infty} w_s e^{rs} \right)^{-1}. \quad (3.2)$$

The assumption of exponential growth is justifiable in an initial epidemic period where the proportion of susceptible individuals in a population is large, but as more individuals get infected and the proportion of uninfected individuals falls, the growth in the number of new cases slows down and becomes sub-exponential. A rule of thumb to identify the time window in which the exponential growth assumption is valid is that the cumulative number of infected individuals does not exceed the square root of the population size. This rule was proposed by Forsberg White and Pagano [2008] and is based on the theoretical work of Ball and Donnelly [1995].

3.3.2 EpiEstim with exponential back-imputation

EpiEstim (Section 2.2.3) is one of the most popular methods to estimate the time-varying reproduction number [Cori et al., 2013], due to its ability to accurately identify R_t and detect abrupt changes in transmissibility [Gostic et al., 2020]. However, the method can suffer from systematic bias in the initial period of estimation [O’Driscoll et al., 2021], for example if the first chains of infections are not observed. To see this, consider the mean of the posterior distribution Eq. (2.10) when $\tau = 1$ and priors are

uninformative (such that $a_{\text{prior}}, b_{\text{prior}}$ tend to 0):

$$\hat{R}_{t,\tau=1} = \frac{I_t}{\Lambda_t} = \frac{I_t}{\sum_{s=0}^{\infty} w_s I_{t-s}}. \quad (3.3)$$

Hence the posterior mean corresponds the ratio between the number of observed infections on day t over the population infectiousness on the same day. When the first chains of infections are not observed, the estimator will tend to attribute all new cases to the first observed cases, overestimating the reproduction number.

In this work, we consider the scenario in which earlier cases would have been reported if the testing capacity had been available. As such, we frame the statistical problem as a missing data issue, where we observe incidence $\underline{I}_T^{\text{obs}} = (I_s)_{0 \leq s \leq T}$ from day 0 up to T , but we are missing the incidence $\underline{I}^{\text{mis}} = (I_s)_{s < 0}$ from the start of the epidemic to the first observed case. Note we are currently not providing a start date of the epidemic; our proposed method will not require this information. The problem lies in quantifying:

$$p\left((R_{t,\tau})_t | \underline{I}_T^{\text{obs}}\right) = \int p\left((R_{t,\tau})_t | \underline{I}_T^{\text{obs}}, \underline{I}^{\text{mis}}\right) p\left(\underline{I}^{\text{mis}} | \underline{I}_T^{\text{obs}}\right) d\underline{I}^{\text{mis}}. \quad (3.4)$$

To perform a fully Bayesian inference of $(R_{t,\tau})$ coherent with EpiEstim's methodology (Section 2.2.3), we should specify a model for $p(\underline{I}^{\text{mis}} | \underline{I}_T^{\text{obs}})$, and successively propagate the uncertainty on the unobserved cases to the posterior distribution of $(R_{t,\tau})$. For instance, we could have imposed a Bayesian linear model on the logged cases parametrized by an intercept, a growth rate and a standard deviation. Then, simple Monte Carlo methods could be used to simulate the unobserved cases, and the posterior of $(R_{t,\tau})$ could be obtained by averaging the posterior of $(R_{t,\tau})$ over the simulated unobserved cases. However, this would result in the loss of a closed form expression for the posterior of $(R_{t,\tau})$, which is one of the main practical advantages of EpiEstim over other packages such as Epidemia (Section 2.2.4, Scott et al. [2020]).

Instead of specifying a probabilistic model for the unobserved cases, we propose a simple back-imputation procedure to account for unobserved initial infections. This is similar in spirit to other extensions of EpiEstim. In Nash et al. [2023], the authors firstly impute daily cases from reported weekly cases through a frequentist expectation-maximisation procedure, and secondly use the imputed cases as an input of the Bayesian EpiEstim algorithm.

We adopt a similar two-step approach, where imputation of unobserved initial infections is based on an exponential growth assumption, similar to Dorigatti et al. [2020]. Specifically, we assume that the epidemic is growing exponentially to back-impute infections in the period prior to the first observation. Note that this assumption is theoretically justified in the early stages of an epidemic [Ma, 2020], and exponential growth inference

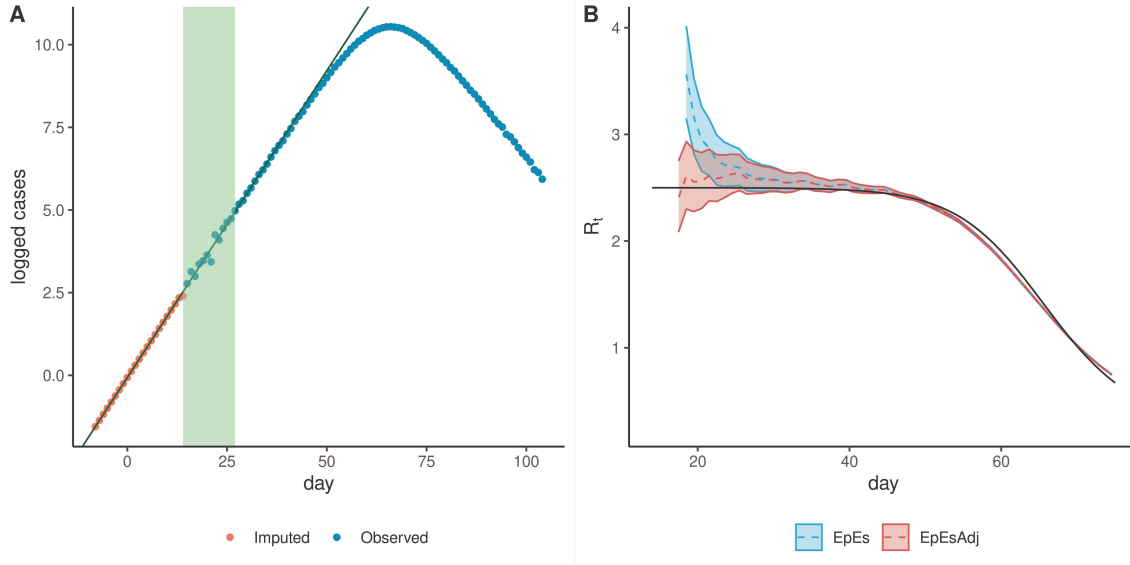


Figure 3.1: Visualisation of the exponential growth adjustment method, applied to simulated data obtained with $R_0 = 2.5$, reporting rate $\rho = 1$ and resulting changes in posterior estimates having assumed 2 unobserved generations. On panel A, the logarithms of the first reported data (those in the green region) are used to fit a linear model (line). The linear model is then used to back-impute unobserved cases (red dots) to complement the available data (blue dots). On panel B, the true R_t value (black solid line) is compared to EpiEstim estimates without (blue) and with (red) adjustment using sliding windows of 7 days. The back-imputation reduces the initial mean estimates (dotted lines) and 95% credible interval widths (ribbons). The adjusted method then converges to the original method as the importance of the imputed datapoints vanishes. Method abbreviations: EpiEstim (EpEs); Adjusted EpiEstim (EpEsAdj). This figure is included in the accepted manuscript provided to *Clinical Infectious Diseases* for further editing.

methods are between the most accurate in this initial period [O’Driscoll et al., 2021]. Our procedure can be summarised in three simple steps and is visualised in Fig. 3.1. First, we fit a linear model on the log-transformed incidence data to estimate the growth rate, as per Eq. (3.1) and as shown in Fig. 3.1A. Second, we use this linear model to back impute incidence data, prior to the time of the first observed case. In particular, we obtain an estimate of the number of cases for S days, where S is the largest possible generation interval length, i.e., the largest value such that $w_s > 0$. We highlight that it is not necessary to round the output of the inferred number of cases, and estimates lower than 1 should not be removed. Finally, we apply EpiEstim to the extended epidemic curve (including the back imputed incidence data).

3.3.3 Simulations

To evaluate the validity of our correction, we compared the performance of EpiEstim with and without the adjustment on simulated data. We additionally compared the results to those obtained by fitting a linear exponential growth model, as the performance of our correction strongly depends on the accuracy of the estimates of the growth rate.

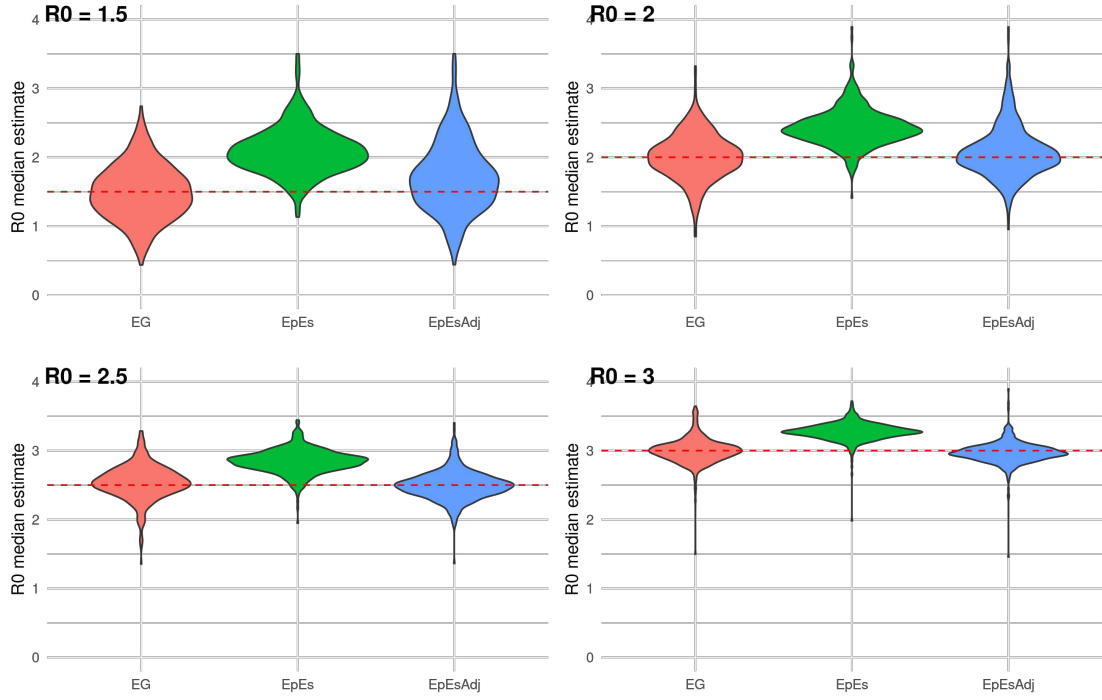


Figure 3.2: Distribution of mean R_0 estimates assuming a fixed reporting rate $\rho=50\%$. Distribution of mean R_0 estimates assuming a fixed reporting rate $\rho=50\%$. Each panel shows the distribution of the mean R_0 estimates obtained using 100 simulations for a given true R_0 value (red dashed line). Method abbreviations: Linear exponential growth rate method (EG); EpiEstim (EpEs); Adjusted EpiEstim (EpEsAdj). This figure is included in the accepted manuscript provided to *Clinical Infectious Diseases* for further editing.

We used a stochastic SEIR simulator to generate 100 epidemic curves for each R_0 values in $\{1.5, 2, 2.5, 3\}$. We considered a large population of 10^6 individuals, and epidemics seeded by 5 initial infections. We assumed 3 days and 3.5 days as the mean times spent in the exposed and infectious compartments respectively, yielding a 6.5 mean generation interval commonly used to model COVID-19 [O’Driscoll et al., 2021]. To account for imperfect reporting, we simulated two types of issues commonly affecting the observed data: underreporting and unobserved initial generations of infections. We simulate unobserved initial generations by considering the simulated data from day 15 to day 28, corresponding to two generation intervals worth of data after having missed the first two generations. Further, we simulated under-reporting by considering a constant reporting rate $\rho \in \{0.15, 0.3, 0.5, 1\}$. Daily observations were sampled as binomial realisations of true incidence with success probability equal to the reporting rate ρ .



Figure 3.3: Bias and coverage comparison by reporting rate. Median (point) and 95% CrI (interval range) of the basic reproduction number (R_0) mean estimates (panel A), mean coverage (panel B) and mean 95% confidence/credible intervals widths (panel C) for different values of R_0 and reporting rates ρ . Each panel represents distinct values of R_0 , while different colours represent different reporting rates. Method abbreviations: Linear exponential growth rate method (EG); EpiEstim (EpEs); Adjusted EpiEstim (EpEsAdj). This figure is included in the accepted manuscript provided to *Clinical Infectious Diseases* for further editing.

3.4 Findings

3.4.1 Method comparison

We fitted the exponential growth, EpiEstim and adjusted EpiEstim methods to the simulated data assuming a Gamma distributed generation interval matching the mean and variance of the generating process (i.e., 6.5 mean generation interval). Fig. 3.2 shows the effect of our adjustment when the reporting rate is 50% and shows that the initial bias observed in the estimates obtained with EpiEstim is strongly reduced by the proposed adjustment, and the mean estimates are comparable with those produced by the exponential growth method. These trends are observed for each value of ρ , implying consistency in the estimates obtained with different reporting rates (Fig. 3.2 and Fig. 3.7). Our adjustment does not only improve the accuracy of the central estimates of R_0 , but also improves uncertainty quantification. Average credible/confidence interval widths and coverage are reported in Fig. 3.3. Coverage remains constant across

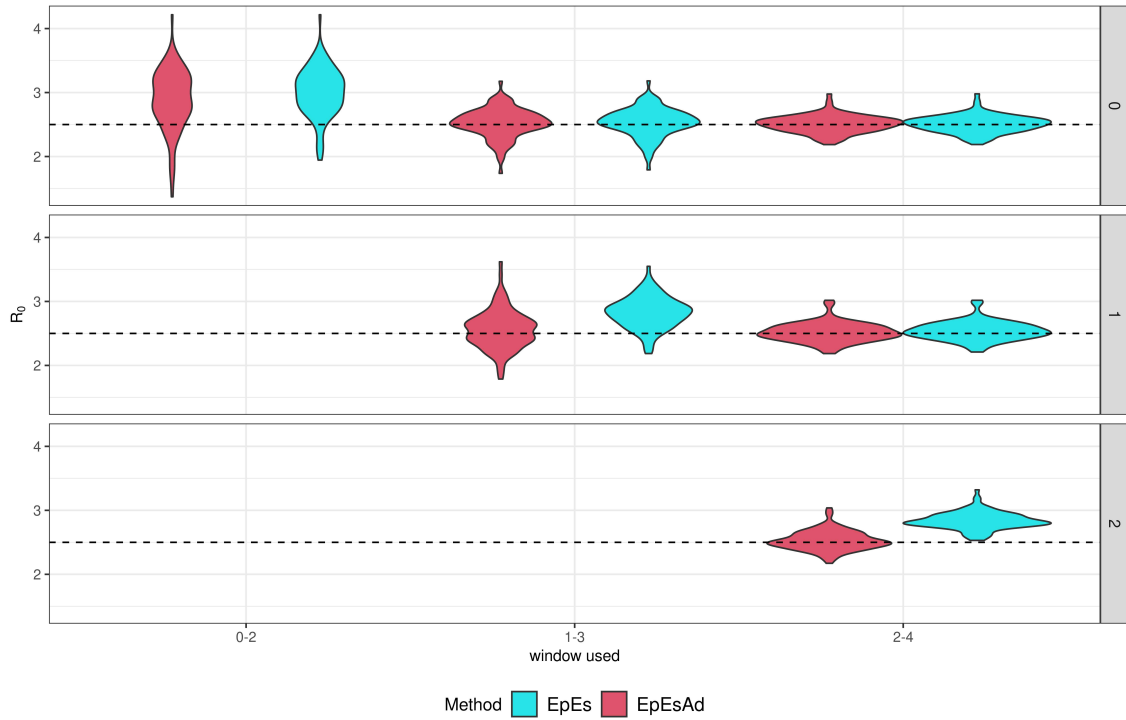


Figure 3.4: Impact of missed generations on R_0 estimates. Distribution of mean R_0 estimates, obtained using biweekly time windows (x-axis) and a variable number of unobserved generations (0, 1 and 2, see rows), assuming $R_0 = 2.5$ and reporting rate $\rho = 1$. EpiEstim estimates are shown in blue, adjusted EpiEstim estimates are shown in red. On the axis, the sliding window used to fit the method is shown. Each window contains 14 consecutive data points, starting at week 0, or week 1, or week 2. Method abbreviations: EpiEstim (EpEs); Adjusted EpiEstim (EpEsAdj). This figure is included in the accepted manuscript provided to *Clinical Infectious Diseases* for further editing.

different values of the reporting rate for the estimates obtained with the exponential growth and adjusted EpiEstim methods. On the other hand, the large bias observed in the estimates obtained with EpiEstim implies that as credible intervals get narrower, coverage decreases dramatically (Fig. 3.3, panels B and C). The EpiEstim adjustment proposed in this paper produces generally narrower credible intervals as compared to the exponential growth method, at the price of a slight decrease in coverage. This trade-off is in favour of our adjustment for values of R_0 larger than 2, when coverage amongst the two methods is similar. Further, we highlight that the proposed adjustment does not influence later estimates of R_t produced by EpiEstim (Fig. 3.1).

3.4.2 Impact of missed generations

Beyond simulating undetected cases, reflecting a surveillance system which may be unprepared or unaware of a newly unfolding epidemic, we investigated the impact of the number of unobserved generations on the estimates obtained with our adjustment using simulated data generated under a scenario with perfect reporting rate ($\rho = 1$)

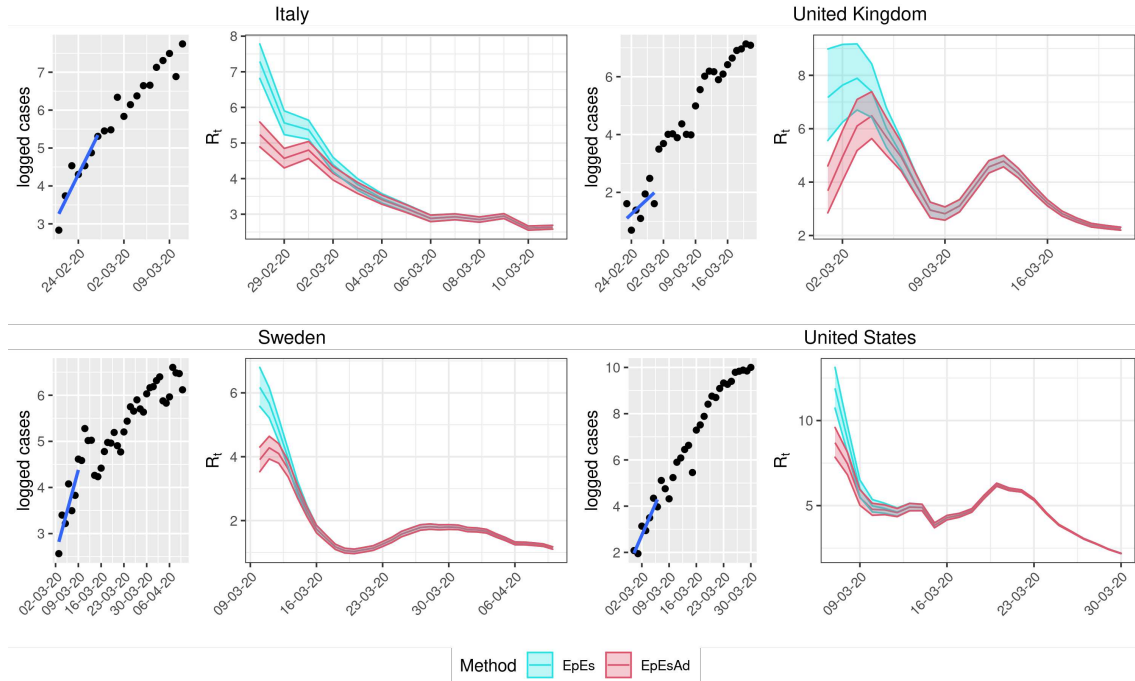


Figure 3.5: Comparison of R_t estimates obtained with (red) and without (blue) adjustment for COVID-19 data in 4 countries: Italy, France, the UK, and the USA. Each quadrant includes a subfigure showing the logarithm of the data and the regression line (left) and the R_t estimates obtained using a sliding window of 7 days (and the data up to that day) and a generation interval of mean 5.7 days and standard deviation of 1.72 days (right). Method abbreviations: EpiEstim (EpEs); Adjusted EpiEstim (EpEsAdj). This figure is included in the accepted manuscript provided to *Clinical Infectious Diseases* for further editing.

and $R_0 = 2.5$. From each epidemic trajectory, we considered 3 different left truncations of the data to account for 0, 1 and 2 unobserved generations. We then applied EpiEstim with and without the proposed adjustment using biweekly time windows starting on weeks 0, 1 and 2. Fig. 3.4 shows the distribution of the mean estimates for 100 simulations. Each row identifies the number of unobserved generations, while on the x-axis we show the time window used to obtain the estimate. Fig. 3.4 also shows that the adjustment is particularly useful when larger numbers of initial generations are unobserved. While the proposed adjustment produces comparable estimates to those obtained with EpiEstim when all generations are observed (top row of Fig. 3.4), its dependence on the exponential growth method introduces a layer of stochasticity that increases the variance of the mean estimates. On the other hand, when 1 or 2 generations are unobserved, the proposed adjustment adequately compensates for EpiEstim's bias and the median estimates become closer to the true R_0 value used to simulate the data.

3.4.3 Application to reported COVID-19 data

In addition to validating the proposed adjustment on simulated data, we applied it to real-world COVID-19 incidence case data reported in the John Hopkins Center for Systems Science and Engineering database (Dong et al. [2020], Center for Systems Science and Engineering (CSSE) [2020]). During the initial phases of the outbreak, surveillance systems of several countries were unprepared to detect infectious individuals and it is likely that the first generations of infections were not observed, justifying our back-imputation adjustment. We selected Italy, Sweden, UK and the US as case studies. For each country, we fitted the log-transformed incidence case count reported for the first sequential 7 days of sustained transmission (no days with 0 new cases in the selected time window) with a linear regression model. We then inferred the number of unobserved cases before the selected time window and applied EpiEstim with and without adjustment to the observed and imputed data, using weekly sliding windows and assuming a generation interval with mean 5.7 days and standard deviation of 1.72 days [Ganyani et al., 2020]. Fig. 3.5 shows that the adjustment lowers the estimates of R_0 in every scenario, suggesting a role for unobserved generations of infections in overestimating the early R_0 estimates of SARS-CoV-2, which in turn also affect the early estimates of R_t . We obtained average R_0 estimates of 3.6 with 95% credible interval (2.8,4.6) in the UK, 5.2 (95% CrI 4.9,5.6) for Italy, 8.7 (95% CrI 7.8,9.6) for the US and 3.9 (95% CrI 3.5, 4.3) in Sweden.

The estimates obtained for the UK show a curious trend, with mean R_t estimates increasing from 3.6 to 6.2 in less than a week time. There are two reasons underlying the trend. Firstly, it could be argued that the growth rate used for back-imputation was underestimated, as shown by the blue line in the left panel of Fig. 3.5. Secondly, a jump can be observed in the number of cases from the start of March. The John Hopkins database reported 33 cases on March first, while the previous three daily reports were 7, 12 and 15. The exact reason for this jump is uncertain, but might be related to increased testing capacity. Official reports on the number of daily tests starts from the 20th of March, but it is likely that a scale up in testing capacity started earlier Department of Health and Social Care [2020].

3.5 Sensitivity analyses

3.5.1 Sensitivity analysis on the generation interval distribution

Estimates of the reproduction number are strongly dependent on the assumption of the underlying generation interval [Wallinga and Lipsitch, 2007]. In practice, especially

during an outbreak, it is impossible to know with certainty the generation interval and therefore it is important to consider different values. Here, we conducted a sensitivity analyses to assess the impact of the assumed length of the mean generation interval on the reproduction number estimates obtained with EpiEstim and our proposed adjustment. We used the **COVID-19** data described in Section 3.4.3 and considered three different possible mean distributions with the same coefficient of variation set to 2, shown in Fig. 3.6, reflecting the heterogeneity in the serial intervals assumptions in the literature [Rai et al., 2021, Flaxman et al., 2020, Ke et al., 2021]. As expected, longer generation intervals produce larger reproduction number estimates. We also observe larger discrepancies between the estimates obtained with the original and adjusted EpiEstim methods as the backimputation procedure plays a larger role when the generation interval is longer, and therefore previous generations of infections are more likely to contribute to transmissions.

3.5.2 Sensitivity analysis on the assumed reporting rate

To determine the impact of reporting rates on inference methods, we here report the results of the simulation analyses described in Section 3.3.3 for different assumed reporting rates $\rho \in \{0.15, 0.30, 1.00\}$. Fig. 3.7 shows that the variance of the R_0 estimators decreases as the reporting rate increases, and that the proposed adjustment consistently reduces the bias in the estimates obtained with EpiEstim.

3.5.3 Sensitivity analysis on exponential growth assumption

To assess the sensitivity of our adjustment to the exponential growth scenario, we further compared the performance of EpiEstim, EpiEstim adjusted and the Exponential Growth method on simulated data featuring sub-exponential growth dynamics. Following Chowell et al. [2016], we simulated an SEIR model with modified transition rates from the Susceptible to the Exposed compartment (Eq. (2.4)) reproducing sub-exponential growth:

$$\frac{dS(t)}{dt} = -\beta S(t)I(t)^\alpha \frac{1}{N}, \quad (3.5a)$$

$$\frac{dE(t)}{dt} = \beta S(t)I(t)^\alpha \frac{1}{N} - \gamma I(t), \quad (3.5b)$$

Where $\alpha \in [0, 1]$ is a constant which quantifies the departure from the exponential growth scenario. Similarly to the main text scenario, we considered rates $\gamma = \frac{1}{3}, \sigma = \frac{1}{3.5}$. We further fixed $I(0) = 5, \beta = 3\gamma$ and $N = 10^6$.

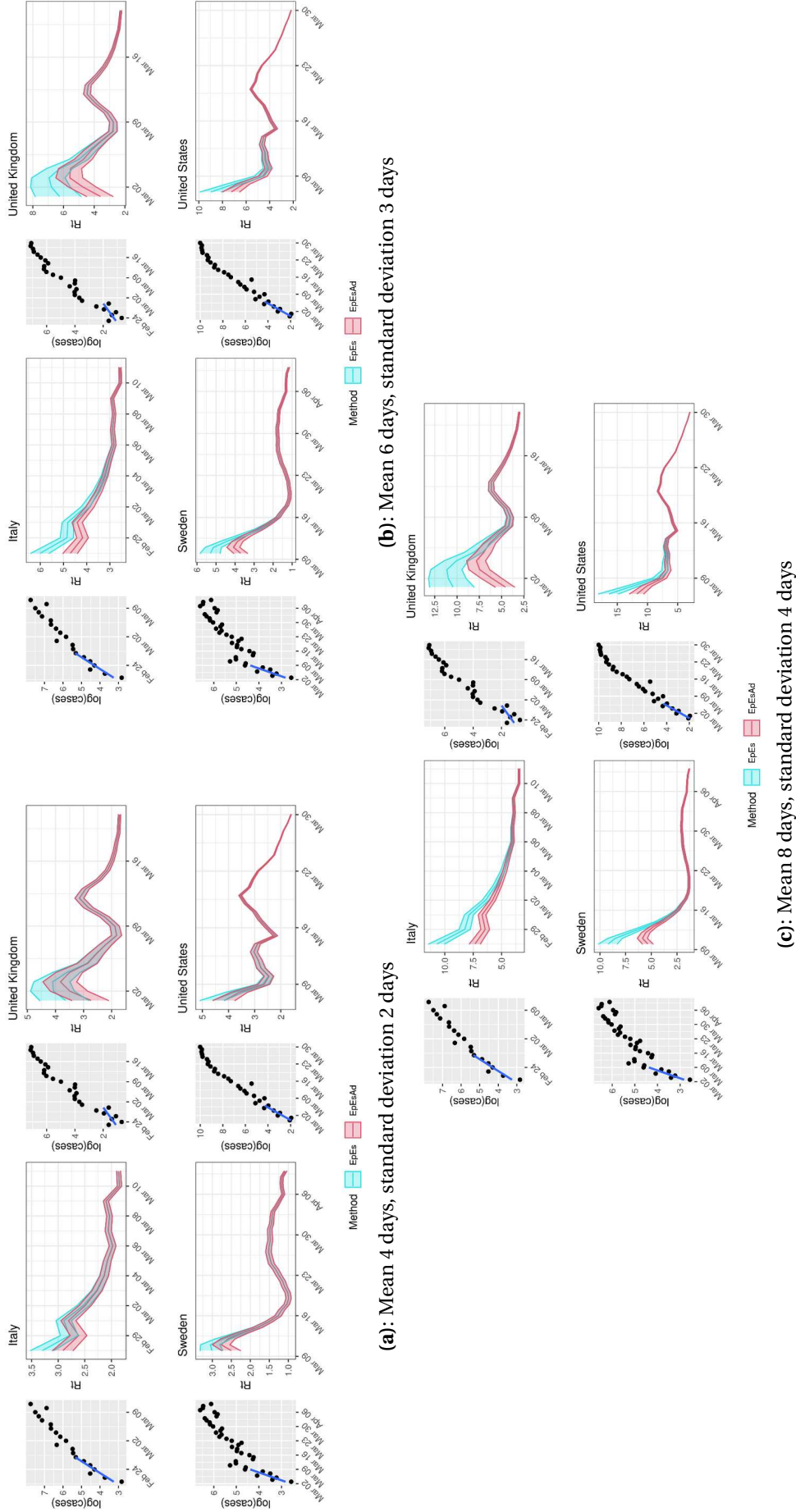


Figure 3.6: Sensitivity analysis on generation interval. Each quadrant includes a subfigure showing the logarithm of the data and the regression line (left) and the R_t estimates obtained using a sliding window of 7 days (and the data up to that day) and different generation interval specifications. Method abbreviations: EpiEstim (EpEs); Adjusted EpiEstim (EpEsAdj). This figure is included in the supplementary materials submitted to *Clinical Infectious Diseases*.

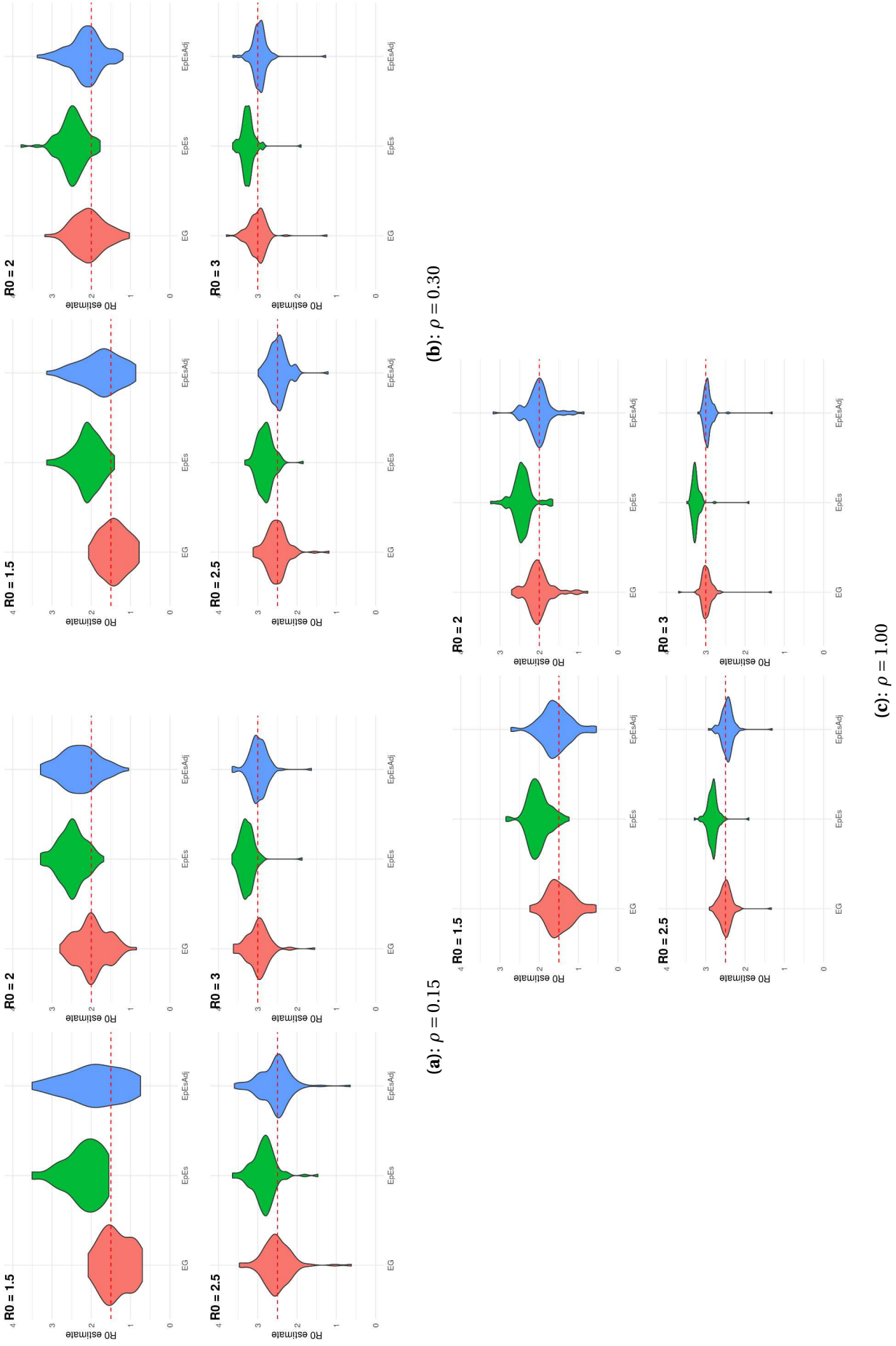


Figure 3.7: Sensitivity analysis on reporting rates. Distribution of mean R_0 estimates assuming fixed reporting rates of $\rho \in \{0.15, 0.30, 1\}$. Each panel shows the distribution of the mean R_0 estimates obtained using 100 simulations for a given true R_0 value (red dashed line). Method abbreviations: Linear exponential growth rate method (EG); EpiEstim (EpiEs); Adjusted EpiEstim (EpiEstimAdj). This figure is included in the supplementary materials submitted to *Clinical Infectious Diseases*.

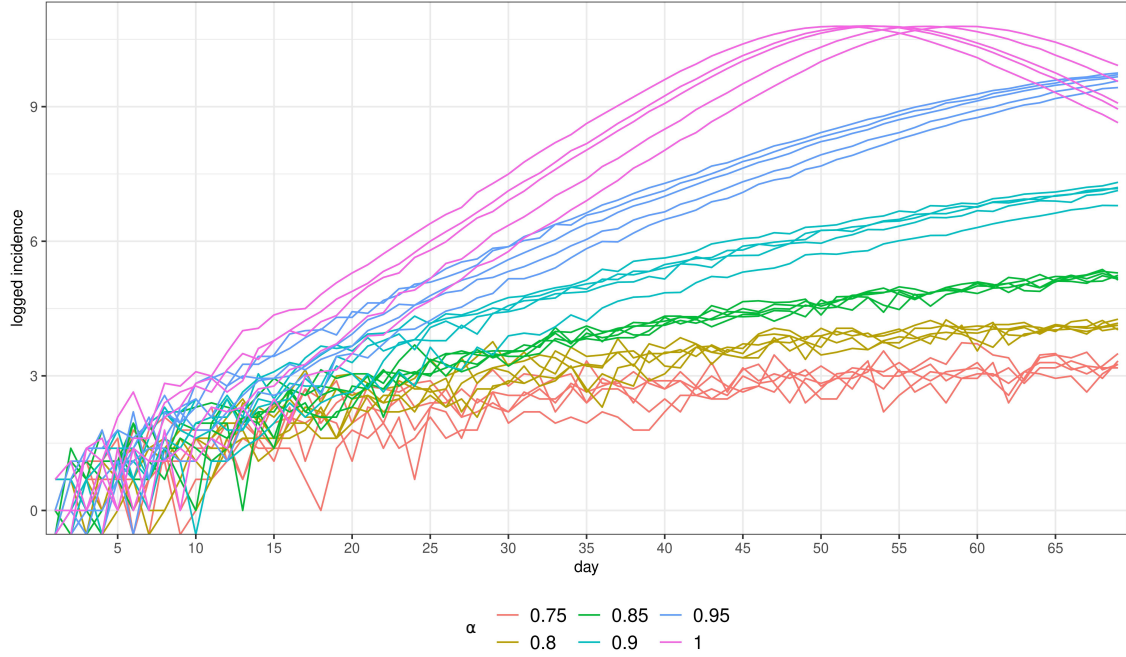


Figure 3.8: Incidence curves simulated via a stochastic SIR model with nonlinear incidence rates, $N = 10^6$, $I(0) = 5$, $\gamma = \frac{1}{3}$, $\sigma = \frac{1}{3.5}$ and $\beta = 3\gamma$. For each value of $\alpha \in \{0.75, 0.8, 0.85, 0.9, 0.95, 1\}$, we show 5 randomly selected incidence curves. Days are indexed on the x-axis while the logarithm of the daily cases is reported on the y-axis. This figure is included in the supplementary materials submitted to *Clinical Infectious Diseases*.

Note that Eq. (3.5) can be rewritten as:

$$\frac{dS(t)}{dt} = -\beta' S(t) \frac{I(t)}{N}, \quad (3.6)$$

where β' corresponds to $\beta I(t)^{(\alpha-1)}$. For this reason, we can define the instantaneous reproduction number as:

$$R_t = \frac{S(t)}{N} \frac{\beta'}{\gamma} = \frac{S(t)}{N} \frac{\beta}{\gamma} I(t)^\alpha. \quad (3.7)$$

In contrast to Eq. (2.3), here the value of the basic reproduction number R_0 , is now a function of the initial number of infections $I(0)$. We run 200 simulations for each $\alpha \in \{0.75, 0.8, 0.85, 0.9, 0.95, 1\}$. Fig. 3.8 shows 5 randomly selected incidence curves for each value of α . In order to simulate missing generations of infections, we truncated the first 14 days of each incidence curve and run the methods to the successive 2 weeks of data. We then treated the first obtained estimate as R_0 estimate.

Fig. 3.9 compares the evolution of the R_t estimates obtained with EpiEstim and our adjustment on a randomly selected incidence curve simulated with $\alpha = 0.7$. As in the previous scenario, the proposed imputation of initial infections leads to smaller initial estimates, which then converge to the ones obtained with EpiEstim. Fig. 3.10

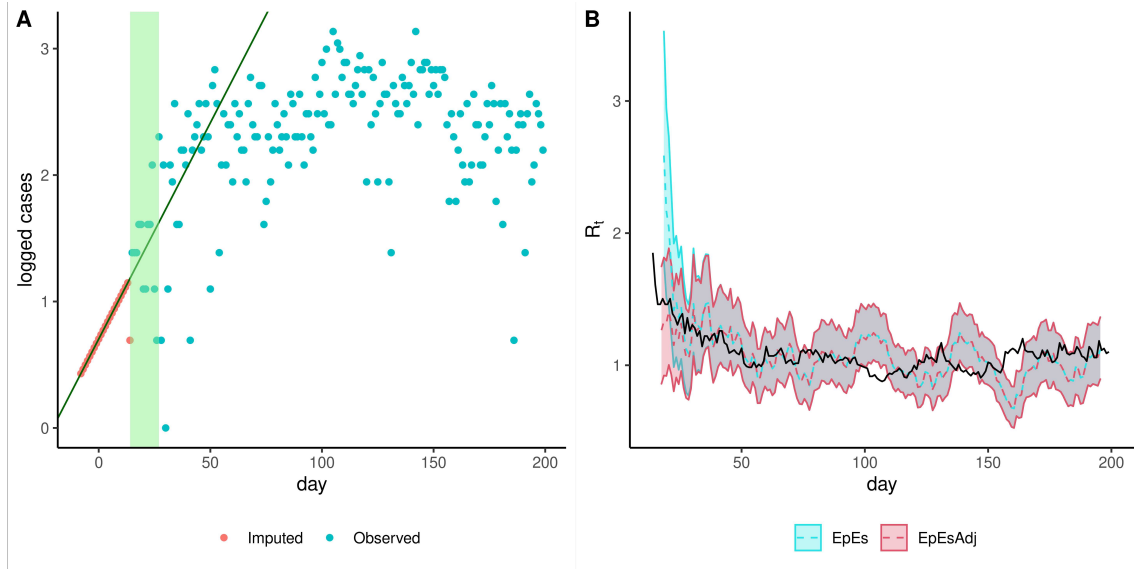


Figure 3.9: Visualisation of the adjustment method applied to simulated data following non-exponential growth. On panel A, the logarithms of the first reported data (those in the green region) are used to fit a linear model (line). The linear model is then used to back-impute unobserved cases (red dots) to complement the available data (blue dots). On panel B, the true R_t values (black solid line) are compared to EpiEstim estimates without (blue) and with (red) adjustment using sliding windows of 7 days. The back-imputation reduces the initial mean estimates (dotted lines) and 95% credible interval widths (ribbons). The adjusted method then converges to the original method as the importance of the imputed datapoints vanishes. Method abbreviations: EpiEstim (EpEs); Adjusted EpiEstim (EpEsAdj). This figure is included in the supplementary materials submitted to *Clinical Infectious Diseases*.

compares the distribution of the mean R_0 estimates obtained across the different α values and demonstrates that all methods based on the exponential growth assumption systematically underestimate the basic reproduction number, but the deviations get smaller as α approaches 1. This behaviour is to be expected: the concavity of the log of the incidence curves shown in Fig. 3.8 suggests that the growth rate inferred in weeks 3 and 4 will be lower than the one that would be estimated in the first 2 weeks. For smaller α values, we observe an apparent improved precision of the estimates obtained with EpiEstim compared with our adjustment (Fig. 3.10), which is due to a trade-off between the underestimation described above with the over-estimation trend described in the main text. With larger values of α , the under-estimation in R_t plays a lesser role, yielding larger estimates of R_0 values obtained with EpiEstim and a better performance of our adjustment.

3.6 Discussion

In this Chapter, we introduced and proposed a correction for the systematic overestimation of R_0 that occurs in the early stages of an epidemic when using EpiEstim and other

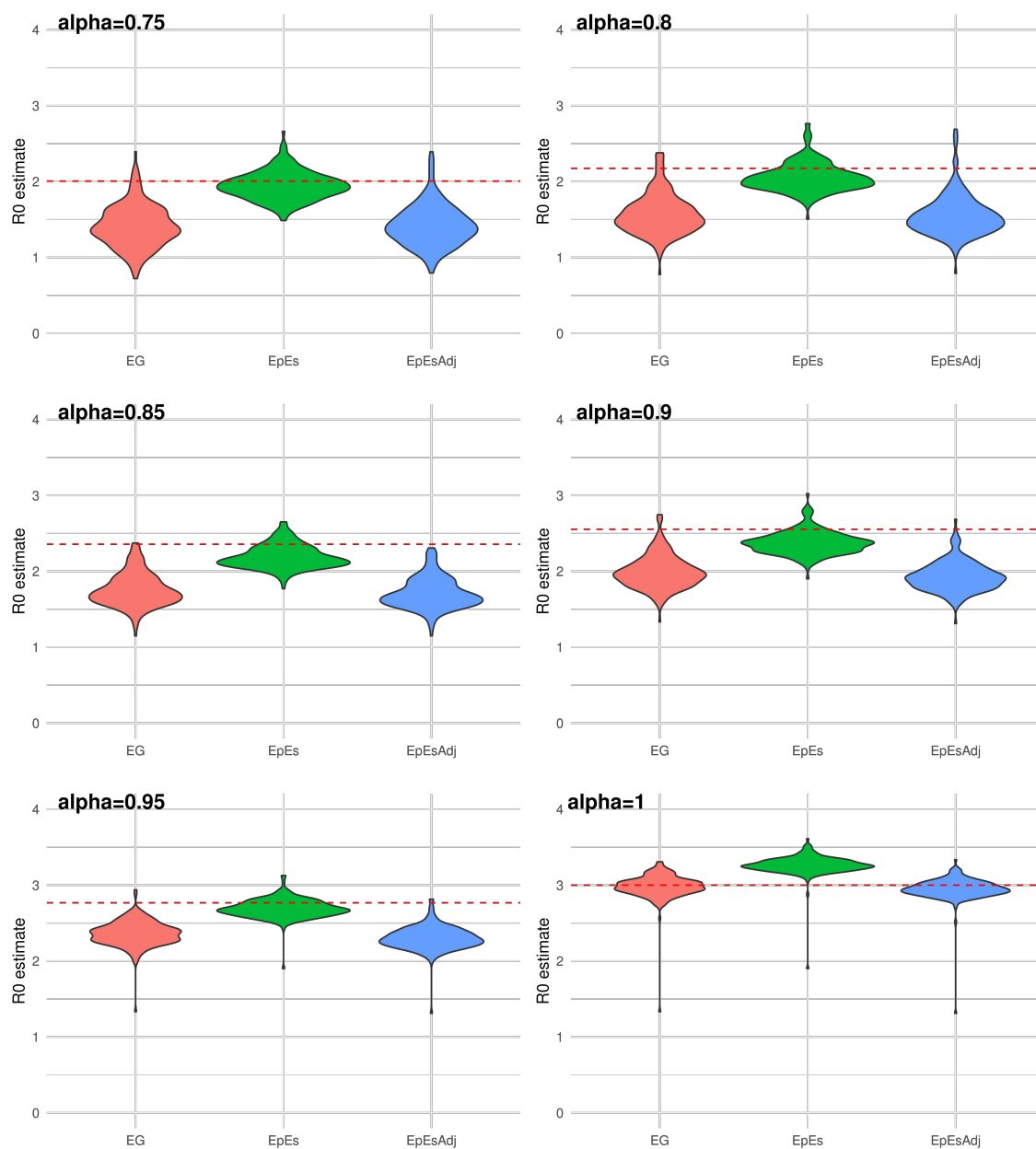


Figure 3.10: Distribution of mean R_0 estimates obtained on simulated data following sub-exponential growth. Each panel shows the distribution of the mean R_0 estimates obtained using 200 simulations for a given true R_0 value (red dashed line). Method abbreviations: Linear exponential growth rate method (EG); EpiEstim (EpEs); Adjusted EpiEstim (EpEsAdj). This figure is included in the supplementary materials submitted to *Clinical Infectious Diseases*.

common inferential methods, utilising a back-imputation procedure. The proposed correction aims to account for unobserved generations of infections. For this reason, we deem it to be most applicable in scenarios where generations of infections may have been missed due to emergency situations and limited testing, and generation intervals are relatively short. In practice, the adjustment will prove especially useful to evaluate transmissibility of diseases that are either asymptomatic or cause mild symptoms for a large proportion of infected individuals, as diagnosing and confirming cases for such diseases may be challenging. Similarly, the adjustment may prove applicable to diseases characterised by long time lags between infection to symptom onset, especially if infectiousness develops significantly earlier than symptoms.

In this work, we demonstrate the application of this adjustment to the EpiEstim method, though it can be applied to other statistical methods where this bias may occur. The resulting adjusted EpiEstim method combines the best features of EpiEstim and the exponential growth method, which is more adequate for the early phases of an epidemic. In particular, the long-run R_t estimates of the original EpiEstim method are preserved, while the initial bias observed in the estimates is reduced by the proposed adjustment. This was confirmed in all scenarios of our simulation experiments, independently of the reproduction number and the reporting rate values used to simulate the data. The adjusted estimates of R_0 strongly outperformed the estimates obtained with the original EpiEstim method in terms of bias and coverage, and produced tighter 95% credible intervals. Our results show that the R_0 estimates obtained with the adjusted EpiEstim and linear exponential growth method are very similar. This is likely due, in part, to the fact that the proposed adjustment relies on linear regression on the logged incidence, which is used for the back imputation of the unobserved generations of infections.

The proposed method currently does not incorporate uncertainty in the growth rate and in the number of imputed cases. A Bayesian approach to imputation would require the specification of a model $p(\underline{I}^{\text{mis}} | \underline{I}^{\text{obs}})$ in Eq. (3.4). Such an approach would reduce the overconfidence in the estimates attributable to treating the imputed cases as known. However, it would most likely result in a loss of a closed-form expression. Alternative Bayesian approaches might include fitting stochastic transmission models as implemented through the R packages Epidemia [Scott et al., 2020] and mcstate [FitzJohn et al., 2024]. These methods allow for a more flexible specification of the transmission model, the observation model and observation times: the user can specify that the epidemic started an arbitrary amount of time prior to the first observation, and overdispersion can be accounted for, for instance by specifying a negative binomial distribution for the number of cases. However, inference based on these models can be sensitive to their initialisations. Namely, the user needs to specify the starting date of

the epidemic as well as the number of seeding cases. These are unknown in practice and can lead to different estimates of the reproduction number.

The effect of our adjustment was evident when working with early COVID-19 data from Italy, the UK, Sweden and the US. The adjusted estimates were significantly lower than the estimates obtained with the original EpiEstim method, and were found to be largely consistent with estimates derived in Ke et al. [2021]. This is especially encouraging considering that our adjusted R_0 estimates were obtained by making use of 7 days' worth of data, while those of Ke et al. [2021] were obtained on both case and death time series data spanning months. This suggests that our method can yield improved precision with limited information, which may prove valuable in emerging epidemics. While several papers and meta-analyses reported estimates of the exponential growth rate for the 4 countries [Billah et al., 2020, Alimohamadi et al., 2020, Yu et al., 2021], the parametrisation of the generation interval used and the reproduction number estimates were often lacking [Spouge, 2021] thus hindering further comparisons. Further, the adjustment alone was not enough to explain the estimated decrease in R_t in the study period considered (see Fig. 3.5), possibly suggesting that changes in individual behaviour and governmental interventions lowered transmissibility of the disease. However, other biases may be playing a role, such as changes in reporting rates, overestimation due to importations [Mercer et al., 2011], or model misspecification.

Critically, the results are sensitive to the choice of EpiEstim parameters, such as the generation interval and the length of the time window used.

Concerning the choice of the sliding window, the larger time windows produce smoother estimates, which reduces the impact of the imputed cases. This means that our adjustment has a much smaller effect when considering sliding windows covering two generation intervals worth of data or more.

4

ESTIMATING FLUCTUATING HOSPITAL FATALITY RATES IN LMIC SETTINGS

4.1 Introduction

THE emergence of novel variants often represents an important shift from the previously understood transmission dynamics of the pandemic, introducing challenges to local health systems, which need new information to update their strategies. For example, although greater infectiousness is almost guaranteed to lead to more cases, it is a priori unknown whether the associated severity of a novel variant is greater than the previous. A timely comparison of variant specific disease severities is therefore crucial to inform public health responses, including the potential necessity for non-pharmaceutical interventions such as lockdowns.

However, comparing disease severities is more complicated than comparing transmissibility. In **HIC**, high testing rates and availability of linked digitalised records allows for the tracking of individual histories from reported infection to hospitalisation, with successive dismissal or death. Further, sequencing capacities allow for the identification of the variant causing the infection, both in surveillance and in clinical settings. This infrastructure allows for observational study designs that can straightforwardly compare variant severities, as exemplified in a multitude of studies in European countries. For example, a cohort study performed between January and March 2021 in Denmark estimates that Alpha is associated with 42% increase in risk of hospitalisation [Bager et al., 2021]. The adjusted Risk Ratios were estimated through Poisson regression with robust standard errors. Funk et al. [2021] used data from the TESSY platform [European Centre for Disease Prevention and Control (ECDC), 2017] from September 2020 to

March 2021 and found higher hospitalisation rates in cases infected with Alpha, Beta, and Gamma relative to wildtype strains in seven European countries. Two parallel analyses were made to estimate odds ratios accounting for confounders: one using stratified logistic regression, and the other using conditional logistic regression on matched cases-controls. A retrospective cohort analysis by [Nyberg et al. \[2023\]](#) used surveillance data from November 2020 to January 2021 to estimate hazard ratios of hospital admissions for Alpha against wildtype variant in England using a stratified Cox proportional hazard model. They found a 50% larger risk of hospitalisation for Alpha cases relative to wildtype. Later on, [Sheikh et al. \[2021\]](#) reported that, between April to June 2021, in Scotland, the risk of hospitalisation was estimated to be double for individuals infected with Delta as compared to Alpha. These results were obtained fitting a Cox proportional hazards model. [Twohig et al. \[2022\]](#) followed a cohort of confirmed COVID-19 cases in England from March to May 2021, and estimated the risk ratio of hospital attendance outcomes through stratified Cox proportional hazard regression, finding a double risk of admission in the first 14 days after a positive test for delta relative to alpha. The same statistical methods were used by [Nyberg et al. \[2022\]](#), who considered a retrospective cohort using data from the UKHSA [\[Clare et al., 2021\]](#) and found a lower risk of hospitalisation within 14 days of confirmed infection for Omicron cases relative to Delta. Even when population testing rates lowered in spring 2022, European protocols for cooperation were established to aggregate country-level results on relative severity of Omicron and Delta through a meta-analysis [\[Nyberg et al., 2023\]](#).

The key to these studies is the availability of linked individual-level data: analysts know how much time passed between a case report to a hospitalisation, the infecting variants, and sometimes her vaccination status. In LMIC, this availability is poor. The infecting variant for an hospitalised patient is rarely known, comorbidities are often not recorded, and health records may not be digitalised. As such, comparing relative severities of variants in LMIC is challenging. [Perez-Guzman et al. \[2024\]](#) use a compartmental model fit on excess mortality data from Kabwe, Zambia. They are able to estimate time-varying infection fatality rates, but are unable to attribute these to specific variants, due to the absence of variant frequency data. On the other hand, the same type of model applied to UK and fitted to multiple data sources was able to rank variants by their basic infection fatality ratio [\[Perez-Guzman et al., 2023\]](#). As such, the lack of data in LMIC explains why most studies in these settings limit themselves to comparing variant transmissibility [\[Tegally et al., 2021, Viana et al., 2022, Dhar et al., 2021\]](#).

In this Chapter we present the work featured in the article “Spatial and temporal fluctuations in COVID-19 fatality rates in Brazilian hospitals”, published in *Nature Medicine* [\[Brizzi et al., 2022\]](#). We describe changes in hospital fatality rates in 14 Brazilian cities following the emergence of Gamma, and assessed whether shocks in fatality

rates were attributable to the novel variant, to resource inequities or shocks in hospital demands. The contributions of the authors of the papers is acknowledged thereafter. Oliver Ratmann, my supervisor and senior author of the paper, conceived the analysis. He, Ester C. Sabino, Nuno R. Faria, and Samir Bhatt oversaw the study. I, Andrea Brizzi, am the first author of the paper. Oliver Ratmann and I developed the Bayesian multi-strain model (Section 4.3.3). I, Luciana M. S. Servo, Iwona Hawryluk, Carlos A. Prete Jr, William M. de Souza, Raphael Sonabend, Leonardo J. T. Araujo, Philippe Lemey, Paula L. C. Fonseca, Victor E. V. Geddes, Diego M. Bonfim, Raphael Sonabend oversaw and performed data collection from primary data sources (Section 4.3.1). Oliver Ratmann, Samir Bhatt conceived the counterfactual analyses (Section 4.3.4) and sensitivity analyses (Section 4.5), which were coded up and run by me and Charles Whittaker. I coded all the Figures and Tables that are not attributed to another co-author. Nuno R. Faria edited Fig. 4.6, Oliver Ratmann coded a first version of Figs. 4.3, 4.4 and 4.8.

This Chapter is structured as follows. Section 4.2 summarises the main contributions of the paper. Section 4.3 describes the collated datasets and the statistical models used for the analysis. Section 4.4 presents the main findings of the paper. Section 4.5 discusses the model validation and sensitivity analyses and Section 4.6 and includes a discussion of the limitations and the implications of our findings.

4.2 Contributions to the literature

In the study presented in this Chapter, we describe extensive geographic and temporal variation in COVID-19 in-hospital fatality rates since the beginning of the epidemic in 14 Brazilian state capitals. We report that the first shocks in fatality rates predated the emergence of the Gamma variant, and were primarily driven by shortages in health-care capacity which, in turn, emerged from a combination of pre-pandemic regional inequities and increased healthcare pressure brought about by epidemic waves of SARS-CoV-2 transmission.

To perform the analysis, we processed and collated data from a variety of publicly available sources, which we made available, together with the analyses code on our Github repository [CADDE-CENTRE, 2022]. For example, we obtained individual-level patient histories following hospitalisation with suspected or confirmed COVID-19 from STV [(2020,1)], we obtained census data from the 2020 National Household Sample Survey COVID-19 [PNA, 2020], and we obtained healthcare-facility-level microdata on personnel and equipment from the National Register of Health Facilities [cne, 2021]. The efforts enabled us to define the pandemic healthcare pressure indices that measure and monitor mismatches between healthcare demand and resource availability in hospital settings. We found that these pandemic healthcare pressure indices are strongly

correlated with variations in in-hospital **COVID-19** fatality rates across 14 Brazilian state capitals.

We developed a Bayesian multi-strain fatality model to disentangle the effects of location-specific inequities [Castro et al., 2019, de Oliveira Andrade, 2020, de Souza et al., 2020, Ranzani et al., 2021a, Li et al., 2021], pandemic healthcare pressures and Gamma specific disease severity on fluctuating **COVID-19** in-hospital fatality rates [de Oliveira Andrade, 2020, de Albuquerque and Ribeiro, 2021, Rocha et al., 2018], while accounting for the substantial cumulated loss of life [Castro et al., 2021a], and age-prioritised vaccine roll-out [Martins et al., 2021, De Oliveira et al., 2021, Freitas et al., 2021]. Briefly, the model reconstructs the replacement dynamics of Gamma in each of the 14 state capitals from weekly **SARS-CoV-2** sequence metadata, and estimates the weekly number of patients that are hospitalised respectively with Gamma and non-Gamma infection in 11 age strata. We modelled age-specific in-hospital fatality rates for each variant through a regression equation that captures non-parametric location effects, fixed non-negative effects associated with the healthcare pressure indices, and non-parametric virus variant effects associated with replacement dynamics of Gamma. In the model, the non-parametric location effects do not vary in time and account for a variety of constant social, economic or healthcare related factors that differentiate one city from another in the weeks when their observed fatality rates were lowest, such as the prevalence of comorbidities that might modify disease severity or pre-existing inequities in both the quality and the quantity of available healthcare resources [de Oliveira Andrade, 2020, Clark et al., 2020]. By contrast, the healthcare pressure indices vary in time and the strength of their association with the fatality rates is estimated through the regression equation. In the model, patients estimated to have been infected with the Gamma variant are considered separately from those estimated to have been infected with non-Gamma variants, which allows us to control for Gamma-specific expansion and in-hospital disease severity when inferring associations between fatality rates and healthcare pressure. This allowed us to overcome the missing individual-level data on missing variants that is available in similar studies from high-income countries [Bager et al., 2021, Funk et al., 2021, Nyberg et al., 2023, Sheikh et al., 2021, Twohig et al., 2022, Nyberg et al., 2022].

We found that the impact of Gamma in Brazil's hospitals has predominantly been indirect and mediated through pre-existing geographic inequities, transient infection waves and concomitant shocks in healthcare demand. To further quantify the impact of these healthcare shortages, we conducted counterfactual analyses projecting that approximately half of Brazil's **COVID-19** deaths in hospitals could have been avoided without pre-pandemic geographic inequities and without the multitude of pandemic healthcare pressures.

Our findings are particularly important in calibrating the risk posed by new highly transmissible virus or by SARS-CoV-2 variants of concern. The conclusions are highly relevant for other countries and show that responding to acute health care shocks is challenging. Without sustained investments in healthcare resources, healthcare optimization, and pandemic preparedness to minimize population wide mortality and morbidity, countries will continue to experience high fatality rates caused by immune-escape SARS-CoV-2 variants or other highly transmissible pathogens.

4.3 Methods

This section presents the data and two models that we considered to analyse geographical and temporal trends in fatality rates in Brazilian hospitals. Section 4.3.1 provides a detailed description of the different sources of data used. Section 4.3.2 presents the model used to describe changes in fatality rates in a way that allows for comparison across different locations after adjusting for differences in the age distribution of hospitalized patients. Section 4.3.3 presents the explanatory model to quantify the contributions of different factors on the observed fluctuations in fatality rates.

4.3.1 Data

Data to characterize hospital mortality

We obtained publicly available individual records of patients with severe acute respiratory infection in public or private hospitals across Brazil that were reported in the SIVEP-Gripe (Sistema de Informação de Vigilância Epidemiológica da Gripe) database, release 31 January 2022 (SIV [(2020)], SIV [(2021)]).

Records were filtered excluding out-of-hospital deaths defined as individuals with a missing hospital admission date or who died on the day of admission, excluding confirmed infections with other respiratory pathogens or with missing diagnosis, excluding non-residents to match city-level population denominators, and excluding patients with at least one vaccine dose prior to hospital admission to ensure trends in fatality rates were not confounded with vaccine status. To avoid bias [de Souza et al., 2020], our data thus included patients with PCR-confirmed, clinically diagnosed, or suspected COVID-19 infection. Alternative inclusion criteria were considered in sensitivity analyses (Section 4.5). The data are available at [inst/data/SIVEP_hospital_31-01-2022-all.rds](#) in the Github repository.

Study location and observation period

We searched [GISAID \[2021\]](#) on June 14, 2021, for sequence data associated with **SARS-CoV-2** Gamma virus genome sequences in the 27 federal units (26 Brazilian states and the Federal District). Gamma genome data was available for 16 federal units, Amapá, Amazonas, Bahia, Goiás, Maranhão, Mato Grosso do Sul, Minas Gerais, Paraíba, Paraná, Rio de Janeiro, Rio Grande do Norte, Rio Grande do Sul, Rondônia, Santa Catarina, São Paulo, and Tocantins. State capitals of these federal units were the focus of our analyses, as key variables such as population denominators, vaccination coverage and healthcare indicators were more directly available at the level of state capitals. Two state capitals, Palmas and Campo Grande, were excluded from further analysis due to limited, weekly age-specific COVID-19 hospital admission and death counts among residents reported to the SIVEP-Gripe database. The cities in our sampling frame were thus Belo Horizonte, Curitiba, Florianópolis, Goiânia, João Pessoa, Macapá, Manaus, Natal, Porto Alegre, Porto Velho, Rio de Janeiro, Salvador, São Luís and São Paulo. Table [A.1](#) reports the observation periods for each capital, reflecting the unsynchronized geographical expansion of COVID-19 through Brazil. To establish a common definition for the start of the epidemic that accounted for the widely varying population sizes across cities, we defined the start date as the Monday following the first date on which at least 2.5 patients per 100,000 population were hospitalized with suspected or confirmed COVID-19, and the end date to 26 July 2021. In our preliminary analyses, lower thresholds resulted in relatively sparse hospitalization data in some cities.

Population denominators data

Population size estimates in each city were retrieved by sex and 1-year age bands from the 2020 National Household Sample Survey COVID-19, Pesquisa Nacional por Amostra de Domicílios COVID-19 ([PNA \[2020\]](#)). The projections were reconciled with available vaccination records of residents in each city, so that at most 99% of residents received at least one vaccine dose during the observation period. We denote the adjusted population estimates in age group a and location l as n_a^l . The data are available at [inst/data/PNADc_vaxadj_population_210802.csv](#) in the Github repository.

Sequence data to characterize Gamma expansion

Viral genome sequence collected between November 1, 2020 to March 31, 2021 across Brazil was downloaded from GISAID on June 28, 2021 [[Shu and McCauley, 2017](#)]. Records with incomplete collection dates were removed, de-duplicated, and classified as Gamma or non-Gamma variants with `pangolin` v.3.0.6, `Pangolearn` v.1.2.12, *scorpio* lineage [[cov-lineages, 2020](#)]. In total, we retained sequences and meta-data

from 7,221 samples, most of which were from São Paulo (1104 viral genomes), with an average of 158 genomes per state. Gamma and non-Gamma frequencies were calculated for each week and location, assuming that the frequency of Gamma in state capitals is similar to the state-level frequency of Gamma in the GISAID data. Throughout, we denote the week index in which Gamma was first detected in each location by $W_l^{\text{detect}-\Gamma}$, the number of sequenced genotypes in location l and collection week w by $s_{l,w}$, and the number sequenced genotypes attributed to the Gamma variant with pangolin by $s_{l,w}^\Gamma$. This data was also used to estimate the date of Gamma's emergence in different cities through phylogenetic analyses carried out by Nuno R. Faria. For technical details, refer to the methods section of the paper.

Healthcare resources data

To further investigate the shocks in in-hospital fatality rates before and after Gamma's emergence, we obtained monthly data on healthcare resources reported by healthcare facilities to the National Register of Health Facilities (Cadastro Nacional de Estabelecimentos de Saúde, CNES) [cne, 2021], release September 15, 2021. Data on healthcare resources are mandatory to report by both public and private healthcare facilities, and comprised personnel and equipment. Data aggregation and processing is detailed in the Methods section of the paper [Brizzi et al., 2022]. Briefly, we obtained the number of physicians, of ICU beds, and ventilators by location and month. The data are available at [inst/data/IPEA_ICUbeds_physicians_210928.csv](#) in the Github repository.

From these data, we calculated healthcare pressure indices to quantify the mismatch between healthcare demand and available resources. Demand was defined in terms of all hospital admissions for severe acute respiratory infections (i. e. including non-residents, cases caused by other pathogens, and vaccinated individuals) and all ICU admissions among patients hospitalised for severe acute respiratory infection. Again, records were filtered to exclude out-of-hospital deaths as defined above. The healthcare pressure indices were calculated by considering demand per available healthcare resource over time. To account for the fact that admitted patients remain in care for several weeks, we considered rolling sums of the form

$$x_{l,w}^{\text{3-wk ICU adm per ICU bed}} = \left(\sum_{i=0}^3 h_{l,w+i}^{\text{ICU}} \right) / r_{l,w}^{\text{ICU beds}}, \quad (4.1)$$

where $h_{l,w}^{\text{ICU}}$ are the number of ICU admissions and $r_{l,w}^{\text{ICU beds}}$ are the number of ICU beds in location l and week w . Data on resources were available per month. In weeks overlapping months, we used weighted averages of the resources in the corresponding months to define the weekly resources, and otherwise we used the monthly values. Table A.2 provides definitions for all healthcare pressure indices considered.

Vaccine administrations data

The individual-level data on vaccine administrations was downloaded from Brazilian Ministry of Health database ([Brazilian Ministry of Health \[2021\]](#)) on 18 June 2021. This was used to adjust the population denominators mentioned above, and to compute the time-varying population at risk of hospitalisation $n_{l,a,w}^R$ in location l , age group a , and week w . In particular, we combined the data with estimates of vaccine effectiveness to estimate the proportion of individuals that were protected from hospitalisation.

4.3.2 Smoothed age-standardised fatality rates model

To characterise longitudinal trends to in-hospital fatality rates, we considered the age strata

$$\mathcal{A} = \left\{ 0 - 15, 16 - 29, 30 - 39, 40 - 49, 50 - 59, 60 - 69, \right. \\ \left. 70 - 74, 75 - 79, 80 - 84, 85 - 89, 90 + \right\}, \quad (4.2)$$

to control for age dependence in fatality rates. To capture temporal shocks in fatality rates across Brazilian hospitals, all data were aggregated to weeks. In COVID-19 attributable patients with unreported outcomes, deaths were predicted independently in each location, week, and age strata based on the fatality rates in COVID-19 attributable patients with observed outcomes in the two previous weeks, assuming data were missing at random over the three weeks $w - 2$, $w - 1$, w in each age group and each location. For every location l , age strata a and week w , we denote by $h_{l,a,w}^{\text{res}}$ the number of COVID-19 attributable hospital admissions in unvaccinated residents, and we denote by $h_{l,a,w}^{\text{res-adj-D}}$ the number of deaths in these admissions, adjusted for under-reporting. Empirical fatality rates $z_{l,a,w}$ were defined as the proportion of these hospital admissions that resulted in death, $h_{l,a,w}^{\text{res-adj-D}} / h_{l,a,w}^{\text{res}}$ except for weeks with no such hospital admissions. Longitudinal trends $\hat{z}_{l,a,w}$ were obtained by fitting the non-parametric loess smoother as implemented in the R `stats` package version 4.0.3 with argument `span=0.3` to the empirical fatality rates $z_{l,a,w}$, separately for each location and age band, and starting from the first week with nonzero hospital admissions in each age band. To compare fatality rates across cities in a simple statistic while accounting for the substantial differences in age demographics, we calculated age-standardised, weekly in-hospital fatality rates with

$$\hat{z}_{l,w} = \sum_a \frac{n_a^{\text{cities}}}{\sum_b n_b^{\text{cities}}} \hat{z}_{l,a,w}, \quad (4.3)$$

where $\hat{z}_{l,a,w}$ are the smoothed fatality rates in location, age band and week, and n_a^{cities} are our 2020 projected population sizes in age band a across the 14 cities (see above).

To avoid extrapolation, the smoothed rates Eq. (4.3) were generated onward from the first week for which the $\hat{z}_{l,a,w}$ were defined for all age groups.

Fig. 4.1 shows the smoothed, age-standardised in-hospital fatality rates over time for Goiânia, Natal, Rio de Janeiro, and São Paulo. Associations between the healthcare pressure indices such as Eq. (4.1) and smoothed, age-standardised in-hospital fatality rates were quantified with Pearson correlation coefficients.

4.3.3 Bayesian multi-strain model

We developed a Bayesian multi-strain fatality model to disentangle the effects of pre-pandemic geographic inequities, pandemic healthcare pressures and Gamma’s in-hospital disease severity on fluctuating COVID-19 in-hospital fatality rates. Overall, the model is structured in three components. In the first model component, a logistic function is fitted to the Gamma variant frequency data to estimate the proportion of Gamma infections in location l and week w , denoted by $\alpha_{l,w}$. Here, we use the phylogenetically derived emergence times of the Gamma variant in each city, and specify that $\alpha_{l,w}$ is essentially zero for all weeks prior to $W_l^{\text{emerge}-\Gamma}$. In the second model component, the proportion of hospitalised patients with Gamma and non-Gamma infections in the weekly hospital admissions are estimated by city, age, and week of admission. The model-based attribution of hospital admissions to Gamma and non-Gamma variants is based on the estimated expansion of the Gamma variant in each city through the $\alpha_{l,w}$. In the third model component, we describe the fatality rates in the Gamma-attributable hospital admissions and non-Gamma-attributable hospital admissions through two corresponding regression equations. The regression equations comprise non-parametric location effects, fixed effects associated with the healthcare pressure indices, and non-parametric virus variant effects of Gamma-specific disease severity in hospitals.

Note that the first two components of the model would be unnecessary were the infecting variants known at the patients level. Such information would allow to calculate risk ratios from standard models, treating the variant as a predictor the same way as age or ethnicity in a logistic regression framework [Adjei et al., 2022].

Variant replacement dynamics component

The first model component describes the temporal expansion of Gamma in hospital admissions. We use a logistic function to reflect the early exponential growth in the

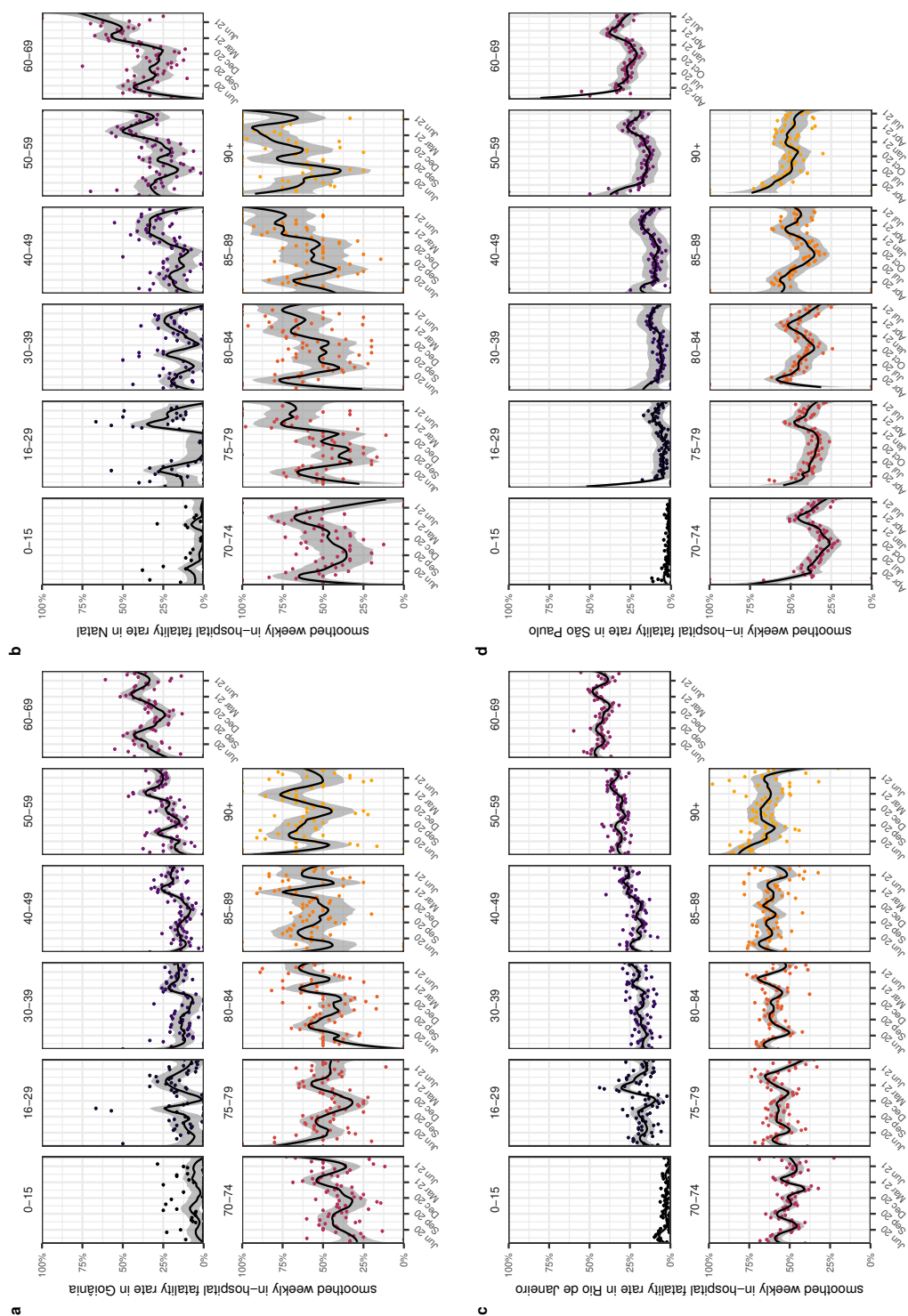


Figure 4.1: Time trends in age-specific COVID-19 in-hospital fatality rates. Weekly, age-specific COVID-19 in-hospital fatality rates are shown as dots, and non-parametric loess estimates of time trends are shown as black solid line along with 95% confidence intervals as grey ribbon. The date of Gamma's first detection is indicated with vertical black lines. Data are shown for Goiânia (a), Natal (b), Rio de Janeiro (c), and São Paulo (d). This figure is included in the accepted manuscript provided to *Nature Medicine* for further editing.

proportion of Gamma sequences which then saturates to a maximum value. We model:

$$s_{l,w}^{\Gamma} \sim \text{Beta-Binomial}(s_{l,w}, \alpha_{l,w}/\theta_1, (1 - \alpha_{l,w})/\theta_1) \quad (4.4a)$$

$$\alpha_{l,w} = \frac{1}{1 + \exp\left(-\alpha_l^{\text{growth}}(w - \alpha_l^{\text{mid}})\right)} \quad (4.4b)$$

$$\alpha_l^{\text{growth}} \sim \mathcal{N}(0, 0.2^2) \quad (4.4c)$$

$$\alpha_l^{\text{mid}} \sim \mathcal{N}(\alpha_l^{\text{mid-mean}}, 3^2) \quad (4.4d)$$

$$\alpha_{l,1:(W_l^{\text{emerge-}\Gamma}-1)} \sim \text{Normal-ccdf}(0, 0.0025^2) \quad (4.4e)$$

$$\theta_1 \sim \text{Exponential}(20), \quad (4.4f)$$

where the Beta-Binomial is specified in terms of the shape-shape parametrisation with mean $s_{l,w}\alpha_{l,w}$ and variance equal to the Binomial variance component $s_{l,w}\alpha_{l,w}(1 - \alpha_{l,w})$ multiplied by $(1 + (s_{l,w} - 1)\frac{1}{\theta_1 + 1})$ to allow for overdispersion in the variant frequency data. Time runs in units of weeks from the start of the first wave until the end of the observation period in each location. The prior mean for the midpoint of the logistic function, $\alpha_l^{\text{mid-mean}}$, was set to the week in which the ratio $s_{l,w}^{\Gamma}/s_{l,w}$ was closest to 0.5 in location l . We force variant frequencies of Gamma to be close to zero before the week of Gamma's emergence in each location, which is implemented through the informative prior Eq. (4.4e), where Normal-ccdf denotes the survival function of a normal density. Intuitively, we put 95% prior probability on the proportion of Gamma sequences being less than 0.5% before the week of Gamma's emergence. The prior in Eq. (4.4f) peaks at zero and thus favours the least complex model with no overdispersion. Fig. 4.2 shows the fit of the logistic function to the Gamma variant frequency data in each location.

Age composition of hospital admissions component

In the second model component, we couple Eq. (4.4b) to decompose the COVID-19 attributable hospital admissions $h_{l,a,w}^{\text{res}}$ by admissions with Gamma and non-Gamma variant. We account for demographic changes in the population at risk of severe infection, either through higher cumulative mortality in older ages, or through prioritised vaccination of older ages, and denote the population that remains at risk of severe infection in location l , age group a and week w by $n_{l,a,w}^R$. Then, we assumed characteristic per-capita rates $\lambda_{l,a}^{\Gamma}$, $\lambda_{l,a}^{\text{non-}\Gamma}$ of severe infection and hospitalisation with Gamma and non-Gamma variants respectively, and considered modelling the expected values of the

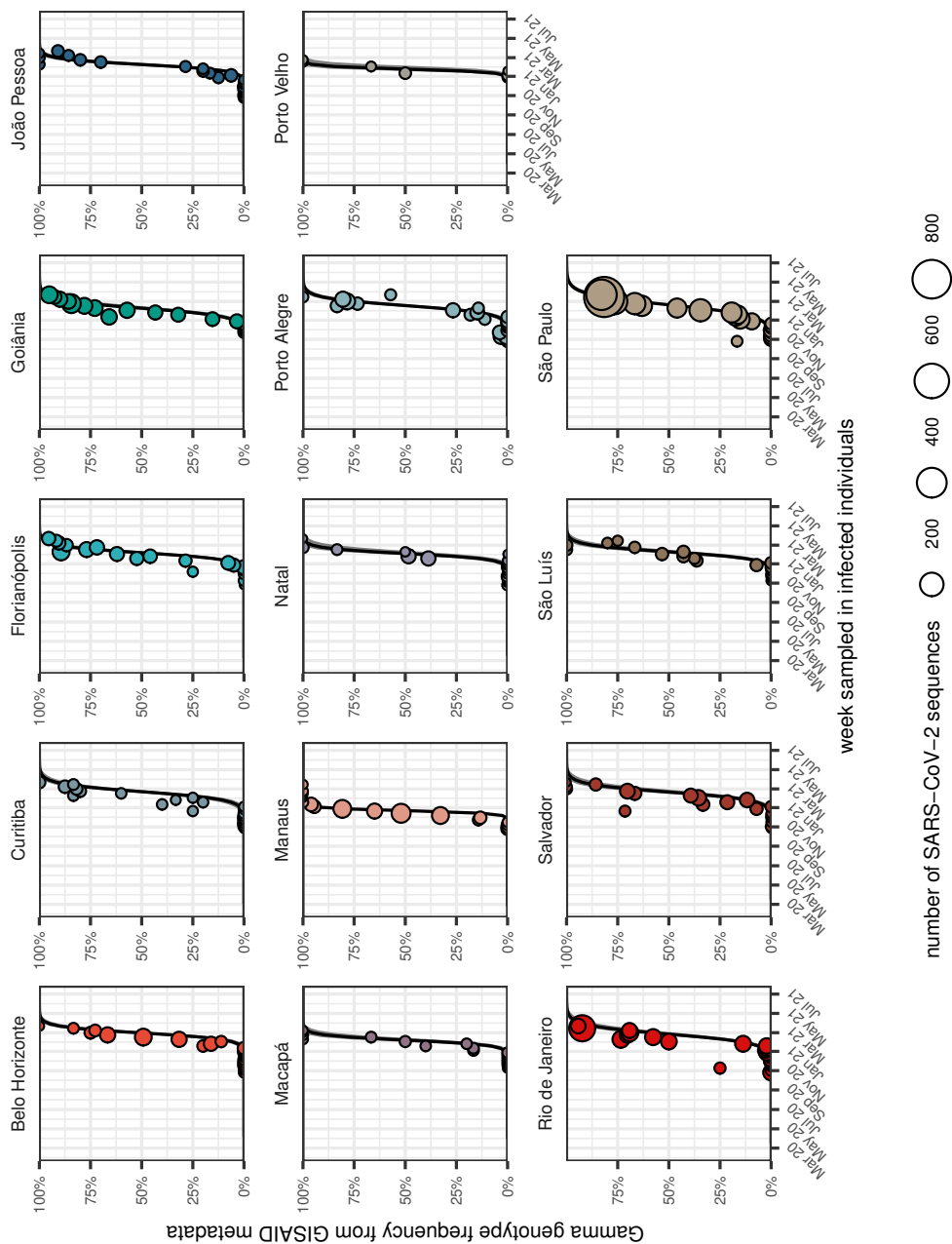


Figure 4.2: Spatiotemporal expansion of the SARS-CoV-2 Gamma variant across Brazil. SARS-CoV-2 genome sequences were obtained from the GISALD repository [Shu and McCauley, 2017] along with confirmed lineage assignments. The frequency of the Gamma variant (dots) in weekly glsars-cov-2 genome sequence counts (size of dots) is shown along with posterior median estimates of Gamma's variant frequencies (black line) under the Bayesian multi-strain fatality model and 95% credible intervals (grey ribbon). This figure is included in the accepted manuscript provided to *Nature Medicine* for further editing.

hospital admissions $h_{l,a,w}^{\text{res}}$ with

$$\mathbb{E}(h_{l,a,w}^{\text{res}}) = (\alpha_{l,w}\pi_{l,a,w}^{\Gamma} + (1 - \alpha_{l,w})\pi_{l,a,w}^{\text{non-}\Gamma}) h_{l,w}^{\text{res}}, \quad (4.5a)$$

$$\pi_{l,a,w}^{\Gamma} = \lambda_{l,a}^{\Gamma} n_{l,a,w}^R / \sum_a \lambda_{l,a}^{\Gamma} n_{l,a,w}^R \quad (4.5b)$$

$$\pi_{l,a,w}^{\text{non-}\Gamma} = \lambda_{l,a}^{\text{non-}\Gamma} n_{l,a,w}^R / \sum_a \lambda_{l,a}^{\text{non-}\Gamma} n_{l,a,w}^R, \quad (4.5c)$$

where $h_{l,w}^{\text{res}}$ is the sum of hospital admissions across age strata, and $\pi_{l,a}^{\text{non-}\Gamma}$ and $\pi_{l,a}^{\Gamma}$ are the time-invariant characteristic age composition of hospital admissions in location l for Gamma and non-Gamma variants respectively.

We specified $n_{l,a,w}^R$ as follows. First, to account for cumulative mortality, we calculated excess deaths based on the all-cause deaths reported by Brazil's Civil Registry [National Civil Registry, 2021], and subtracted from the population size projections the larger of cumulated excess deaths or all reported COVID-19 deaths in the SIVEP-Gripe database. Second, to account for increasing vaccine protection, we calculated vaccination coverage by vaccine administered in each location, age group and week, based on vaccine administration records from the Brazilian Ministry of Health [Brazilian Ministry of Health, 2022]. We then calculated the population at risk of severe infection, $n_{l,a,w}^R$, by subtracting further a proportion of the individuals vaccinated with 1 or 2 doses for at least 2 weeks according to the vaccine-specific efficacy values [Ranzani et al., 2021b, Voysey et al., 2021, Madhi et al., 2021, Clemens et al., 2021, Khoury et al., 2021] listed in Table A.3.

Following the mean structure of Eq. (4.5), the second model component is given by

$$\mathbf{h}_{l,w}^{\text{res}} \sim \text{Dirichlet-Multinomial}(h_{l,w}^{\text{res-sum}}, \phi_{l,w} \boldsymbol{\pi}_{l,w}) \quad (4.6a)$$

$$\pi_{l,a,w} = \alpha_{l,w}\pi_{l,a,w}^{\Gamma} + (1 - \alpha_{l,w})\pi_{l,a,w}^{\text{non-}\Gamma} \quad (4.6b)$$

$$\pi_{l,a,w}^{\Gamma} = \text{softmax}(\log \lambda_{l,a}^{\Gamma} + \log n_{l,a,w}^R) \quad (4.6c)$$

$$\pi_{l,a,w}^{\text{non-}\Gamma} = \text{softmax}(\log \lambda_{l,a}^{\text{non-}\Gamma} + \log n_{l,a,w}^R) \quad (4.6d)$$

$$\log \lambda_{l,a}^{\Gamma} \sim \mathcal{N}(0, 1) \quad (4.6e)$$

$$\log \lambda_{l,a}^{\text{non-}\Gamma} \sim \mathcal{N}(0, 1) \quad (4.6f)$$

$$\phi_{l,w} = (h_{l,w}^{\text{res-sum}} - 1) / (\theta_2 + 1) \quad (4.6g)$$

$$\theta_2 \sim \text{Exponential}(20), \quad (4.6h)$$

where we denote the vector of age-specific hospital admissions in location l and week w by $\mathbf{h}_{l,w}^{\text{res}} = (h_{l,a,w}^{\text{res}})_{a \in \mathcal{A}}$ and the vector of the age composition of hospital admissions in location l and week w by $\boldsymbol{\pi}_{l,w} = (\pi_{l,a,w})_{a \in \mathcal{A}}$, again such that $\sum_a \pi_{l,a,w} = 1$. The Dirichlet-Multinomial is in standard sample size-scale parametrisation such that the means

are given by $h_{l,w}^{\text{res-sum}} \pi_{l,a,w}$ as in Eq. (4.4b). The softmax transformations (Eq. (4.6c)-Eq. (4.6d)) allow for convenient prior specifications of the log hospital admission rates on the real line, and run over age bands a for fixed location l and fixed week w . The scale parameter $\phi_{l,w}$ is conditional on the known $h_{l,w}^{\text{res-sum}}$ re-parametrised into the overdispersion parameter θ_2 , with $\theta_2 > 0$, and the prior on θ_2 in Eq. (4.6h) peaks at zero and favours the least complex model with no overdispersion.

As a discussion to this model component, we note that hospital admissions are decomposed through Eq. (4.5) into infections with Gamma and non-Gamma variants according to the replacement dynamics in each city. It is clear that $\pi_{l,a}^{\text{non-}\Gamma}$ can be inferred from patient records when $\alpha_{l,w} \approx 0$, i. e. from before Gamma's expansion, and $\pi_{l,a}^{\Gamma}$ can be inferred from patient records when $\alpha_{l,w} \approx 1$, i. e. from after Gamma's expansion. Across time, the proportion of hospital admissions in age band a is described as a mixture of the Gamma and non-Gamma proportions that is determined by the increasing weight of Gamma's variant frequency $a_{l,w}$. Good model fit can only be obtained if the model reflects the observed age distributions of hospitalised patients shift over time according to Gamma's expansion, the heavier cumulative loss of lives in older age groups and age-prioritised vaccinations. Initial versions of the model did not account for age-prioritised vaccinations in $n_{l,a,w}^R$ which resulted in very poor fit against the observed data in locations where Gamma's expansion coincided with rapidly increasing vaccination coverage.

In-hospital fatality rates component

In the third model component, we quantify and model the fatality rates in the Gamma- and non-Gamma- attributable hospital admissions separately. We model in-hospital fatality rates in week w , location l and age group a that are respectively infected with non-Gamma and Gamma variants with the decomposition

$$\zeta_{l,a,w}^{\text{non-}\Gamma} = \text{logit}^{-1} \left(\eta_{l,a}^{\text{non-}\Gamma} + X_{l,w} \beta_l \right) \quad (4.7a)$$

$$\zeta_{l,a,w}^{\Gamma} = \text{logit}^{-1} \left(\eta_{l,a}^{\text{non-}\Gamma} + \eta_{l,a}^{\Gamma\text{-random-effect}} + X_{l,w} \beta_l \right), \quad (4.7b)$$

where $X_{l,w} \in \mathbb{R}^{1 \times p}$ are p standardised healthcare pressure indices in location l and week w and $\beta_l \in \mathbb{R}^p$ are constrained to be non-negative. For each location, we standardised the healthcare pressure indices to zero for the week in which the empirical fatality rates Eq. (4.3) are lowest prior to Gamma's detection, which was typically at the beginning of the pandemic and observation period. Specifically, we standardised each index according to

$$\tilde{x}_{l,p,w} = (x_{l,p,w} - x_{l,p,\tilde{w}_l}) / \text{sd}(x_{l,p,w}), \quad (4.8)$$

where $\tilde{w}_l = \operatorname{argmin}_{w=1:(W_l^{\text{detect}}-\Gamma-1)} \hat{z}_{l,w}$ and sd denotes the standard deviation of the indicator across all weeks in location l , using the notation in (Eq. (4.3)-Eq. (4.1)). The p regression coefficients β_l were constrained to be positive so that each standardised index of increasing hospital demand pressure could only have an increasing effect on fatality rates when $\beta_l > 0$, or no effect when $\beta_l = 0$, and not offset each other and artifactually increase the effect of another index.

Thus we can interpret the $\operatorname{logit}^{-1}(\eta_{l,a}^{\text{non-}\Gamma})$ as the lowest observed fatality rate in each location and age group, and take ratios of these terms across locations to quantify the impact of geographic inequities (*location effect*). These ratios measure baseline differences in fatality rates across the state capitals that are not captured by our healthcare pressure indices. Next, for each location we can interpret the ratio of $\operatorname{logit}^{-1}(\eta_{l,a}^{\text{non-}\Gamma} + X_{l,w}\beta_l)$ divided by $\operatorname{logit}^{-1}(\eta_{l,a}^{\text{non-}\Gamma})$ as the multiplier to in-hospital fatality rates in week w and age group a that is associated with the increasing healthcare pressure indices (*healthcare pressure effect*). In Eq. (4.7b), $\eta_{l,a}^{\Gamma\text{-random-effect}}$ are non-parametric random effects that measure age-specific deviations in fatality rates in hospitalised patients attributed to Gamma infection. So, for each location we can interpret the ratio of $\operatorname{logit}^{-1}(\eta_{l,a}^{\text{non-}\Gamma} + \eta_{l,a}^{\Gamma\text{-random-effect}})$ divided by $\operatorname{logit}^{-1}(\eta_{l,a}^{\text{non-}\Gamma})$ as the multiplier to fatality rates in hospitalised patients that is attributable to Gamma infection versus non-Gamma infection (*variant effect* in hospitals), controlling for differences in healthcare pressure in the weeks before and after emergence of Gamma (because the $X_{l,w}\beta_l$ are zeroed), as well as cumulative mortality and vaccination in the full model (see below). We note that the regression model is measuring associations and is not guaranteed to identify causal relationships. Indeed, causal relationships are difficult to establish in observational data, even if the variant status of the patients were known. Compared to Clinical Trials, it is impossible and unethical to randomise patients to different variants, and therefore analysts are limited to adjusting models by accounting for potential confounders [Olier et al., 2023].

We describe the expected fatality rates in the COVID-19 attributable hospital admissions according to the modelled Gamma and non-Gamma fatality rates Eq. (4.7), and their corresponding contributions to the overall fatality rates according to Gamma's

increasing frequency Eq. (4.9b), which is given by

$$h_{l,a,w}^{\text{res-adj-D}} \sim \text{Beta-Binomial}\left(h_{l,a,w}^{\text{res}}, \zeta_{l,a,w}/\theta_3, (1 - \zeta_{l,a,w})/\theta_3\right) \quad (4.9a)$$

$$\zeta_{l,a,w} = \frac{(1 - \alpha_{l,w})\pi_{l,a,w}^{\text{non-}\Gamma}}{\alpha_{l,w}\pi_{l,a,w}^{\Gamma} + (1 - \alpha_{l,w})\pi_{l,a,w}^{\text{non-}\Gamma}} \zeta_{l,a,w}^{\text{non-}\Gamma} + \quad (4.9b)$$

$$\frac{\alpha_{l,w}\pi_{l,a,w}^{\Gamma}}{\alpha_{l,w}\pi_{l,a,w}^{\Gamma} + (1 - \alpha_{l,w})\pi_{l,a,w}^{\text{non-}\Gamma}} \zeta_{l,a,w}^{\Gamma} \quad (4.9c)$$

$$\text{logit } \zeta_{l,a,w}^{\text{non-}\Gamma} = \eta_{l,a}^{\text{non-}\Gamma} + X_{l,w}\beta_l \quad (4.9d)$$

$$\text{logit } \zeta_{l,a,w}^{\Gamma} = \eta_{l,a}^{\text{non-}\Gamma} + \eta_{l,a}^{\Gamma\text{-random-effect}} + X_{l,w}\beta_l \quad (4.9e)$$

$$\eta_{l,a}^{\text{non-}\Gamma} \sim \mathcal{N}(-0.25, 1.5^2) \quad (4.9f)$$

$$\eta_{l,a}^{\Gamma\text{-random-effect}} \sim \mathcal{N}(0, \sigma_{\zeta}^2) \quad (4.9g)$$

$$\beta_{l,i} \sim \mathcal{N}_{[0,\infty]}(0, \kappa_{l,i}^2 \tau^2) \quad (4.9h)$$

$$\kappa_{l,i} \sim \text{Half-Cauchy}(0, 1) \quad (4.9i)$$

$$\tau \sim \text{Half-Cauchy}(0, 0.01) \quad (4.9j)$$

$$\sigma_{\zeta} \sim \text{Exponential}(2) \quad (4.9k)$$

$$\theta_3 \sim \text{Exponential}(100), \quad (4.9l)$$

where $w = 1, \dots, W_l$ and $h_{l,a,w}^{\text{res-adj-D}}$ is the number of under-reporting adjusted deaths among the $h_{l,a,w}^{\text{res}}$ COVID-19 attributable hospital admissions from residents in week w and age group a in location l . The Beta-Binomial is in the shape-shape parameterisation with mean $h_{l,a,w}^{\text{res}} \zeta_{l,a,w}$ and variance equal to the Binomial variance component $h_{l,a,w}^{\text{res}} \zeta_{l,a,w} (1 - \zeta_{l,a,w})$ multiplied by $(1 + (h_{l,a,w}^{\text{res}} - 1) \frac{1}{\theta_3^{-1} + 1})$ to allow for overdispersion. The priors in Eq. (4.9f) were chosen to place the non-Gamma in-hospital fatality rate around the empirically observed range. In Eq. (4.9g), we model the Gamma in-hospital fatality rate as a random effect around the non-Gamma in-hospital fatality rate. In the model we conservatively sought to favour no dependence of in-hospital fatality rates on the hospital pressure indices, which we implemented using the horseshoe-type shrinkage prior in (Eq. (4.9h)-Eq. (4.9j)). The prior in Eq. (4.9l) peaks at zero and favours the least complex model with no overdispersion. The overdispersion encoded in the Beta-Binomial distribution allows for the model to account for the variability in the number of hospitalised patients deaths that is not captured by our parametrisation of the mean Eq. (4.9b). In fact, there might be other time-varying factors affecting fatality rates which are not captured by our predictors.

Computational inference

The model described through equations (Eq. (4.4)-Eq. (4.9)) was implemented in the Stan probabilistic computing language, is available at [inst/stan-models/age_hfr_210719d](https://github.com/inst/stan-models/tree/main/age_hfr_210719d) in the Github repository, and was independently fitted to data from each location using cmdstanr version 0.3.0.9 [Carpenter et al., 2017, Homan and Gelman, 2014b]. As the model was run separately for each location, the estimates from each location provide independent support into the inferred relationships between in-hospital fatality rates, healthcare inequities, and healthcare pressure. Alternatively, a hierarchical model could have been used to pool information across locations, but this would have required a more complex model and would have been computationally more demanding. We found the slowest model to take 6 hours to fit. Further, each of the 14 fitted objects required almost 3 GB of storage space. Loading all of them in memory simultaneously would have made local explorations and postprocessing difficult.

In total, the model comprised 75 parameters, although some were interrelated through hierarchical components. The variant replacement subcomponent included 3 parameters, the age composition of hospital admissions subcomponent included 23 parameters, and the in-hospital fatality rates subcomponent included 49 parameters. Each inference was conducted in 4 Hamiltonian Monte Carlo chains, each over 500 warmup iterations, and 50,500 sampling iterations. The model converged well, and the smallest bulk effective sample size was 1687, (see Table A.7, Fig. A.12).

4.3.4 Key Epidemiological quantities uncovered

From the Monte Carlo samples of the joint posterior distribution of the fitted Bayesian multi-strain fatality model, we generated a collection of quantities of interest. We calculate the expected, COVID-19 attributable hospital admissions among residents in location l , age band a , and week w for non-Gamma and Gamma variants that had no evidence of vaccination prior to hospitalisation respectively by

$$\begin{aligned} h_{l,a,w}^{\text{res-non-}\Gamma} &= (1 - \alpha_{l,w}) \pi_{l,a,w}^{\text{non-}\Gamma} h_{l,a,w}^{\text{res}} \\ h_{l,a,w}^{\text{res-}\Gamma} &= \alpha_{l,w} \pi_{l,a,w}^{\Gamma} h_{l,a,w}^{\text{res}}, \end{aligned} \quad (4.10)$$

where $h_{l,a,w}^{\text{res}}$ are observed and $\alpha_{l,w}$, $\pi_{l,a,w}^{\text{non-}\Gamma}$, $\pi_{l,a,w}^{\Gamma}$ are from the joint posterior. The expected, COVID-19 attributable hospital admissions among residents in location l , age band a , and week w across all variants are

$$h_{l,a,w}^{\text{res-all}} = h_{l,a,w}^{\text{res-non-}\Gamma} + h_{l,a,w}^{\text{res-}\Gamma}. \quad (4.11)$$

We thus have that the expected share of age group a among hospital admissions among residents of location l in week w with non-Gamma variants is $\frac{h_{l,a,w}^{\text{res-non-}\Gamma}}{\sum_b h_{l,b,w}^{\text{res-non-}\Gamma}} = \pi_{l,a,w}^{\text{non-}\Gamma}$, and similarly for Gamma. The expected share of age group a among hospital admissions among residents of location l in week w across all variants is $\pi_{l,a,w} = (1 - \alpha_{l,w})\pi_{l,a,w}^{\text{non-}\Gamma} + \alpha_{l,w}\pi_{l,a,w}^{\Gamma}$. We calculate the expected, COVID-19 attributable deaths among hospital admissions in residents in location l , age band a , and week w for non-Gamma and Gamma variants respectively by

$$\begin{aligned} h_{l,a,w}^{\text{res-non-}\Gamma\text{-D}} &= (1 - \alpha_{l,w})\pi_{l,a,w}^{\text{non-}\Gamma} h_{l,a,w}^{\text{res}} \zeta_{l,a,w}^{\text{non-}\Gamma} \\ h_{l,a,w}^{\text{res-}\Gamma\text{-D}} &= \alpha_{l,w}\pi_{l,a,w}^{\Gamma} h_{l,a,w}^{\text{res}} \zeta_{l,a,w}^{\Gamma}, \end{aligned} \quad (4.12)$$

where $h_{l,a,w}^{\text{res}}$ are observed and $\alpha_{l,w}$, $\pi_{l,a,w}^{\text{non-}\Gamma}$, $\pi_{l,a,w}^{\Gamma}$, $\zeta_{l,a,w}^{\text{non-}\Gamma}$, $\zeta_{l,a,w}^{\Gamma}$ are from the joint posterior. The expected, COVID-19 attributable deaths among hospital admissions in residents in location l , age band a , and week w for all variants are

$$h_{l,a,w}^{\text{res-all-D}} = h_{l,a,w}^{\text{res-non-}\Gamma\text{-D}} + h_{l,a,w}^{\text{res-}\Gamma\text{-D}}. \quad (4.13)$$

The estimated COVID-19 attributable hospital admissions and deaths by SARS-CoV-2 variant are shown in Fig. 4.3 and Fig. 4.4 respectively.

To compare COVID-19 in-hospital fatality rates across locations, we define the overall in-hospital fatality rate in location l and week w in an age-standardised population that adjust for differences in age composition across locations. Specifically, we calculate

$$\zeta_{l,w}^{\text{age-std}} = \sum_a \frac{n_a^{\text{cities}}}{\sum_b n_b^{\text{cities}}} \zeta_{l,a,w} \quad (4.14)$$

where n_a^{cities} is the population size in age band a across all cities considered, and $\zeta_{l,a,w}$ is from the joint posterior.

Location effect on COVID-19 in-hospital fatality rates

For each location, we define the week w_l^* with lowest in-hospital fatality rate as the week that minimises the posterior median of Eq. (4.14),

$$w_l^* = \operatorname{argmin}_{w \in 1:W_l^{\text{detect-}\Gamma}} (\text{posterior median of } \zeta_{l,w}^{\text{age-std}}), \quad (4.15)$$

and calculate the lowest, age standardised COVID-19 in-hospital fatality rates prior to first detection of Gamma through

$$\zeta_l^{\text{lowest}} = \zeta_{l,w_l^*}^{\text{age-std}}. \quad (4.16)$$

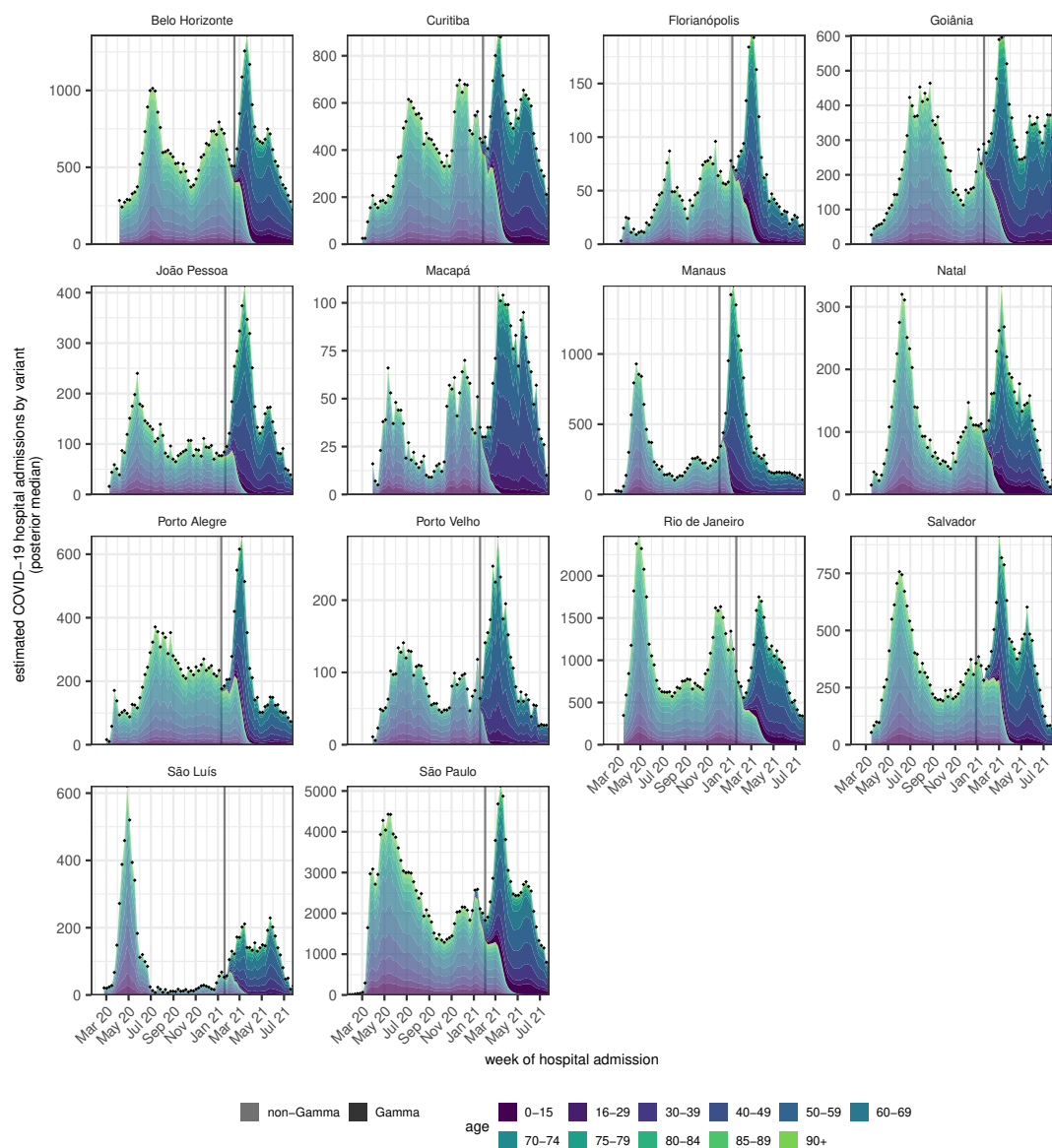


Figure 4.3: Estimated COVID-19 attributable hospital admissions by SARS-CoV-2 variant. Posterior median estimates of hospital admissions among residents in each location, that are attributed to non-Gamma variants are shown for each age band (color) in lighter shades, while those for the Gamma variant are shown in darker shades. Estimates are derived using the Bayesian multi-strain fatality model. Locations are shown across facets. The date of Gamma's detection is shown as a grey vertical line. Observed weekly hospital admissions in residents are shown as black dots. This figure is included in the accepted manuscript provided to *Nature Medicine* for further editing.

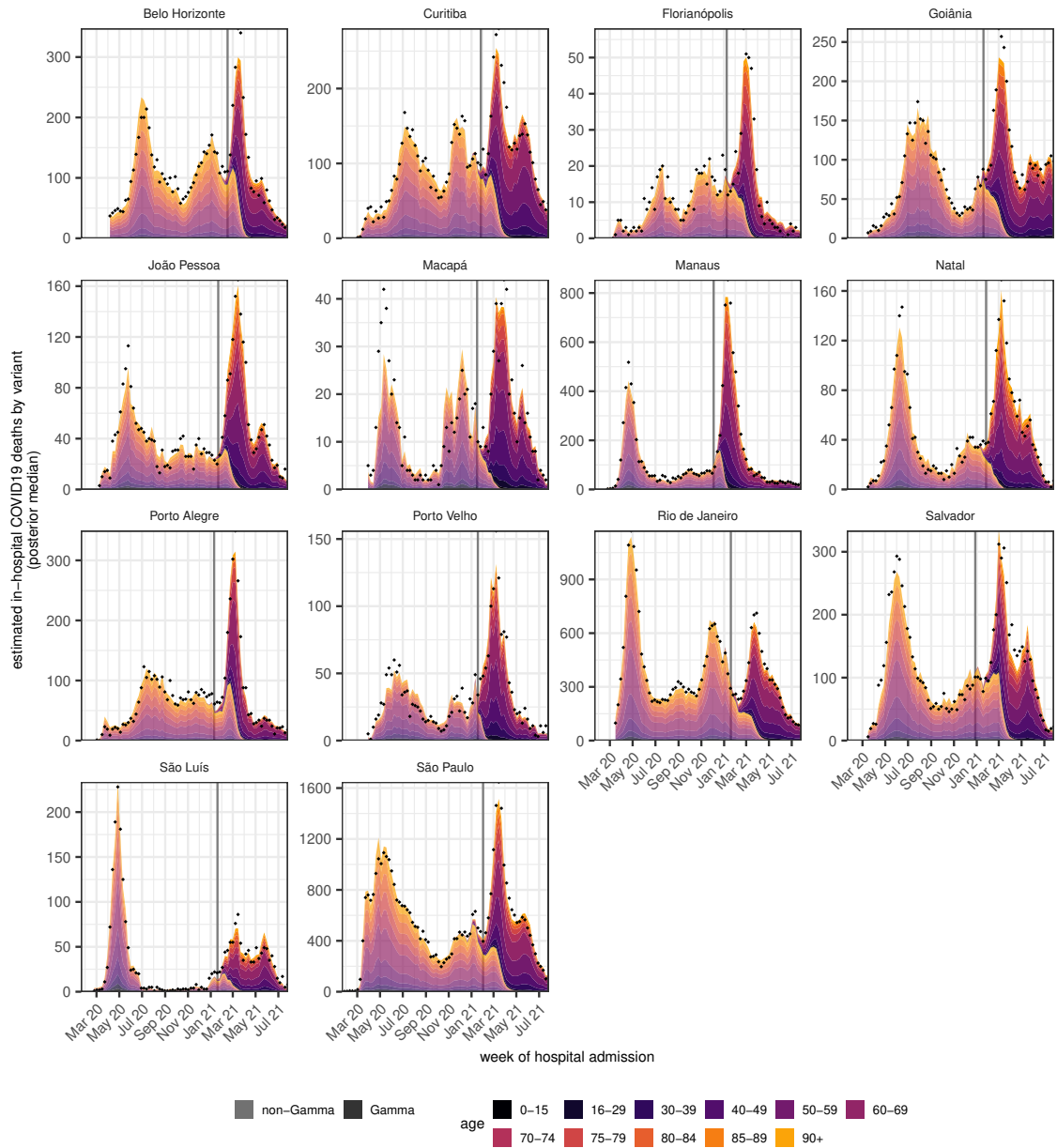


Figure 4.4: Estimated COVID-19 attributable deaths in hospitals by SARS-CoV-2 variant. Posterior median estimates of deaths following hospital admissions of residents in each location, that are attributed to non-Gamma variants are shown for each age band (color) in lighter shades, while those for the Gamma variant are shown in darker shades. Estimates are derived using the Bayesian multi-strain fatality model. Locations are shown across facets. The date of Gamma's detection is shown as a grey vertical line. Observed weekly deaths following hospital admissions in residents are shown as black dots. This figure is included in the accepted manuscript provided to *Nature Medicine* for further editing.

The estimated Gamma frequencies $\alpha_{l,w}$ are of course also very small prior to first detection of Gamma, and so the lowest, age standardised COVID-19 in-hospital fatality rate are approximately

$$\zeta_l^{\text{lowest}} \approx \sum_a \frac{n_a^{\text{cities}}}{\sum_b n_b^{\text{cities}}} \text{logit}^{-1}(\eta_{l,a}^{\text{non-}\Gamma}), \quad (4.17)$$

where $\eta_{l,a}^{\text{non-}\Gamma}$ are the intercept terms in our regression Eq. (4.7) and from the joint posterior. Then, to compare Eq. (4.16) across locations, we find the location with overall lowest age-standardised in-hospital fatality rate by

$$l^* = \text{argmin}_l \zeta_l^{\text{lowest}}, \quad (4.18)$$

and compute for all other locations l the ratio

$$\zeta_l^{\text{lowest-ratio}} = \zeta_l^{\text{lowest}} / \zeta_{l^*}^{\text{lowest}}. \quad (4.19)$$

We interpret Eq. (4.19) as the location effect on COVID-19 in-hospital fatality rates. Because of Eq. (4.17), the location effect does not include contributions attributable to the healthcare pressure indices $(X_{l,w}\beta_l)$, nor any contributions attributable to the non-parametric Gamma effects $(\eta_{l,a}^{\Gamma})$ on in-hospital fatality rates.

Healthcare pressure effect on COVID-19 in-hospital fatality rates

To compare time trends to in-hospital fatality rates in each location, we calculate for each location the multiplicative effect of changes in healthcare demand and resources in week w in location l by

$$\zeta_{l,w}^{\text{multiplier}} = \sum_a \frac{n_a^{\text{cities}}}{\sum_b n_b^{\text{cities}}} \frac{\zeta_{l,a,w}^{\text{non-}\Gamma}}{\zeta_{l,a,w_l^*}^{\text{non-}\Gamma}}, \quad (4.20)$$

where $\zeta_{l,a,w}^{\text{non-}\Gamma}$ are from the joint posterior. Recalling Eq. (4.7a), we have

$$\zeta_{l,w}^{\text{multiplier}} \approx \sum_a \frac{n_a^{\text{cities}}}{\sum_b n_b^{\text{cities}}} \frac{\text{logit}^{-1}(\eta_{l,a}^{\text{non-}\Gamma} + X_{l,w}\beta_l)}{\text{logit}^{-1}(\eta_{l,a}^{\text{non-}\Gamma})}. \quad (4.21)$$

Here, the approximation is because the Gamma frequencies $\alpha_{l,w}$ are very small but not exactly zero, and because the week in which the loess-smoothed fatality rates are lowest may not exactly coincide with the week w_l^* in which the model-based fatality rates are lowest. We interpret Eq. (4.20) as the healthcare pressure effect on COVID-19 in-hospital fatality rates. Because of Eq. (4.21), the healthcare pressure effect does not include

contributions attributable to the non-parametric Gamma effects ($\eta_{l,a}^\Gamma$) on in-hospital fatality rates, and corresponds to the multiplier to the minimum fatality rates in each location that is associated with the healthcare pressure indices ($X_{l,w}\beta_l$).

Variant effect on COVID-19 in-hospital fatality rates

Finally we describe the effect of Gamma on in-hospital fatality rates in the model. Standardising across age bands, we calculate the ratio in Gamma versus non-Gamma in-hospital fatality rates in location l by

$$\zeta_l^{\Gamma\text{-ratio}} = \sum_a \frac{n_a^{\text{cities}}}{\sum_b n_b^{\text{cities}}} \zeta_{l,a,w_l^*}^\Gamma / \zeta_{l,a,w_l^*}^{\text{non-}\Gamma}, \quad (4.22)$$

where $\zeta_{l,a,w_l^*}^\Gamma, \zeta_{l,a,w_l^*}^{\text{non-}\Gamma}$ are from the joint posterior. We interpret Eq. (4.22) as the Gamma effect on in-hospital fatality rates. It is again helpful to recall Eq. (4.7), which shows that we can approximate the Gamma effect through

$$\zeta_l^{\Gamma\text{-ratio}} \approx \sum_a \frac{n_a^{\text{cities}}}{\sum_b n_b^{\text{cities}}} \frac{\text{logit}^{-1}(\eta_{l,a}^{\text{non-}\Gamma} + \eta_{l,a}^\Gamma)}{\text{logit}^{-1}(\eta_{l,a}^{\text{non-}\Gamma})}. \quad (4.23)$$

The ratio $\zeta_l^{\Gamma\text{-ratio}}$ does not include contributions attributable to the healthcare pressure indices ($X_{l,w}\beta_l$), and corresponds to the multiplier to the minimum fatality rates in each location that is associated with the non-parametric Gamma effects ($\eta_{l,a}^{\text{non-}\Gamma}$).

In summary, the definition of the location, healthcare pressure and variant effects are an attempt to obtain interpretable parameter-specific effects that mimic odds ratios in logistic regression models. As with odds ratio, we compute the ratio between the probabilities of the same outcome occurring at two different levels of a factor. We additionally obtain summary age-standardised versions to ease comparisons across locations and time.

Counterfactual analyses

To quantify the impact that the observed fluctuations in COVID-19 attributable in-hospital rates had on the death toll in the 14 state capitals, we performed two counterfactual analyses. The aim of the first counterfactual (Scenario 1) was to estimate how many COVID-19 attributable deaths could have been avoided with sufficient healthcare resources so that healthcare pressures would not have resulted in shocks in in-hospital fatality rates in each city. We implemented Scenario 1 by predicting deaths in each city under the minimum in-hospital fatality rate that were observed in each age group in that city. Specifically, for each location l , we considered the estimated age-specific

in-hospital fatality rates Eq. (4.9b) in the week Eq. (4.15), and computed

$$h_l^{\text{hyp1-res-D}} = \sum_{w=1}^{W_l} \sum_{a \in \mathcal{A}} h_{l,a,w}^{\text{res}} \zeta_{l,a,w_l^*} \quad (4.24)$$

where the week indices range over the entire observation period. We define the expected COVID-19 attributable deaths that could have been avoided in the absence of health-care pressures in location l during the observation period relative to this hypothetical scenario by

$$h_l^{\text{hyp1-avoidable-D}} = \left(\sum_{w=1}^{W_l} \sum_{a \in \mathcal{A}} h_{l,a,w}^{\text{res-adj-D}} \right) - h_l^{\text{hyp1-res-D}}. \quad (4.25)$$

Similarly, we define the expected percentage reduction in COVID-19 attributable deaths relative to this hypothetical scenario by

$$p_l^{\text{hyp1-avoidable-D}} = 1 - \frac{h_l^{\text{hyp1-res-D}}}{\sum_{w=1}^{W_l} \sum_{a \in \mathcal{A}} h_{l,a,w}^{\text{res-adj-D}}}. \quad (4.26)$$

The aim of the second counterfactual (Scenario 2) was to estimate how many COVID-19 attributable deaths could have been avoided with sufficient healthcare resources and without healthcare inequities across the cities. We implemented Scenario 2 by predicting deaths in each city under the minimum in-hospital fatality rate seen across all cities, which we observed in Belo Horizonte. Specifically, we considered the location l^{**} in which we found the lowest age-standardised in-hospital fatality rate prior to Gamma's detection in each location,

$$l^{**} = \operatorname{argmin}_l \left(\min_{w \in 1:W_l^{\text{detect-}\Gamma}} (\text{posterior median of } \xi_{l,w}^{\text{age-std}}) \right), \quad (4.27)$$

and computed

$$h_l^{\text{hyp2-res-D}} = \sum_{w=1}^{W_l} \sum_{a \in \mathcal{A}} h_{l,a,w}^{\text{res}} \zeta_{l^{**},a,w^*} \quad (4.28)$$

where again the week indices range over the entire observation period. In analogy to the first counterfactual, we then calculate

$$h_l^{\text{hyp2-avoidable-D}} = \left(\sum_{w=1}^{W_l} \sum_{a \in \mathcal{A}} h_{l,a,w}^{\text{res-adj-D}} \right) - h_l^{\text{hyp2-res-D}}, \quad (4.29)$$

and the expected percentage reduction

$$p_l^{\text{hyp2-avoidable-D}} = 1 - \frac{h_l^{\text{hyp2-res-D}}}{\sum_{w=1}^{W_l} \sum_{a \in \mathcal{A}} h_{l,a,w}^{\text{res-adj-D}}}. \quad (4.30)$$

4.4 Findings

4.4.1 Fluctuations of in-hospital fatality rates around Gamma's emergence

Fig. 4.5 summarises the data selection process. 641,618 patients with COVID-19 attributable infection were admitted to hospitals across the 14 cities and reported to Brazil's SIVEP-Gripe (Sistema de Informação da Vigilância Epidemiológica da Gripe) database (Fig. 4.5 and Table A.1). To match these data to city-level transmission dynamics, we retained for analysis the 495,100 COVID-19 attributable hospital admissions among residents in each state capital. We further excluded 32,734 resident patients who were administered at least one vaccine dose before hospitalisation to avoid confounding of time trends in fatality rates with vaccine rollout, which occurred during the same period. Thus, our study population comprised the remaining 462,366 COVID-19 attributable hospital admissions in residents. Among those, 120,288 (26.02%) outcomes were fatal. However, 32,585 (7%) admissions had unreported clinical outcomes, which occurred primarily after Gamma's detection, and so it is important to account for under-reporting of COVID-19 deaths [de Souza et al., 2020, Fellows et al., 2021, Veiga E Silva et al., 2020]. We estimate 10,029 additional fatalities by considering the proportion of deaths among patients with known outcomes, stratified by age and observation week (Table A.4).

Fig. 4.6c shows the resulting, empirical COVID-19 in-hospital fatality rates for Manaus, defined as the proportion of under-reporting adjusted deaths in weekly COVID-19-attributable hospital admissions among residents with no evidence of vaccination prior to admission (see Fig. 4.1 for 4 other state capitals). The observed, age-dependent patterns in COVID-19 in-hospital fatality rates are consistent with earlier observations that COVID-19 fatality rates increase with age [Brazeau et al., 2020]. Moreover, we observe across all age bands marked increases in COVID-19 in-hospital fatality rates after detection of Gamma, a pattern that is consistent across the 14 cities. Yet, importantly, the within age band variation reveals stark geographical differences and temporal fluctuations since the beginning of the pandemic rather than just since Gamma's emergence, reinforcing previous findings on geographical heterogeneity in the first three months of the pandemic [Ranzani et al., 2021a]. For example, in Belo Horizonte no age group experienced shocks of COVID-19 in-hospital fatality rates above 50% that lasted at least four consecutive weeks, whereas in Macapá all age groups of age 40 or older experienced such fatality shocks (Table A.5). We also find that the increases in COVID-19 in-hospital fatality rates after detection of Gamma were largely temporary, declining in tandem with fewer admissions, as shown in Fig. 4.6c for Manaus.

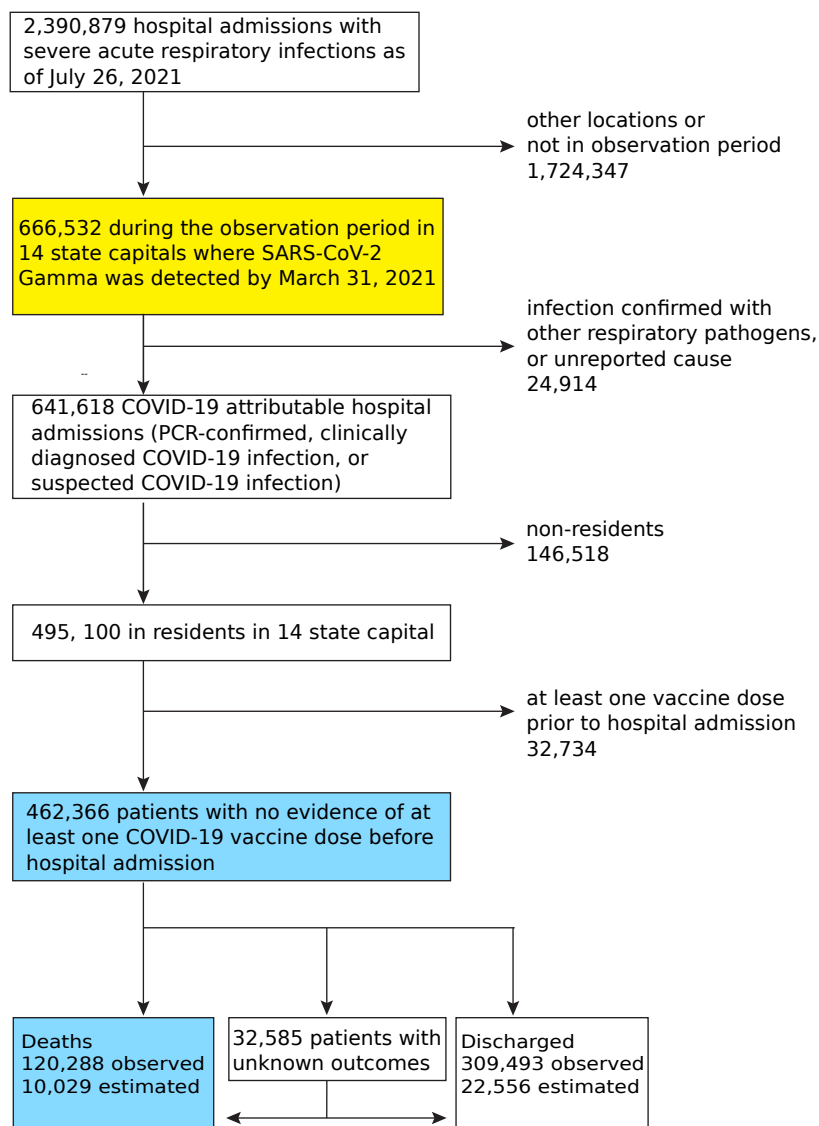


Figure 4.5: Analysis flowchart. Individual-level records on hospital admissions with severe acute respiratory infection across Brazil are mandatory to report to the SIVEP-Gripe (Sistema de Informação da Vigilância Epidemiológica da Gripe) database, and publicly available records between 20 January 2020 and 26 July 2021 were downloaded on 31 January 2022. Data used to derive **COVID-19** in-hospital fatality rates are shown in blue, and data used to derive the healthcare pressure indices are shown in yellow. This figure is included in the accepted manuscript provided to *Nature Medicine* for further editing.

Fig. 4.7a shows the age-standardised **COVID-19** in-hospital fatality rates (Eq. (4.3)) since the beginning of the pandemic across locations in black lines, and Table 4.1 reports the minimum and maximum observed values in each location. The minimum age-standardised in-hospital fatality rates occurred either at the start of the observation period or between waves of hospital admissions in each location for all cities except Belo Horizonte, where the lowest fatality rates were seen at the end of the observation period. We performed additional analyses suggesting that this further drop in fatality rates in Belo Horizonte at the end of the observation period is likely related to missing

Table 4.1: Temporal fluctuations in COVID-19 attributable in-hospital fatality rate, and avoidable COVID-19 attributable deaths in hospitals.

Location	Observation Period	COVID-19 attributable hospital admissions in unvaccinated residents		Age-standardised weekly COVID-19 in-hospital fatality rate (HFR)			Estimated, avoidable COVID-19 deaths in hospitals assuming the lowest HFR	
		Total	Deaths*	Lowest [†]	Highest [†]	Fold increase [†]	in each city Percent [‡]	across all 14 cities Percent [‡]
Belo Horizonte	06/04/20-26/07/21	42,113	7,692	7.7%	12.7%	1.64	29.1% (24.6%-33.7%)	29.1% (24.6%-33.7%)
Curitiba	02/03/20-26/07/21	33,015	7,535	8.1%	18.1%	2.23	39.6% (34.2%-44.7%)	47.3% (43.9%-50.8%)
Florianópolis	09/03/20-26/07/21	4,032	893	7.8%	17.1%	2.20	22.7% (12.6%-32.5%)	44.5% (40.9%-48.1%)
Goiânia	16/03/20-26/07/21	20,044	6,246	11.6%	25.9%	2.23	44.7% (39.0%-50.0%)	61.2% (58.6%-63.7%)
João Pessoa	09/03/20-26/07/21	10,552	3,636	15.3%	29.7%	1.94	18.8% (13.0%-25.1%)	65.9% (63.7%-68.1%)
Macapá	30/03/20-26/07/21	3,169	1,011	11.0%	41.5%	3.78	41.6% (30.5%-51.8%)	68.4% (66.2%-70.6%)
Manaus	24/02/20-26/07/21	26,260	10,168	16.4%	33.4%	2.04	39.3% (35.9%-42.3%)	70.9% (69%-72.9%)
Natal	16/03/20-26/07/21	9,344	3,512	11.4%	32.9%	2.87	35.4% (30.2%-40.5%)	66.9% (64.8%-69%)
Porto Alegre	02/03/20-26/07/21	16,106	5,266	8.6%	27.1%	3.16	41.8% (37.7%-45.7%)	59.6% (57.1%-62.3%)
Porto Velho	30/03/20-26/07/21	6,795	2,473	11.5%	32.2%	2.79	39.1% (32.2%-45.5%)	70.3% (68.3%-72.4%)
Rio de Janeiro	16/03/20-26/07/21	73,139	28,053	19.2%	26.0%	1.35	10.2% (8.5%-11.9%)	66.9% (64.8%-69%)
Salvador	16/03/20-26/07/21	26,964	8,509	9.7%	26.9%	2.76	20.8% (16.2%-25.2%)	61.3% (58.8%-63.9%)
São Luís	24/02/20-26/07/21	8,545	2,547	8.2%	26.5%	3.25	35.7% (28.4%-42.4%)	60.5% (57.9%-63.1%)
São Paulo	20/01/20-26/07/21	182,288	42,769	8.8%	19.8%	2.24	36.2% (33.9%-38.4%)	50.1% (46.9%-53.4%)
All 14 cities	20/01/20-26/07/21	462,366	130,317				29.8% (28.7%-30.9%)	57.1% (54.3%-59.9%)
* Observed deaths plus expected deaths in COVID-19 attributable hospital admissions with unknown outcome. [†] Age-specific COVID-19 attributable in-hospital fatality rates were estimated from linked data on underreporting-adjusted deaths in COVID-19 attributable hospital admissions. Non-parametric loess estimates were obtained and weighted by the population age composition across cities. Lowest fatality rates were calculated in the period prior to Gamma's first detection in each location, and highest fatality rates were calculated including the time after Gamma's first detection. The lowest fatality rates in the period prior to Gamma's first detection agreed with those observed across the entire study period for all cities except Belo Horizonte [‡] Estimates are based on hypothetical scenarios evaluated under the Bayesian multi-strain fatality model, assuming the lowest observed in-hospital fatality rates seen in the periods prior to Gamma's first detection in each location.								

data on vaccine status in hospitalised patients in Belo Horizonte (Fig. A.10). For this reason, and to guard our analyses against other potential confounders such as improved treatment with Dexamethasone, we evaluated the minimum fatality rates over the period until detection of Gamma detection. In Belo Horizonte, the minimum age-standardised COVID-19 in-hospital fatality rate before detection of Gamma was 7.7%, and the maximum value over the entire study period was 12.7%, a 1.64-fold increase. We observed higher fold increases in all other state capitals, apart from Rio de Janeiro where age-standardised in-hospital fatality rates were very high throughout the study period. Rates tended to reach similar baseline values between shock periods in each location, and were lowest in state capitals from the South and Southeast regions of Brazil. The maximum rates occurred in most locations after first detection of Gamma, except for

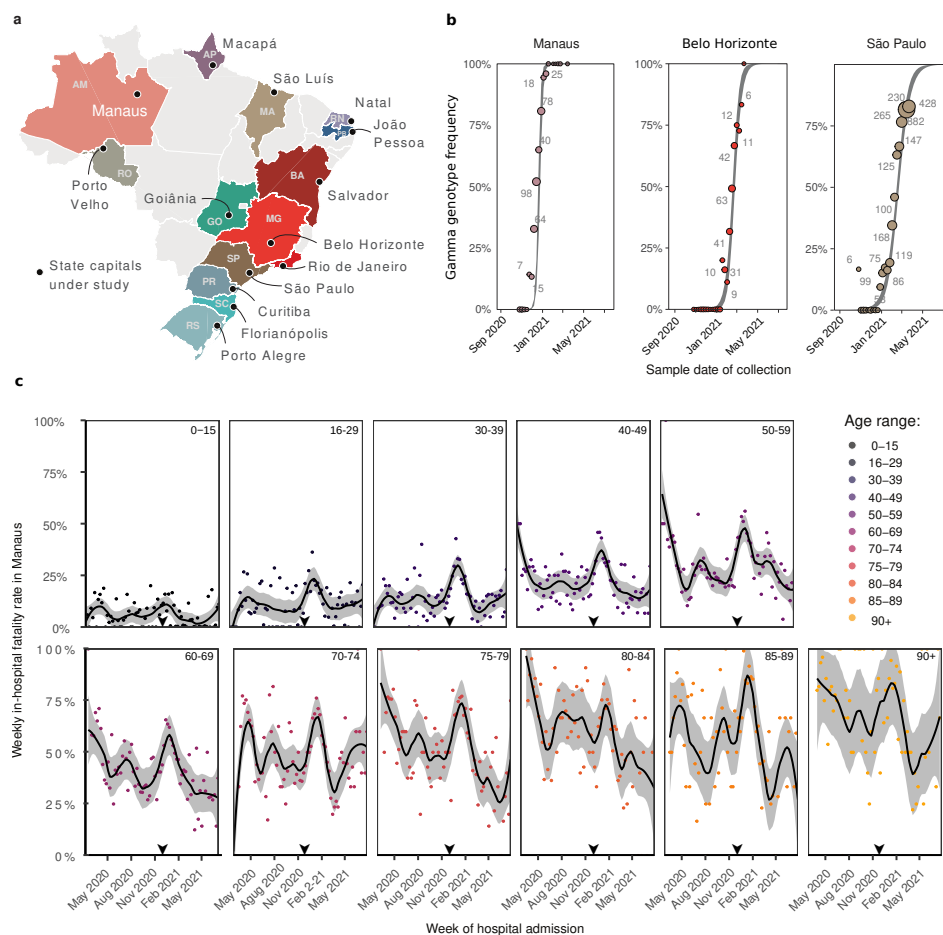


Figure 4.6: Spatio-temporal expansion of SARS-CoV-2 Gamma in Brazil and associated shocks in COVID-19 fatality rates in hospitals. (a) The 14 states and state capitals in which Gamma was detected by March 31, 2021 and which were included in the analysis (b) Time evolution of SARS-CoV-2 Gamma variant frequencies in three locations, suggesting rapid expansion. Data from GISAID [Shu and McCauley, 2017] (dots) are shown along with the number of sequenced SARS-CoV-2 samples (text), and posterior median model fits (line) and associated 95% credible intervals (CrI) (grey ribbon). (c) Weekly COVID-19 in-hospital fatality rates among hospitalised residents in Manaus with no evidence of vaccination prior to admission (dots), by age group (facets). Non-parametric loess mean estimates of time trends are shown as block solid line along with 95% confidence intervals as grey ribbon. The date of Gamma's first detection is indicated as grey dotted vertical line. This figure is included in the accepted manuscript provided to *Nature Medicine* for further editing.

João Pessoa, Macapá and Rio de Janeiro where they occurred before first detection of Gamma. Overall, rates tended to be highest in the North, Northeast, and Center-West regions of Brazil.

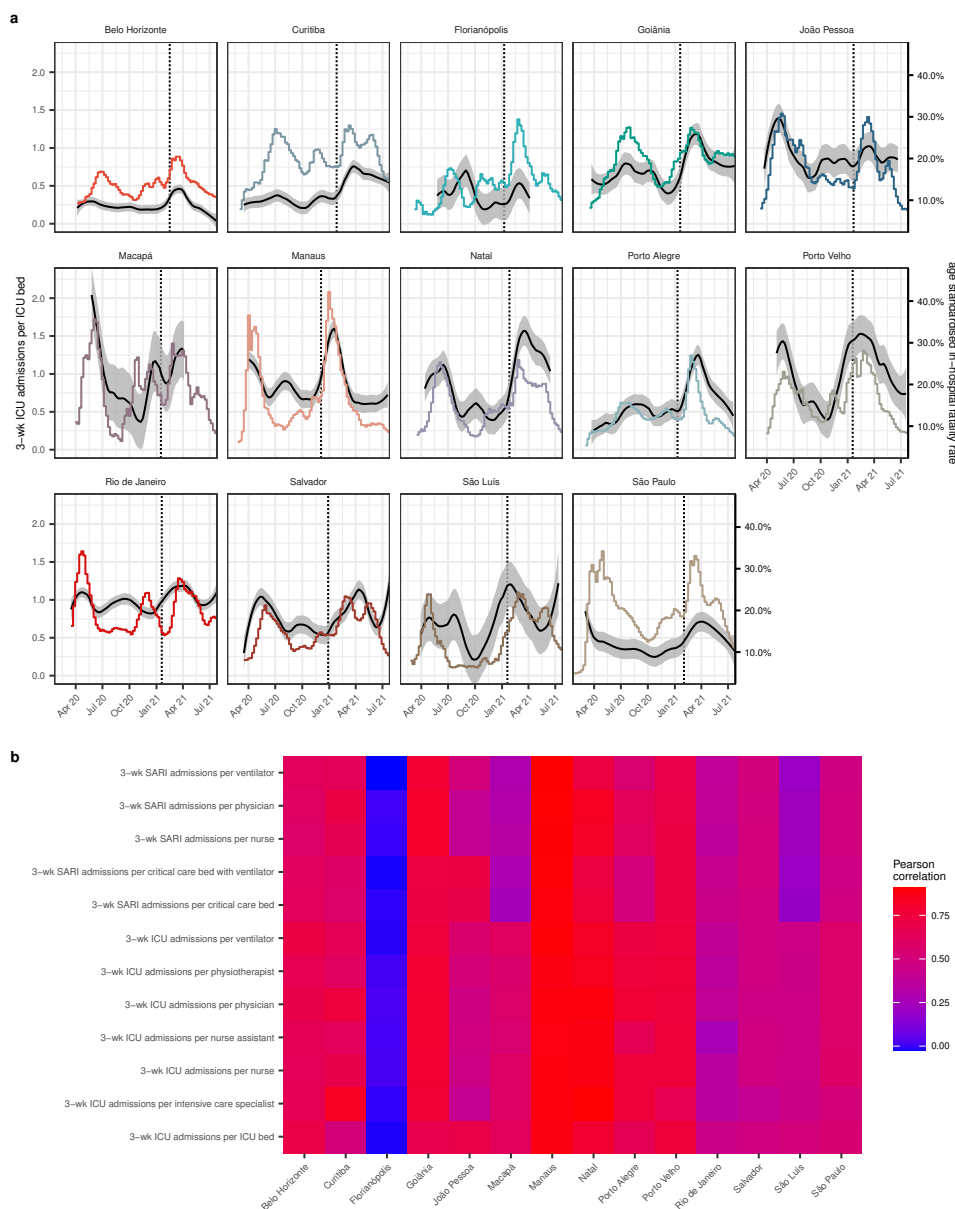


Figure 4.7: Time trends in age-standardised COVID-19 in-hospital fatality rates and pandemic healthcare pressure. (a) Non-parametric median estimates (line) and 95% CI (ribbons) of age-standardised COVID-19 in-hospital fatality rates (black, right hand side axis) are shown against the healthcare pressure index of ICU admissions over three weeks per available ICU bed in each city (colour, left hand side axis). The date of Gamma's first detection is added as a vertical dashed line. (b) Heatmap of Pearson correlation coefficients between age-standardised in-hospital fatality rates and each pandemic healthcare pressure index. This figure is included in the accepted manuscript provided to *Nature Medicine* for further editing.

4.4.2 In-hospital fatality rates are tracked by healthcare pressure indices

We find large inequities in healthcare resources across Brazil. In March 2020 the number of available ventilators per 100,000 population ranged from 21.7 in Macapá to 102 in

Porto Alegre, and the number of physicians per 100,000 population ranged from 124 in Macapá to 631 in Belo Horizonte (Table A.6). From these data we constructed healthcare pressure indices that capture changes in hospital demand per available resource, with demand comprising all hospitalised patients with severe acute respiratory infection (SARI), including non-residents and individuals with vaccine breakthrough infections. One such index, the moving sum of ICU admissions over three weeks per ICU bed is shown in Fig. 4.7a, and four further indices are shown in Fig. A.1-Fig. A.2. We find that all healthcare pressure indices are strongly correlated with the age-standardised, weekly COVID-19 in-hospital fatality rates in most cities (Fig. 4.7b). The only clear exception is Florianópolis, where the first peak in in-hospital fatality rates lagged the peak in healthcare pressure indices by more than one month, while the second peak in the time series coincided. This explains why no single lag led to high cross-correlation measures in this city (Fig. A.11). There may be different potential explanations for this anomaly, for instance an investment in healthcare resources during the first wave that was not immediately effective.

4.4.3 In-hospital fatality rates are mostly driven by regional and healthcare inequities

Fig. 4.8a compares the fitted age-standardised COVID-19 in-hospital fatality rates across Brazil's macroregions, the North, Northeast, Central-West, Southeast and South, revealing considerable geographical heterogeneity. Before first detection of Gamma in each location, the fitted age-standardised in-hospital fatality rate was lowest in Belo Horizonte and highest in Rio de Janeiro (shown as dotted horizontal lines in Fig. 4.8a). The high estimates for Rio de Janeiro prompted us to compare excess deaths derived from Brazilian's Civil Registry to the COVID-19 attributable in hospital deaths, which suggested that a smaller than expected proportion of hospitalised patients with unreported clinical outcomes may have died. Yet, Rio de Janeiro's age-standardised COVID-19 fatality rates remained the highest even when we assume that all patients with unreported clinical outcomes were successfully treated and survived COVID-19 (Fig. A.4).

Fig. 4.8b-d compare the estimated location, Gamma and healthcare pressure contributions to in-hospital fatality rates (Section 4.3.4). We find that the marked increase in COVID-19 in-hospital fatality rates is better explained by changes in healthcare pressures rather than a direct effect of Gamma on fatality rates in hospitalized patients. This result is inferred from three observations. First, the empirical in-hospital fatality rates also fluctuated prior to the detection of Gamma, second the rates largely decreased after the initial shock post-Gamma detection, and third the rates were strongly associated with the time-varying pandemic healthcare pressure indices (Fig. A.3). Relative to Belo

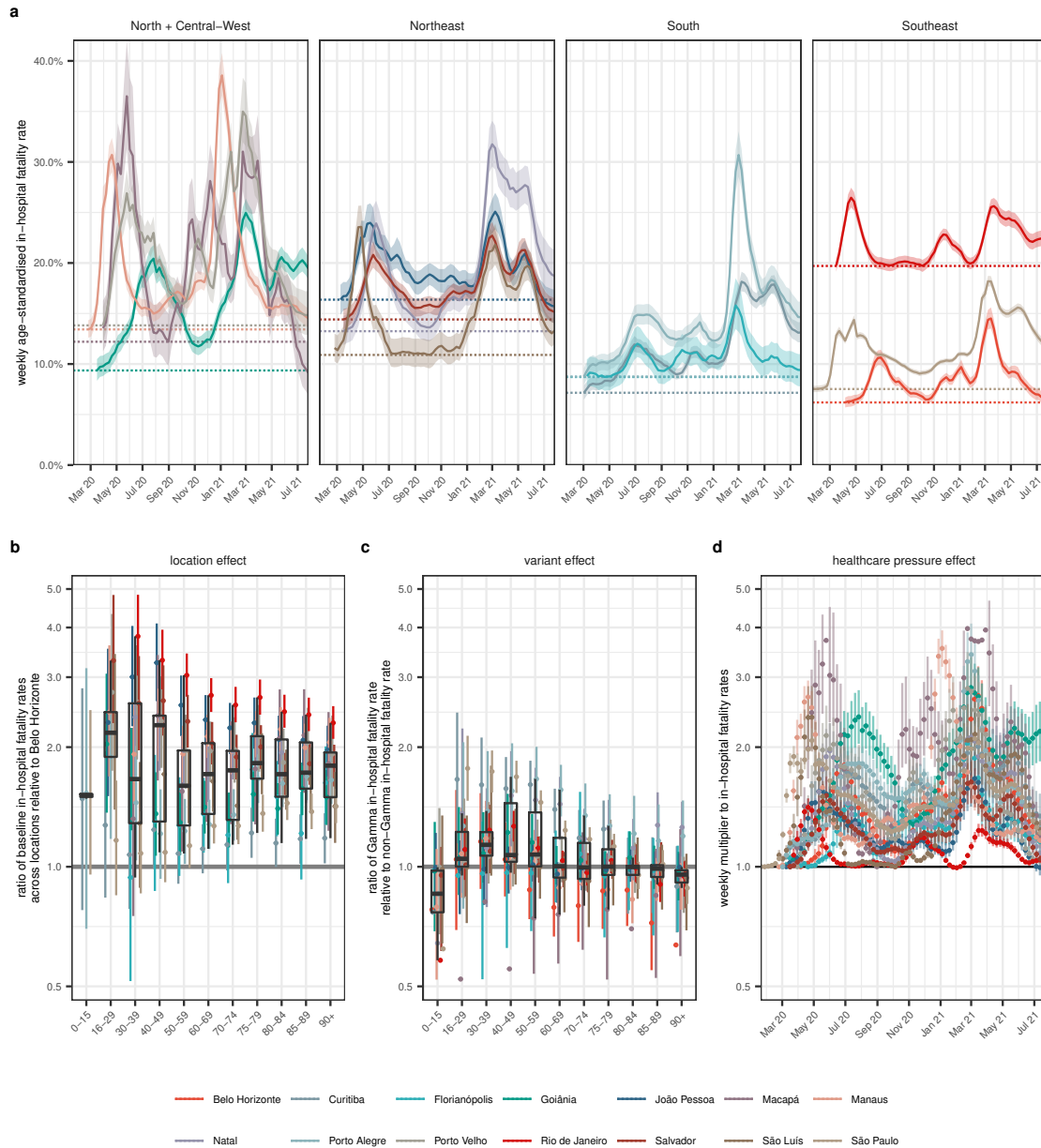


Figure 4.8: Estimated contribution of location effects, Gamma's infection severity, and pandemic healthcare pressure on COVID-19 in-hospital fatality rates. (a) Estimated, weekly age-standardised COVID-19 in-hospital fatality rates, averaged across SARS-CoV-2 variants. Posterior median estimates (line) are shown with 95%CrIs (ribbon), and the lowest estimated fatality rates prior to Gamma's detection in each state capital (dotted horizontal line). (b) Estimated ratio in lowest in-hospital fatality rates in each location relative to that seen in Belo Horizonte. (c) Estimated ratio in in-hospital fatality rates for Gamma versus non-Gamma lineages. (d) Estimated multiplier to the lowest age-standardised fatality rates prior to Gamma's detection shown in subplot a, that is associated with the pandemic healthcare pressure indices. In each plot, posterior median estimates are shown as dots, 95%CrIs as linerange, and summaries across locations as boxplots. Multipliers and ratios in panels b-d are reported on a logarithmic scale. This figure is included in the accepted manuscript provided to *Nature Medicine* for further editing.

Horizonte, the lowest fatality rates were an estimated 1.95-fold (95% CrI: 1.11-3.25) higher in all other state capitals, a location effect that was consistent across all age groups. Although our model does not provide insights into the exact drivers of this location effect, it is likely due to a combination of socio-economic factors unique to each city, including pre-existing inequities in the availability, quality, and accessibility of healthcare [Castro et al., 2021b, de Souza Santos et al., 2021]. At peak times, pandemic healthcare pressures as measured by our indices were associated with a 2.33-fold (95% CrI: 1.39-4.27) multiplicative effect across the 14 cities, and this contribution was stronger in places that already had high in-hospital fatality rates at baseline. Gamma was not associated with a statistically significant effect on in-hospital fatality rates after controlling for healthcare pressures (posterior p-value of $p=0.14$ that Gamma-specific in-hospital disease severity is lower or equal to that of non-Gamma variants in patients aged 30-39 compared to non-Gamma variants of SARS-CoV-2, and larger those younger than 30 years and older than 39 years).

4.4.4 Number of deaths attributable to resource limitations

To quantify the impact that the observed fluctuations in COVID-19 attributable in-hospital rates had on the death toll in the 14 state capitals, we considered counterfactual simulations where infection waves and resource limitations did not result in surging in-hospital fatality rates. We consider this counterfactual conservative as it assumes for each location achievable in-hospital fatality rates that are based on the lowest values observed prior to first detection of Gamma in each state capital, and thus prior to improvements in clinical management, patient triage, treatments (e. g. dexamethasone). Across the 14 state capitals, we find deaths could have been, in the absence of pandemic resource limitations, reduced by an estimated 29.8% (95 CrI: 28.7%-30.9%) (Table 4.1).

Using additional counterfactual simulations, we further evaluated how many in-hospital deaths could have been avoided in the absence of regional inequities and pandemic resource limitations, i. e. if all 14 state capitals would have experienced throughout the pandemic the lowest observed in-hospital fatality rates across all 14 cities, which were found in Belo Horizonte. We consider these projections less conservative, since the observed, low in-hospital fatality rates in Belo Horizonte may reflect a large range of factors that may not be translatable to other locations in Brazil. We find in-hospital deaths could have been reduced by an estimated 57.1% (95% CrI: 54.3-59.9%) in the 14 state capitals (Table 4.1)

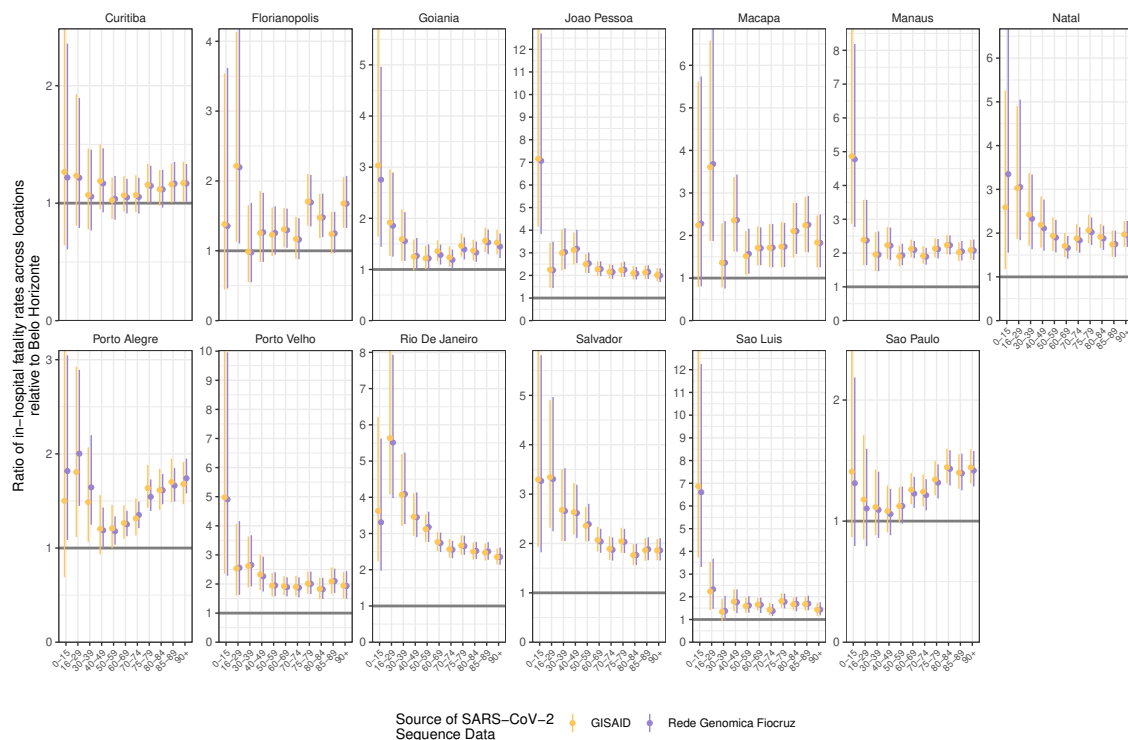


Figure 4.9: Estimated effect of location on in-hospital fatality rates when using sequence data from Rede Genômica FioCruz. Posterior median estimates (dots) are shown along with 95% credible intervals (liners) for the central analysis (yellow, $n=7,221$ samples) and sensitivity analysis (purple, $n=17,070$ samples).

4.5 Model validation and sensitivity analyses

Concerns on SARS-CoV-2 sequence sample representativeness [WHO, 2021] prompted us to re-evaluate our findings using monthly state-level variant frequency data reported by Rede Genômica FioCruz [Rede Genômica FioCruz, 2021]. Data on monthly Gamma variant frequencies in each Brazilian federal unit or state between November 2020 and May 2021 were retrieved from the dashboard on June 15, 2021. Again, we assumed that the variant frequencies reported at state level were representative of the variant frequencies in state capitals. We also considered weekly city-level Gamma variant frequency data from Manaus, Belo Horizonte, and São Paulo obtained under controlled sampling frames. In Manaus, samples from PCR-positive residents testing in two private laboratories through nasal and oropharyngeal swabs were selected at random for sequencing regardless of Ct values. The samples were sequenced and processed using the ARTIC bioinformatics pipeline [artic-network, 2021] as described in Faria et al. [2021]. Viral genomes recovered from 147 samples collected between November 1, 2020 and January 10, 2021 had sufficient genome coverage enabling lineage classification with *pangolin* version 2.2.1 [O’Toole et al., 2021, Rambaut et al., 2020].

We found only minor differences in our primary findings depending on what SARS-CoV-2 variant frequency data were used (Fig. 4.9 and Figs. A.5 to A.7)

Given long-term underfunding of Brazil's public healthcare system, we investigated if geographic inequities could be modulated by different proportions of patients in public and private hospitals, and if the strong fluctuations in COVID-19 in-hospital fatality rates are also observed in private hospitals. For São Paulo, we could classify most hospitals as either private or public. We found consistently lower in-hospital fatality rates in private hospitals that however fluctuated in synchrony with those in public hospitals, in line with our primary findings (Fig. A.8).

In Belo Horizonte, the age-standardised COVID-19 in-hospital fatality rates declined over the summer months of 2021 to levels well below those seen during earlier time periods (Fig. 4.8a). We hypothesised that larger numbers of patients hospitalised in Belo Horizonte may have been vaccinated, with no vaccination record reported to SIVEP-Gripe. To address this possibility, we performed a sub-analysis in which we excluded patients with unreported vaccination status in the SIVEP-Gripe data set (Fig. A.10). In the figure, fatality rate estimates are not shown for weeks in which patient denominators were too small. In this sub-analysis, we found larger discrepancies in fatality rates compared to the main analysis only for Belo Horizonte, suggesting that missing data on vaccination status has likely no substantial impact on our overall findings. However for Belo Horizonte, it is possible that our in-hospital fatality rate estimates are confounded with unreported vaccination status, which in turn would explain the estimated negative Gamma effects in this city (Fig. 4.8c).

The highest age-standardised COVID-19 in-hospital fatality rates were observed in Rio de Janeiro, against the national trend of declining rates in Brazil's South and Southeast macroregions. It is possible that smaller proportions of hospitalised patients with unreported clinical outcomes may have died, and analyses were repeated assuming that all patients with unreported clinical outcomes survived. We found in comparison to other cities, that Rio de Janeiro's baseline in-hospital fatality rates were in this sensitivity analysis considerably lower than in the main analysis, but remained overall largest and showed the same strong fluctuations as in the main analysis (Fig. A.4).

In-hospital fatality rates also depend on who, and under what circumstances, severely ill patients are admitted to hospitals. This prompted us to investigate if the observed fluctuations in COVID-19 fatality rates could in part be the result of concomitant changes in the profile of admitted patients. There is limited data on disease severity at time of hospital admission available in SIVEP-Gripe, however one indicator that can be readily calculated is the time between hospital admission and death in patients with a fatal outcome (Fig. A.9). In Macapá, Manaus, Porto Alegre, and Porto Velho, we find sub-

stantially shorter times to death when the number of **COVID-19** attributable hospital admissions peaked, suggesting admitted patients may already have been at a more severe clinical stage when admitted, or that during times of peak demand healthcare pressure in hospitals both increased fatality rates and led to faster progression to death. These data suggest that increased healthcare pressure likely acts to shape in-hospital fatality rates through distinct mechanisms, through a combination of both a reduced ability to provide adequate care and an increase in the average severity of admitted patients.

4.6 Discussion

In this Chapter, we describe temporal fluctuations in **COVID-19** in-hospital fatality rates across 14 Brazilian state capitals. In summary, we found that the impact of Gamma in Brazil's hospital fatality rates has been mainly indirect, influenced by pre-existing geographic inequities and pandemic-related healthcare challenges. Through counterfactual analyses, we additionally estimated that around half of Brazil's **COVID-19** hospital deaths could have been prevented without these factors, highlighting the importance of addressing healthcare inequities and ensuring adequate healthcare resources to reduce **COVID-19** mortality.

Our multi-strain Bayesian model used inferred variant replacement dynamics to estimate the proportion of **COVID-19** hospitalisations with either variant. This modelling approach is not necessary in other studies conducted in **HICs** aimed at comparing variant-specific severities, for a couple of reasons. Firstly, as Brazilian hospital records were not linked to case surveillance data, we could not compute the time delay between case reporting to hospitalisation, nor the fraction of cases that were hospitalised. As such survival analyses estimating hazard ratios were impossible [Twohig et al., 2022]. Secondly, as we did not have available data on the infecting variant of patients, it was impossible to compare hospital outcomes across different groups of individuals through, for example, logistic regression [Funk et al., 2021]. Further, such analyses typically account for a variety of comorbidities to adjust for confounding. In the case of the SIVEP data, comorbidities were sparsely reported, and probably unreliable for such analyses.

The findings presented in this chapter should be considered in the context of the following limitations. First, we here delineate fluctuations in **COVID-19** in-hospital fatality rates at city level to focus on the extensive spatio-temporal heterogeneity in fatality rates across Brazil. We recognise this broader geographic focus masks important differences within cities, and thus we cannot identify the exact factors determining the substantial location effects that we measure. However, these are likely related to (i) differences in catchment populations as vulnerable populations with poor healthcare

access are highly clustered across Brazil's largest cities [Pereira et al., 2021, Nicolelis et al., 2021], (ii) underfunding of the public healthcare system [Castro et al., 2019, Funcia, 2019] and emerging discrepancies in healthcare resources compared to private hospitals [Ranzani et al., 2021a, Castro et al., 2021b, Nicolelis et al., 2021, Noronha et al., 2020], or (iii) inequities in the quality and capabilities of healthcare systems which are well-documented, for example, for pre-pandemic sepsis survival rates [Machado et al., 2017]. Additional relevant factors could include (iv) differences in demand reflecting variation in epidemic magnitude [Castro et al., 2021b], and (v) the timing and extent of public health measures aimed at controlling spread and preventing infection in vulnerable groups [de Albuquerque and Ribeiro, 2021, de Souza Santos et al., 2021, Rocha et al., 2021]. As such, the location effects could also reflect healthcare pressures that are present already at the lowest incidences of COVID-19 hospital admissions in each location. Importantly, we observe substantial fluctuations in in-hospital fatality rates even in private hospitals of São Paulo city (Fig. A.8), suggesting that large effects of healthcare pressure on in-hospital fatality rates are common across Brazil.

Second, the pandemic healthcare pressure indices that we derive are based on data reported to CNES, which does not capture all resource limitations such as the acute shortages in oxygen supply that were experienced in January 2021 in Manaus [Malta et al., 2021], and is prone to potential reporting biases [Rocha et al., 2018]. In our view, the inferred associations between healthcare pressure indices and fatality rates demonstrate that where resources are limited, real-time monitoring of available resources is especially important to identify critical resource limitations and avoid the lamentable shocks in death rates that we describe for many state capitals across Brazil.

Third, in characterising hospital fatality rates in a wide Brazilian context, we are ignoring historical policy aspects that can be a source of heterogeneity. The standard of healthcare per unit staff member can be improved by federal coordinated actions and policies on training, as well as unambiguous governmental medical policies to prevent disinformation [Leite et al., 2021]. These federal policies also extend to financing to ensure that the underlying structure of primary care and emergency services all present a standardised quality of care [Mendes et al., 2018].

Finally, our analyses start with hospitalisation, which is a limitation because in-hospital fatality rates also depend on who, and under what circumstances, severely ill patients are admitted to hospitals. Thus, it is possible that healthcare demand was amplified due to an increased risk of hospitalisation following infection with a variant of concern with higher infection severity, as has been shown for SARS-CoV-2 Alpha [Sheikh et al., 2021, Grint et al., 2021]. We also find evidence that in several locations, the time to death following admission tended to be shorter during peak demand (Fig. A.9). These observa-

tions indicate that increased healthcare pressure acts to shape in-hospital fatality rates through distinct mechanisms, likely through a combination of both the reduced ability to provide adequate care and an increase in the average severity of admitted patients (with only the most severely ill admitted during periods of highest pressure). In this context, if less severe COVID-19 cases could also have been admitted, we expect that the estimated proportion of avoidable COVID-19 in-hospital deaths in the absence of healthcare pressure effects would be lower than what is reported here. At the same time, the projected numbers are likely an underestimate of COVID-19 deaths that could have been avoided in the absence of healthcare pressure effects because we did not account for deaths in severely ill individuals that were not cared for in hospitals.

5

PHYLOGENETIC ESTIMATION OF HIV TIME SINCE INFECTION: SOFTWARE AND APPLICATION IN UGANDA

5.1 Introduction

HIV transmission dynamics are complex, and highly structured. That is, the risk of HIV transmission depends on the characteristics of individuals at risk and those already infected, such as patterns in sexual behaviour [Anderson et al., 1991], as well as the varying prevalence of HIV across different age groups, genders, and locations [Dwyer-Lindgren et al., 2019]. Understanding which population subgroups are over-represented in transmission events allows for tailored interventions to individuals who most need them. In turn, interventions affect transmission dynamics and potentially change the subpopulations at highest need of interventions. This creates a feedback loop in which transmission patterns need to be routinely analysed to reflect the changing dynamics introduced by epidemiological interventions.

Scientific methods aiming to describe transmissions need to overcome a variety of data issues. Due to long incubation periods, infrequent testing, and stigma associated to HIV, it is practically infeasible to 1) observe all HIV infections occurring within a population, 2) accurately date infections, and 3) determine with certainty the source of infection for a given case. These challenges are particularly pronounced in the context of generalized epidemics, such as in Africa, where tracking all possible sources of infections is infeasible due to the higher prevalence of HIV. In this context, phylogenetic approaches emerge as a natural solution, as they aim to reconstruct the genetic history of related viral

sequences to approximate transmission history [Romero-Severson et al., 2014]. For example, these methods have been used to detect linkage among PLHIV [Campbell et al., 2011], directionality of transmission [Romero-Severson et al., 2016], and to describe the geographic movement and history of the virus [Faria et al., 2014]. A comprehensive review of phylogenetic approaches with applications to HIV epidemiology is reported in Grabowski and Redd [2014].

Phylogenetic analyses are tailored to, and are limited by, the sequencing technologies available. *Sanger* sequencing [Sanger et al., 1977] was the first commercially available method to summarize the viral diversity in an HIV infected host through a single *consensus* sequence, which reports the most common nucleotide found at each position in a sequence alignment. Although phylogenies of consensus sequences aren't sufficient to infer direction of transmission [Volz and Frost, 2013, Romero-Severson et al., 2014], they are affected by transmission dynamics [Frost and Volz, 2013, Colijn and Gardy, 2014] and were used to estimate “global” epidemiological parameters, such as reproduction number [Volz et al., 2009]. A prominent approach in modelling structured transmission dynamics involved positing a compartmental transmission model parametrized by systems of ODEs [Volz, 2012] and fitting it to a reconstructed phylogeny. In Volz et al. [2009] these two steps are performed sequentially for computation ease, while Rasmussen et al. [2014] jointly infer epidemiological parameters and phylogenies.

The advent of *Next-Generation Sequencing* (NGS) technologies, which produce multiple viral variants reads for a given sample, widely increased the scope of phylogenetic analyses. The unprecedented resolution of within-host viral diversity made it possible to address the challenge of inferring transmission direction. In fact, transmission direction may be reflected in the topology of a phylogeny [Romero-Severson et al., 2016], specifically in the configuration of the clades containing the viral reads of two different hosts. The impact of these advancements were profound: for example, confirmatory direction analyses were essential in proving the effectiveness of ART in two clinical trials [Eshleman et al., 2011, Campbell et al., 2011].

From a statistical modelling perspective, having access to phylogenetically likely source-recipient pairs [Ratmann et al., 2019] allows epidemiological models to be fitted to *partially observed transmission networks*. This approach differs to fitting models to entire phylogenies, which are generally more complex objects to interpret and model. The benefits of the source-recipient approach include ease in computational complexity, enhanced flexibility and transparency when specifying how transmission dynamics affect the modelled object, and the ability to account for selection bias [Xi, 2022]. In practice, these models have been used to describe transmission flows between different

population subgroups, for example by location [Ratmann et al., 2020, Bbosa et al., 2020], age and gender [Xi et al., 2022, Magosi et al., 2022, Hall et al., 2024, Bu et al., 2024].

However restricting analyses to likely source-recipient pairs comes with limitations. For example it is not possible to infer the exact date of a transmission event without external sources of information. As a consequence, the previously cited studies typically assume constant transmission dynamics, under the assumption that the observed infections happened in a short time frame relative to the time scale of the epidemic. As such, studies based on partially observed transmission networks are better suited at describing recent transmission dynamics with high resolution, as opposed to identifying long-term trends, as for example in phylogeographic studies [Faria et al., 2014].

In this Chapter, we present some of the work included in the article “Longitudinal population-level HIV epidemiologic and genomic surveillance highlights growing gender disparity of HIV transmission in Uganda”, published in *Nature Microbiology*. Here, we used recent phylogenetic methods to estimate the *Time Since Infection (TSI)* [Golubchik et al., 2022] of recipients in observed transmission pairs and subsequently infer temporal changes in HIV transmission dynamics in South Western Uganda. The contributions of the authors of the papers is acknowledged thereafter. Oliver Ratmann, my supervisor, and M. Kate Grabowski conceived the analysis. Mélodie Monod is the first author of the paper, she is responsible for the statistical modelling summarised in Section 5.3.5. She coded most of Fig. 5.5 using the time-since infection estimates of my model, and I additionally modified the figure to satisfy publication requirements. Further, she coded Fig. 5.6. I, Andrea Brizzi, am the second author of the paper. I coded the phylogenetics pipeline for time since infection in Section 5.3.3, ran it on the RCCS transmission cohort data (Section 5.3.1), matched the outputs against the source-recipient pair outputs in Section 5.3.2, and developed the infection dating model in Section 5.3.4. In Section 5.3.2, I modified how the direction of transmission scores are typically computed ([Ratmann et al., 2019]) to account for serohistory data. I took over the main project after Mélodie Monod completed her PhD, contributing to the code for preprocessing, postprocessing, visualisation, and computation of selection probabilities in Section 5.3.5. Ronald M. Galiwango, Robert Ssekubugu, Edward Nelson Kankaka, Victor Ssempijja, Lucie Abeler-Dörner, David Bonsall, Larry W. Chang, Christophe Fraser, Tanya Golubchik, Ronald H. Gray, Julio Croda, Godfrey Kigozi, Lisa A. Mills, Oliver Laeyendecker, Thomas C. Quinn, Steven J. Reynolds, John Santelli, Nelson K. Sewankambo, Joseph Ssekasanvu, Laura Thomson, Maria J. Wawer, David Serwadda, Joseph Kagaayi and M. Kate Grabowski oversaw and performed data collection (Section 5.3.1). Xiaoyue Xi contributed to Section 5.3.2. Specifically, she wrote and run the phylogenetic pipeline to identify potential networks and confirm source-recipient pairs.

This Chapter is structured as follows. Section 5.2 summarises my main contributions on the paper. Section 5.3 describes the sequence data, the phylogenetic pipelines and the statistical models used for the analysis. Section 5.4 presents our main findings and Section 5.5 includes a discussion of the limitations and potential research direction.

5.2 Contributions to the literature

In this Chapter, we describe the pipeline I coded up to estimate HIV TSI from sequence data, and a method to use these estimates to date transmission events which are consistent with individual-level testing histories and known transmission networks. To the best of our knowledge, this is the only open-source pipeline that takes individual-level NGS data and outputs TSI estimates. From this point of view, we empower a wider range of researches to apply HIV-phylo-TSI, Golubchik et al. [2022] for their own analyses. We then describe the implications of dating infection events on the statistical modelling of partially observed transmission networks as well as practical application on surveillance data from the RCCS. Importantly, we reported temporal changes in transmission dynamics in South Western Uganda, consisting of a shift in the age-profile of sources of HIV transmission.

A major novelty of this study consists in the comparison of HIV transmission flows over time. This was made possible by dating the transmission events, and successively fitting the same epidemiological model to different time windows. In particular, our work is an extension of the framework described in Xi et al. [2022], where age at infection is implicitly defined as the age at the midpoint of the study period, possibly yielding bias and inconsistencies. For instance, if most transmission events occurred far into the past, the model would overestimate the contributions to HIV transmission of older age groups. Our approach, presented in Section 5.3.4, not only allows estimation of age at infection which is consistent with individual-level testing histories, but also allows to impute other characteristics at the date of infection, such as location of residence. The code for these analyses is included in the Github repository <https://github.com/MLGlobalHealth/phyloSI-RakaiAgeGender> and the data made openly available at <https://zenodo.org/records/8412741>.

In the context of the African response and surveillance of the HIV epidemic, which makes strong use of sequence data and phylogenetic analyses through organizations such as *Phylogenetics And Networks for Generalised Epidemics in Africa* (PANGEA) [Pillay et al., 2015, Abeler-Dörner et al., 2019] and IAVI [Africa, 2024], our pipeline opens novel research opportunities to tackle important problems, such as the estimation of HIV incidence [Golubchik et al., 2022]. We made the code of our pipeline openly available at <https://github.com/abriz97/HIV-phyloTSI-workshops>, and together with PANGEA,

IAVI, and Dr. Moyo, we organized workshops in Nairobi and Gaborone to train local researchers on how to use it, with an example application on a small dataset from Zambia.

5.3 Methods

5.3.1 RCCS transmission cohort data 2010-2018

Within the RCCS, we performed population-based HIV deep-sequencing spanning survey rounds 14 to 18, that is from January 2010 to April 2018. The primary purpose of viral deep sequencing was to reconstruct transmission networks and identify the population-level sources of infections. The secondary purpose was to estimate the TSI for individuals living with HIV, providing information on the incidence of HIV infection in the population, and allowing for a comparison of transmission patterns over time.

The RCCS viral phylogenetic transmission cohort includes participants with HIV for whom at least one HIV deep sequence sample satisfying minimum quality criteria for deep-sequence phylogenetic analysis is available. For survey rounds 14 to 16 (PANGEA-HIV 1), viral sequencing was performed on plasma samples from participants with HIV who had no viral load measurement and self-reported being ART-naïve at the time of the survey, or who had a viral load measurement above 1,000 copies/mL. For survey rounds 17 to 18 (PANGEA-HIV 2), viral load measurements were available for all infected participants and viral sequencing was performed on plasma samples of individuals who had not yet been sequenced and who had a viral load measurement above 1,000 copies/mL.

We restricted our analysis to samples from 2,172 RCCS participants that satisfied minimum criteria on read length and depth for phylogeny reconstruction and subsequent inferences. We further included samples from the spatially neighbouring MRC/UVRI/LSHTM surveillance cohorts to enhance the genetic background and improve the resolution of the phylogenetic analyses.

For each participant in the transmission cohort, we had access to the birthdate, the date of first positive test, and sometimes the date of last negative test. These last two dates provided a range of possible infection dates, and are referred to as “serohistory” in the next sections.

5.3.2 Identification of plausible source-recipient pairs

The HIV deep-sequencing pipeline provided sequence fragments that capture viral diversity within individuals, which enables phylogenetic inference into the direction

of transmission from sequence data alone [Wymant et al., 2018b, Ratmann et al., 2019, Zhang et al., 2021]. First, potential transmission networks were identified, and in the second step transmission networks were confirmed and the transmission directions in the networks were characterized as possible.

In this study, the first step was modified from previous protocols [Ratmann et al., 2019] to ease computational burden, while the second step remained unchanged, and was performed with *phyloscanner*, using version 1.8.1. Specifically, deep-sequencing reads were assembled with the *shiver* sequence assembly software, version 1.5.7 [Wymant et al., 2018a]. Next, *phyloscanner* version 1.8.1 [Wymant et al., 2018b] was used to merge paired-end reads, and only merged reads of at least 250 bp in length were retained in order to generate 250bp deep-sequence alignments as established in earlier work Ratmann et al. [2019].

Phylogenetically likely source-recipient pairs with linkage scores ≥ 0.5 and direction scores ≥ 0.5 were extracted, and only the most likely source-recipient pair with highest linkage score was retained if multiple likely sources were identified for a particular recipient. The resulting source-recipient pairs were checked further against serohistory data from both individuals where available. If serohistory data indicated the opposite direction of transmission, the estimated likely direction of transmission was set to that indicated by serohistory data.

In total, we identified 539 source-recipient pairs that involved participants from the 36 survey communities and further individuals from the background dataset. In 13 of the 539 source-recipient pairs, available dates of last negative tests indicated that only the opposite transmission direction was possible and in these cases the inferred direction of transmission was set to the opposite direction. The resulting pairs included 501 unique recipient partners, and for each we retained the most likely transmitting partner. To identify pairs capturing transmission events within the RCCS inland communities, we restricted analysis initially to 236 heterosexual source-recipient pairs in whom both individuals were ever resident in the 36 survey communities. Of these, 142 pairs were from men to women and 94 from women to men. Infection times were estimated for all sampled individuals and refined for the recipient partners in the 236 heterosexual source-recipient pairs, as described in the following two subsections.

5.3.3 Phylogenetic pipeline for HIV time-since-infection

To evaluate changes in transmission dynamics through phylogenetic analyses, it is necessary to estimate when, and under which circumstances, transmission events occurred. Golubchik et al. [2022] showed that it is possible to estimate the TSI of individuals living with HIV using deep-sequencing data. Briefly, the authors trained a

Random Forest model, HIV-phyloTSI to predict TSI from a set of seroconverters with known serohistories, using outputs of phyloscanner [Wymant et al., 2018b] and minor allele frequencies as predictors. Our aim was thus to integrate the predictions from HIV-phyloTSI into our source-recipient pair analysis to date transmission events.

Even though the code for HIV-phyloTSI is publicly available, the authors did not provide code to obtain the necessary inputs to run the model. The reasons are twofold. Firstly, NGS phylogenetic analyses require massive parallel computations, as alignment and tree construction are computed across multiple windows of the HIV genome. Parallel computations are typically performed on high-performance computing clusters, which are not always available to all researchers, and may differ in their configuration. Secondly, the authors claim the algorithm to be robust to small differences in the arguments used to generate the input data. These arguments may include the number of sequences used, and arguments controlling the behaviour of alignment programs, tree construction programs, and phyloscanner.

In this section, we describe our own implementation of a phylogenetic pipeline which starts from individual-level deep-sequence data and ends with TSI estimates. The pipeline described here is based on the one outlined in Section 3 of Xi [2022] to find source-recipient pairs through phyloscanner [Wymant et al., 2018b], but with important modifications in functionality and improvements in automation. Regarding functionality, finding source-recipient pairs requires the construction of a “potential transmission network”. This initial step is not necessary for TSI estimation, and is therefore omitted. Also the last steps differ, as our pipeline terminates with the application the HIV-phyloTSI algorithm [Golubchik et al., 2022]. Regarding automation, our software performs intermediate output checks and, after completion of one step, it automatically launches the following. This functionality is achieved through a bash script, `runall_TSI.sh` which acts a controller, and a set of R scripts that are able to interact with the controller and queue follow-up steps. Overall, these changes represent important improvements over the original pipeline, which required manual intervention at each step.

A brief skeleton of the pipeline is reported in Fig. 5.1, which demonstrates how different steps are linked. Note that these steps are not subcomponents of a model, and do not include any parameter tuning or quantification of uncertainty. As in [Wymant et al., 2018b, Ratmann et al., 2019], robustness to phylogenetic signal is achieved by averaging results over hundreds of windows across each gene.

The following subsections describe each of the reported steps in more detail.

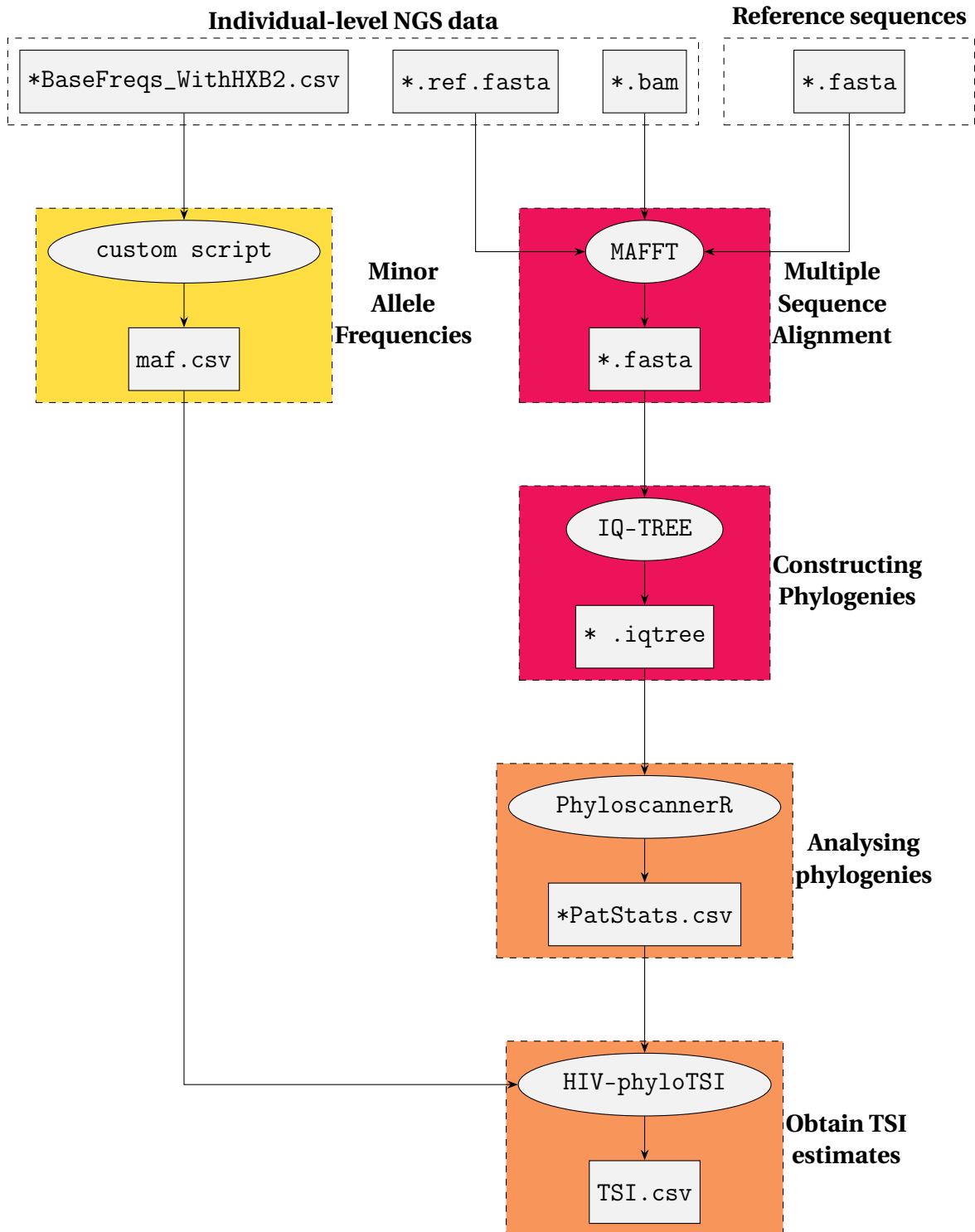


Figure 5.1: Pipeline for estimating TSI from NGS data. The programs (in ellipses) and the data (in rectangles) are shown together with arrows indicating the order of execution. Steps run for each group and window are shown in red, steps run for each group are shown in orange and steps run once for the entire dataset are shown in yellow.

Grouping of sequences

NGS phylogenetic analyses can be extremely computationally intensive, due to the large number of sequences that need to be aligned and compared. As such, it is often necessary to partition the set of all sequences available for analysis into smaller groups, perform group-level analyses and then aggregate the results at the end. This is the case for most phylogenetic analyses built on top `phyloscanner`, as this program considers phylogenies built on multiple overlapping windows of the **HIV** genome. For instance, [Xi \[2022\]](#) uses a clustering algorithm that groups sequences into potential transmission networks, and suggests avoiding clusters containing data for more than 50 individuals. Groups of 50 sequences per groups are consistent with [Golubchik et al. \[2022\]](#) suggestions, and therefore we used this number in our pipeline. Further, only one sample per individual is used. We selected the least recent one to minimize the risk of obtaining a **TSI** estimate inconsistent with the known date of first positive test.

Multiple Sequence Alignment

A **Multiple Sequence Alignment (MSA)** is an arrangement of a set of nucleotide sequences in which each column of the alignment represents homologous (evolutionarily related) positions across the sequences. We perform **MSAs** at a group-window level exploiting `phyloscanner_make_trees.py` script, which in turns calls MAFFT [\[Kato, 2002\]](#). Briefly, we divide the **HIV** genome into overlapping windows of 250 bp, positioned at a distance of 25bp. Then, for each group, and if possible, we align the sequences in each window using MAFFT [\[Kato, 2002\]](#) with flags `-globalpair -maxiterate 1000`. When applied to the RCCS transmission cohort, single alignments required up to 1 day to complete, even though the majority were finished in less than 8 hours. However the alignment step as a whole lasted 6 days due to bottlenecks in the throughput of the cluster, which meant only 500 windows could be aligned in parallel.

Phylogenetic trees construction

Once the deep-sequences are aligned, it is possible to construct phylogenies representing the evolutionary relationships between them. For each group and window, phylogenetic trees are constructed from the **MSAs** using IQ-TREE [\[Nguyen et al., 2015\]](#) with a generalised time reversible model with FreeRate with six categories of rate variation (i.e. GTR+F+R6 model). In the RCCS case study, a single tree reconstruction required up to 3 days to complete, even though the majority required less than 8 hours. However, due to throughput bottlenecks, the entire step lasted 4 weeks. The following steps of the pipeline were much shorter, requiring just a few hours.

Phylogenetic trees analyses

Phylogenetic trees are then processed and summarised through phyloscanner [Wymant et al., 2018b], using the same settings as reported by Golubchik et al. [2022], namely: `-sks -ow -rda -swt 0.5 -amt -sat 0.33 -rcm -blr -pbk 15 -rtt 0.005 -rwt 3 -m 1E-5`. Among the outputs, we only require the `*PatStats.csv` files, which contain summary statistics for each window.

Minor Allele Frequencies

Minor Allele Frequency (MAF) calculations are an extra step beyond the source-recipient pairs pipeline outlined in Xi [2022]. MAF quantifies within-host diversity [Carlisle et al., 2019, Kouyos et al., 2011] by assessing, for each sequence and position, the fraction of observed nucleotides that differ from the predominant base. Our calculations were performed within a custom script, specifically through the R function `generate.sample`.

TSI estimation

The final step of the pipeline involves the estimation of individual-level TSI estimates based on the intermediate outputs produced by the previous steps. Briefly, summary statistics are extracted from the `PatStats.csv` files obtained with `phyloscannerR` and from the `maf.csv` file. These are then aggregated by calculating averages over all windows across different genes and they are successively fed to HIV-phyloTSI. Short descriptions of the predictors used are reported in Table 5.1.

5.3.4 Estimation of infection data for source-recipient pairs

As described in the previous subsection, HIV-phyloTSI's TSI estimates are obtained using exclusively individual-level deep-sequence data. This raises important considerations when dating transmission events, i. e. infection time of the recipient in the source recipient pairs. Briefly:

- Although phylogenetic signal can be strong, there is wide uncertainty in the estimates obtained from the Random Forest model, which are characterised by potentially very wide prediction intervals.
- As serohistory is not used as a model predictor, some TSI estimates may be inconsistent with the known serohistory: the algorithm may infer an infection date which predated the last negative test.
- Finally, the TSI estimates may be inconsistent with the inferred direction of transmission of a source-recipient pair: the algorithm may infer the recipient's infection date to predate the source's infection date.

Predictor	Description	Genes	Codes
Number of Tips	Number of viral sequences in the phylogenetic tree coming from the same sample.	gag gp41 gp120	gag_tips gp41_tips gp120_tips
Largest Subgraph Root to Tip Distance	Longest distance from the root of the tree to any of the terminal nodes (tips), calculated on the largest subgraph obtained after excluding the tips from other samples. This distance measures the maximum genetic divergence between the ancestral state at the root of the subtree and its most divergent descendant sequences.	gag pol gp120	gag_lrtt pol_lrtt gp120_lrtt
Minor Allele Frequencies	The fraction of observed nucleotides that differ from the predominant base at each position in the alignment. It is a measure of cumulated genetic divergence over time. The predictor name is differently suffixed depending on whether this frequency is computed using nucleotides on the first two bases (12c) or the base (3c) of each codon	gag gp41	gag_maf3c gp41_maf12c gp41_3c

Table 5.1: **Description of HIV-phyloTSI predictors.**

A naive imputation of the infection date for the source-recipient pairs would therefore result in a variety of inconsistencies. Here, we present our approach to avoid these and “refine” infection date estimates. We first describe the mathematical formulation of the problem, together with a rejection sampling algorithm to sample from the posited distribution of plausible infection dates.

Mathematical formulation

Consider a transmission network $\mathcal{N} = (\mathcal{V}, E)$ (Section 2.1.5). For every individual $i \in \mathcal{V}$, we denote by t_i the (unknown) infection date. As each source must be infected before the recipient, we have the following constraint:

$$t_s \leq t_r \quad \forall (s, r) \in E. \quad (5.1)$$

We define a *plausible range* of infection dates $[m_i, M_i]$, as any interval such that $t_i \in [m_i, M_i]$ with certainty. As this framework is motivated by HIV applications, we can assume that we have access to an initial plausible range, denoted by the superscript “hist”. In practice, m_i^{hist} may correspond to the last negative test date, or the date of sexual debut, and M_i^{hist} to the earliest date of a positive test. The inference of the infection dates $(t_i)_{i \in \mathcal{V}}$ is divided into two tasks:

- First determine the domain of possible infection dates. That is, finding the largest subset of $D = \mathbb{R}^{|\mathcal{V}|}$ such that $(t_i)_{i \in \mathcal{V}} \in D$ implies consistency with the network structure and the plausible ranges.
- Second, impose a probability distribution p^D on D whose aim is to reflect that some combination of infection dates may be more likely than others.

Note that the two tasks are conceptually separate. The first consists in a mathematical derivation subject to a list of constraints: its solution is unique. Instead, the second task requires modelling choices to define a probability distribution that best reflects the modeller's beliefs on how infection dates may be related to each other.

To find the largest domain of possible infection D our first aim is to determine the shortest possible plausible ranges, which satisfy a set of constraints listed below.

- The infection date of an individual cannot precede the infection date of their source:

$$m_i \geq m_{\text{source}(i)}. \quad (5.2)$$

- Analogously, the infection date of a source cannot follow the infection date of any of the recipients:

$$M_i \leq \min_{j \in \text{recipients}(i)} M_j. \quad (5.3)$$

- Infection dates must be consistent with the testing history of an individual, and as such plausible ranges must be a subset of the interval $[m_i^{\text{hist}}, M_i^{\text{hist}}]$:

$$m_i^{\text{hist}} \leq m_i \leq M_i \leq M_i^{\text{hist}}. \quad (5.4)$$

Exploiting the acyclic nature of the transmission network, we can efficiently propagate information through the networks through “tree-traversal” algorithms [Valiente, 2021]. This family of algorithms are characterised by the fact that each node is visited at most once, and that the visiting order is determined by the structure of the network. Information on the lower bound of infection dates $(m_i)_{i \in \mathcal{V}}$ propagates from sources to recipients (Eq. (5.2)) which suggests a “top-down” traversal as implemented in Algorithm 1. On the other hand, information on the upper bound of infection dates $(M_i)_{i \in \mathcal{V}}$ propagates from recipients to sources (Eq. (5.3)) which suggests a “bottom-up” traversal as implemented in Algorithm 2. The outputs of these traversal algorithms, which are $(m_v^*)_{v \in \mathcal{V}}, (M_v^*)_{v \in \mathcal{V}}$, are guaranteed to:

- Terminate in less than $|\mathcal{V}|$ iterations as every node is attached to the queue at most once.

Algorithm 1: Minimum Infection Date Propagation**Input :**

- A graph $\mathcal{N} = (V, E)$ representing the connected network.
- Array m^{hist} representing the prior minimum infection times of each individual.

Output:

- Array $(m_i)_{i \in E}$ representing the minimum infection dates of each node consistent with the network

```

1 Find the root node  $i_0$  representing the root node, that is the only element in
    $V \setminus \{r : (s, r)\}$ 
2 Initialize an array  $m$  of size  $|V|$  with elements set  $m = m^{\text{hist}}$ .
3 Initialize a queue  $Q$  and enqueue  $i_0$ .
4 while  $|Q| \neq 0$  do
5   Dequeue a node  $i$  from  $Q$ .
6   for  $j \in \text{recipients}(i)$  do
7      $m_j = \max(m_j, m_i)$ .
8   Enqueue  $j$  to  $Q$ .
```

- Provide the longest plausible ranges consistent with the network structure (Eqs. (5.2) and (5.3)) and the initial plausible ranges (Eq. (5.4)). See Appendix C for a proof of this for Algorithm 1.

The algorithms rely on the assumption that the transmission network is correctly specified. If this assumption is not met, Eq. (5.4) may not be satisfied, hence indicating that the transmission network is not consistent with the initial plausible infection ranges.

The previously obtained plausible ranges summarize consistency constraints on the individual-level infection dates t_i . More generally, we are interested in the collection of infection dates $\mathbf{t} = (t_i)_{i \in \mathcal{V}} \in \mathbb{R}^{|\mathcal{V}|}$ for all individuals in the network. As such, we define the *plausible region* of infection dates $D \subseteq \mathbb{R}^{|\mathcal{V}|}$ as the intersection of the Cartesian product of the $|\mathcal{V}|$ individual-level plausible ranges and the $|E|$ half-hyperplanes defined by the direction of transmission of each transmission event $t_i \geq t_{\text{source}(i)}$:

$$D = \times_{i \in \mathcal{V}} [m_i^*, M_i^*] \cap \bigcap_{(s,r) \in E} \{\mathbf{t} \in \mathbb{R}^{|\mathcal{V}|} : t_r \geq t_s\}. \quad (5.5)$$

Having obtained the region of all infection dates consistent with our data, the second set consists in specifying a probability distribution p^D over it. Below we detail a natural choice for this distribution, which we define iteratively over the network structure, proceeding “breadth-first” from the root of the network. Specifically, let i_0 be the root of the network $\mathcal{N}$, and let $d(i, i_0)$ denote the length of the path between i and i_0 . We can

Algorithm 2: Maximum Infection Date Propagation**Input :**

- A graph $\mathcal{N} = (V, E)$ representing the connected network.
- Array M^{hist} representing the prior minimum infection times of each individual.

Output:

- Array $(M_i)_{i \in E}$ representing the maximum infection dates of each node consistent with the network.

- 1 Find the root node r_0 representing the root node, that is the only element in $V \setminus \{r : (s, r) \in E\}$
- 2 For each node in $V \setminus i_0$, define the priority $\pi(i)$ as the distance between i and i_0 , that is the number of edges in the shortest path connecting i_0 and i .
- 3 Initialize an array M of size $|V|$ with elements set $M = M^{\text{hist}}$.
- 4 Initialize a priority queue containing nodes that are not sources $Q = V \setminus r : (s, r) \in E$, with priority given by π
- 5 **while** $|Q| \neq 0$ **do**
- 6 Dequeue the node i from Q with highest priority.
- 7 **for** $j \in \text{source}(i)$ **do**
- 8 $M_j = \min(M_j, M_i)$.
- 9 If $j \notin Q$, enqueue j to Q .

decompose our model as:

$$p^D(\mathbf{t}|\mathcal{N}, D) = \prod_{l=0}^L \prod_{\substack{i \in \mathcal{V}: \\ d(i, i_0)=l}} p(t_i | t_{\text{source}(i)}, \mathcal{N}, D). \quad (5.6)$$

Assuming a constant hazard of infection λ , we can model the individual-level infection dates as truncated exponential distributions:

$$p^D(\mathbf{t}|\mathcal{N}, D) \propto \prod_{l=0}^L \prod_{\substack{i \in \mathcal{V}: \\ d(i, i_0)=l}} \exp(-\lambda t_i) \cdot \mathbb{1}(\max(m_i^*, t_{\text{source}(i)}) \leq t_i \leq M_i^*). \quad (5.7)$$

Notice such a model has closed form expression for the full conditional distributions:

$$p^D(t_i | \mathbf{t}_{-i}, \mathcal{N}, D) \propto \exp(-\lambda t_i) \cdot \mathbb{1}\left(\max(m_i^*, t_{\text{source}(i)}) \leq t_i \leq \min\left(\min_{r \in \text{recipients}(i)} (t_r), M_i^*\right)\right), \quad (5.8)$$

where the normalising constant can be computed analytically. Although the above is a natural and simple way to model the infection dates, the mathematical formulation was completed after submission of the manuscript, and in the paper we use two different distributions. The first one assumes a uniform distribution over the plausible region D , and it can be considered as a limit of Eq. (5.7) as $\lambda \rightarrow 0$. The use of an uniform distribution to model infection dates constrained to a range is common in the literature, especially regarding estimation of HIV incidence [Skar et al., 2000, Amornkul et al.,

Algorithm 3: Rejection Sampling for a connected transmission network**Input :**

- Connected transmission network $\mathcal{N} = (V, E)$.
- Lower bound for date of infection m_v for all $v \in V$.
- Upper bound for date of infection M_v for all $v \in V$.
- Proposal distributions $p_v(\cdot)$ for all $v \in V$, supported in $[m_v, M_v]$.
- Target p^D known up to a normalizing constant α .
- $C := \inf_{x \in \text{supp}(p^D)} \frac{\alpha p^D(x)}{\prod_{v \in V} p_v(x_v)}$
- Number of samples N .

Output: N samples from the target p^D .

```

1 Set  $i \leftarrow 1$ 
2 while  $i \leq N$  do
3   Sample  $\hat{y}_v \sim p_v$  for all  $v \in V$ 
4   for (source, recipient)  $\in E$  do
5     if  $\hat{y}_{\text{source}} > \hat{y}_{\text{recipient}}$  then
6       Reject the sample
7     continue from line 2
8   Sample  $u \sim \text{Uniform}(0, 1)$ 
9   if  $u < \frac{\alpha p^D(\hat{y})}{C \prod_{v \in V} p_v(\hat{y}_v)}$  then
10    Accept the sample  $(\hat{y}_v)_{v \in V}$ 
11     $i \leftarrow i + 1$ 

```

2013, Kharsany et al., 2019]. We also considered a second distribution: a piecewise constant distribution with assigned higher probability to infection events occurring in early infection stages of the sources.

Imputation of infection dates can then be performed by either sampling from p^D or obtaining summary statistics such as the marginal medians. For either choice of p^D , we can sample from the posterior distribution of the infection dates using rejection sampling, as per Algorithm 3. A natural choice for the proposal is the product of $|\mathcal{V}|$ uniform distributions supported in the individual-level plausible ranges, which was sufficient for reasonable acceptance probabilities in our experiments. If full conditional distributions are available, as in Eq. (5.8), a Gibbs samplers could be used instead.

Application to source-recipient pairs

The source-recipient pairs obtained in Section 5.3.2 defined a partially observed transmission network, which we assumed to be correct, and consistent with the individual-level serohistories. To apply the method described just above, we first partitioned the network into connected components and then applied the rejection sampling algorithm to each of the resulting sub-networks.

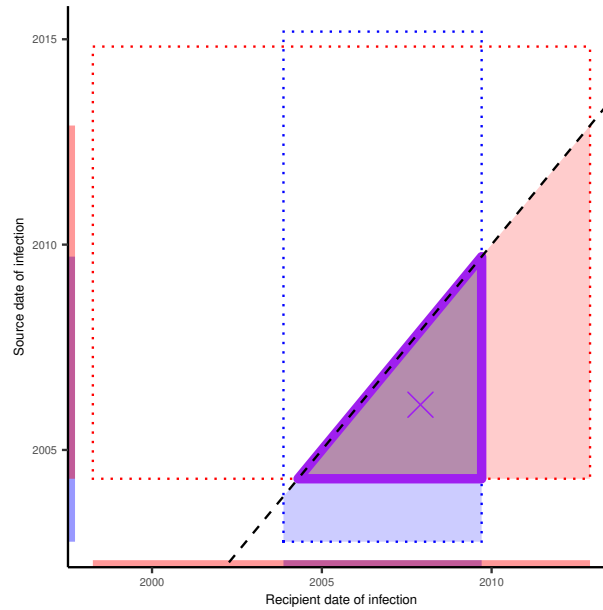


Figure 5.2: Schema illustrating the refinement of phylogenetic time since infection estimates. Combinations of infection dates for recipient (x-axis) and source (y-axis) which are consistent with direction of transmission (below black dotted line), with HIV-phyloTSI prediction range (red dotted rectangle), and with serohistory (blue rectangle). The intersection of this region (purple triangle) is consistent with all of the above, and we imputed its centre of mass as the refined transmission date. This figure is included in the accepted manuscript provided to *Nature Microbiology* for further editing.

We considered multiple ways to define the initial plausible ranges for the source-recipient pairs, which used the HIV-phyloTSI estimates, the serohistory data, or a combination of both. The HIV-phyloTSI plausible ranges were simply defined using the 2.5% and 97.5% quantiles of the prediction intervals, which can be extracted from the algorithm's output. On the other hand, the serohistory-based plausible ranges were defined as follows. The upper bound of the infection date was set as the 30th day prior to the first positive test of the participant [Cohen et al., 2010]. The lower bound of the infection date was set to the largest of the following dates: the date of last negative test if available, the 15th birthday, or the date corresponding to 15 years prior the upper bound [Pantazis et al., 2021]. We then considered a combination of the two approaches, where the initial plausible ranges were defined as the intersection of the two sets of ranges, and if these intersection were empty, we discarded the HIV-phyloTSI estimates.

The schema in Fig. 5.2 summarizes the process of date of infection imputation for a transmission network which only includes a single source-recipient pair, as most of the networks we considered.

Fig. 5.3 compares the refined infection date against the original HIV-phyloTSI estimates. It is clear that the refinement process strongly reduces the uncertainty around the estimates, as can be noted by the typically much narrower 95% credible intervals.

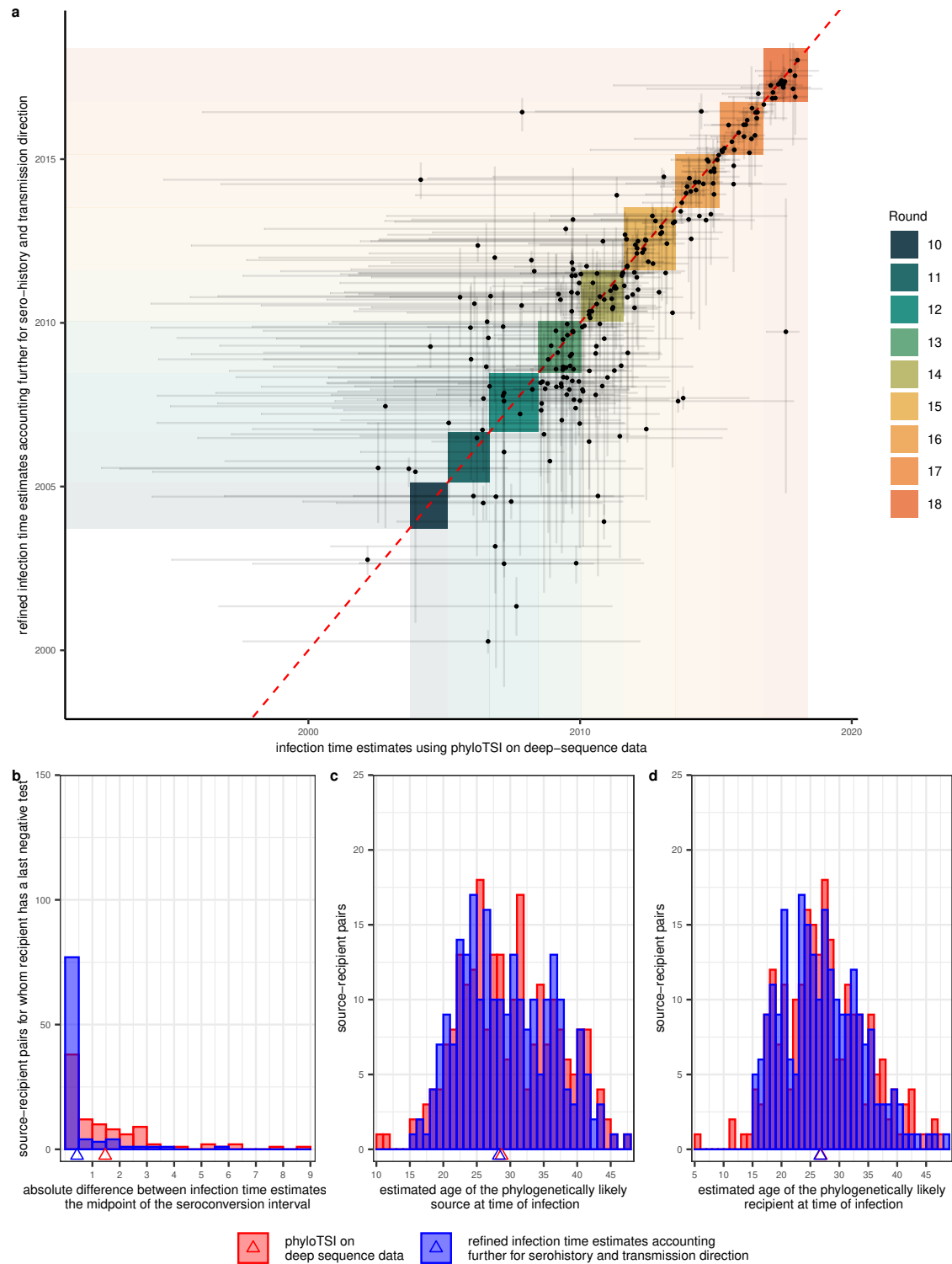


Figure 5.3: Comparison of estimated infection dates in reconstructed source-recipient pairs. (a) Estimated infection dates of the recipient in the 236 source-recipient pairs from HIV-phyloTSI (x-axis) against refined estimates accounting for serohistory and inferred direction of transmission (y-axis). Medians (dots) are shown along 95% uncertainty ranges (lines). (b) Histogram and mean (triangle) of absolute differences between infection date estimates and the midpoint of seroconversion intervals in 98 source-recipient pairs in which the recipient had a last negative test, across methods (color). (c) Histogram and median (triangle) age of the phylogenetically likely recipient, across methods (colors). (d) Histogram (bars) and median (triangle) age of the phylogenetically likely source, across methods (colors). This figure is included in the accepted manuscript provided to *Nature Microbiology* for further editing.

5.3.5 Modeling changes in transmission flow

In the previous subsections, we described how to obtain individual-level **TSI** estimates, and how we can combine these to obtain transmission date estimates for source-recipient pairs which are consistent with the serohistory and the inferred direction of transmission. In this section, we describe how source-recipient pairs with known transmission dates can be used to estimate changes in **HIV** transmission flow, i. e. the proportion of transmissions that occur between different population subgroups. The model in this section was developed by Mélodie Monod, building on top of [Xi et al. \[2022\]](#). For this reason, I will only summarise the key modelling concepts, highlighting how treating infection time as known can be used to describe changes in transmission patterns over time.

As in [Xi et al. \[2022\]](#), we consider a population divided in two genders $\mathcal{G} = \{\text{male}, \text{female}\}$, and multiple age groups $\mathcal{A} = \{15, 16, \dots, 49\}$. A susceptible-infected (SI) model is assumed, where the expected number of transmission events from one group to another is proportional to a transmission intensity, the number of susceptible individuals in the recipient group and the number of infected individuals in the source group. Whereas [Xi et al. \[2022\]](#) considered transmission intensities which were constant in time, we considered time-varying transmission intensities, which depended on the round r in which the infection date was attributed. As such, for each round r , we defined the expected number of transmission events from individuals of gender g and age i to individuals of gender h and age j as:

$$\lambda_{r,i,j}^{g \rightarrow h} = \beta_{r,i,j}^{g \rightarrow h} \times S_j^h \times I_i^g \times |t_r^{\text{end}} - t_r^{\text{start}}|, \quad (5.9)$$

where $\beta_{r,i,j}^{g \rightarrow h}$ is the transmission rate exerted from one virally unsuppressed individual of gender g aged i to an individual of gender h and age j , S_j^h is the number of susceptible individuals of gender h and age j , I_i^g is the number of virally unsuppressed individuals of gender g and aged i , and the last term refers to the length of the round r . An implicit assumption in Eq. (5.9) is that the changes in the number of susceptible and infected individuals are negligible over the length of the round, which is reasonable given the short duration of the rounds, and the low incidence rate of **HIV** in the population.

We then model the unobserved number of transmission events $Z_{r,i,j}^{g \rightarrow h}$ from individuals of gender g and age i to individuals of gender h and age j in round r as:

$$Z_{ij}^{g \rightarrow h} \sim \text{Poisson}(\lambda_{ij}^{g \rightarrow h}). \quad (5.10)$$

Phylogenetically detected source-recipient pairs may not be representative of all the transmission events, as the detection of such a transmission event relies on both indi-

viduals being sequenced, and sequencing probabilities vary over time, age and gender. For this reason, we assume that both sources and recipients are sampled at random within each strata, with sampling probabilities $\xi_{r,g,i}^1$ and $\xi_{r,h,j}^2$ respectively. By thinning properties of the Poisson distribution, the observed number of transmission events $Y_{r,i,j}^{g \rightarrow h}$ is then distributed as:

$$Y_{r,i,j}^{g \rightarrow h} \sim \text{Poisson}(\lambda_{i,j}^{g \rightarrow h}, \xi_{g,i,j}^1 \xi_{h,i,j}^2). \quad (5.11)$$

To make inference computationally manageable, $\beta_{r,i,j}^{g \rightarrow h}$ are modelled as efficient 2D-B-splines Gaussian processes [Monod et al., 2023a]. Technical details on prior choice and definition of the sampling probabilities $\xi_{r,g,i}^1$ and $\xi_{r,h,j}^2$ are provided in the methods section of Monod et al. [2023b].

5.3.6 Ethical consideration

Our research group is part of the PANGAEA HIV consortium [Abeler-Dörner et al., 2019], and as such follows the consortium’s ethical guidelines [Coltart et al., 2018]. Crucial among these are the protection of the rights and interests of study participants, provision of informed consent, and communication of the results to the communities involved.

Phylogenetic analyses are performed on anonymised data, with identification keys different from the ones provided by the RHSP. The identifiers were further changed in the shared data, to ensure participants cannot be identified from this data set. For instance, age of the participants is not provided, and figures such as Fig. 5.5 report age at estimated infection, which depends on the inferences of time since infection.

This specific study was independently reviewed and approved by the Ugandan Virus Research Institute, Scientific Research and Ethics Committee, protocol GC/127/13/01/16; the Ugandan National Council of Science and Technology; and the Western Institutional Review Board, protocol 200313317. All study participants provided written informed consent at baseline and follow-up visits using institutional review board approved forms. This project was reviewed in accordance with CDC human research protection procedures and was determined to be research, but CDC investigators did not interact with human subjects or have access to identifiable data or specimens for research purposes. Participants in the RCCS received 10,000UGX (approximately 2.50USD) in compensation for the baseline and follow-up surveys.

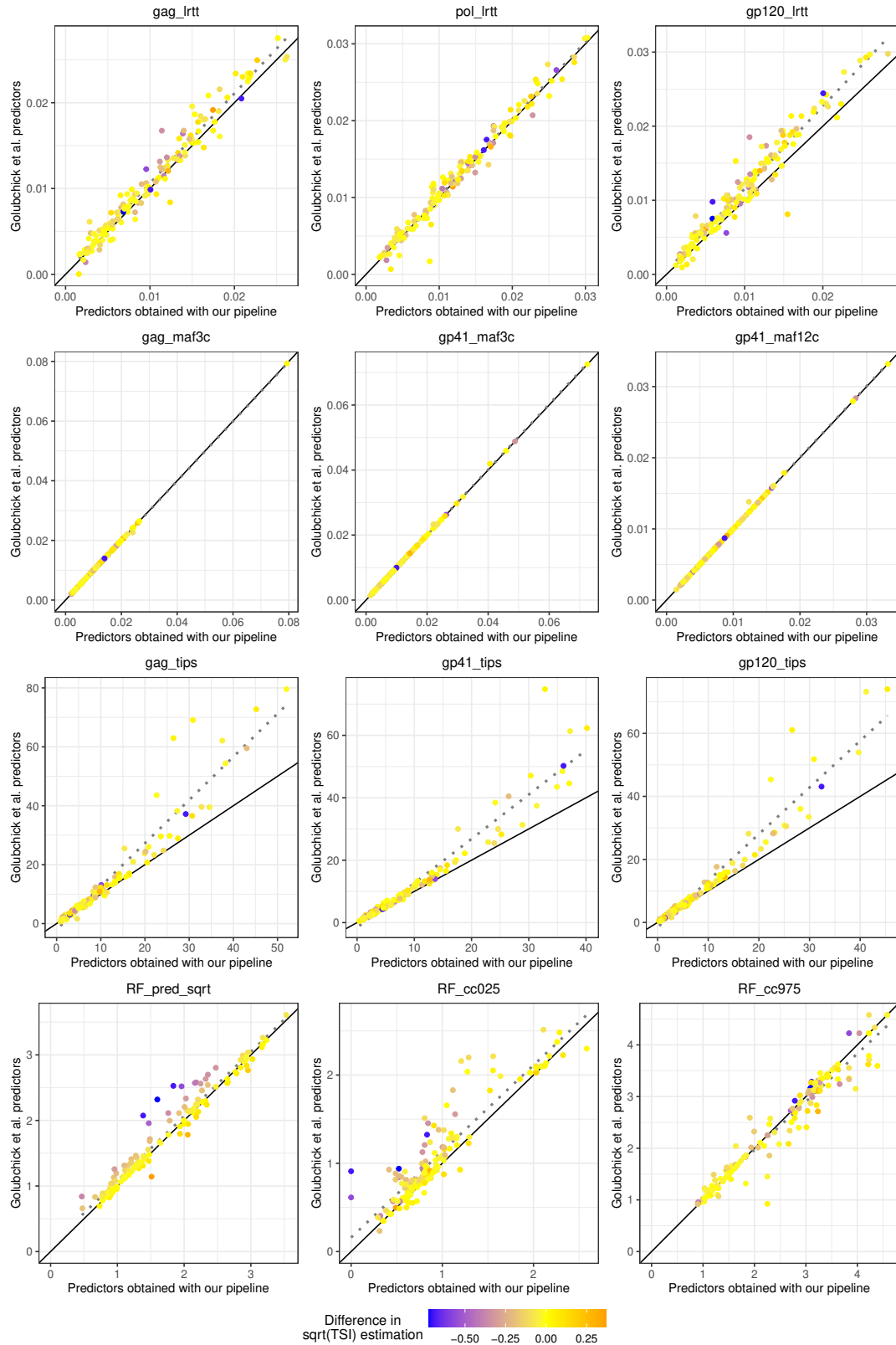


Figure 5.4: Comparison of pipeline output for a set of 152 sequences appearing both in HIV-phyloTSI’s training data and in the RCCS. For each of the predictors (first 3 rows) and HIV-phyloTSI outputs (central prediction and 2.5th and 97.5th percentiles in the last row), we report the values obtained with our pipeline (x-axis) against those obtained in Golubchick et al. [2022] (y-axis). The continuous line represents the identity line, while the dotted line represents the linear regression fit.

5.4 Findings

Our TSI pipeline is consistent with the original one.

To test whether the implementation of the pipeline was successful, we compared our outputs against the ones from [Golubchik et al. \[2022\]](#) on a subset of the sequences. Briefly, 152 sequences of the RCCS sequences appeared in the training data of HIV-phyloTSI, and we contacted the author to share the values of the predictors and the TSI estimates they obtained. The comparison is reported in Fig. 5.4, showing consistent predictions (see yellow colour, regardless of slight changes in predictors value) with the original study. The computed MAFs were identical among predictors, which is expected as they only depend on deterministic calculations. Some differences were observed in the computed number of tips and largest root-to-tip distances, which are strongly dependent on the grouping of sequences and stochastic phylogeny reconstruction. In particular, there was a light under-estimation of the number of tips, which however did not seem to affect the TSI estimates.

Identification and dating of source-recipient pairs in the RCCS

When applying our phylogenetic analyses to the RCCS transmission cohort data, we identified 236 heterosexual source-recipient pairs in whom both individuals were ever resident in the 36 survey communities. Of these, 142 pairs were from men to women and 94 from women to men. Infection times were estimated for all sampled individuals and refined for the recipient partners in the 236 heterosexual source-recipient pairs as per Section 5.3.4. For 4 recipient partners, the HIV-phyloTSI estimates were ignored as they were incompatible with inferred transmission direction and survey data, and was based on serohistory data only. The phylogenetically most likely location of both individuals at time of transmission was estimated as their location at the RCCS visit date that was closest to the posterior median infection time estimate. Using this location estimate, 233 of the 236 heterosexual source-recipient pairs were estimated to capture transmission events in RCCS inland communities and were retained for further analysis. A further six recipient partners had posterior median infection time estimates outside the observation period from September 2003 to May 2018 and were excluded, leaving for analysis 227 heterosexual source-recipient pairs that captured transmission events in RCCS inland communities during the observation period. This excluded 88 potential source-recipient pairs from our study due to ethical considerations and prior analyses suggesting these pairs most likely represent partially sampled transmission chains (i.e., “false positives”) [\[Ratmann et al., 2019\]](#). Fig. 5.5 shows the number of heterosexual

source-recipient pairs by the date of infection of the recipient, the age of the recipient at infection, and transmission direction.

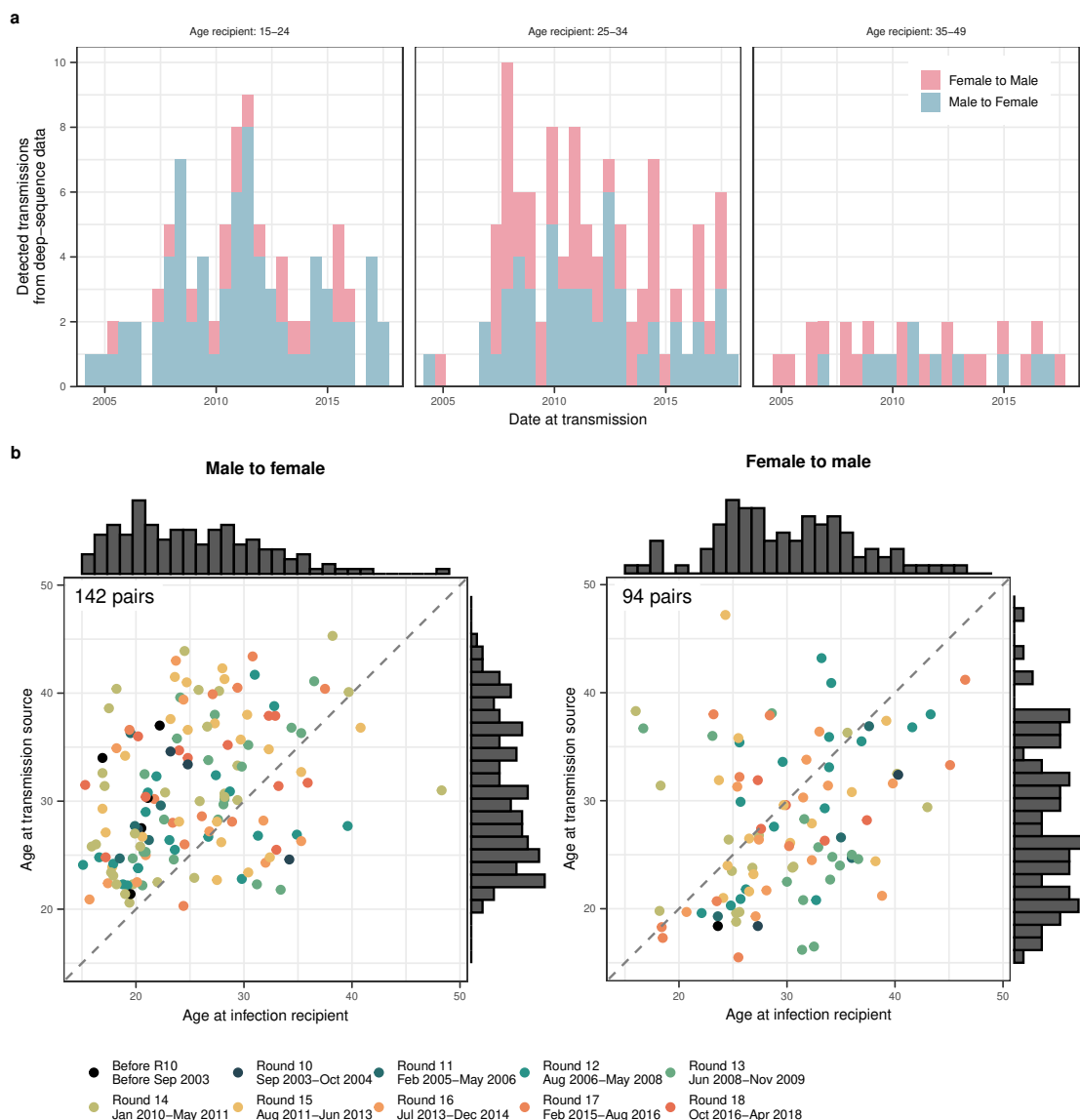


Figure 5.5: Phylogenetically reconstructed source-recipient pairs. (a) Number of heterosexual source-recipient pairs by the date of infection of the recipient (x-axis), the age of the recipient at infection (panel), and transmission direction (color). (b) Heterosexual source-recipient pairs by the age of the recipient (x-axis) and the age of the source (y-axis) at the median infection time estimate by the round (color) in which transmission was estimated to have occurred. The number of phylogenetically reconstructed source-recipient pairs is indicated in the top-left corner. This figure is included in the accepted manuscript provided to *Nature Microbiology* for further editing.

Temporal changes in transmission dynamics

Fig. 5.6 shows the age profile of the estimated male and female sources of infection, such that the male plus the female sources sum to 100% for each survey round. Overall, we



Figure 5.6: Time trends in age-specific sources of HIV infections in women and men. (a) Estimated age distributions of transmitting partners (posterior median: line, 95% credible interval: ribbon), along with the median age at transmission (posterior median: cross, 95% credible interval: linebar). Age contributions sum to 100% for each round, summing over men and women. This figure is included in the accepted manuscript provided to *Nature Microbiology* for further editing.

found that the contribution of men to onward transmission increased progressively from 57.9% [56.2-59.6] in 2003 to 62.8% [60.2-65.2] in 2018, indicating that HIV transmission is now more disproportionately driven by men than has been the case previously.

We also found that sources of HIV infection are increasingly older. The age of transmitting male partners progressively increased from a median age of 28.5 [27.1-30.1] years in 2003 to 33.5 [31.0-36.0] years in 2018 (Fig. 5.6), and this increase in the age of transmitting male partners was largest in transmissions to women aged 20-24. In contrast, the median age of female transmitting partners remained similar (from 25.0 [23.0-27.0] years in 2003 to 26.0 [24.0-28.0] years in 2018).

Over time, substantially fewer infections occurred in adolescent girls and young women aged 15-24 years, and the largest share of transmission flows shifted to women aged 25-34 years, from male partners 0-6 years older.

5.5 Discussion

In this Chapter, we described our own implementation of a high-throughput phylogenetic pipeline to estimate HIV TSI. The pipeline exclusively makes use of individual-level deep sequence data, and as such may report prediction ranges which are inconsistent with the known serohistory of a patient, or with a known or inferred transmission direction. As such, we developed a method to refine the date of transmission for phylo-

genetically inferred source-recipient pairs, which is consistent with all of the available sources of information. When we applied these methods to the **RCCS** transmission cohort, we dated infection events for 236 heterosexual source-recipient pairs, and were able to estimate changes in the transmission flows by age and gender over time.

HIV-phyloTSI is not the only method to estimate **HIV TSI**, but previous approaches rely on data, such as longitudinal CD4 counts and viral load measurements, that are not commonly available in **LMIC** settings. For instance, [Van Sighem et al. \[2015\]](#) use a multi-state model including CD4 count as predictors to estimate the history of **HIV** incidence, changes in time to diagnosis in **MSMs** and size of the unaware population in the Netherlands from 1980 to 2011. [Stirrup and Dunn \[2018\]](#) developed a Bayesian, biomarkers-based survival model to estimate **TSI** in a cohort of cases diagnosed in London between 2009 and 2013. [Pantazis et al. \[2019\]](#) applied another Bayesian approach by modelling the joint evolution of CD4 counts and viral loads through a bivariate linear mixed model. Their goal was to determine the place of **HIV** acquisition for a cohort of **HIV**-positive migrants diagnosed in Europe.

The implementation of our pipeline, together with its openly available code, represents a step forward towards the democratisation of the phylogenetic analyses to estimate **TSI**. However, there are a couple important limitations that need to be addressed to guarantee other research groups can use our pipeline. Firstly, the pipeline is highly computationally intensive and necessitates access to a high-performance computing cluster, even for small datasets. We have only managed to implement our pipeline on Imperial College's HPC cluster, and therefore job submission scripts may require adjustments to function on other clusters. Nonetheless, the modular structure of the pipeline implies that only a few functions would need tweaking to enable its operation on different clusters. Our hope is that, with the continuous improvements in computational power, the pipeline will be able to be run on cloud-based services at a limited cost in the future years. Secondly, even when utilizing a high-performance computing cluster, the pipeline may necessitate prolonged execution times, particularly for extensive datasets. The primary bottleneck of in the high-throughput required to parallelize the alignment and tree construction steps across multiple groups of sequences and hundreds of windows across the **HIV** genome. To address this limitation, a potential solution involves altering our sequence grouping strategy: as alignment and phylogeny reconstruction scale at least quadratically with the number of sequences in each group [[Baichoo and Ouzounis, 2017](#)], reducing the number of sequences in each group may lead to a substantial reduction in the time to completion.

Through our application on data from the **RCCS** transmission cohort, we demonstrated that **TSI** estimates allow for novel insights into the dynamics of **HIV** transmission. In

particular, dating transmission events allows to compare source-recipient pairs deemed to have been infected at different times, and to impute individual-level characteristics at the time of transmission, such as age or residence. However, we believe TSI estimation may open up further research directions, and we list some of these below. Firstly, as per HIV-phyloTSI original motivations, TSI estimates may be used to estimate HIV incidence in a population. The rationale is that although the predictions may not always be accurate at the individual-level, population-level averages may more accurately depict the proportion of recent infections at the population level [Golubchik et al., 2022]. For these applications, it would be necessary to take into account the differences between the training data used in Golubchik et al. [2022] and adjust for false-positive and false-negative rates accordingly [Saerens et al., 2002, González et al., 2017]. Secondly, incorporating information on infection dates of PLHIV may allow for the reconstruction of bigger transmission networks than previously possible [Ratmann et al., 2019]. Here, we presented a method to impute the date of transmission events t conditional on a known transmission network $\mathcal{N}$ and with the known serohistory of each individual. Namely, we developed simple tree-traversal algorithms [Valiente, 2021] which exploited the network structure to efficiently exclude combinations of infection dates inconsistent with the data. We then imposed a probabilistic model supported on the remaining dates, and proposed rejection sampling or importance sampling algorithms to sample from, and summarise, the distribution of t .

This work could be generalised to jointly model the infection dates and the network $p(\mathcal{N}, t)$, for example taking inspiration from the literature on influenza or foot and mouth disease [Cauchemez et al., 2009, 2011, Cauchemez and Ferguson, 2012], but with careful consideration of the underlying assumptions. In particular, contrary to typical HIV scenarios, the papers assume that all potential sources of infection are followed, and that either the infection dates are known, or an individual's infectiousness is a function of symptom onset only. Careful consideration would also be needed to encode structure in HIV transmission, such that infections occur only among individuals of different genders. In our attempts, we did not manage to find a convincing solution, and found it especially challenging to either model $P(\mathcal{N})$ or $P(\mathcal{N}|t)$. Specifically, we did not know how to account for unobserved HIV cases which could act as missing links in the transmission network.

Our pipeline provides a base to iterate on the hard problem of estimating TSI. The lack of a large dataset including individuals with precisely known infection dates makes the training of machine learning models difficult. This results in some model limitations. For instance, HIV-phyloTSI random forest's 95% prediction intervals do not achieve the nominal coverage, and the width of these intervals on the linear scale can be very small for infections predicted to be recent. If prediction ranges are deemed more important

than point estimates, it could be possible to modify the random forest algorithm to perform quantile regression [[Meinshausen, 2006](#)]. This could be easily implemented by replacing the model in the last step of our pipeline.

6

AGE AND GENDER PROFILES OF HIV INFECTION BURDEN AND VIRAEMIA

6.1 Introduction

WITH the expansion of HIV testing and treatment services, many countries in ESA have either attained or are close to attaining the “95-95-95” targets at a population level [Mine et al., 2024, Maheu-Giroux and Mishra, 2024]. Thus, determining novel effective methods to measure HIV epidemic control and direct future efforts to accelerate epidemic decline in such contexts is required. While stratification of “95-95-95” targets by age and gender represents one approach, current evidence suggests that these metrics may not be predictive of incidence declines [Miller et al., 2013, Solomon et al., 2016, Tanser et al., 2017, Larmarange et al., 2023]. Indeed, although populations stratifications of existing targets are useful in highlighting important inequalities in treatment coverage [Frescura et al., 2022], they do not account for underlying demographic structures or distributions of HIV burden in the population. Consequently, they may not provide a complete understanding of the people most at risk of HIV-related morality and onward transmission, which is necessary to tailor and implement effective prevention interventions.

Motivated by the UNAIDS suppression goals, previous studies have focused on analysing relationships between viral suppression outcomes and other factors. The majority of studies take individuals as experimental units. Gaolathe et al. [2016] use Poisson Generalized Estimating Equation on data from a household survey in Botswana to examine associations between individual socio-demographic factors and outcomes including viral suppression. Pathmanathan et al. [2021] compute the prevalence of HIV and viral

suppression in a longitudinal cohort of performing home based testing. They compute average across population strata and assume the study population is homogeneous with respect to the source population to generalize their estimates of prevalence. Pillay et al. [2020] perform multivariable logistic regression to explain associations with respect to recent viral load testing and suppressed viral load in South Africa. They find female sex and older age to be associated with higher levels of suppression. In Grabowski et al. [2021], “visit-pairs” of consecutive visits from the same RCCS participant are used as units of analysis. Viral load test outcomes are used as response to a multivariate Poisson regression, with the aim of understanding patterns in individual viral load histories. The study found younger age, male sex and mobility to be associated with consecutive unsuppressed viral load measurements. In these studies, age is treated as a categorical variable, with bespoke breakpoints in each study. Other studies have used communities as units of analysis. Farahani et al. [2021] used logistic regression to estimate the probability of one recent infection occurring in an enumeration area, finding strong associations with the prevalence of viral suppression. Similar results were obtained in an study of MSM and people who inject drugs in India. Patel et al. [2021] use linear regression to assess associations between HIV incidence and community-level measures of viraemia.

In this Chapter, we present the work submitted to *International Journal of Epidemiology* as a manuscript titled “Age and gender profiles of HIV infection burden and viraemia: novel metrics for HIV epidemic control in African populations with high antiretroviral therapy coverage” and presented at the 2024 “Conference on Retroviruses and Opportunistic Infections” as an Oral Abstract. Using longitudinal data from the RCCS, we describe current limitations in the use of the “95-95-95” targets to assess HIV epidemic control in ESA. We then propose alternative metrics that characterize the age and gender profiles of the population burdens of HIV infection and of viraemia. The contributions of the authors acknowledged thereafter. My supervisor Oliver Ratmann, M. Kate Grabowski, Steven J. Reynolds, Robert Ssekubugu designed the study. Ronald H. Gray, Robert Ssekubugu, Godfrey Kigozi, Oliver Laeyendecker, Thomas C. Quinn, Fred Nalugoda, Larry W. Chang, Steven J. Reynolds, Joseph Kagaayi, Gertrude Nakigozi, M. Kate Grabowski oversaw and performed data collection. I, Andrea Brizzi, performed most of the analysis, and wrote code for the statistical models, figures and tables. Imogen Kyle, Alexandra Blenkinsop, Mélodie Monod, M. Kate Grabowski, contributed to the statistical analysis. I wrote the first draft together with M. Kate Grabowski and Oliver Ratmann.

This Chapter is structured as follows: Section 6.2 summarises the contributions of this study. Section 6.3, describes the study population and the statistical models developed to estimate prevalence of HIV and viraemia. Section 6.4, summarises the findings of this

study. In Section 6.5, we discuss the implications of the findings and the limitations of the study.

6.2 Contributions to the literature

The contributions of the work presented in this Chapter are two-fold. Our first contribution concerns the detailed description of the evolution of HIV burden by age and gender over time across the RCCS study population. This description involves the group-specific estimation of three types of prevalences, namely the prevalence of HIV, the prevalence of viral suppression among PLHIV and the prevalence of viraemia among all individuals. The groups are defined by the location of residence, the gender and the age. Compared to most similar studies, our estimates are reported at the age-specific level, and we do not group the population into 5 or 10 year age groups [Lebelonyane et al., 2021]. This is achieved through the use of Gaussian processes Section 2.3.6 to model variation in the prevalence across ages, in such a way that similar ages have similar prevalence estimates. We prefer this approach to the standard grouping as inference for parameters of different age groups is done independently, and because grouping could mask differences in the characteristics and behaviours of individuals within each group. For instance, HIV prevalence may vary between individuals aged 15 and 19 years and those aged 20 and 24 years. By comparing the prevalence of suppression among PLHIV, we evaluate the progress towards the “95-95-95” UNAIDS targets.

Our second contribution concerns the development of metrics to accompany the “95-95-95” targets. In simple terms, we argue that if the majority of PLHIV are middle-aged, then a greater focus should be placed on these population subgroups compared to a scenario in which the majority of PLHIV are young adults. The current “95-95-95” targets are unable to account for differences in the demographic structure of the population, and ignoring this limitation may lead to suboptimal treatment and prevention strategies. To address this limitation, we propose the *age and gender profiles* of PLHIV and of people exhibiting viraemia, which describe the share of the burden of HIV and viraemia that is borne by individuals in a certain subgroup of the population. This is achieved by combining our posteriors on prevalence with population sizes to estimate the number of individuals in each subgroup that are HIV positive or viraemic. In summary, we show that the population subgroups that contribute the most to population viraemia are not necessarily the ones with the lowest levels of viral suppression.

The code necessary to reproduce all analyses was made freely available on GitHub at the repository https://github.com/MLGlobalHealth/longi_viral_loads. Data necessary to reproduce the analyses are available on Zenodo (<https://doi.org/10.5281/zenodo.12663850>) as open access dataset under the CC-BY-4.0 license.

6.3 Methods

6.3.1 Population size and participation rates

Between July 2013 and May 2019, four survey rounds (survey rounds 16, 17, 18 and 19) were conducted in four Lake Victoria fishing and 36 inland agrarian and semi-urban communities as part of the **RCCS** (see Section 1.3). A total of 37,743 individuals participated in the study at least once (see Table B.1 for participation rates and Fig. 6.1 for study flowchart). Of these, 22,215 (58.86%) participated for the first time (Table B.2).

To account for population structures, we used **RCCS**'s census data which reported census-eligible population sizes by community type, survey rounds, gender and 1-year age bands. The ages reported by household heads during the census (see Section 1.3) reflected grouping patterns towards ages which were multiples of 5. As such, we obtained smoothed population sizes $\hat{N}_{g,a}$ for gender g across ages $a \in \mathcal{A} \in \{15, 16, \dots, 49\}$ years, stratifying by round and community type using **LOESS** as implemented in the R package `stats` version 4.2.2 with `span=0.5` [Monod et al., 2023b]. We analogously obtained smooth estimates of participation rates $\hat{\theta}_{g,a}^{\text{loess}}$ of census-eligible populations in each community type, survey round, gender and 1-year age band stratum. Here, the **LOESS** span argument was set to `span=0.25`.

The data and resulting fits are illustrated in Fig. B.1. Inland communities were characterized by expansive population pyramids with a predominantly young population, consistent with rapid population growth across Uganda and **ESA** [Pyramid, 2023]. In contrast, fishing communities were characterized by constrictive population pyramids comprising primarily working-age adults (20-40 years) and substantially fewer children or families than in inland communities (Fig. 6.5b). Gender distributions also differed, with women accounting for 52.2% of the population in inland communities and 44.3% in fishing communities in 2018 (Table 6.1).

6.3.2 Measurement of HIV serostatus and viraemia

After being interviewed on demographic characteristics (including age), behavioural and healthcare behaviour, all **RCCS** participants were offered free **HIV** testing. Testing was performed through a combination of three rapid tests with confirmation of positive, faintly positive, and discordant results by at least two EIAs and Western Blot or DNA PCR [Galiwango et al., 2013]. Overall, 99.7% of participants accepted **HIV** testing across survey rounds. Among participants living with **HIV**, **HIV**-1 viral load was measured on stored serum samples using the Abbott real-time m2000 assay (Abbott Laboratories, IL, USA). Overall, viral load measurements were obtained from nearly all (99.4%) partici-

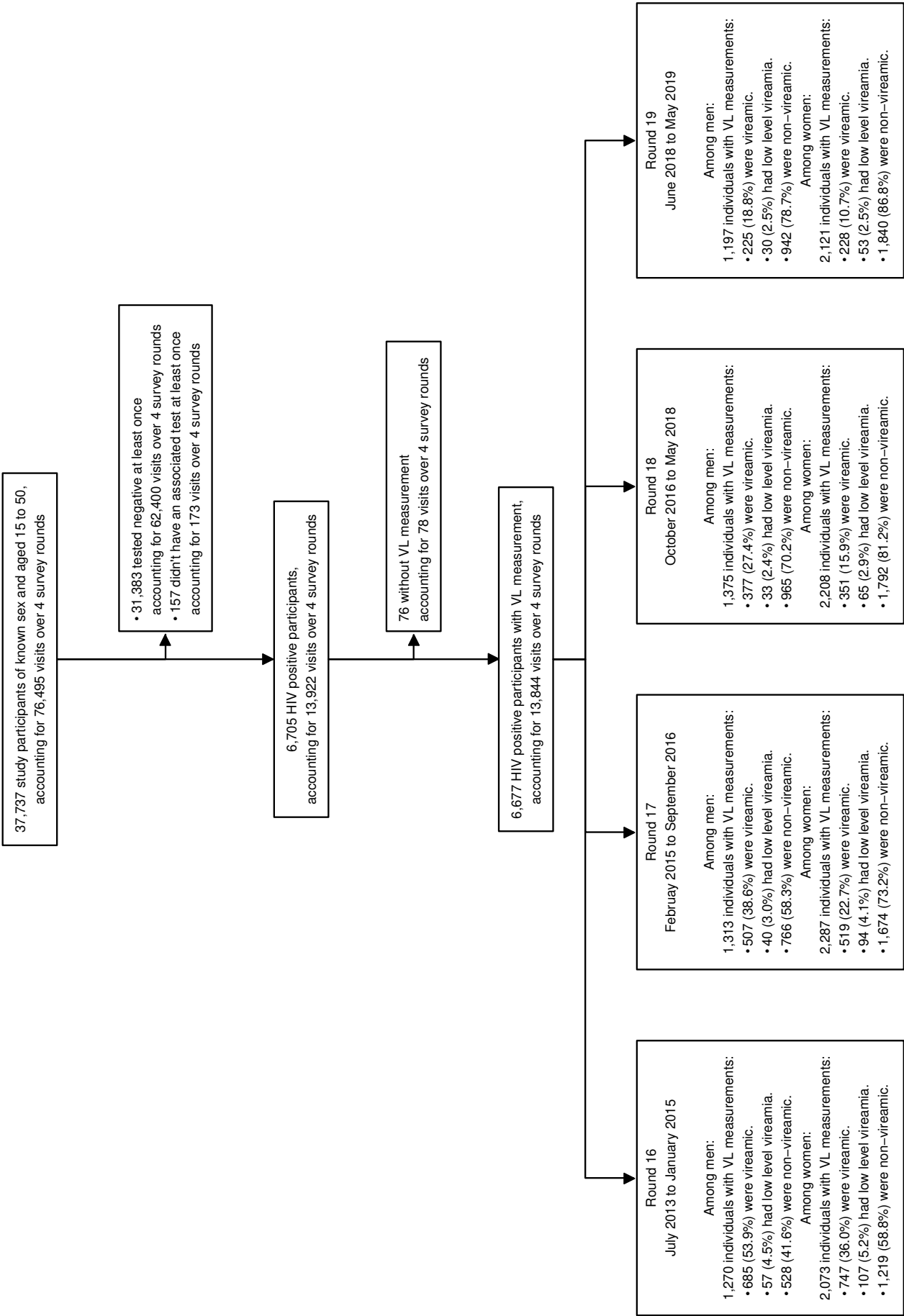


Figure 6.1: Study flowchart from round 16 to round 19 (July 2013 to May 2019). Number of **HIV** and viral load test results across four survey rounds spanning July 2013 to May 2019. Viral load measurement are reported by round and gender, and are classified as viraemic (≥ 1000 copies/mL), non-viraemic (< 200 copies/mL) or with low-level viraemia (> 200 and < 1000 copies/mL). Abbreviations: VL, viral load.

pants living with HIV [Ssempijja et al., 2018, 2020, Grabowski et al., 2021]. Participants who previously tested positive could potentially refuse to test again and drop out for viral load measurement. If all missed measurements were from participants living with HIV, we would have missed at most 1.66% of measurements Fig. 6.1. Viral suppression was defined as a viral load measurement below 1,000 HIV RNA copies/mL plasma, following recommendations of the WHO [WHO, 2016] at the time. People who did not achieve viral suppression were defined as exhibiting viraemia.

6.3.3 Population-level HIV prevalence estimates by age and gender

Our primary objective was to obtain smooth estimates of the proportion of the census-eligible population in each community type, survey round, gender and 1-year age band stratum who have HIV, and similarly to obtain smooth estimates of the age and gender profile of the census-eligible population with HIV in each community type and survey round. We obtained smooth age-specific estimates independently for each gender, survey round and community type combination, and below drop notation relevant to gender, survey round and community type. Throughout, we fitted models to HIV test data from two subsets of participants: all participants (denoted by $(r = 1)$), and first-time participants (denoted by $(r = 2)$). We nonetheless treated all data points independent and did not consider first-time and repeat participants separately because in young age groups all participants were first-timers and this posed challenges in obtaining robust non-parametric prevalence estimates under this alternative approach. The rationale for such division concerns the generalizability of the results to the entire source population. Similarly as McGovern et al. [2015], Palma et al. [2022], our aim was to obtain prevalence estimates representative of the entire population, and we wanted to avoid overly optimistic estimates which do not account for selection bias. As we believe that first-timers are more representative of non-participants than all participants, we used them as a proxy for non-participants. For each participant type r and age group a , we denote the number of study participants with $Z_{a,r}$ and the subset of participants with HIV by $Y_{a,r}$. We modelled age-specific HIV prevalence rates with the non-parametric Bayesian logit model [Monod et al., 2023b].

$$Y_{a,r} \sim \text{Binomial}\left(Z_{a,r}, \boldsymbol{\pi}_{a,r}^{\text{hiv}}\right) \quad (6.1a)$$

$$\text{logit}(\boldsymbol{\pi}_{a,r}^{\text{hiv}}) = \nu_r + f_r(a) \quad (6.1b)$$

$$\nu_r \sim \text{Normal}(0, 10^2) \quad (6.1c)$$

$$f_r \sim \text{Gaussian-Process}(0, k_{\sigma, \rho}) \quad (6.1d)$$

$$\sigma \sim \text{Half-Normal}(0, 1^2) \quad (6.1e)$$

$$\rho \sim \text{Inverse-Gamma}(\rho_{\text{shape}}, \rho_{\text{scale}}), \quad (6.1f)$$

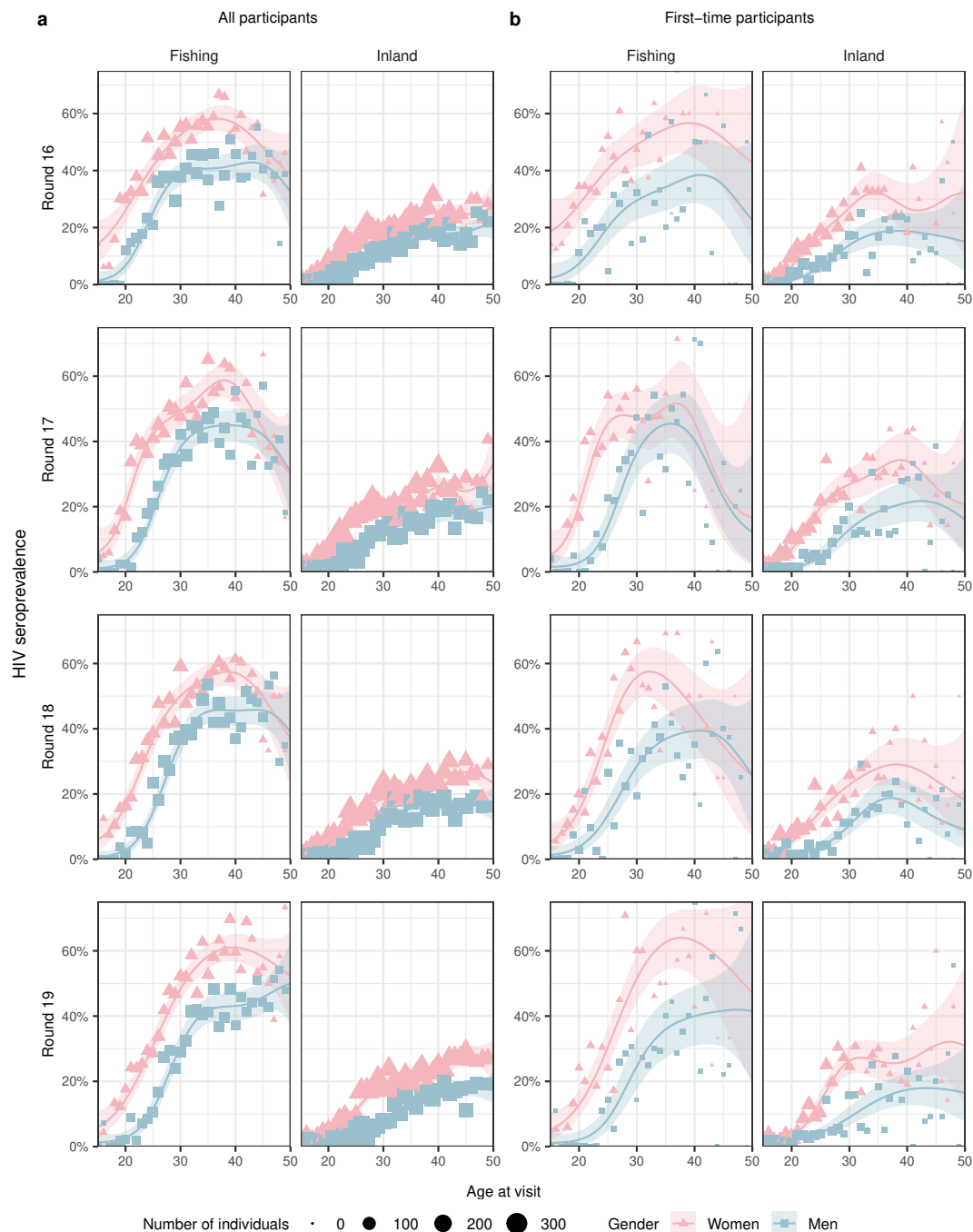


Figure 6.2: Age- and gender-specific HIV prevalence estimates among study participants and first-time participants, 2013-2019. (a) Posterior median estimates of the proportion of all participants in each age and gender group who have HIV (lines) and 95% CrI (ribbons), shown along empirical proportions (squares and triangles for men and women respectively). Participant numbers in each stratum are indicated by the size of squares or triangles. (b) Same for the proportion of first-time participants in each age and gender group who have HIV.

where $a \in \mathcal{A}$, and $r \in \{1, 2\}$. Age-specific HIV prevalence in all participants and first-time participants was modelled on the logit scale with fixed, independent baseline effects v_r plus 1-dimensional Gaussian processes priors $f_r(a)$ specified by squared exponential kernels $k_{\sigma, \rho}$ with variance parameter σ and length-scale ρ . Age inputs were translated to the interval $[1, 35]$. The same hyperparameters σ, ρ were used for both participant types, under the assumption that the age-dependent fluctuations in HIV prevalence were similar for the two groups. The hyperparameters ρ_{shape} and ρ_{scale} were chosen to address identifiability issues relating to minimum and maximum covariate distances, and set such that approximately 99% of the prior mass for the length-scale ρ is contained within $[1, 35]$, i.e. $\rho_{\text{shape}} = 2.21$ and $\rho_{\text{scale}} = 7.05$ [Betancourt, 2020]. In simple terms, the informative prior is necessary to avoid functions $f_r(a)$ which are very “wiggly” between two age groups (low ρ), or that are very “flat” over the entire range (large ρ) (recall Fig. 2.2). A simpler approach could have involved fitting a Bayesian binomial regression model with age as a categorical covariate. However, this would have resulted in prevalence estimates that are not continuous in age, potentially masking important dynamics between age groups. To enforce continuity, a Generalized Additive Model (GAM) could have been used, incorporating penalized splines for the age variable. We believe this approach would have yielded similar inferences, given the comparable expressiveness of GAMs and Gaussian processes in modeling smooth, non-linear relationships.

Model Eq. (6.1) was fitted repeatedly to all gender, survey round and community type combinations with the R package `cmdstanr` version 0.5.3, using four separate HMC chains with 500 warmup iterations and 10,000 sampling iterations each [Gabry et al., 2023, Carpenter et al., 2017]. Convergence and mixing diagnostics are reported in Table B.4, and the model fitted all the available data well (Figs. 6.2 and B.2).

From the fitted models, we generated smooth age-specific estimates of the proportion of the census-eligible population in each community type (denoted by c), survey round (denoted by t), gender g and 1-year age band a stratum who have HIV with

$$\pi_{c,t,g,a}^{\text{hiv-pop}} = \hat{\theta}_{c,t,g,a}^{\text{loess}} \times \pi_{c,t,g,a,r=1}^{\text{hiv}} + (1 - \hat{\theta}_{c,t,g,a}^{\text{loess}}) \times \pi_{c,t,g,a,r=2}^{\text{hiv}}, \quad (6.2)$$

where $\hat{\theta}_{c,t,g,a}^{\text{loess}}$ are the smooth participation rate estimates from Section 6.3.1, $\pi_{c,t,g,a,r=1}^{\text{hiv}}$ are the posterior HIV prevalence estimates among all participants and $\pi_{c,t,g,a,r=2}^{\text{hiv}}$ are the posterior HIV prevalence estimates among first-time participants from model Eq. (6.1). In Eq. (6.2), we assume HIV prevalence in the subpopulations who did not participate is the same as HIV prevalence among first-time participants.

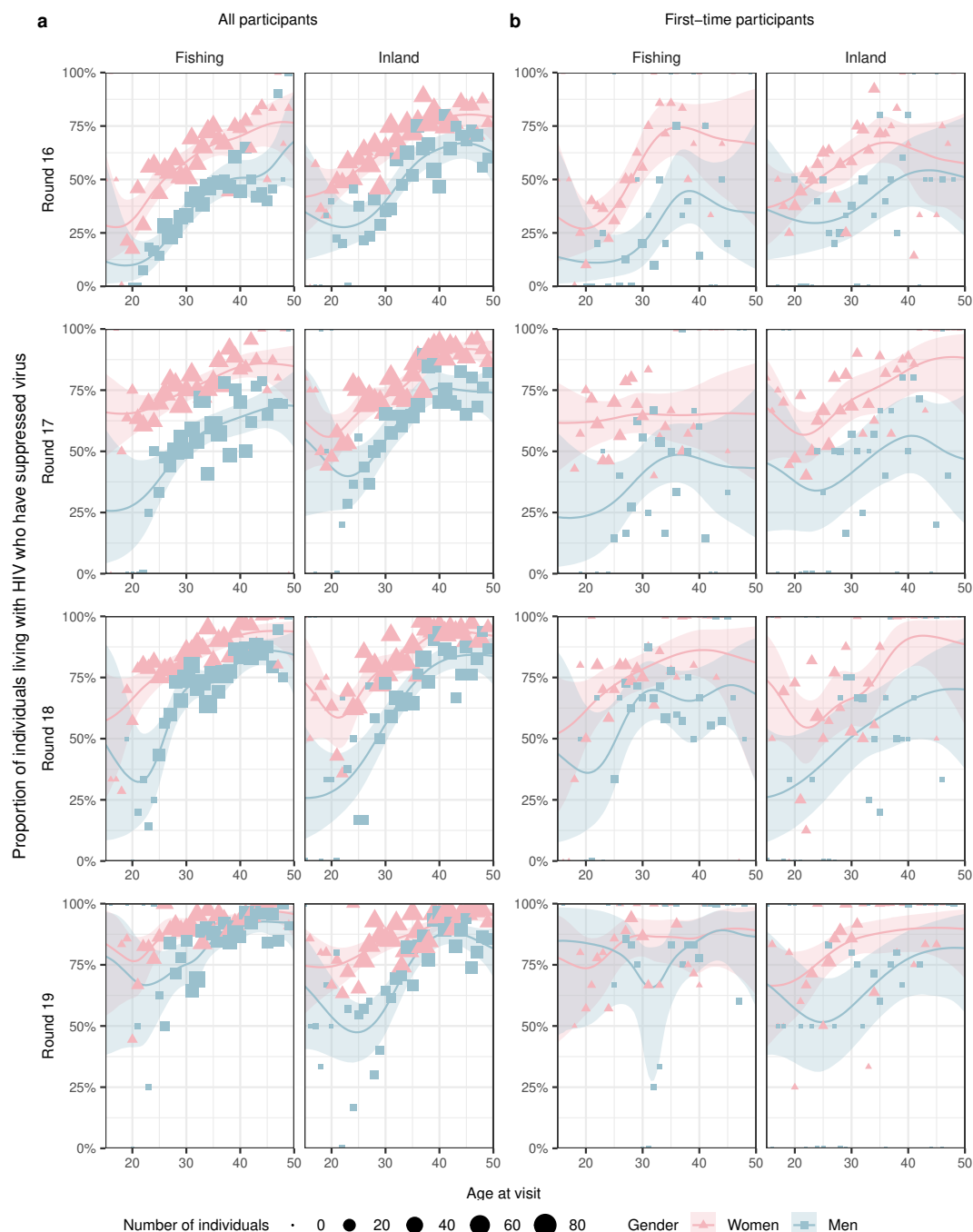


Figure 6.3: Age- and gender-specific viral load suppression rate estimates among study participants and first-time participants with HIV, 2013-2019. (a) Posterior median estimates of the proportion of all participants with HIV in each age and gender group who have suppressed virus (lines) and 95% CrI (ribbons), shown along empirical proportions (squares and triangles for men and women respectively). Participant numbers in each stratum are indicated by the size of squares or triangles. **(b)** Same for the proportion of first-time participants with HIV in each age and gender group who have suppressed virus.

6.3.4 Age and gender profiles of PLHIV

To investigate the age and gender profile of PLHIV in each community type (denoted by c) and survey round (denoted by t), we first estimated the number of PLHIV of gender g and age a as:

$$I_{c,t,g,a} = \hat{N}_{c,t,g,a} \times \pi_{c,t,g,a}^{\text{hiv-pop}}, \quad (6.3)$$

where $\pi_{c,t,g,a}^{\text{hiv-pop}}$ are from Eq. (6.2), $\hat{N}_{c,t,g,a}$ are the LOESS-smoothed population size estimates from Section 6.3.1. Next, we generated the age and gender profile of PLHIV in each community type c and survey round t , by normalising the age and gender specific number of PLHIV by the total over all age groups and genders:

$$\gamma_{c,t}^{\text{hiv}}(a, g) = I_{c,t,g,a} / \left(\sum_{a' \in \mathcal{A}, g' \in \mathcal{G}} I_{c,t,g',a'} \right), \quad (6.4)$$

where $\mathcal{G} = \{\text{men, women}\}$. In Eq. (6.4), the profile sums to 1 over all age and gender groups in each subpopulation specified by community type and survey round. The age and gender profile of HIV prevalence also accounts for population structure between men and women, as for example in fishing communities the gender ratio is far from 50:50.

6.3.5 Prevalence of viral suppression among PLHIV by age and gender

Our objective was to obtain smooth estimates of the proportion of the census-eligible population with HIV in each community type, survey round, gender and 1-year age band stratum who had suppressed virus. We obtained smooth age-specific estimates independently for each gender, survey round and community type combination using the same model as for estimating population-level HIV prevalence. Specifically, and suppressing again notation for community type, survey round and gender, we fitted model Eq. (6.1) to the number $Z_{a,r}^{\text{hiv}}$ of study participants with HIV in age band a and of participant type r , and the subset $Y_{a,r}^{\text{supp}}$ of participants with HIV in age band a and of participant type r who had suppressed virus.

$$Y_{a,r}^{\text{supp}} \sim \text{Binomial} \left(Z_{a,r}^{\text{hiv}}, \pi_{a,r}^{\text{supp-hiv}} \right), \quad (6.5)$$

where $\pi_{a,r}^{\text{supp-hiv}}$ was modeled analogously as $\pi_{a,r}^{\text{hiv}}$ in Eq. (6.1). Convergence and mixing diagnostics are reported in Table B.4, and the model fitted all the available data well (Figs. 6.3 and B.2).

From the fitted model, we obtained smooth age-specific estimates $\pi_{c,t,g,a}^{\text{supp-hiv}}$ of the proportion of the census-eligible population with HIV in each community type (denoted

by c), survey round (denoted by t), gender g and 1-year age band a stratum who had suppressed virus in analogy to Eq. (6.2).

6.3.6 Age and gender profile of people exhibiting viraemia

Finally, we sought to obtain smooth estimates of the proportion of the census-eligible population in each community type, survey round, gender and 1-year age band stratum exhibiting viraemia, and similarly to obtain smooth estimates of the age and gender profile of the census-eligible population exhibiting viraemia in each community type and survey round. We again obtained smooth age-specific estimates independently for each gender, survey round and community type combination using the same model as for estimating population-level HIV prevalence. Specifically, and suppressing again notation for community type, survey round and gender, we fitted model Eq. (6.1) to the number $Z_{a,r}$ of study participants in age band a and of participant type r , and the subset $Z_{a,r}^{\text{hiv}} - Y_{a,r}^{\text{supp}}$ of participants exhibiting viraemia in age band a and of participant type r .

$$Z_{a,r}^{\text{hiv}} - Y_{a,r}^{\text{supp}} \sim \text{Binomial}\left(Z_{a,r}, \pi_{a,r}^{\text{vir-pop}}\right) \quad (6.6)$$

where $\pi_{a,r}^{\text{vir-pop}}$ was modeled analogously as $\pi_{a,r}^{\text{hiv}}$ in Eq. (6.1). Convergence and mixing diagnostics are reported in Table B.4, and the model fitted all the available data well (Figs. 6.4 and B.2). Due to the small sample sizes, results for first-time participants display wider credible intervals for the proportion of individuals exhibiting viraemia compared to the results for all participants. From the fitted model, we obtained smooth age-specific estimates $\pi_{c,t,g,a}^{\text{vir-pop}}$ of the proportion of the census-eligible population in each community type (denoted by c), survey round (denoted by t), gender g and 1-year age band a stratum who had suppressed virus in analogy to Eq. (6.2).

Further, we obtained estimates of the number of individuals exhibiting viraemia $V_{c,t,g,a}$ analogously to Eq. (6.3) and computed the age profile $\gamma_{c,t}(a)$ of individuals exhibiting viraemia in each community type and survey round as in Eq. (6.4), replacing $I_{c,t,g,a}$ by $V_{c,t,g,a}$. Namely, the number $V_{c,t,g,a}$ of individuals exhibiting viraemia in community type c , at round t of gender g and age a was defined as:

$$V_{c,t,g,a} = \hat{N}_{c,t,g,a} \times \pi_{c,t,g,a}^{\text{vir-pop}}, \quad (6.7)$$

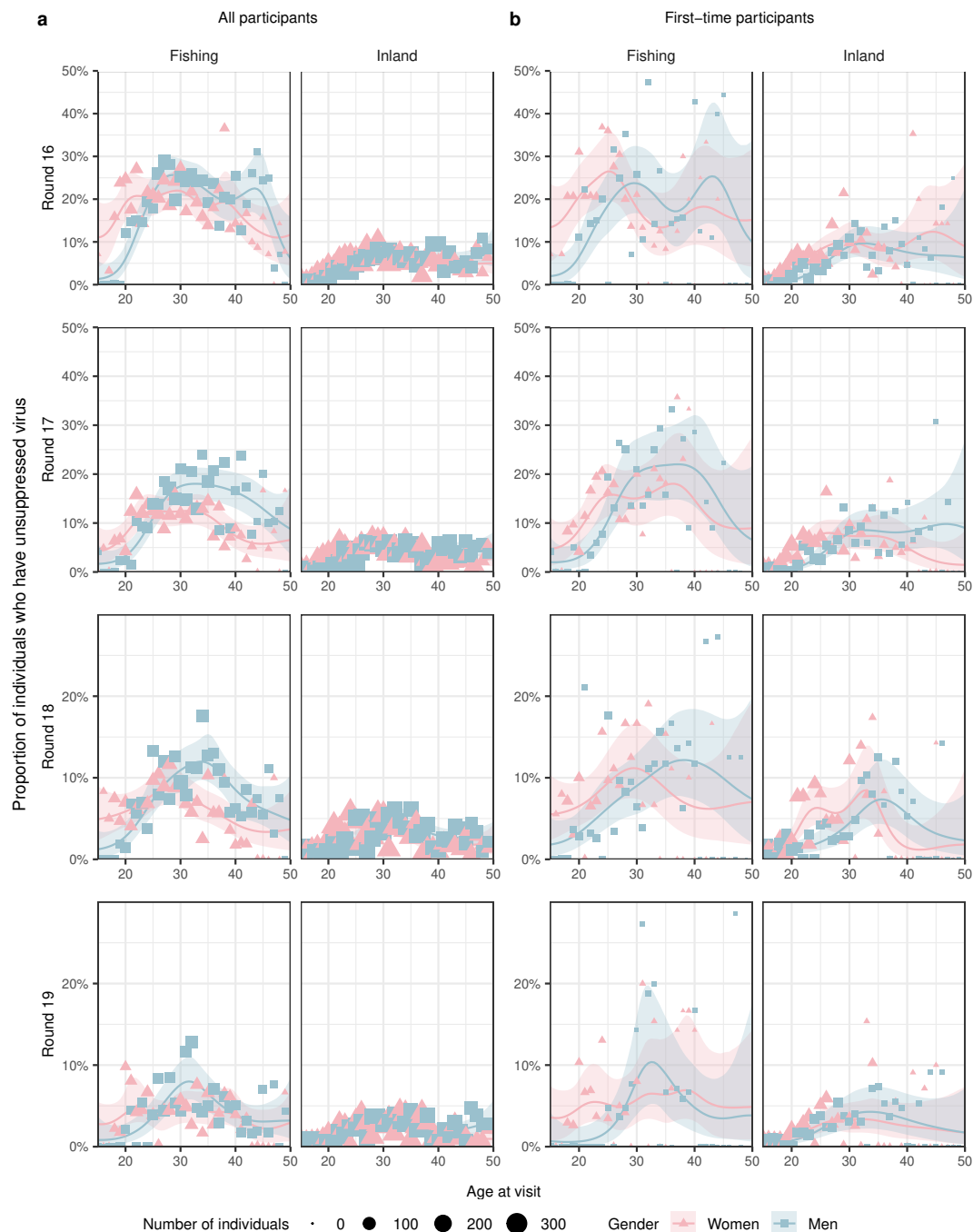


Figure 6.4: Age- and gender-specific viral load suppression rate estimates among census-eligible people with HIV. Posterior median estimates of the proportion of census-eligible people with HIV in each age and gender group exhibiting viraemia (lines) and 95% CrI (ribbons), shown along empirical proportions (squares and triangles for men and women respectively). Participant numbers in each stratum are indicated by the size of squares or triangles.

while the age and gender profile of individuals exhibiting viraemia in each community type c and survey round t was defined as:

$$\gamma_{c,t}^{\text{vir}}(a, g) = V_{c,t,g,a} / \left(\sum_{a' \in \mathcal{A}, g' \in \mathcal{G}} V_{c,t,g',a'} \right). \quad (6.8)$$

6.3.7 Age quantiles for burden profiles by gender

We decided to capture and summarize changes in the burden profiles of **PLHIV** by calculating the p quantiles $q_{c,t,g}^{\text{hiv}-p}$ of the age profiles by survey round t , community type c , and gender g for $p \in \{25\%, 50\%, 75\%\}$. Specifically, for every Monte Carlo iteration, we used the profiles defined in Eq. (6.4) to define a continuous probability distribution $g_{c,t,g}^{\text{hiv}}$ over the ages for every gender, and evaluated the quantiles for these distributions. We simply assumed $g_{c,t,g}^{\text{hiv}}$ to be a step function such that for every age integer a , the distribution is constant on $[a, a + 1)$, and equals $\gamma_{t,c}^{\text{hiv}}(g, a)$. Then, for every $p \in \{25\%, 50\%, 75\%\}$, we solve:

$$p = \int_{15}^{q_{c,t,g}^{\text{hiv}-p}} g_{c,t,g}^{\text{hiv}}(x) dx, \quad (6.9)$$

in terms of $q_{c,t,g}^{\text{hiv}-p}$.

We analogously calculated p quantiles $q_{c,t,g}^{\text{vir}-p}$ for the age profile of people exhibiting viraemia. Specifically, for every Monte Carlo iteration we defined a continuous probability distribution $g_{c,t,g}^{\text{vir}}$ from the burden profile defined in Eq. (6.8), and solved an analogous equation to Eq. (6.9) for every community type c , survey round t , and gender g .

We note that the age quantiles $q_{c,t,g}^{\text{hiv}-p}, q_{c,t,g}^{\text{vir}-p}$ only partially capture the change in the age composition of the population living with **HIV** and the population exhibiting viraemia, as the age of the study population is restricted to 15-49 years.

6.4 Findings

6.4.1 Age and gender burden profiles of **HIV** infection

To describe the burden of **HIV** infection, we initially followed the typical approach to reporting **HIV** seroprevalence and estimated the proportion of individuals with **HIV** infection (i. e. , positive **HIV** serostatus) in each 1-year age band (Fig. 6.5c, orange). Despite close geographical proximity, **HIV** seroprevalence was higher in fishing communities along Lake Victoria than in the inland communities [Chang et al., 2016]. In 2019, **HIV** seroprevalence was 32.8% (95% posterior credible interval: 31.5%-34.2%) in

Table 6.1: Characteristics of populations with and without HIV in the **RCCS**, in round 19 (June 2018-May 2019)

Location	Gender	Age group	Census-eligible, individuals in each age group		Proportion of PLHIV, in each age group		Proportion of PLHIV, exhibiting viraemia in each age group		Proportion of people exhibiting viraemia in each age group		Age profile of people with HIV		Age profile of people exhibiting viraemia	
			+	±	+	±	+	±	+	±	+	±	+	±
Inland	Women	Total	12,893 (52.35%)	13.9% (13.2%-14.6%)	13.2% (11.5%-15.0%)	1.8% (1.6%-2.1%)	66.2% (63.8%-68.4%)	49.7% (43.6%-55.8%)						
		15-19	3,418 (13.88%)	1.6% (1.2%-2.1%)	29.6% (20.2%-40.9%)	0.7% (0.4%-1.0%)	2.0% (1.5%-2.6%)	4.9% (3.0%-7.3%)						
		20-24	2,372 (9.63%)	6.6% (5.8%-7.6%)	24.8% (19.8%-31.2%)	1.7% (1.3%-2.2%)	5.9% (5.1%-6.7%)	8.4% (6.5%-11.2%)						
		25-29	1,927 (7.82%)	17.3% (15.7%-18.9%)	18.2% (14.4%-21.7%)	2.8% (2.2%-3.5%)	12.2% (11.2%-13.3%)	11.3% (9.0%-14.0%)						
		30-34	1,619 (6.57%)	22.0% (20.2%-23.7%)	13.9% (11.3%-16.8%)	3.0% (2.3%-3.8%)	13.4% (12.4%-14.4%)	10.6% (8.2%-13.0%)						
		35-39	1,527 (6.20%)	22.6% (20.6%-24.7%)	10.3% (8.1%-13.7%)	2.5% (1.8%-3.2%)	12.7% (11.6%-13.7%)	7.9% (6.0%-10.2%)						
		40-44	1,234 (5.01%)	26.0% (23.8%-28.2%)	6.8% (4.8%-9.0%)	1.6% (1.1%-2.2%)	11.7% (10.8%-12.7%)	4.1% (2.9%-5.7%)						
		45-49	796 (3.23%)	28.1% (25.0%-31.3%)	5.3% (3.2%-8.2%)	1.2% (0.7%-2.0%)	8.2% (7.3%-9.1%)	2.0% (1.1%-3.3%)						
		Total	11,736 (47.65%)	7.8% (7.1%-8.5%)	26.3% (22.2%-30.8%)	2.0% (1.7%-2.4%)	33.8% (31.6%-36.2%)	50.3% (44.2%-56.4%)						
		15-19	3,070 (12.46%)	1.1% (0.8%-1.5%)	39.2% (25.2%-54.8%)	0.5% (0.3%-0.8%)	1.3% (0.9%-1.7%)	3.0% (1.7%-5.0%)						
Fishing	Women	20-24	2,013 (8.17%)	2.5% (2.0%-3.1%)	49.1% (38.0%-60.6%)	1.2% (0.9%-1.8%)	1.9% (1.5%-2.3%)	5.4% (3.9%-7.4%)						
		25-29	1,719 (6.98%)	5.8% (4.9%-7.1%)	47.3% (37.9%-57.4%)	2.9% (2.1%-3.9%)	3.7% (3.1%-4.4%)	10.5% (8.0%-13.8%)						
		30-34	1,523 (6.18%)	10.8% (9.3%-12.7%)	33.8% (26.6%-41.6%)	3.6% (2.7%-4.9%)	6.0% (5.2%-6.9%)	11.6% (8.9%-14.9%)						
		35-39	1,367 (5.55%)	15.3% (13.4%-17.5%)	20.5% (14.9%-27.0%)	3.0% (2.1%-4.3%)	7.6% (6.8%-8.5%)	8.6% (6.2%-11.6%)						
		40-44	1,156 (4.69%)	17.4% (15.2%-19.7%)	15.0% (10.3%-20.9%)	2.3% (1.5%-3.4%)	7.6% (6.7%-8.5%)	5.9% (4.0%-8.3%)						
		45-49	888 (3.61%)	17.6% (14.7%-20.8%)	16.6% (10.9%-24.1%)	2.5% (1.6%-3.9%)	5.7% (4.8%-6.7%)	4.7% (3.0%-7.1%)						
		Total	24,629 (100.00%)	11.0% (10.5%-11.5%)	17.6% (15.8%-19.6%)	1.9% (1.7%-2.2%)	100.00%	100.00%						
		Total	3,236 (44.71%)	40.5% (38.6%-42.5%)	11.4% (9.3%-13.8%)	4.6% (3.7%-5.6%)	55.2% (53.1%-57.3%)	43.7% (36.5%-51.1%)						
		15-19	474 (6.55%)	9.0% (6.5%-12.2%)	24.0% (13.1%-40.7%)	3.7% (1.8%-5.7%)	1.8% (1.3%-2.4%)	5.1% (2.5%-7.8%)						
		20-24	605 (8.36%)	21.8% (19.1%-24.6%)	21.0% (15.0%-28.7%)	4.9% (3.6%-7.2%)	5.6% (5.0%-6.3%)	9.0% (6.6%-12.8%)						
Men	Men	25-29	657 (9.08%)	40.6% (37.4%-44.0%)	12.2% (8.0%-16.8%)	4.5% (3.0%-6.1%)	11.3% (10.4%-12.1%)	8.7% (5.9%-11.5%)						
		30-34	586 (8.10%)	55.2% (51.8%-58.6%)	10.5% (7.6%-14.0%)	5.1% (3.8%-7.4%)	13.7% (12.9%-14.5%)	8.9% (6.7%-12.6%)						
		35-39	467 (6.45%)	61.2% (57.4%-64.8%)	9.6% (6.5%-13.9%)	5.1% (3.6%-7.7%)	11.6% (11.0%-12.3%)	6.7% (4.9%-10.1%)						
		40-44	313 (4.32%)	60.1% (55.5%-64.6%)	6.4% (3.7%-10.1%)	3.7% (2.1%-5.4%)	7.9% (7.4%-8.5%)	3.4% (2.0%-4.9%)						
		45-49	134 (1.85%)	55.3% (47.2%-63.0%)	5.4% (1.8%-10.8%)	3.1% (1.1%-5.2%)	3.2% (2.7%-3.6%)	1.3% (0.4%-2.1%)						
		Total	4,002 (55.29%)	26.5% (24.7%-28.5%)	18.1% (14.5%-22.3%)	4.8% (3.8%-5.9%)	44.8% (42.7%-46.9%)	56.3% (48.9%-63.5%)						
		15-19	428 (5.91%)	1.7% (0.7%-3.5%)	21.2% (8.1%-44.2%)	0.8% (0.2%-2.4%)	0.3% (0.1%-0.6%)	1.0% (0.3%-3.0%)						
		20-24	563 (7.78%)	5.5% (3.8%-7.6%)	25.2% (14.6%-40.7%)	1.7% (0.9%-3.0%)	1.4% (1.0%-1.9%)	3.0% (1.6%-5.2%)						
		25-29	796 (11.00%)	18.6% (15.5%-22.5%)	26.5% (18.3%-36.1%)	5.2% (3.6%-7.5%)	6.2% (5.2%-7.3%)	12.2% (8.8%-16.4%)						
		30-34	776 (10.72%)	34.5% (30.5%-39.0%)	25.7% (18.3%-38.0%)	8.6% (6.1%-12.3%)	11.1% (10.0%-12.3%)	19.6% (14.8%-25.9%)						
Total	Total	35-39	666 (9.20%)	40.9% (36.7%-45.0%)	15.0% (10.2%-20.4%)	5.8% (4.0%-8.3%)	11.4% (10.3%-12.4%)	11.4% (7.9%-15.3%)						
		40-44	469 (6.48%)	42.9% (38.6%-47.3%)	9.7% (5.8%-14.3%)	3.6% (2.3%-5.5%)	8.7% (7.8%-9.5%)	5.2% (3.3%-7.6%)						
		45-49	304 (4.20%)	45.8% (39.3%-52.8%)	9.6% (5.2%-16.3%)	3.5% (1.8%-6.4%)	5.7% (4.9%-6.6%)	3.0% (1.6%-5.5%)						
		Total	7,238 (100.00%)	32.8% (31.5%-34.2%)	14.4% (12.4%-16.7%)	4.7% (4.0%-5.5%)	100.0%	100.0%						
		Total	7,238 (100.00%)	32.8% (31.5%-34.2%)	14.4% (12.4%-16.7%)	4.7% (4.0%-5.5%)	100.0%	100.0%						

† (n, (% of total population))

‡ Posterior median estimate, (95% CrI)

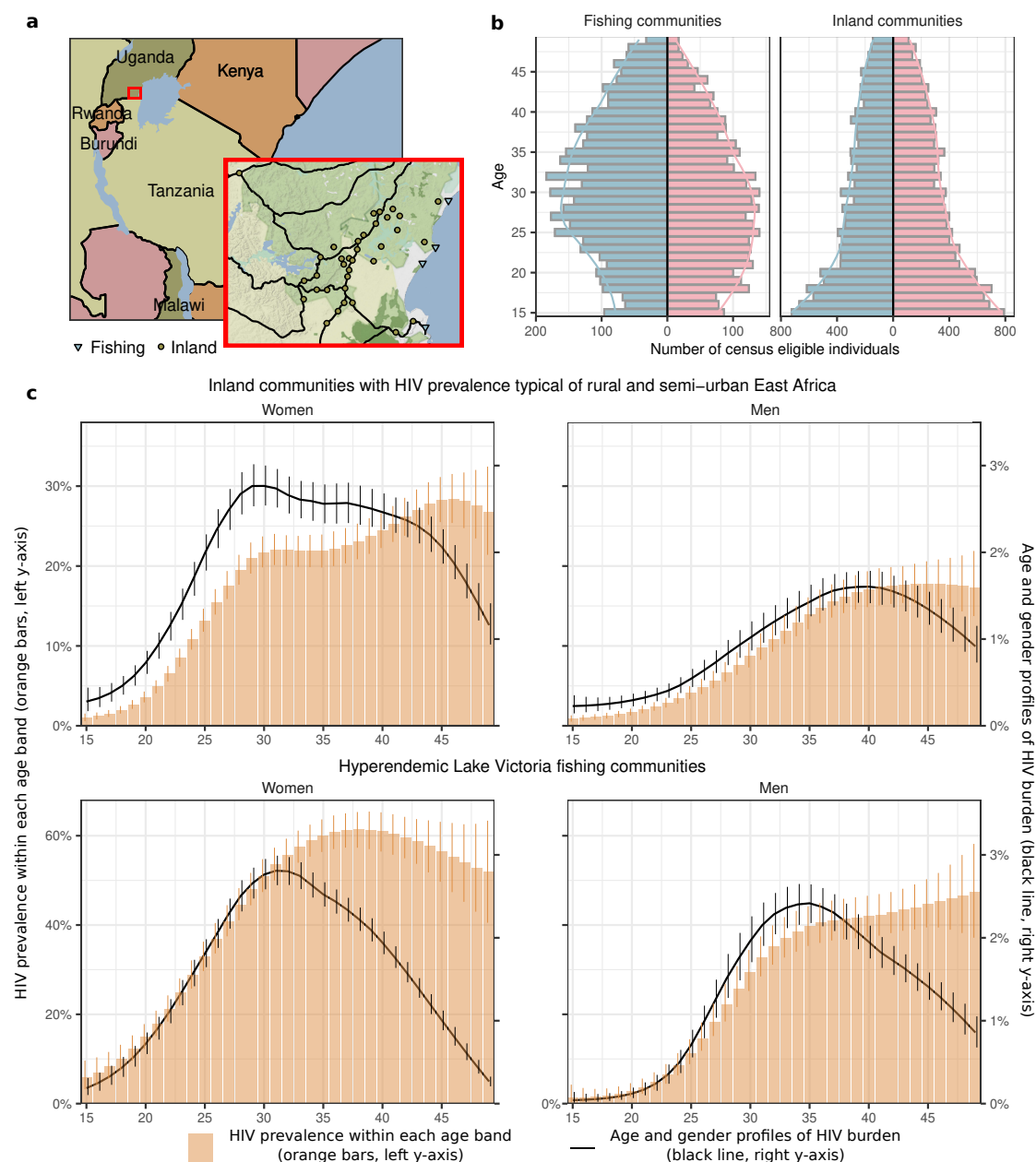


Figure 6.5: Age and gender profiles of HIV burden in the Rakai Community Cohort Study, June 2018-May 2019 (round 19). (a) Map of the fishing (triangles) and inland (circles) study communities within the Rakai region in south-central Uganda, produced using the ggmap package version 3.0.2 [Kahle and Wickham, 2013] in R (©Stadia Maps ©Stamen Design ©OpenMapTiles © OpenStreetMap contributor) (b) Population pyramids in fishing and inland communities. For each 1-year age band, the number of census-eligible individuals (colored bars) are shown for men (blue) and women (pink). Colored lines represent smoothed population size estimates. (c) The profile for the burden of HIV in the population is shown in black (line: posterior median estimate, linerange: 95% credible interval) and values shown on the right y-axis, for both community types (top, bottom) and both genders (left, right). For comparison, HIV seroprevalence within each 1-year age band is shown in orange (bars: posterior median estimates, error bar: 95% credible interval), and values shown on the left y-axis.

fishing communities, versus 11.0% (10.5%-11.5%) in inland communities (Table 6.1), and it was larger among first-time participants (Fig. 6.2).

The distinct population pyramids of these communities also prompted us to characterize the HIV infection burden profile (Fig. 6.5c). We found that this profile centered more strongly on younger populations than statistics of HIV seroprevalence within each 1-year age band might suggest, such that some younger populations with relatively lower HIV seroprevalence accounted for a higher fraction of the HIV infection burden. For example, by 2019, in inland communities half (51.4% (49.3%-53.5%)) of all women with HIV were 27-39 years old and half (53.1% (49.5%-56.3%)) of all men with HIV were 34-45 years old.

Differences in the gender burden of HIV infection across community types depended both on differences in seroprevalence and population structures. In 2019, HIV seroprevalence among women was 1.53 (1.40-1.66) fold higher than among men in fishing communities and 1.78 (1.61-1.98) fold higher among women than among men in inland communities. Further, as women accounted for a smaller proportion of the population in fishing communities (44.3%) as opposed to in inland communities (52.2%), women made up a smaller fraction of the PLHIV in fishing communities (55.2% (53.1%-57.3%)) rather than in inland communities (66.2% (63.8%-68.4%)).

6.4.2 “95-95-95” targets are achievable in Rakai

We next characterized progress towards “95-95-95” targets between 2013 and 2019 in the high-prevalence fishing communities that have since 2013 been prioritized for TasP, and in the inland communities where UTT was implemented beginning in late 2016 (see Table 1.1). The number of people with viraemia decreased substantially from 2013 to 2019 (Table 6.2). Specifically, in inland communities, the proportion of women with HIV exhibiting viraemia decreased from 34.7% (32.3%-37.1%) in 2013 to 13.2% (11.5%-15.0%) in 2019, while the proportion of men with HIV exhibiting viraemia decreased from 50.7% (46.5%-54.8%) in 2013 to 26.3% (22.2%-30.8%) in 2019. In fishing communities, the proportion of women with HIV exhibiting viraemia decreased from 44.0% (40.9%-47.0%) in 2013 to 11.4% (9.3%-13.8%) in 2019, while the proportion of men with HIV exhibiting viraemia decreased from 64.7% (60.8%-68.5%) in 2013 to 18.1% (14.5%-22.3%) in 2019.

Considering the number of people exhibiting viraemia in the entire population, by 2019, there were 233 (200-270) women and 236 (193-286) men with viraemia in inland communities, corresponding to 1.8% (1.6%-2.1%) and 2.0% (1.7%-2.4%) of the census-eligible female and male populations, respectively. In fishing communities, there were 148 (120-180) women and 190 (152-237) men exhibiting viraemia in 2019, corresponding to

Table 6.2: Changes in number of people exhibiting viraemia in the **RCCS**, from round 16 (July 2013-January 2015) to round 19 (June 2018-May 2019)

Location	Gender	Round	Census eligible individuals		Number of PLHIV		Number of people exhibiting viraemia		Proportion of PLHIV exhibiting viraemia		Proportion of people exhibiting viraemia		Percent reduction in the number of people exhibiting viraemia relative to R16		Male-female difference in the prop. of PLHIV exhibiting viraemia relative to R16	
			(n)	†	(n)	†	(n)	†	(%)	†	(%)	†	(%)	†	(%)	†
Fishing	Women	Round 16	2,757	1,225	537	44.0%	19.4%	–	–	–	–	–	–	–	–	–
		Round 17	3,224	1,315	363	27.8%	11.2%	32.4%	–	–	–	–	–	–	–	–
		Round 18	3,359	1,347	240	18.0%	7.2%	55.2%	–	–	–	–	–	–	–	–
		Round 19	3,236	1,312	148	11.4%	4.6%	72.4%	–	–	–	–	–	–	–	–
	Men	Round 16	3,546	1,052	673	64.7%	18.9%	–	–	–	–	–	–	–	–	20.7% (15.7, 25.5)
		Round 17	4,068	1,187	573	49.2%	14.0%	14.8%	–	–	–	–	–	–	–	21.3% (16.3, 26.3)
		Round 18	4,224	1,197	347	29.6%	8.2%	48.4%	–	–	–	–	–	–	–	11.6% (7.6, 15.6)
		Round 19	4,002	1,064	190	18.1%	4.8%	71.7%	–	–	–	–	–	–	–	6.7% (2.4, 11.4)
Inland	Women	Round 16	11,346	1,729	588	34.7%	5.2%	–	–	–	–	–	–	–	–	–
		Round 17	11,990	1,876	451	24.0%	3.8%	23.2%	–	–	–	–	–	–	–	–
		Round 18	12,193	1,728	336	19.3%	2.8%	42.7%	–	–	–	–	–	–	–	–
		Round 19	12,893	1,789	233	13.2%	1.8%	60.3%	–	–	–	–	–	–	–	–
	Men	Round 16	10,541	1,003	490	50.7%	4.7%	–	–	–	–	–	–	–	–	16.0% (11.0, 20.8)
		Round 17	10,939	995	399	39.9%	3.7%	18.5%	–	–	–	–	–	–	–	15.9% (11.3, 20.5)
		Round 18	11,076	912	306	34.3%	2.8%	37.4%	–	–	–	–	–	–	–	15.0% (10.2, 19.9)
		Round 19	11,736	915	236	26.3%	2.0%	51.8%	–	–	–	–	–	–	–	13.1% (39.4%-61.8k)

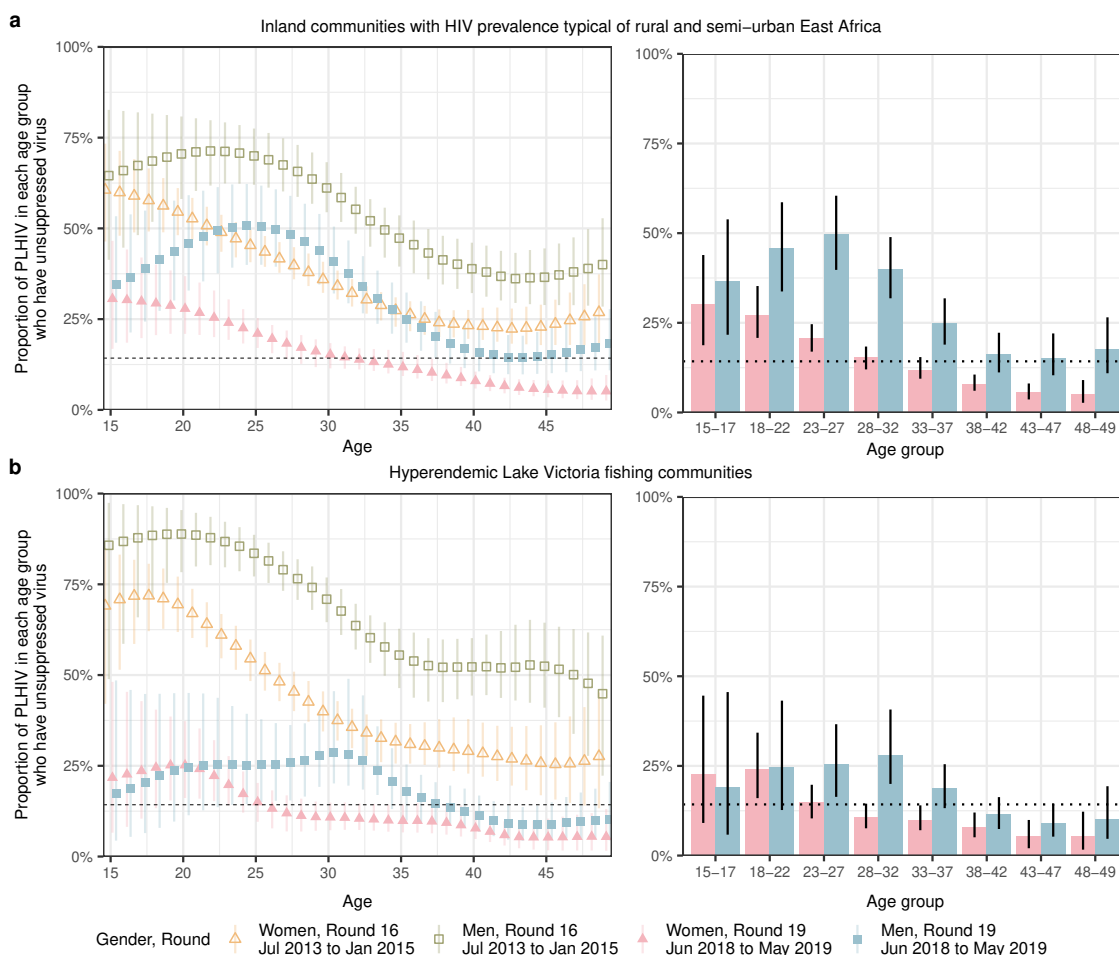


Figure 6.6: Trends in the proportions of PLHIV who have unsuppressed viraemia in the Rakai Community Cohort Study. (a) Estimated proportions of people with HIV in 1-year age bands who have unsuppressed viraemia, in men and women of inland communities. On the left, time trends in the proportion of PLHIV with unsuppressed viraemia from round 16 (July 2013–January 2015) to round 19 (June 2018 to May 2019) are shown for men and women by 1-year age bands (points: posterior median estimates, linerange: 95% CrIs). On the right, posterior median estimates in the proportion of PLHIV with unsuppressed viraemia in round 19 are shown for men and women by 5-year age bands (bars: posterior median estimates, linerange: 95% CrIs). The horizontal dotted line represents the 14% suppression threshold implied by 95-95-95 targets. (b) Same for PLHIV in fishing communities.

4.6% (3.7%-5.6%) and 4.8% (3.8%-5.9%) of census-eligible women and men, respectively (Tables 6.1 and 6.2). Thus, while suppression levels among PLHIV were higher in fishing communities compared to inland communities in both men and women, suppression levels in the population, i. e. among census-eligible individuals, were lower in fishing communities rather than inland communities due to the overall higher HIV prevalence.

6.4.3 Age and gender profiles of the burden of viraemia

To characterize the age and gender-specific populations in which viraemia concentrates, we first considered the proportion of PLHIV exhibiting viraemia stratified by age and gender. By 2019, in inland communities, the proportion of PLHIV exhibiting viraemia was highest in men aged 23-27 years (49.8% (39.9%-60.3%)) (Fig. 6.6), showing also that conventional 5-year age banding (15-19 years, 20-24 years ...) can mask the actual age groups in which prevalence of viraemia peaks. In fishing communities, the proportion of PLHIV exhibiting viraemia in 2019 was highest in men aged 28-32 years (28.2% (20.0%-41.2%)) (Fig. 6.6). Across all age groups, the proportion of PLHIV exhibiting viraemia was larger in first-time participants (Figs. 6.3 and 6.4), and lower, or equal in women than men. By 2019, among women, this proportion was highest in women aged 15-17 years in inland communities (30.3% (18.7%-44.0%)) and in women aged 18-22 years in fishing communities (23.9% (16.2%-34.0%)).

Then, we accounted for the distinct population structure in inland and fishing communities to characterize the population subgroups in which viraemia is concentrating. Fig. 6.7 compares the age and gender specific prevalence of viraemia among PLHIV to the viraemia burden profile in the population. We found that viraemia did not center in the age groups currently furthest from achieving “95-95-95” targets. Specifically, in inland communities, women aged 15-19 years were furthest from achieving the “95-95-95” targets in 2019, but we estimate they contributed only 4.9% (3.0%-7.3%) to people exhibiting viraemia. Instead, women aged 20-24 years and 25-29 years each contributed approximately 10% to people exhibiting viraemia in 2019, although these age groups were either close to or had already exceeded the “95-95-95” targets. Similarly, in inland communities, men aged 20-24 years were furthest from achieving “95-95-95” targets in 2019, but we estimate they contributed only 5.4% (3.9%-7.4%) to the burden of viraemia. Rather, men aged 25-29 years and 30-34 years each contributed approximately 10% to the burden of viraemia in 2019. Similar patterns were observed in fishing communities. In 2019, women aged 15-19 were one of the few groups that had not yet met “95-95-95” targets, but this group only contributed 5.1% (2.5%-7.8%) to the burden of viraemia. Instead, women aged 30-34 had met the targets but contributed 8.9% (6.7%-12.6%) to the burden of viraemia. Among men, age groups from 20 to 34 years exhibited similar levels of viral suppression among PLHIV. However, there were more men aged 30-34

exhibiting viraemia as compared to men aged 20-24. The former amounted to 19.6% (14.8%-25.9%) of the burden of viraemia as compared to 3.0% (1.6%-5.2%) in the latter group.

Fig. 6.7c depicts temporal changes in absolute suppression levels, as well as shifts in the profile of the burden of viraemia. Even though we found a steady increase in the median age of PLHIV by gender and community strata, the age profile of people exhibiting viraemia remained relatively stable over time (Table B.3).

6.5 Discussion

In this Chapter, we estimated the age and gender burden profiles of HIV infection and HIV viraemia in south-central Uganda using data from communities with HIV risk and demographic profiles typical of rural and semi-urban ESA and data from Lake Victoria fishing communities with high HIV seroprevalence and atypical demographic profiles. Overall, the burden profiles revealed important differences by geographic location, gender and age group. Following the earlier expansion of TasP programs, fishing communities witnessed faster increases in suppression levels among PLHIV than in inland communities, with overall levels surpassing the “95-95-95” targets by 2019. However, the burden of HIV viraemia remained larger in fishing communities, where the proportion of individuals exhibiting viraemia was approximately 2.5 times larger than in inland communities, reflecting higher HIV prevalence in fishing communities. When analyzing the burden of viraemia by gender, we found that in both community types, most individuals exhibiting viraemia were men, although fewer men have HIV than women. When also considering age, we found that the age groups that contributed most to the burden of viraemia were not necessarily the ones that were furthest from achieving the “95-95-95” targets. For example, in inland communities, individuals aged 25-34 contributed to approximately 40% of the burden of viraemia but had higher suppression levels than younger age groups.

Overall, our results show that the metrics considered by the “95-95-95” targets are insufficient to characterize the population exhibiting viraemia, and hence at risk of transmitting HIV. For example, by 2019, fishing communities had an overall prevalence of suppression among PLHIV that surpassed “95-95-95” targets, while inland communities had not yet achieved them. However, individuals exhibiting viraemia represented a larger proportion of the population in fishing communities as opposed to inland communities, consistent with HIV incidence remaining higher in the former communities despite additional interventions such as Pro-Exposure Prophylaxis [Ssempijja et al., 2022]. This is largely explained by differences in HIV prevalence and implies that reaching “95-95-95” by 2025 may be insufficient to lower incidence below elimi-

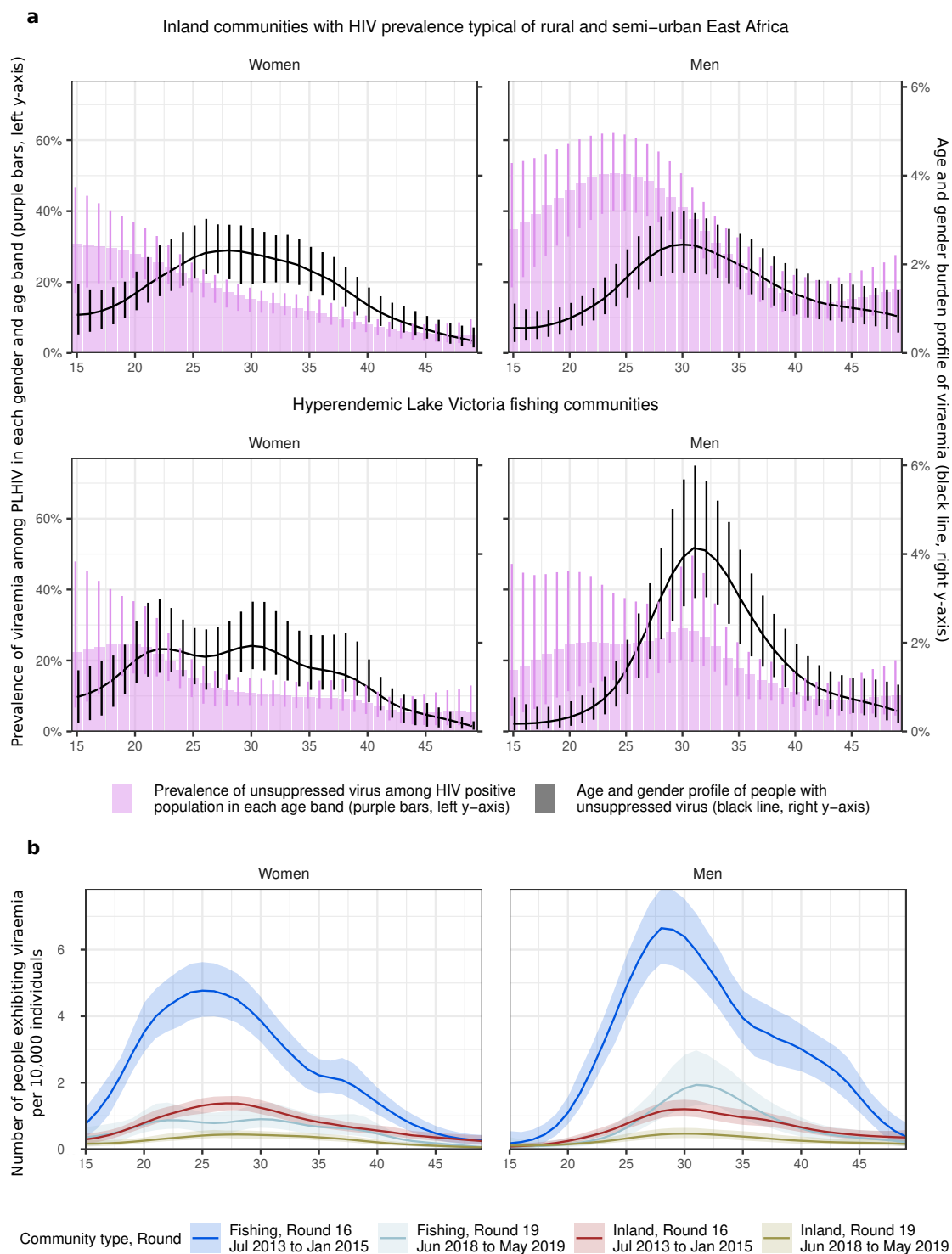


Figure 6.7: Age and gender metrics for people exhibiting viraemia. (a) Estimates for the prevalence of viraemia among **PLHIV** (median: purple bars, lines: 95% CrI, right y-axis) in each gender and age band are compared to the burden profile of viraemia (median: black curve, lines: 95% CrI, left y-axis) for inland (top row) and fishing communities (bottom row). (b) Time trends in the burden profiles of viraemia for both men and women (facets), colored by community type (inland or fishing) and survey round (16 or 19) colors (line: posterior median estimate, ribbon: 95% credible interval). The profiles were multiplied by the population prevalence of viral suppression to visualize shifts in the size of the population exhibiting viraemia.

nation threshold [Granich et al., 2009], especially in countries with elevated burden of HIV [Dimitrov et al., 2020].

Moreover, since 2020 the updated UNAIDS targets recommend the “95-95-95” goals to be met within subpopulations stratified by age and gender [UNAIDS, 2015]. In practice, this supports the prioritization of interventions to those subgroups that are furthest from achieving the targets and improving barriers and inequalities in treatment coverage [Frescura et al., 2022]. However, these subgroups may not disproportionately contribute to the burden of HIV viraemia. For example, in inland communities, we found that young individuals aged 15-24 constituted approximately two fifths of the census-eligible population, but only one fifth of the population exhibiting viraemia. In fact, the lower suppression levels commonly reported among young PLHIV [Mine et al., 2024, Lesotho Ministry of Health LB of S, 2020, Eswatini Ministry of Health and Central Statistical Office, 2022, Malawi Ministry of Health, 2022, Mozambique Ministry of Health (MISAU), 2022, Uganda Ministry of Health, 2022, Pathmanathan et al., 2021, Pillay et al., 2020, Nyaboke et al., 2023] are counterbalanced by the small absolute number of young PLHIV. This observation is typically missing from studies supporting prioritization of younger age-groups in ESA based on the UNAIDS targets [Hall et al., 2024]. However, there are instances in which higher suppression levels and HIV seroprevalence lead to balanced profiles. For instance, although most PLHIV are women, there are similar numbers of men and women exhibiting viraemia in inland communities. Taken together, these findings support the need to reach and treat a wide subset of the population: men aged 25-34 and women aged 20-39, who disproportionately contribute to unsuppressed viraemia.

While we assessed both age and gender burden profiles of HIV infection and HIV viraemia in this study, the latter metric provides a cross-sectional description of individuals at most at risk of HIV mortality and onward transmission. Hence, we propose the viraemia burden profile as a new metric for HIV control. Critically we observed that the burden profiles of viraemia revealed similar relative contributions of men and women, and larger relative contributions of middle-aged individuals (25-34) than younger individuals (15-24). We further show that these important observations are obscured by age-gender stratified “95-95-95” targets. Furthermore, unlike the “95-95-95” targets [Miller et al., 2013, Solomon et al., 2016], population burden of HIV viraemia has been shown to be strongly predictive of population HIV incidence [Lamarange et al., 2023, Patel et al., 2021]. Of course, the viraemia burden profile cannot be considered in isolation when designing interventions, as sexual behaviour of PLHIV also contributes crucially to transmission risk. For instance, younger women with HIV have been reported to have a higher exposure to pregnancy [Marston et al., 2018] and higher rates of partner acquisition [Calvert et al., 2023].

The statistical modelling choices made in this study have to be considered in the context of the studies aims. Due to the growing epidemiological interest in high-resolution, age-specific patterns in HIV and viraemia burden, we chose to use models which could provide prevalence estimates at the one-year bands level. To achieve this, it was necessary for the models to borrow information across similar age groups, prompting the use of Gaussian Processes. Additional information borrowing could have included other covariates, such as location and interview round, instead of treating these as independent strata. For instance, the Gaussian process kernel could have been extended as a product of kernels for age, location and round [Flaxman et al., 2015]. Inference may then have required low-rank approximations of the resulting Gaussian process, for example through the use of Hilbert space approximations [Riutort-Mayol et al., 2022], or random Fourier features [Miller and Reich, 2022]. The advantages may have then included tighter credible intervals but at the cost of increased computational complexity and, more importantly, harder generalizability to other settings in ESA. To elaborate on the last point, our study aimed to develop cross-sectional measures to describe the burden of HIV and viraemia, thus necessitating a model that accurately reflects a cross-sectional design. For these reasons we avoided borrowing information across rounds.

This study has several limitations. First, due to the RCCS definition of census-eligible individuals in the study period, our study population is limited to individuals aged 15-49 years. Accordingly, we are not able to quantify the growing seroprevalence of HIV among individuals aged 50 or more, nor viral suppression outcomes in children (<15 years) living with HIV. Importantly, the reported age and gender specific percentages in the burden profiles in the census-eligible population aged 15-49 years should not be interpreted as fractions of the entire population, as they would be overestimates. Further, an extension of our burden profiles to include children would provide more insights into the contributions of vertical transmission to the HIV epidemic, which is estimated to account for one fifth of new infections in Uganda CDC [2024]. Second, our study only considered data from Uganda, and our results may not directly generalize to other settings. However epidemiological patterns reported here are typical of ESA, such as: expansive population pyramids [Pyramid, 2023], patterns in HIV prevalence by age and gender [Dwyer-Lindgren et al., 2019], and gender gaps in suppression levels [Mine et al., 2024, Lesotho Ministry of Health LB of S, 2020, Eswatini Ministry of Health and Central Statistical Office, 2022, Malawi Ministry of Health, 2022, Mozambique Ministry of Health (MISAU), 2022, Uganda Ministry of Health, 2022, Pathmanathan et al., 2021, Pillay et al., 2020, Nyaboke et al., 2023]. These observations suggests that the major contributions of individuals aged 25-34 to the burden of viraemia may generalize across the subcontinent. Third, to generalize our results to the entire census-eligible population, we assumed that first-time study participants were representative of non-participants in

terms of age and gender-specific prevalence and viral suppression levels. Despite similar **HIV** seroprevalence estimates, our estimates of viral suppression among first-time participants were lower than for all participants. Generalizing the results to the census-eligible population avoided overestimation of viral suppression levels and allowed us to match our analyses' denominators to the ones defined in the **UNAIDS** goals. Fourth, the age profile of the population exhibiting viraemia may not match the age profile of the sources of **HIV** infections, but rather the potential sources of **HIV** transmission. The difference between the two populations can be explained by heterogeneities in transmission risk among individuals exhibiting viraemia. For example, phylogenetic studies in Uganda [Monod et al., 2023b], Zambia [Hall et al., 2024], and Botswana [Magosi et al., 2022] have estimated highest transmission risks in men aged 30-34 and women aged 20-24, and much lower for example among individuals aged more than 50. These results hence reinforce the importance of the subpopulations contributing most strongly to the burden of viraemia described in this Chapter.

7

DISCUSSION

The aim of this thesis consisted in developing statistical models and analyses to inform public health response to epidemics. The results of this work are demonstrated through four case studies on **COVID-19** and **HIV**, which reflect the rapidly changing questions and interests as an epidemic unfolds and evolves. Under this perspective, the Chapters aims can be reformulated as follows. Chapter **3** focuses on the very early stages of an outbreak, when the interest lies in understanding the growth of the cases. In Chapter **4**, the interest concerns reassessing disease severity following mutations in the pathogen. Other changes can be brought about by public health response and interventions; in Chapter **5** we aim to improve the detection of changes in transmission patterns over time. Finally, in Chapter **6** we provide novel metrics to assess progress towards the end of **AIDS** as a public health threat. Our work was mostly focused on **LMIC** as these countries are the ones that suffer most from infectious diseases. We also made an effort to make our code and data publicly available, in line with the open science movement. In the following, we discuss the motivations, implications, strengths, limitations, and future research directions for each piece of work.

Chapter 3: Refined R_0 estimates in emerging epidemics

In the early stages of an epidemic outbreak, one of the most pressing public health issues concerns the prediction of the short-term burden of infections. The predicted burden has to be compared to available resources, such as hospital personnel, beds, and ventilators, whose absence can turn a curable disease into a deadly one. Early projections of disease incidence are based on cases reported by surveillance systems, which typically suffer from a variety of data-collection biases. In the case of **COVID-19**, these biases were accentuated by the fact that many cases were asymptomatic, and the disease

could be confused with others, such as influenza. As a consequence, in most countries around the world, SARS-CoV-2 was circulating before the first cases were reported to the national surveillance systems. Not accounting for these unobserved infections can lead to confounded estimation of the time-varying reproduction number [O’Driscoll et al., 2021] which in turn can be used to attribute changes in transmissibility to early measures whose real effect may be limited, or overestimated.

In our work, we addressed this limitation for EpiEstim, one of the most popular statistical packages to infer changes in the reproduction number R_t . The statistical challenge concerned missing data: in this instance, unreported infections that took place before the first reported case. Assuming that early cases follow an exponential growth, we propose a two-step procedure to avoid overestimation of the reproduction number: first, we impute cases in the period prior to the setup of surveillance systems, and secondly we fit the EpiEstim model to both the imputed and the observed cases. We evaluated this approach on simulated data, showing a reduction in bias in R_0 estimates, and an increase in coverage of the 95% credible intervals. We then applied this method to early COVID-19 data from Italy, the UK, Sweden, and the US, showing that the adjusted estimates were closer to those obtained by others on longer time series [Flaxman et al., 2020, Ke et al., 2021].

The main strengths of our approach are simplicity and generalisability. Alternative approaches to estimate the time-varying reproduction number exist [Bhatt et al., 2023, Scott et al., 2020], but due to their flexibility these require more work to set up, and take longer to run. On the other hand, our method runtime is negligible. Moreover, the two-step procedure involving back-imputation could be applied to other statistical methods, and is not exclusive to EpiEstim.

One limitation of our approach is that we do not quantify the uncertainty in the imputed cases. A fully Bayesian approach could be used to propagate this uncertainty in the estimate of the reproduction number, but this would imply the inability to express the posterior of EpiEstim in closed form. Instead, our imputation procedure requires the inference of two parameters for a linear model, and then considers the imputed cases as known, and fixed. We believe that, due to the relatively (exponentially) smaller magnitude of imputed cases compared to the observed cases, this simplification is likely to have limited impact on the final estimates. Secondly, the limitations of EpiEstim still apply. Most notably, the estimates of the reproduction number are dependent on the assumed generation/serial interval. In practice, this is generally set to be equal to serial interval of similar diseases as it is unlikely to obtain this information from very early outbreak data. Future work may focus on evaluating tools that do not depend on

such estimates, and determine whether a time-varying exponential growth rate r_t can replace the reproduction number R_t for early-outbreak public health decision-making.

Chapter 4: Estimating Fluctuating Hospital Fatality Rates in LMIC settings

Rapidly spreading pathogens, such as SARS-CoV-2, have the potential to rapidly mutate. Most of the mutations are typically neutral or deleterious for the pathogen, but some of these can confer a selective advantage and hence increase transmission risk, and potentially change the associated disease's severity. Changes in transmissibility are relatively easy to detect: these are typically followed by an increased in the estimated reproduction number with a simultaneous replacement of the dominant strain in the pathogen's population. The former is typically detected through an increase in reported cases, while the latter is typically detected through genomic surveillance. However, changes in disease severity are harder to evaluate, as they require comparison of health outcomes for a cohort of individuals whose infecting strain is known. These studies are easier to perform in HIC, where high quality, linked genomic surveillance and hospital records are available [Funk et al., 2021, Twohig et al., 2022, Nyberg et al., 2021, 2022, 2023]. Nonetheless, viral mutations are likely to appear in LMIC, as these comprise a large share of the world's population. In summary, changes in disease severity associated with pathogen mutations may only be assessed once the pathogen has reached HIC, which may be too late to inform public health responses in either settings. Exploratory studies on the associations between viral mutations and disease severity in LMIC are therefore crucial to improve strategic planning in epidemic control.

An example of this problem concerned the emergence of the Gamma variant in Brazil, whose appearance coincided with an increase in the number of cases and large fluctuations in in-hospital fatality rates across the country. To disentangle contributions to fatality rates from the viral variants, from mismatches in hospital resources and from other locations-specific factors, we built a Bayesian model estimating the proportion of hospitalised patients with either variant, and used this to evaluate the effect of the variant on fatality rates in 14 Brazilian state capitals. Our model synthesised data from different sources: genomic surveillance data was used to reconstruct replacement dynamics, census and vaccination data were used to quantify the amount of people at risk of hospitalisation and hospital records were used to estimate changes in fatality rates by age. The model for each city comprised 75 parameters in total. The variant replacement subcomponent included 3 parameters, the model for the age-composition of hospital admission included 23 parameters, and the in-hospital fatality rates subcomponent included 49 parameters. Model diagnostics involved checking $\hat{R}$ statistics, trace plots,

effective sample sizes as well as visual assessment of different submodel fits. We further evaluated the robustness of our inference with respect to different data sources or modelling choices through the sensitivity analysis presented in the Chapter.

Our results suggest that conditional on hospitalisation, the Gamma variant was not associated with higher mortality. Rather, the observed fluctuations in fatality rates over time were largely explained by the mismatch of hospital resources and hospital demand, which likely increased due to the rapid spread of the variant. For example, the city of Belo Horizonte did not witness changes in hospital-fatality rates despite the variant's presence, likely due to better pandemic preparedness and higher hospital resources. In our counterfactual analysis, we showed that one out two deaths could have been averted had the fatality rates in Belo Horizonte been preserved across all state capitals.

There are two wider implications: Firstly, although most studies assessing disease severity are performed in **HIC**, analysing data from **LMIC** can provide crucial insights to inform global public health response. Secondly, our results suggest that investment in pandemic preparedness and hospital resources are critical to avoid extreme shocks in health systems brought about by changes in pathogen transmissibility.

Nonetheless, limitations due to data quality and availability are present. Our model did not account for the effect of the variant on the probability of hospitalisation. It could be the case that Gamma had an increased severity largely mediated through an increase in the risk of hospitalisation. Although our analysis considered data from different locations in Brazil, we could not fit a unique hierarchical model fitting to all locations. Borrowing information across different states may allow for more precise estimation of variant and location effects, and may be helpful to generalise similar analyses to multiple countries.

Chapter 5: HIV-TSI: pipeline and applications

Following public health interventions, the transmission dynamics of a pathogen can change. For instance, school closures can reduce the number of contacts between children yielding decreases in the number of **COVID-19** cases [Wu et al., 2022]. In the case of **HIV**, transmission dynamics have been changing over time as observed through world-wide decreases in incidence [Frank et al., 2019] following the implementation of different interventions, for example **TasP**. Changes in transmission patterns also mean that the populations at highest risk of infection and at highest risk of transmission may change, highlighting the need for continuous monitoring of transmission dynamics.

The work in this Chapter aimed to extend existing models for estimation of transmission flows to account for changes in transmission dynamics over time. In summary,

we completed three tasks. Firstly we wrote and shared a phylogenetic pipeline to obtain **TSI** estimates from **HIV** sequences through the machine learning algorithm HIV-phyloTSI [Golubchik et al., 2022]. Secondly, to guarantee consistency between a patient's estimated **TSI** estimates, her testing history and a known transmission network, we developed a method to refine the date of transmission events. This method proceeds by excluding all combinations of infection dates inconsistent with the known data, and successively imposing a probability distribution over the remaining dates. Finally, we applied these methods to the **RCCS** transmission cohort to date infection events, and then fed the dated transmission events into a Bayesian model to estimate transmission flows over two periods of time. This transmission flow model did not take into account uncertainty in the **TSI** estimates. The reasons are twofold. Firstly, the model used estimated dates of infection to classify pairs into two periods. In its current formulation, the model is unable to account for uncertainty in the classification of pairs, although it might be possible to extend the framework to account for such uncertainty, analogously to Bu et al. [2024], where preclassification of transmission status is not necessary. Secondly, although HIV-phyloTSI provides prediction ranges (as quantile of each individual regression tree), it is not a Bayesian method and its training does not guarantee quality of uncertainty quantification. One of the simplest approaches to improve the quality of uncertainty quantification may concern the use of model-free prediction intervals [Salnikov et al., 2024].

In summary, from 2003 to 2018, we showed increases in the average age of men who are sources of **HIV** infections, and that women aged 15-24 make part a decreasing proportion of new infections. Instead, women aged 25-34 are increasingly becoming the main recipients infections. These findings suggest that **PEPFAR**'s focus on prevention to young and adolescent women may need to be extended to middle-aged women [United States Department of State, 2022] to further reduce the burden of new infections.

However, the applications of our phylogenetic pipeline for estimation of **HIV TSI** may be more general. We highlight two possibilities for future research. Firstly, the time since infection predictions obtained from sequence data may lead to novel methods to estimate **HIV** incidence without requiring expensive longitudinal studies. These approaches would require adjustments to account for differences in the training data used in Golubchik et al. [2022] and sequence data from surveillance cohorts, which likely includes fewer long-term infections. Secondly, we introduced a method to impute infection dates for a network of source-recipient pairs, while assuming that the network is known. Further extension may include a joint model for the estimation of the network and the infection dates, which could be adapted from previous literature. Specifically, these extension may look especially promising to analyse **HIV** contact tracing data [McGowan et al., 2024], which may become a more widespread intervention in **ESA**, and

would allow to analyse bigger transmission networks than previously possible [Ratmann et al., 2019].

Chapter 6: Profiles of HIV infection burden and viraemia

The successful scale up of ART around the world has led to notable decreases in the number of deaths caused by HIV and in the number of new infections. Despite this, much work needs to be done to end the AIDS epidemic as a public health threat by 2030. For instance, incidence still remains higher than the estimated elimination threshold of 0.1% [Granich et al., 2009], and many key populations are still left behind due to stigma and poorer access to health services [UNAIDS, 2020]. As such, population-level statistics such as incidence only paint a partial picture of a much more complex epidemic, and metrics additional to the UNAIDS “95-95-95” targets are needed to 1. define progress towards the end of the HIV epidemic [Galvani et al., 2018, Ghys et al., 2018] and 2. to identify which groups would benefit most from targeted interventions.

In Chapter 6, we assessed progress towards the UNAIDS’ “95-95-95” targets in the RCCS from 2013 to 2019. The statistical challenges consisted in obtaining fine-grained estimates of different types of prevalence for every one-year age group, which we addressed through the use of GPs. In particular, estimates for the different measures of prevalence relied on learning hyperparameters of a squared exponential kernel by fitting it to empirical data. Model diagnostics included controlling the number of divergent transitions, the effective sample size, and the $\hat{R}$ statistics. Model fit was also assessed through posterior predictive checks. We found that the prevalence of HIV in the RCCS decreased from 2013 to 2019, and that the prevalence of viraemia decreased in all age groups. Further, we found that although younger groups were furthest from achieving the “95-95-95” targets, they were not the group that made up the largest proportion of individuals exhibiting viraemia. To better describe which population subgroups make up the largest share of the people living with HIV, or exhibiting viraemia, we introduced the age- and gender- profiles for the burden of HIV and viraemia

Our results suggest that although women constitute the majority of people living with HIV, men constitute the majority of people exhibiting viraemia. As such we suggest strengthening prevention programs providing ART to men. Our results are in line with other literature, such as the *Population-based HIV Impact Assessment (PHIA)* surveys [Lesotho Ministry of Health LB of S, 2020, Uganda Ministry of Health, 2022, Malawi Ministry of Health, 2022, Mozambique Ministry of Health (MISAU), 2022], which however only focus on suppression rates among different groups of PLHIV, but without providing a comprehensive overview of the burden of HIV and viraemia.

Future research is needed to assess whether the gap in suppression levels between men and women living with HIV will shrink. Also, it might be interesting to consider burden profiles of viraemia which are not grouped by age and sex, but also by other characteristics, such as income, education or residence. Such statistics are relatively easy to compute as they rely on typically known information: group-specific estimates of HIV prevalence, prevalence of viraemia, and population sizes.

7.1 Conclusions

In this thesis, we presented four cases studies on how Bayesian statistical methods can be used to inform public health response to epidemics. We typically chose a Bayesian approach due to its flexibility, the ability to incorporate prior information in settings where data is scarce, and, importantly, the interpretability of the results. In the last years however, the rapid expansion of AI has raised important discussions on the relative importance of accuracy and interpretability: deep learning models have been used to solve problems that were intractable with traditional methods [Jumper et al., 2021, Liu et al., 2019], but they are often black boxes without guarantees on the quality of out-of-distribution prediction [Guérin et al., 2022, Rössli et al., 2022]. In epidemiological applications, it is hence crucial to understand in which applications the benefits of increased prediction accuracy outweigh the need for interpretability. The answer to the question depends on the sensitivity of the data, on the model predictive power and its assessment, on the availability of high quality data, on the ability to incorporate information from data sources which are currently neglected, or unusable and on the ability to perform counterfactuals.

One of the most promising applications of AI in the field regards its integration to existing mechanistic models of transmission [Ye et al., 2025]. For instance, integrating non-traditional data sources such as social media trends [Qin et al., 2020], search queries, satellite imagery [Pley et al., 2021] and wastewater analyses [Abdeldayem et al., 2022] to traditional surveillance system could enhance early outbreak detection. Such approaches could extend our framework in Chapter 3, where the number of missed generations of infections and its impact on transmission could be informed by multiple data sources.

Models of transmissions could further be extended to model economic and behavioral factors, providing more comprehensive counterfactuals to inform public health intervention. For instance, a major criticism of the non-pharmaceutical intervention against the spread COVID-19 regarded the lack of consideration of the economic consequences [Yamaka et al., 2022]. Capturing economic factors could provide weaker or stronger support for the implementation of interventions. For instance, early assessment of

a highly transmissible variant could avoid a later need for measures hindering economic activity [Shami and Lazebnik, 2022]. In the context of HIV, agent-based models [Pickles et al., 2021] could capture a more dynamic evolution of individuals behaviour following infection, diagnosis and treatment. In summary, the resulting simulations and counterfactuals could provide more detailed insights, extending the scope of our counterfactuals in Chapter 4, or potentially describing the evolution of HIV incidence after updating global guidelines and targets, as suggested in Chapter 6.

Finally, in recent years there have been significant breakthroughs steps in the development of generative models, such as large language models [Brown et al., 2020], or protein folding models [Jumper et al., 2021]. Their ability to simulate complex structure-mirroring the complexity of real-world language and proteins-has sparked optimism in their wider applications, potentially in phylogenetics [Smith and Hahn, 2023]. Augmenting the collected sequence data with synthetic sequences could improve training of models such as HIV-phyloTSI, and increase the amount of information that could be learnt from phylogenetic analyses. Nonetheless, training of these models typically requires large amounts of high quality data, and their validation may be challenging. For this reason, more immediate opportunities might involve leveraging existing tools to help healthcare providers and researchers. For instance, meta-analyses could be significantly streamlined by models capable of processing and parsing vast volumes of literature, while code completion could improve the development and maintenance of reproducible code and reproducible analyses [Husein et al., 2025].

Despite these opportunities, the integration of AI into epidemiology is not without challenges. Rigorous evaluation of model robustness, sensitivity, and generalizability is essential in real-world settings. Similarly, data privacy and algorithmic biases raise important ethical considerations that need to be addressed. Collaboration between epidemiologists, data scientists, and policymakers will be crucial to ensure that AI models are grounded in domain expertise and address real-world public health priorities.

BIBLIOGRAPHY

- O. M. Abdeldayem, A. M. Dabbish, M. M. Habashy, M. K. Mostafa, M. Elhefnawy, L. Amin, E. G. Al-Sakkari, A. Ragab, and E. R. Rene. Viral outbreaks detection and surveillance using wastewater-based epidemiology, viral air sampling, and machine learning techniques: A comprehensive review and outlook. *Science of The Total Environment*, 803:149834, Jan. 2022. ISSN 00489697. doi: 10.1016/j.scitotenv.2021.149834. URL <https://linkinghub.elsevier.com/retrieve/pii/S0048969721049093>.
- L. Abeler-Dörner, M. K. Grabowski, A. Rambaut, D. Pillay, C. Fraser, and o. b. o. t. P. Consortium. PANGEA-HIV 2: Phylogenetics And Networks for Generalised Epidemics in Africa. *Current Opinion in HIV and AIDS*, 14(3):173, May 2019. ISSN 1746-630X. doi: 10.1097/COH.0000000000000542. URL https://journals.lww.com/co-hivandaids/fulltext/2019/05000/pangea_hiv_2__phylogenetics_and_networks_for.5.aspx.
- S. Adjei, K. Hong, N.-A. M. Molinari, L. Bull-Otterson, U. A. Ajani, A. V. Gundlapalli, A. M. Harris, J. Hsu, S. S. Kadri, J. Starnes, K. Yeoman, and T. K. Boehmer. Mortality Risk Among Patients Hospitalized Primarily for COVID-19 During the Omicron and Delta Variant Pandemic Periods — United States, April 2020–June 2022. *MMWR. Morbidity and Mortality Weekly Report*, 71(37):1182–1189, Sept. 2022. ISSN 0149-2195, 1545-861X. doi: 10.15585/mmwr.mm7137a4. URL http://www.cdc.gov/mmwr/volumes/71/wr/mm7137a4.htm?s_cid=mm7137a4_w.
- R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. Hinton. Neural Additive Models: Interpretable Machine Learning with Neural Nets, 2020. URL <https://arxiv.org/abs/2004.13912>.
- Y. Alimohamadi, M. Taghdir, and M. Sepandi. Estimate of the Basic Reproduction Number for COVID-19: A Systematic Review and Meta-analysis. *Journal of Preventive Medicine and Public Health*, 53(3):151–157, May 2020. ISSN 1975-8375, 2233-4521. doi: 10.3961/jpmph.20.076. URL <http://jpmph.org/journal/view.php?doi=10.3961/jpmph.20.076>.
- L. J. Allen. A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling*, 2(2):128–142, Mar. 2017. ISSN 2468-2152. doi: 10.1016/j.idm.2017.03.001. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6002090/>.

- P. N. Amornkul, E. Karita, A. Kamali, W. N. Rida, E. J. Sanders, S. Lakhi, M. A. Price, W. Kilembe, E. Cormier, O. Anzala, M. H. Latka, L.-G. Bekker, S. A. Allen, J. Gilmour, P. E. Fast, and f. t. I. A. H. P. Partnership. Disease progression by infecting HIV-1 subtype in a seroconverter cohort in sub-Saharan Africa. *AIDS*, 27(17):2775, Nov. 2013. ISSN 0269-9370. doi: 10.1097/QAD.0000000000000012. URL https://journals.lww.com/aidsonline/fulltext/2013/11130/disease_progression_by_infecting_hiv_1_subtype_in.12.aspx.
- R. M. Anderson, R. M. May, M. C. Boily, G. P. Garnett, J. T. Rowley, and R. M. May. The spread of HIV-1 in Africa: Sexual contact patterns and the predicted demographic impact of AIDS. *Nature*, 352(6336):581–589, Aug. 1991. ISSN 1476-4687. doi: 10.1038/352581a0. URL <https://www.nature.com/articles/352581a0>.
- V. Andreasen. The Final Size of an Epidemic and Its Relation to the Basic Reproduction Number. *Bulletin of Mathematical Biology*, 73(10):2305–2321, 2011. ISSN 0092-8240. doi: 10.1007/s11538-010-9623-3. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7088810/>.
- S. Attia, M. Egger, M. Müller, M. Zwahlen, and N. Low. Sexual transmission of HIV according to viral load and antiretroviral therapy: Systematic review and meta-analysis. *AIDS*, 23(11):1397–1404, July 2009. ISSN 0269-9370. doi: 10.1097/QAD.0b013e32832b7dca. URL <https://journals.lww.com/00002030-200907170-00013>.
- B. Auvert, D. Taljaard, E. Lagarde, J. Sobngwi-Tambekou, R. Sitta, and A. Puren. Randomized, Controlled Intervention Trial of Male Circumcision for Reduction of HIV Infection Risk: The ANRS 1265 Trial. *PLOS Medicine*, 2(11):e298, Oct. 2005. ISSN 1549-1676. doi: 10.1371/journal.pmed.0020298. URL <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020298>.
- P. Bager, J. Wohlfahrt, J. Fonager, M. Rasmussen, M. Albertsen, T. Y. Michaelsen, C. H. Møller, S. Ethelberg, R. Legarth, M. S. F. Button, et al. Risk of hospitalisation associated with infection with SARS-CoV-2 lineage B.1.1.7 in Denmark: An observational cohort study. *The Lancet Infectious Diseases*, 2021. doi: 10.1016/S1473-3099(21)00290-5.
- S. Baichoo and C. A. Ouzounis. Computational complexity of algorithms for sequence comparison, short-read assembly and genome alignment. *Biosystems*, 156–157:72–85, June 2017. ISSN 0303-2647. doi: 10.1016/j.biosystems.2017.03.003. URL <https://www.sciencedirect.com/science/article/pii/S0303264717300461>.
- R. C. Bailey, S. Moses, C. B. Parker, K. Agot, I. Maclean, J. N. Krieger, C. F. Williams, R. T. Campbell, and J. O. Ndinya-Achola. Male circumcision for HIV prevention in young men in Kisumu, Kenya: A randomised controlled trial. *The Lancet*, 369(9562):

- 643–656, Feb. 2007. ISSN 01406736. doi: 10.1016/S0140-6736(07)60312-2. URL <https://linkinghub.elsevier.com/retrieve/pii/S0140673607603122>.
- F. Ball and P. Donnelly. Strong approximations for epidemic models. *Stochastic Processes and their Applications*, 55(1):1–21, Jan. 1995. ISSN 0304-4149. doi: 10.1016/0304-4149(94)00034-Q. URL <https://www.sciencedirect.com/science/article/pii/030441499400034Q>.
- F. Barré-Sinoussi, J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dautoguet, C. Axler-Blin, F. Vézinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. Isolation of a T-Lymphotropic Retrovirus from a Patient at Risk for Acquired Immune Deficiency Syndrome (AIDS). *Science*, 220(4599):868–871, May 1983. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.6189183. URL <https://www.science.org/doi/10.1126/science.6189183>.
- N. Bbosa, D. Ssemwanga, A. Ssekagiri, X. Xi, Y. Mayanja, U. Bahemuka, J. Seeley, D. Pillay, L. Abeler-Dörner, T. Golubchik, C. Fraser, P. Kaleebu, O. Ratmann, and on behalf of the MRC/UVRI & LSHTM Uganda Research Unit and The PANGAEA Consortium. Phylogenetic and Demographic Characterization of Directed HIV-1 Transmission Using Deep Sequences from High-Risk and General Population Cohorts/Groups in Uganda. *Viruses*, 12(3):331, Mar. 2020. ISSN 1999-4915. doi: 10.3390/v12030331. URL <https://www.mdpi.com/1999-4915/12/3/331>.
- M. Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo, July 2018. URL <http://arxiv.org/abs/1701.02434>.
- L. M. A. Bettencourt and R. M. Ribeiro. Real Time Bayesian Estimation of the Epidemic Potential of Emerging Infectious Diseases. *PLOS ONE*, 3(5):e2185, May 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0002185. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0002185>.
- S. Bhatt, N. Ferguson, S. Flaxman, A. Gandy, S. Mishra, and J. A. Scott. Semi-mechanistic Bayesian modelling of COVID-19 with renewal processes. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(4):601–615, Oct. 2023. ISSN 0964-1998. doi: 10.1093/jrssa/qnad030. URL <https://doi.org/10.1093/jrssa/qnad030>.
- M. A. Billah, M. M. Miah, and M. N. Khan. Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence. *PLOS ONE*, 15(11):e0242128, Nov. 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0242128. URL <https://dx.plos.org/10.1371/journal.pone.0242128>.
- E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman. Pyro: Deep universal probabilistic

- programming. *Journal of Machine Learning Research*, 20(28):1–6, 2019. URL <http://jmlr.org/papers/v20/18-403.html>.
- N. Brazeau, R. Verity, S. Jenks, H. Fu, C. Whittaker, P. Winskill, I. Dorigatti, P. Walker, S. Riley, R. P. Schnekenberg, H. Hoeltgebaum, T. A. Mellan, S. Mishra, J. T. Unwin, O. J. Watson, Z. M. Cucunubá, M. Baguelin, L. Whittles, S. Bhatt, A. C. Ghani, N. M. Ferguson, and L. C. Okell. Report 34: COVID-19 Infection Fatality Ratio: Estimates from Seroprevalence. *Imperial College London*, Oct. 2020. doi: 10.25561/83545.
- A. Brizzi, C. Whittaker, L. M. S. Servo, I. Hawryluk, C. A. Prete, W. M. De Souza, R. S. Aguiar, L. J. T. Araujo, L. S. Bastos, A. Blenkinsop, L. F. Buss, D. Candido, M. C. Castro, S. F. Costa, J. Croda, A. A. De Souza Santos, C. Dye, S. Flaxman, P. L. C. Fonseca, V. E. V. Geddes, B. Gutierrez, P. Lemey, A. S. Levin, T. Mellan, D. M. Bonfim, X. Miscouridou, S. Mishra, M. Monod, F. R. R. Moreira, B. Nelson, R. H. M. Pereira, O. Ranzani, R. P. Schnekenberg, E. Semenova, R. Sonabend, R. P. Souza, X. Xi, E. C. Sabino, N. R. Faria, S. Bhatt, and O. Ratmann. Spatial and temporal fluctuations in COVID-19 fatality rates in Brazilian hospitals. *Nature Medicine*, 28(7):1476–1485, July 2022. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-022-01807-1. URL <https://www.nature.com/articles/s41591-022-01807-1>.
- T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners, 2020. URL <https://arxiv.org/abs/2005.14165>.
- F. Bu, J. Kagaayi, M. K. Grabowski, O. Ratmann, and J. Xu. Inferring HIV transmission patterns from viral deep-sequence data via latent typed point processes. *Biometrics*, 80(1):ujad015, Jan. 2024. ISSN 0006-341X, 1541-0420. doi: 10.1093/biomtc/ujad015. URL <https://academic.oup.com/biometrics/article/doi/10.1093/biomtc/ujad015/7610191>.
- C. Calvert, R. Scott, M. Palmer, A. Dube, M. Marston, K. Wellings, and E. Slaymaker. Rates of sexual partner acquisition from nationally representative surveys: Variation between countries and by age, sex, wealth, partner and HIV status. *Sexual Health*, 21(1):NULL–NULL, Dec. 2023. ISSN 1449-8987. doi: 10.1071/SH23134. URL <https://www.publish.csiro.au/sh/SH23134>.
- M. S. Campbell, J. I. Mullins, J. P. Hughes, C. Celum, K. G. Wong, D. N. Raugi, S. Sorensen, J. N. Stoddard, H. Zhao, W. Deng, E. Kahle, D. Panteleeff, J. M. Baeten, F. E. McCutchan,

- J. Albert, T. Leitner, A. Wald, L. Corey, J. R. Lingappa, and for the Partners in Prevention HSV/HIV Transmission Study Team. Viral Linkage in HIV-1 Seroconverters and Their Partners in an HIV-1 Prevention Clinical Trial. *PLoS ONE*, 6(3):e16986, Mar. 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0016986. URL <https://dx.plos.org/10.1371/journal.pone.0016986>.
- L. A. Carlisle, T. Turk, K. Kusejko, K. J. Metzner, C. Leemann, C. D. Schenkel, N. Bachmann, S. Posada, N. Beerenwinkel, J. Böni, S. Yerly, T. Klimkait, M. Perreau, D. L. Braun, A. Rauch, A. Calmy, M. Cavassini, M. Battegay, P. Vernazza, E. Bernasconi, H. F. Günthard, R. D. Kouyos, Swiss HIV Cohort Study, A. Anagnostopoulos, M. Battegay, E. Bernasconi, J. Böni, D. L. Braun, H. C. Bucher, A. Calmy, M. Cavassini, A. Ciuffi, G. Dollenmaier, M. Egger, L. Elzi, J. Fehr, J. Fellay, H. Furrer, C. A. Fux, H. F. Günthard, D. Haerry, B. Hasse, H. H. Hirsch, M. Hoffmann, I. Hösli, M. Huber, C. Kahlert, L. Kaiser, O. Keiser, T. Klimkait, R. D. Kouyos, H. Kovari, B. Ledergerber, G. Martinetti, B. Martinez De Tejada, C. Marzolini, K. J. Metzner, N. Müller, D. Nicca, P. Paioni, G. Pantaleo, M. Perreau, A. Rauch, C. Rudin, A. U. Scherrer, P. Schmid, R. Speck, M. Stöckle, P. Tarr, A. Trkola, P. Vernazza, G. Wandeler, R. Weber, and S. Yerly. Viral Diversity Based on Next-Generation Sequencing of HIV-1 Provides Precise Estimates of Infection Recency and Time Since Infection. *The Journal of Infectious Diseases*, 220(2):254–265, June 2019. ISSN 0022-1899, 1537-6613. doi: 10.1093/infdis/jiz094. URL <https://academic.oup.com/jid/article/220/2/254/5369731>.
- B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76:1–32, Jan. 2017. ISSN 1548-7660. doi: 10.18637/jss.v076.i01. URL <https://doi.org/10.18637/jss.v076.i01>.
- A. Carrasco-Montalvo, I. Armendáriz-Castillo, C. Tello, D. Morales, R. Armas-Gonzalez, D. Guizado-Herrera, A. León-Sosa, D. Ramos-Sarmiento, B. Fuertes, and L. Patino. First detection of SARS-CoV-2 variant B.1.1.529 (Omicron) in Ecuador. *New Microbes and New Infections*, 45:100951, Jan. 2022. ISSN 2052-2975. doi: 10.1016/j.nmni.2022.100951. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8739794/>.
- J. W. Carswell and G. Lloyd. Rise in prevalence of HIV antibodies recorded at an antenatal booking clinic in Kampala, Uganda. *AIDS (London, England)*, 1(3):192–193, Sept. 1987. ISSN 0269-9370.
- M. C. Castro, A. Massuda, G. Almeida, N. A. Menezes-Filho, M. V. Andrade, K. V. M. de Souza Noronha, R. Rocha, J. Macinko, T. Hone, R. Tasca, et al. Brazil’s unified health system: The first 30 years and prospects for the future. *The Lancet*, 394:345–356, 2019. doi: 10.1038/s41591-021-01437-z.

- M. C. Castro, S. Gurzenda, C. M. Turra, S. Kim, T. Andrasfay, and N. Goldman. Reduction in life expectancy in Brazil after COVID-19. *Nature Medicine*, 27:1629–1635, 2021a. doi: 10.1038/s41591-021-01437-z.
- M. C. Castro, S. Kim, L. Barberia, A. F. Ribeiro, S. Gurzenda, K. B. Ribeiro, E. Abbott, J. Blossom, B. Rache, and B. H. Singer. Spatiotemporal pattern of COVID-19 spread in Brazil. *Science (New York, N.Y.)*, 372:821–826, 2021b. doi: 10.1126/science.abh1558.
- S. Cauchemez and N. M. Ferguson. Methods to infer transmission risk factors in complex outbreak data. *Journal of the Royal Society Interface*, 9(68):456–469, Mar. 2012. ISSN 1742-5689. doi: 10.1098/rsif.2011.0379. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3262428/>.
- S. Cauchemez, N. M. Ferguson, C. Wachtel, A. Tegnell, G. Saour, B. Duncan, and A. Nicoll. Closure of schools during an influenza pandemic. *The Lancet Infectious Diseases*, 9(8):473–481, Aug. 2009. ISSN 14733099. doi: 10.1016/S1473-3099(09)70176-8. URL <https://linkinghub.elsevier.com/retrieve/pii/S1473309909701768>.
- S. Cauchemez, A. Bhattarai, T. L. Marchbanks, R. P. Fagan, S. Ostroff, N. M. Ferguson, D. Swerdlow, the Pennsylvania H1N1 working group, S. V. Sodha, M. E. Moll, F. J. Angulo, R. Palekar, W. R. Archer, and L. Finelli. Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proceedings of the National Academy of Sciences*, 108(7):2825–2830, Feb. 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1008895108. URL <https://pnas.org/doi/full/10.1073/pnas.1008895108>.
- Centers for Disease Control (CDC). Pneumocystis carinii pneumonia among persons with hemophilia A. *MMWR. Morbidity and mortality weekly report*, 31(27):365–367, July 1982. ISSN 0149-2195.
- L. W. Chang, M. K. Grabowski, R. Ssekubugu, F. Nalugoda, G. Kigozi, B. Nantume, J. Lessler, S. M. Moore, T. C. Quinn, S. J. Reynolds, R. H. Gray, D. Serwadda, and M. J. Wawer. Heterogeneity of the HIV epidemic in agrarian, trading, and fishing communities in Rakai, Uganda: An observational epidemiological study. *The Lancet HIV*, 3(8):e388–e396, Aug. 2016. ISSN 23523018. doi: 10.1016/S2352-3018(16)30034-0. URL <https://linkinghub.elsevier.com/retrieve/pii/S2352301816300340>.
- L. W. Chang, I. Mbabali, X. Kong, H. Hutton, K. R. Amico, C. E. Kennedy, F. Nalugoda, D. Serwadda, R. C. Bollinger, T. C. Quinn, S. J. Reynolds, R. Gray, M. Wawer, and G. Nakigozi. Impact of a community health worker HIV treatment and prevention intervention in an HIV hotspot fishing community in Rakai, Uganda (mLAKE): Study

- protocol for a randomized controlled trial. *Trials*, 18(1):494, Oct. 2017. ISSN 1745-6215. doi: 10.1186/s13063-017-2243-6. URL <https://doi.org/10.1186/s13063-017-2243-6>.
- N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei, J. Xia, T. Yu, X. Zhang, and L. Zhang. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *The Lancet*, 395(10223):507–513, Feb. 2020. ISSN 0140-6736. doi: 10.1016/S0140-6736(20)30211-7. URL <https://www.sciencedirect.com/science/article/pii/S0140673620302117>.
- G. Chowell, C. Viboud, L. Simonsen, and S. M. Moghadas. Characterizing the reproduction number of epidemics with early subexponential growth dynamics. *Journal of The Royal Society Interface*, 13(123):20160659, Oct. 2016. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2016.0659. URL <https://royalsocietypublishing.org/doi/10.1098/rsif.2016.0659>.
- T. Clare, K. A. Twohig, A.-M. O’Connell, and G. Dabrera. Timeliness and completeness of laboratory-based surveillance of COVID-19 cases in England. *Public Health*, 194: 163–166, May 2021. ISSN 1476-5616. doi: 10.1016/j.puhe.2021.03.012.
- A. Clark, M. Jit, C. Warren-Gash, B. Guthrie, H. H. Wang, S. W. Mercer, C. Sanderson, M. McKee, C. Troeger, K. L. Ong, et al. Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: A modelling study. *The Lancet Global Health*, 8:1003–1017, 2020.
- S. A. C. Clemens, P. M. Folegatti, K. R. Emary, L. Y. Weckx, J. Ratcliff, S. Bibi, A. V. De Almeida Mendes, E. P. Milan, A. Pittella, A. V. Schwarzbald, et al. Efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine against SARS-CoV-2 lineages circulating in brazil. *Nature communications*, 12(1):1–10, 2021.
- M. S. Cohen, C. L. Gay, M. P. Busch, and F. M. Hecht. The Detection of Acute HIV Infection. *The Journal of Infectious Diseases*, 202(S2):S270–S277, Oct. 2010. ISSN 0022-1899, 1537-6613. doi: 10.1086/655651. URL <https://academic.oup.com/jid/article-lookup/doi/10.1086/655651>.
- M. S. Cohen, Y. Q. Chen, M. McCauley, T. Gamble, M. C. Hosseinipour, N. Kumarasamy, J. G. Hakim, J. Kumwenda, B. Grinsztejn, J. H. Pilotto, S. V. Godbole, S. Mehendale, S. Chariyalertsak, B. R. Santos, K. H. Mayer, I. F. Hoffman, S. H. Eshleman, E. Piwowar-Manning, L. Wang, J. Makhema, L. A. Mills, G. De Bruyn, I. Sanne, J. Eron, J. Gallant, D. Havlir, S. Swindells, H. Ribaudo, V. Elharrar, D. Burns, T. E. Taha, K. Nielsen-Saines, D. Celentano, M. Essex, and T. R. Fleming. Prevention of HIV-1 Infection with Early Antiretroviral Therapy. *New England Journal of Medicine*, 365(6):493–505, Aug. 2011.

- ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa1105243. URL <http://www.nejm.org/doi/abs/10.1056/NEJMoa1105243>.
- C. Colijn and J. Gardy. Phylogenetic tree shapes resolve disease transmission patterns. *Evolution, Medicine, and Public Health*, 2014(1):96–108, Jan. 2014. ISSN 2050-6201. doi: 10.1093/emph/eou018. URL <https://academic.oup.com/emph/article/2014/1/96/1845716>.
- C. E. M. Coltart, A. Hoppe, M. Parker, L. Dawson, J. J. Amon, M. Simwanga, G. Geller, G. Henderson, O. Laeyendecker, J. D. Tucker, P. Eba, V. Novitsky, A.-M. Vandamme, J. Seeley, G. Dallabetta, G. Harling, M. K. Grabowski, P. Godfrey-Faussett, C. Fraser, M. S. Cohen, D. Pillay, J. J. Amon, R. Baggaley, E. J. Bernard, D. Burns, M. S. Cohen, C. C. Coltart, G. Dallabetta, L. Dawson, N. Dedes, V. Delpech, P. M. Eba, C. Fraser, G. Geller, D. German, P. Godfrey-Faussett, M. K. Grabowski, I. Hall, G. Harling, G. Henderson, A. Hoppe, Z. Kozlakidis, O. Laeyendecker, F. Mwanza, V. Novitsky, M. Parker, D. Pillay, A. Reis, J. Seeley, M. Simwanga, J. D. Tucker, A.-M. Vandamme, J. O. Wertheim, and R. Zimmerman. Ethical considerations in global HIV phylogenetic research. *The Lancet HIV*, 5(11):e656–e666, Nov. 2018. ISSN 23523018. doi: 10.1016/S2352-3018(18)30134-6. URL <https://linkinghub.elsevier.com/retrieve/pii/S2352301818301346>.
- A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of Epidemiology*, 178(9):1505–1512, Nov. 2013. ISSN 0002-9262. doi: 10.1093/aje/kwt133. URL <https://doi.org/10.1093/aje/kwt133>.
- D. Cucinotta and M. Vanelli. WHO Declares COVID-19 a Pandemic. *Acta Bio Medica : Atenei Parmensis*, 91(1):157–160, 2020. ISSN 0392-4203. doi: 10.23750/abm.v91i1.9397. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7569573/>.
- S. Dattani and M. Roser. What were the death tolls from pandemics in history? *Our World in Data*, Dec. 2023. URL <https://ourworldindata.org/historical-pandemics>.
- N. G. Davies, A. J. Kucharski, R. M. Eggo, A. Gimma, W. J. Edmunds, T. Jombart, K. O'Reilly, A. Endo, J. Hellewell, E. S. Nightingale, B. J. Quilty, C. I. Jarvis, T. W. Russell, P. Klepac, N. I. Bosse, S. Funk, S. Abbott, G. F. Medley, H. Gibbs, C. A. B. Pearson, S. Flasche, M. Jit, S. Clifford, K. Prem, C. Diamond, J. Emery, A. K. Deol, S. R. Procter, K. van Zandvoort, Y. F. Sun, J. D. Munday, A. Rosello, M. Auzenberg, G. Knight, R. M. G. J. Houben, and Y. Liu. Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: A modelling study. *The Lancet Public Health*, 5(7):e375–e385, July 2020. ISSN 2468-2667. doi: 10.1016/S2468-2667(20)30133-X. URL [https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667\(20\)30133-X/fulltext](https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(20)30133-X/fulltext).

- N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. B. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne, T. Wenseleers, A. Gimma, W. Waites, K. L. M. Wong, K. van Zandvoort, J. D. Silverman, CMMID COVID-19 Working Group, COVID-19 Genomics UK (COG-UK) Consortium, K. Diaz-Ordaz, R. Keogh, R. M. Eggo, S. Funk, M. Jit, K. E. Atkins, and W. J. Edmunds. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science (New York, N.Y.)*, 372:3055, Apr. 2021.
- M. V. de Albuquerque and L. H. L. Ribeiro. Inequality, geographic situation, and meanings of action in the COVID-19 pandemic in Brazil. *Cadernos de Saúde Pública*, 36, 2021. doi: 10.1590/0102-311X00208720.
- D. De Angelis, A. M. Presanis, P. J. Birrell, G. S. Tomba, and T. House. Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics*, 10: 83–87, Mar. 2015. ISSN 1755-4365. doi: 10.1016/j.epidem.2014.09.004. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383805/>.
- M. H. S. De Oliveira, G. Lippi, and B. M. Henry. Sudden rise in COVID-19 case fatality among young and middle-aged adults in the south of Brazil after identification of the novel B.1.1.28.1 (P.1) SARS-CoV-2 strain: Analysis of data from the state of Parana, Mar. 2021. URL <http://medrxiv.org/lookup/doi/10.1101/2021.03.24.21254046>.
- R. de Oliveira Andrade. COVID-19 is causing the collapse of Brazil’s national health service. *BMJ (Clinical research ed.)*, 370, 2020. doi: 10.1136/bmj.m3032. URL <https://www.bmj.com/content/370/bmj.m3032>.
- W. M. de Souza, L. F. Buss, D. da Silva Candido, J.-P. Carrera, S. Li, A. E. Zarebski, R. H. M. Pereira, C. A. Prete, A. A. de Souza-Santos, K. V. Parag, et al. Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. *Nature Human Behaviour*, 4:856–865, 2020. doi: 10.1038/s41562-020-0928-4.
- A. A. de Souza Santos, D. da Silva Candido, W. M. de Souza, L. Buss, S. L. Li, R. H. Pereira, C.-H. Wu, E. C. Sabino, and N. R. Faria. Dataset on SARS-CoV-2 non-pharmaceutical interventions in Brazilian municipalities. *Scientific Data*, 8:73, 2021. doi: 10.1038/s41597-021-00859-1.
- S. N. DeWitte. Mortality Risk and Survival in the Aftermath of the Medieval Black Death. *PLoS ONE*, 9(5):e96513, May 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0096513. URL <https://dx.plos.org/10.1371/journal.pone.0096513>.
- M. S. Dhar, R. Marwal, R. VS, K. Ponnusamy, B. Jolly, R. C. Bhoyar, V. Sardana, S. Naushin, M. Rophina, T. A. Mellan, S. Mishra, C. Whittaker, S. Fatihhi, M. Datta, P. Singh,

- U. Sharma, R. Ujjainiya, N. Bhatheja, M. K. Divakar, M. K. Singh, M. Imran, V. Senthivel, R. Maurya, N. Jha, P. Mehta, V. A, P. Sharma, A. VR, U. Chaudhary, N. Soni, L. Thukral, S. Flaxman, S. Bhatt, R. Pandey, D. Dash, M. Faruq, H. Lall, H. Gogia, P. Madan, S. Kulkarni, H. Chauhan, S. Sengupta, S. Kabra, The Indian SARS-CoV-2 Genomics Consortium (INSACOG), R. K. Gupta, S. K. Singh, A. Agrawal, and P. Rakshit. Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. *Science*, 374(6570):995–999, Nov. 2021. doi: 10.1126/science.abj9932. URL <https://www.science.org/doi/10.1126/science.abj9932>.
- J. Diamond. *Guns, Germs & Steel – The Fates of Human Societies*. W. W. Norton & Company, New York, first edition in this form. edition, Apr. 1999. ISBN 978-0-393-31755-8.
- D. Dimitrov, J. R. Moore, D. J. Donnell, and Marie-Claude Boily. Achieving 95-95 may not be enough to end the AIDS epidemic in South Africa, Mar. 2020. URL croiconference.org/wp-content/uploads/sites/2/posters/2020/1430_6_Dimitrov_01095.pdf.
- E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, May 2020. ISSN 1473-3099, 1474-4457. doi: 10.1016/S1473-3099(20)30120-1. URL [https://www.thelancet.com/journals/laninf/article/PIIS1473-3099\(20\)30120-1/fulltext](https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30120-1/fulltext).
- I. Dorigatti, L. Okell, A. Cori, N. Imai, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg, R. Fitzjohn, H. Fu, K. Gaythorpe, A. Hamlet, W. Hinsley, N. Hong, M. Kwun, D. Laydon, G. Nedjati Gilani, S. Riley, S. Van Elsland, E. Volz, H. Wang, C. Walters, X. Xi, C. Donnelly, A. Ghani, and N. Ferguson. Report 4: Severity of 2019-novel coronavirus (nCoV). Technical report, [object Object], Feb. 2020. URL <http://spiral.imperial.ac.uk/handle/10044/1/77154>.
- D. B. Dunson. Commentary: Practical Advantages of Bayesian Analysis of Epidemiologic Data. *American Journal of Epidemiology*, 153(12):1222–1226, June 2001. ISSN 0002-9262. doi: 10.1093/aje/153.12.1222. URL <https://doi.org/10.1093/aje/153.12.1222>.
- D. Duong. Alpha, Beta, Delta, Gamma: What’s important to know about SARS-CoV-2 variants of concern? *CMAJ: Canadian Medical Association Journal*, 193(27):E1059–E1060, July 2021. ISSN 0820-3946. doi: 10.1503/cmaj.1095949. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8342008/>.
- L. Dwyer-Lindgren, M. A. Cork, A. Sligar, K. M. Steuben, K. F. Wilson, N. R. Provost, B. K. Mayala, J. D. VanderHeide, M. L. Collison, J. B. Hall, M. H. Biehl, A. Carter, T. Frank, D. Douwes-Schultz, R. Burstein, D. C. Casey, A. Deshpande, L. Earl, C. El Bcheraoui,

- T. H. Farag, N. J. Henry, D. Kinyoki, L. B. Marczak, M. R. Nixon, A. Osgood-Zimmerman, D. Pigott, R. C. Reiner, J. M. Ross, L. E. Schaeffer, D. L. Smith, N. Davis Weaver, K. E. Wiens, J. W. Eaton, J. E. Justman, A. Opio, B. Sartorius, F. Tanser, N. Wabiri, P. Piot, C. J. L. Murray, and S. I. Hay. Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. *Nature*, 570(7760):189–193, June 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1200-9. URL <https://www.nature.com/articles/s41586-019-1200-9>.
- S. H. Eshleman, S. E. Hudelson, A. D. Redd, L. Wang, R. Debes, Y. Q. Chen, C. A. Martens, S. M. Ricklefs, E. J. Selig, S. F. Porcella, S. Munshaw, S. C. Ray, E. Piwowar-Manning, M. McCauley, M. C. Hosseinipour, J. Kumwenda, J. G. Hakim, S. Chariyalertsak, G. de Bruyn, B. Grinsztejn, N. Kumarasamy, J. Makhema, K. H. Mayer, J. Pilotto, B. R. Santos, T. C. Quinn, M. S. Cohen, and J. P. Hughes. Analysis of Genetic Linkage of HIV From Couples Enrolled in the HIV Prevention Trials Network 052 Trial. *The Journal of Infectious Diseases*, 204(12):1918–1926, Dec. 2011. ISSN 0022-1899. doi: 10.1093/infdis/jir651. URL <https://doi.org/10.1093/infdis/jir651>.
- Eswatini Ministry of Health and Central Statistical Office. Eswatini Summary Sheet 2021, 2022. URL https://phia.icap.columbia.edu/wp-content/uploads/2022/12/53059_14_SHIMS3_Summary-sheet-Web.pdf.
- M. Farahani, E. Radin, S. Saito, K. K. Sachathap, W. Hladik, A. C. Voetsch, A. F. Auld, S. Balachandra, B. A. Tippet Barr, A. Low, T. F. Smart, G. Musuka, S. Jonnalagadda, A. J. Hakim, N. W. Wadonda-Kabondo, A. Jahn, O. Mugurungi, D. B. Williams, D. T. Barradas, D. Payne, B. Parekh, H. Patel, L. Wiesner, D. Hoos, and J. E. Justman. Population Viral Load, Viremia, and Recent HIV-1 Infections: Findings From Population-Based HIV Impact Assessments (PHIAs) in Zimbabwe, Malawi, and Zambia. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 87(1):S81–S88, Aug. 2021. ISSN 1525-4135. doi: 10.1097/QAI.0000000000002637. URL <https://journals.lww.com/10.1097/QAI.0000000000002637>.
- N. R. Faria, A. Rambaut, M. A. Suchard, G. Baele, T. Bedford, M. J. Ward, A. J. Tatem, J. D. Sousa, N. Arinaminpathy, J. Pépin, D. Posada, M. Peeters, O. G. Pybus, and P. Lemey. The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205):56–61, Oct. 2014. doi: 10.1126/science.1256739. URL <https://www.science.org/doi/10.1126/science.1256739>.
- N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. d. S. Candido, S. Mishra, M. A. E. Crispim, F. C. S. Sales, I. Hawryluk, J. T. McCrone, R. J. G. Hulswit, L. A. M. Franco, M. S. Ramundo, J. G. de Jesus, P. S. Andrade, T. M. Coletti, G. M. Ferreira, C. A. M. Silva, E. R. Manuli, R. H. M. Pereira, P. S. Peixoto, M. U. G. Kraemer, N. Gaburo, C. d. C. Camilo, H. Hoeltgebaum, W. M. Souza, E. C. Rocha, L. M. de Souza, M. C. de Pinho, L. J. T.

- Araujo, F. S. V. Malta, A. B. de Lima, J. d. P. Silva, D. A. G. Zauli, A. C. d. S. Ferreira, R. P. Schnekenberg, D. J. Laydon, P. G. T. Walker, H. M. Schlüter, A. L. P. dos Santos, M. S. Vidal, V. S. Del Caro, R. M. F. Filho, H. M. dos Santos, R. S. Aguiar, J. L. Proença-Modena, B. Nelson, J. A. Hay, M. Monod, X. Miscouridou, H. Coupland, R. Sonabend, M. Vollmer, A. Gandy, C. A. Prete, V. H. Nascimento, M. A. Suchard, T. A. Bowden, S. L. K. Pond, C.-H. Wu, O. Ratmann, N. M. Ferguson, C. Dye, N. J. Loman, P. Lemey, A. Rambaut, N. A. Fraiji, M. d. P. S. S. Carvalho, O. G. Pybus, S. Flaxman, S. Bhatt, and E. C. Sabino. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in manaus, brazil. *Science (New York, N.Y.)*, 372:815–821, 2021. ISSN 0036-8075. doi: 10.1126/science.abh2644.
- M. Fellows, V. Paye, A. Alencar, M. Nicácio, I. Castro, M. E. Coelho, C. V. J. Silva, M. Bandeira, R. Lourival, and P. C. Basta. Under-reporting of COVID-19 cases among indigenous peoples in Brazil: A new expression of old inequalities. *Frontiers in Psychiatry*, 12:352, 2021.
- N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, ZULMA. Cucunuba Perez, G. Cuomo-Dannenburg, A. Dighe, I. Dorigatti, H. Fu, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, L. Okell, S. Van Elsland, H. Thompson, R. Verity, E. Volz, H. Wang, Y. Wang, P. Walker, P. Winskill, C. Whittaker, C. Donnelly, S. Riley, and A. Ghani. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. Technical report, Imperial College London, Mar. 2020. URL <http://spiral.imperial.ac.uk/handle/10044/1/77482>.
- R. FitzJohn, M. Baguelin, E. Knock, L. Whittles, J. Lees, and R. Sonabend. *Mcstate: Monte Carlo Methods for State Space Models*, 2024. URL <https://github.com/mrc-ide/mcstate>.
- S. Flaxman, A. G. Wilson, D. B. Neill, H. Nickisch, and A. J. Smola. Fast Kronecker Inference in Gaussian Processes with non-Gaussian Likelihoods. 2015.
- S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584:257–261, 2020. doi: 10.1038/s41586-020-2405-7.
- A. Fontanet, B. Autran, B. Lina, M. P. Kieny, S. S. A. Karim, and D. Sridhar. SARS-CoV-2 variants and ending the COVID-19 pandemic. *The Lancet*, 397(10278):952–954, Mar. 2021. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(21)00370-6. URL [https://www.thelancet.com/journals/lancet/article/piiS0140-6736\(21\)00370-6/fulltext](https://www.thelancet.com/journals/lancet/article/piiS0140-6736(21)00370-6/fulltext).

- L. Forsberg White and M. Pagano. A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Statistics in Medicine*, 27(16):2999–3016, July 2008. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.3136. URL <https://onlinelibrary.wiley.com/doi/10.1002/sim.3136>.
- T. D. Frank, A. Carter, D. Jahagirdar, M. H. Biehl, D. Douwes-Schultz, S. L. Larson, M. Arora, L. Dwyer-Lindgren, K. M. Steuben, H. Abbastabar, L. J. Abu-Raddad, D. M. Abyu, M. Adabi, O. M. Adebayo, V. Adekanmbi, O. O. Adetokunboh, A. Ahmadi, K. Ahmadi, E. Ahmadian, E. Ahmadpour, M. B. Ahmed, C. G. Akal, F. Alahdab, N. Alam, S. B. Albertson, B. T. T. Alemnew, K. A. Alene, V. Alipour, N. Alvis-Guzman, S. Amini, Z. Anbari, N. H. Anber, M. Anjomshoa, C. A. T. Antonio, J. Arabloo, O. Aremu, H. A. Areri, E. T. Asfaw, A. F. Ashagre, D. Asmelash, A. A. Asrat, E. F. G. A. Avokpaho, A. Awasthi, N. Awoke, M. A. Ayanore, S. Azari, A. Badawi, M. Bagherzadeh, M. Banach, A. Barac, T. W. Bärnighausen, S. Basu, N. Bedi, M. Behzadifar, B. B. Bekele, S. A. Belay, Y. B. Belay, Y. M. Belayneh, A. Berhane, A. G. Bhat, K. Bhattacharyya, B. Biadgo, A. Bijani, M. S. B. Sayeed, H. Bitew, A. Blinov, K. A. Bogale, H. A. Bojia, S. B. N. B. Nagaraja, Z. A. Butt, L. Cahuana-Hurtado, J. C. C. Rincon, F. Carvalho, V. K. Chattu, D. J. Christopher, D.-T. Chu, R. Crider, T. Dahiru, L. Dandona, R. Dandona, A. Daryani, J. das Neves, J.-W. D. Neve, L. Degenhardt, F. M. Demeke, A. B. Demis, D. B. Demissie, G. T. Demoz, K. Deribe, D. D. Jarlais, G. P. Dhungana, D. Diaz, S. Djalalinia, H. P. Do, L. P. Doan, H. Duber, M. Dubey, E. Dubljanin, E. E. Duken, B. D. Adema, A. Effiong, A. Eftekhari, M. E. S. Zaki, S. I. El-Jaafary, Z. El-Khatib, A. Elsharkawy, A. Y. Endries, S. Eskandarieh, O. Eyawo, F. Farzadfar, B. Fatima, N. Fentahun, E. Fernandes, I. Filip, F. Fischer, M. O. Folayan, M. Foroutan, T. Fukumoto, N. Fullman, A. L. Garcia-Basteiro, R. T. Gayesa, K. B. Gebremedhin, G. G. G. Gebremeskel, K. K. Gebreyohannes, G. A. Gedefaw, B. K. Gelaw, H. A. Gesesew, B. Geta, K. E. Gezae, K. Ghadiri, A. Ghashghae, T. T. G. Ginindza, H. C. Gugnani, R. A. Guimarães, M. T. Haile, G. B. Hailu, A. Haj-Mirzaian, A. Haj-Mirzaian, S. Hamidi, S. Handanagic, D. W. Handiso, L. K. Hanfore, A. Hasanzadeh, H. Hassankhani, H. Y. Hassen, S. I. Hay, A. Henok, C. L. Hoang, H. D. Hosgood, M. Hosseinzadeh, M. Hsairi, S. E. Ibitoye, B. Idrisov, K. S. Ikuta, O. S. Ilesanmi, S. S. N. Irvani, C. J. Iwu, K. H. Jacobsen, S. L. James, E. Jenabi, R. P. Jha, J. B. Jonas, Z. J. Shushtari, A. Kabir, Z. Kabir, R. Kadel, A. Kasaeian, B. Kassa, G. M. Kassa, T. D. Kassa, G. A. Kayode, M. M. Kebede, A. T. Kefale, A. P. Kengne, Y. S. Khader, M. A. Khafaie, N. Khalid, E. A. Khan, G. Khan, J. Khan, Y.-H. Khang, K. Khatab, S. Khazaei, A. T. Khoja, A. A. Kiadaliri, Y. J. Kim, A. Kisa, S. Kisa, S. Kochhar, H. Komaki, P. A. Koul, A. Koyanagi, B. K. Defo, G. A. Kumar, M. Kumar, D. Kuupiel, D. K. Lal, J. J.-H. Lee, T. L. Lenjebo, C. T. Leshargie, E. R. K. Macarayan, E. R. Maddison, H. M. A. E. Razek, C. Magis-Rodriguez, P. W. Mahasha, M. Majdan, A. Majeed, R. Malekzadeh, N. Manafi, C. C. Mapoma, F. R. Martins-Melo, A. Masaka, E. N. L. Mayenga, V. Mehta,

G. G. Meles, H. G. Meles, A. Melese, M. Melku, P. T. N. Memiah, Z. A. Memish, A. T. Mena, W. Mendoza, D. T. Mengistu, G. Mengistu, T. J. Meretoja, T. Mestrovic, T. R. Miller, B. Moazen, B. Mohajer, A. Mohamadi-Bolbanabad, K. A. Mohammad, Y. Mohammad, A. M. Darwesh, N. M. G. Mezerji, M. Mohammadi, R. Mohammadibakhsh, M. Mohammadoo-Khorasani, J. A. Mohammed, S. Mohammed, F. Mohebi, A. H. Mokdad, Y. Moodley, M. Moossavi, G. Moradi, M. Moradi-Lakeh, M. M. Moschos, T. B. Mossie, S. M. Mousavi, K. F. Muchie, A. G. Muluneh, M. K. Muriithi, G. Mustafa, S. Muthupandian, A. J. Nagarajan, G. Naik, F. Najafi, J. Nazari, D. E. Ndwandwe, C. T. Nguyen, H. L. T. Nguyen, S. H. Nguyen, T. H. Nguyen, D. N. A. Ningrum, M. R. Nixon, C. A. Nnaji, M. Noroozi, J. J. Noubiap, M. N. Shiadeh, M. S. Obsa, E. A. Odame, R. Ofori-Asenso, F. A. Ogbo, A. Okoro, O. Oladimeji, A. T. Olagunju, T. O. Olagunju, S. Olum, K. O. A. O. Asante, E. Oren, S. S. Otstavnov, M. Pa, J. R. Padubidri, S. Pakhale, A. H. Pakpour, S. K. Patel, K. Paulos, V. C. F. Pepito, E. K. Peprah, B. Piroozi, A. Pourshams, M. Qorbani, M. Rabiee, N. Rabiee, A. Radfar, A. Rafay, A. Rafiei, F. Rahim, A. Rahimi-Movaghar, V. Rahimi-Movaghar, S. ur Rahman, C. L. Ranabhat, S. Rawaf, C. Reis, V. Renjith, M. A. Reta, M. S. Rezai, C. M. R. González, E. M. Roro, A. Rostami, S. Rubino, S. S. Moghaddam, S. Safari, R. Sagar, M. A. Sahraian, M. R. R. Salem, Y. Salimi, J. A. Salomon, E. Z. Sambala, A. M. Samy, B. Sartorius, M. Satpathy, M. Sawhney, M. Sayyah, A. E. Schutte, S. G. Sepanlou, S. Seyedmousavi, H. Shabaninejad, A. A. Shaheen, M. A. Shaikh, S. A. Shallo, M. Shamsizadeh, H. Sharifi, K. Shibuya, J. I. Shin, R. Shirkoohi, D. A. S. Silva, D. G. A. Silveira, J. A. Singh, M. M. M. Sisay, M. Sisay, S. Sisay, A. E. Smith, A. Sokhan, R. Somayaji, S. Soshnikov, D. J. Stein, M. B. Sufiyan, B. F. Sunguya, B. L. Sykes, B. T. Tadesse, D. B. Tadesse, K. S. Tamirat, N. Taveira, S. W. Tekelemedhin, H. D. Temesgen, F. H. Tesfay, M. Y. Teshale, S. Thapa, K. G. Tlaye, S. M. Topp, M. R. Tovani-Palone, B. X. Tran, K. B. Tran, I. Ullah, B. Unnikrishnan, O. A. Uthman, Y. Veisani, S. K. Vladimirov, F. W. Wada, Y. Waheed, K. G. Weldegwergs, G. T. T. Welde-samuel, R. Westerman, T. Wijeratne, H. F. Wolde, D. Z. Wondafrash, T. E. Wonde, B. Y. Wondmagegn, A. G. Yeshanew, M. T. Yilma, E. M. Yimer, N. Yonemoto, M. Yotebieng, Y. Youm, C. Yu, Z. Zaidi, A. Zarghi, Z. M. Zenebe, T. A. Zewale, A. Ziapour, S. Zodpey, M. Naghavi, S. E. Vollset, H. Wang, S. S. Lim, H. H. Kyu, and C. J. L. Murray. Global, regional, and national incidence, prevalence, and mortality of HIV, 1980–2017, and forecasts to 2030, for 195 countries and territories: A systematic analysis for the Global Burden of Diseases, Injuries, and Risk Factors Study 2017. *The Lancet HIV*, 6(12):e831–e859, Dec. 2019. ISSN 2352-3018. doi: 10.1016/S2352-3018(19)30196-1. URL [https://www.thelancet.com/journals/lanhiv/article/PIIS2352-3018\(19\)30196-1/fulltext](https://www.thelancet.com/journals/lanhiv/article/PIIS2352-3018(19)30196-1/fulltext).

C. Fraser. Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. *PLOS ONE*, 2(8):e758, Aug. 2007. ISSN 1932-6203. doi: 10.1371/journal.pone.0000758. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000758>.

pone.0000758.

- A. R. R. Freitas, O. A. Beckedorff, L. P. de Góes Cavalcanti, A. M. Siqueira, D. B. de Castro, C. F. da Costa, D. R. Q. Lemos, and E. N. Barros. The emergence of novel SARS-CoV-2 variant P. 1 in Amazonas (Brazil) was temporally associated with a change in the age and sex profile of COVID-19 mortality: A population based ecological study. *The Lancet Regional Health-Americas*, 1:100021, 2021.
- L. Frescura, P. Godfrey-Faussett, A. Feizzadeh A., W. El-Sadr, O. Syarif, P. D. Ghys, and on and behalf of the 2025 testing treatment target Working Group. Achieving the 95 95 95 targets for all: A pathway to ending AIDS. *PLOS ONE*, 17(8):e0272405, Aug. 2022. ISSN 1932-6203. doi: 10.1371/journal.pone.0272405. URL <https://dx.plos.org/10.1371/journal.pone.0272405>.
- S. D. W. Frost and E. M. Volz. Modelling tree shape and structure in viral phylodynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614):20120208, Mar. 2013. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2012.0208. URL <https://royalsocietypublishing.org/doi/10.1098/rstb.2012.0208>.
- T. Fujino, H. Nomoto, S. Kutsuna, M. Ujiie, T. Suzuki, R. Sato, T. Fujimoto, M. Kuroda, T. Wakita, and N. Ohmagari. Novel SARS-CoV-2 variant in travelers from Brazil to Japan. *Emerging Infectious Diseases*, 27, Apr. 2021.
- F. R. Funcia. Underfunding and federal budget of SUS: Preliminary references for additional resource allocation. *Ciencia & Saude Coletiva*, 24:4405–4415, 2019. doi: 10.1590/1413-812320182412.25892019.
- T. Funk, A. Pharris, G. Spiteri, N. Bundle, A. Melidou, M. Carr, G. Gonzalez, A. Garcia-Leon, F. Crispie, L. O'Connor, N. Murphy, J. Mossong, A. Vergison, A. K. Wienecke-Baldacchino, T. Abdelrahman, F. Riccardo, P. Stefanelli, A. D. Martino, A. Bella, A. L. Presti, P. Casaca, J. Moreno, V. Borges, J. Isidro, R. Ferreira, J. P. Gomes, L. Dotsenko, H. Suija, J. Epstein, O. Sadikova, H. Sepp, N. Ikonen, C. Savolainen-Kopra, S. Blomqvist, T. Möttönen, O. Helve, J. Gomes-Dias, C. Adlhoch, and o. b. o. C. study Groups. Characteristics of SARS-CoV-2 variants of concern B.1.1.7, B.1.351 or P.1: Data from seven EU/EEA countries, weeks 38/2020 to 10/2021. *Eurosurveillance*, 26(16):2100348, Apr. 2021. ISSN 1560-7917. doi: 10.2807/1560-7917.ES.2021.26.16.2100348. URL <https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2021.26.16.2100348>.
- J. Gabry, R. Češnovar, and A. Johnson. *Cmdstanr: R Interface to 'CmdStan'*, 2023.
- R. M. Galiwango, R. Musoke, L. Lubyayi, R. Ssekubugu, S. Kalibbala, V. Ssekweyama, V. Mirembe, G. Nakigozi, S. J. Reynolds, D. Serwadda, R. H. Gray, and G. Kigozi. Evalua-

- tion of current rapid HIV test algorithms in Rakai, Uganda. *Journal of Virological Methods*, 192(1-2):25–27, Sept. 2013. ISSN 01660934. doi: 10.1016/j.jviromet.2013.04.003. URL <https://linkinghub.elsevier.com/retrieve/pii/S0166093413001079>.
- R. C. Gallo, S. Z. Salahuddin, M. Popovic, G. M. Shearer, M. Kaplan, B. F. Haynes, T. J. Palker, R. Redfield, J. Oleske, B. Safai, G. White, P. Foster, and P. D. Markham. Frequent Detection and Isolation of Cytopathic Retroviruses (HTLV-III) from Patients with AIDS and at Risk for AIDS. *Science*, 224(4648):500–503, May 1984. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.6200936. URL <https://www.science.org/doi/10.1126/science.6200936>.
- A. P. Galvani, A. Pandey, M. C. Fitzpatrick, J. Medlock, and G. E. Gray. Defining control of HIV epidemics. *The Lancet HIV*, 5(11):e667–e670, Nov. 2018. ISSN 23523018. doi: 10.1016/S2352-3018(18)30178-4. URL <https://linkinghub.elsevier.com/retrieve/pii/S2352301818301784>.
- T. Ganyani, C. Kremer, D. Chen, A. Torneri, C. Faes, J. Wallinga, and N. Hens. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance*, 25(17):2000257, Apr. 2020. ISSN 1560-7917. doi: 10.2807/1560-7917.ES.2020.25.17.2000257. URL <https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.17.2000257>.
- T. Gaolathe, K. E. Wirth, M. P. Holme, J. Makhema, S. Moyo, U. Chakalisa, E. K. Yankinda, Q. Lei, M. Mmalane, V. Novitsky, L. Okui, E. Van Widenfelt, K. M. Powis, N. Khan, K. Bennett, H. Bussmann, S. Dryden-Peterson, R. Lebelonyane, S. el-Halabi, L. A. Mills, T. Marukutira, R. Wang, E. J. T. Tchetgen, V. DeGruttola, M. Essex, and S. Lockman. Botswana's progress toward achieving the 2020 UNAIDS 90-90-90 antiretroviral therapy and virological suppression goals: A population-based survey. *The Lancet HIV*, 3(5):e221–e230, May 2016. ISSN 23523018. doi: 10.1016/S2352-3018(16)00037-0. URL <https://linkinghub.elsevier.com/retrieve/pii/S2352301816000370>.
- A. Gelman and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, Nov. 1992. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1177011136. URL <https://projecteuclid.org/journals/statistical-science/volume-7/issue-4/Inference-from-Iterative-Simulation-Using-Multiple-Sequences/10.1214/ss/1177011136.full>.
- A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák. Bayesian Workflow, Nov. 2020. URL <http://arxiv.org/abs/2011.01808>.

- P. D. Ghys, B. G. Williams, M. Over, T. B. Hallett, and P. Godfrey-Faussett. Epidemiological metrics and benchmarks for a transition in the HIV epidemic. *PLOS Medicine*, 15(10):e1002678, Oct. 2018. ISSN 1549-1676. doi: 10.1371/journal.pmed.1002678. URL <https://dx.plos.org/10.1371/journal.pmed.1002678>.
- W. R. Gilks, A. Thomas, and D. J. Spiegelhalter. A Language and Program for Complex Bayesian Modelling. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 43(1):169–177, 1994. ISSN 0039-0526. doi: 10.2307/2348941. URL <https://www.jstor.org/stable/2348941>.
- T. Golubchik, L. Abeler-Dörner, M. Hall, C. Wymant, D. Bonsall, G. Macintyre-Cockett, L. Thomson, J. M. Baeten, C. L. Celum, R. M. Galiwango, B. Kosloff, M. Limbada, A. Mujugira, N. R. Mugo, A. Gall, F. Blanquart, M. Bakker, D. Bezemer, S. H. Ong, J. Albert, N. Bannert, J. Fellay, B. Gunsenheimer-Bartmeyer, H. F. Günthard, P. Kivelä, R. D. Kouyos, L. Meyer, K. Porter, A. Van Sighem, M. Van Der Valk, B. Berkhout, P. Kellam, M. Cornelissen, P. Reiss, H. Ayles, D. N. Burns, S. Fidler, M. K. Grabowski, R. Hayes, J. T. Herbeck, J. Kagaayi, P. Kaleebu, J. R. Lingappa, D. Ssemwanga, S. H. Eshleman, M. S. Cohen, O. Ratmann, O. Laeyendecker, C. Fraser, and the BEEHIVE collaboration and the PANGAEA consortium. HIV-phyloTSI: Subtype-independent estimation of time since HIV-1 infection for cross-sectional measures of population incidence using deep sequence data. Preprint, HIV/AIDS, May 2022. URL <http://medrxiv.org/lookup/doi/10.1101/2022.05.15.22275117>.
- P. González, A. Castaño, N. V. Chawla, and J. J. D. Coz. A Review on Quantification Learning. *ACM Computing Surveys*, 50(5):74:1–74:40, Sept. 2017. ISSN 0360-0300. doi: 10.1145/3117807. URL <https://dl.acm.org/doi/10.1145/3117807>.
- I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks, 2014. URL <https://arxiv.org/abs/1406.2661>.
- K. M. Gostic, L. McGough, E. B. Baskerville, S. Abbott, K. Joshi, C. Tedijanto, R. Kahn, R. Niehus, J. A. Hay, P. M. D. Salazar, J. Hellewell, S. Meakin, J. D. Munday, N. I. Bosse, K. Sherratt, R. N. Thompson, L. F. White, J. S. Huisman, J. Scire, S. Bonhoeffer, T. Stadler, J. Wallinga, S. Funk, M. Lipsitch, and S. Cobey. Practical considerations for measuring the effective reproductive number, Rt. *PLOS Computational Biology*, 16(12):e1008409, Dec. 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008409. URL <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008409>.
- N. Gozzi, M. Chinazzi, N. E. Dean, I. M. Longini Jr, M. E. Halloran, N. Perra, and A. Vespignani. Estimating the impact of COVID-19 vaccine inequities: A modeling study. *Nature Communications*, 14:3272, June 2023. ISSN 2041-1723. doi:

- 10.1038/s41467-023-39098-w. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10241610/>.
- M. K. Grabowski and A. D. Redd. Molecular tools for studying HIV transmission in sexual networks. *Current Opinion in HIV and AIDS*, 9(2):126, Mar. 2014. ISSN 1746-630X. doi: 10.1097/COH.0000000000000040. URL https://journals.lww.com/co-hivandaids/fulltext/2014/03000/molecular_tools_for_studying_hiv_transmission_in.6.aspx.
- M. K. Grabowski, D. M. Serwadda, R. H. Gray, G. Nakigozi, G. Kigozi, J. Kagaayi, R. Ssekubugu, F. Nalugoda, J. Lessler, T. Lutalo, R. M. Galiwango, F. Makumbi, X. Kong, D. Kabatesi, S. T. Alamo, S. Wiersma, N. K. Sewankambo, A. A. Tobian, O. Laeyendecker, T. C. Quinn, S. J. Reynolds, M. J. Wawer, and L. W. Chang. HIV Prevention Efforts and Incidence of HIV in Uganda. *New England Journal of Medicine*, 377(22): 2154–2166, Nov. 2017. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa1702150. URL <http://www.nejm.org/doi/10.1056/NEJMoa1702150>.
- M. K. Grabowski, E. U. Patel, G. Nakigozi, V. Ssempijja, R. Ssekubugu, J. Ssekasanvu, A. Ndyababo, G. Kigozi, F. Nalugoda, R. H. Gray, S. Kalibbala, D. M. Serwadda, O. Laeyendecker, M. J. Wawer, L. W. Chang, T. C. Quinn, J. Kagaayi, A. A. R. Tobian, and S. J. Reynolds. Prevalence and Predictors of Persistent Human Immunodeficiency Virus Viremia and Viral Rebound After Universal Test and Treat: A Population-Based Study. *The Journal of Infectious Diseases*, 223(7):1150–1160, Apr. 2021. ISSN 0022-1899, 1537-6613. doi: 10.1093/infdis/jiab021. URL <https://academic.oup.com/jid/article/223/7/1150/6104546>.
- R. M. Granich, C. F. Gilks, C. Dye, K. M. De Cock, and B. G. Williams. Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: A mathematical model. *The Lancet*, 373(9657):48–57, Jan. 2009. ISSN 01406736. doi: 10.1016/S0140-6736(08)61697-9. URL <https://linkinghub.elsevier.com/retrieve/pii/S0140673608616979>.
- R. H. Gray and M. J. Wawer. Reassessing the hypothesis on STI control for HIV prevention. *The Lancet*, 371(9630):2064–2065, June 2008. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(08)60896-X. URL <https://www.thelancet.com/journals/lancet/article/PIIS014067360860896X/fulltext>.
- R. H. Gray, F. Wabwire-Mangen, G. Kigozi, N. K. Sewankambo, D. Serwadda, L. H. Moulton, T. C. Quinn, K. L. O'Brien, M. Meehan, C. Abramowsky, M. Robb, and M. J. Wawer. Randomized trial of presumptive sexually transmitted disease therapy during pregnancy in Rakai, Uganda. *American Journal of Obstetrics and Gynecology*, 185(5):1209–1217, Nov. 2001. ISSN 0002-9378. doi: 10.1067/mob.2001.118158. URL <http://dx.doi.org/10.1067/mob.2001.118158>.

- R. H. Gray, G. Kigozi, D. Serwadda, F. Makumbi, S. Watya, F. Nalugoda, N. Kiwanuka, L. H. Moulton, M. A. Chaudhary, M. Z. Chen, N. K. Sewankambo, F. Wabwire-Mangen, M. C. Bacon, C. F. M. Williams, P. Opendi, S. J. Reynolds, O. Laeyendecker, T. C. Quinn, and M. J. Wawer. Male circumcision for HIV prevention in men in Rakai, Uganda: A randomised trial. *Lancet (London, England)*, 369(9562):657–666, Feb. 2007. ISSN 1474-547X. doi: 10.1016/S0140-6736(07)60313-4.
- S. Gregson, G. P. Garnett, and R. M. Anderson. Is HIV-1 likely to become a leading cause of adult mortality in sub-Saharan Africa? *Journal of Acquired Immune Deficiency Syndromes*, 7(8):839–852, Aug. 1994. ISSN 0894-9255.
- V. Grimm, A. S. A. Johnston, H.-H. Thulke, V. E. Forbes, and P. Thorbek. Three questions to ask before using model outputs for decision support. *Nature Communications*, 11(1):4959, Sept. 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17785-2. URL <https://www.nature.com/articles/s41467-020-17785-2>.
- D. J. Grint, K. Wing, C. Houlihan, H. P. Gibbs, S. J. W. Evans, E. Williamson, H. I. McDonald, K. Bhaskaran, D. Evans, A. J. Walker, G. Hickman, E. Nightingale, A. Schultze, C. T. Rentsch, C. Bates, J. Cockburn, H. J. Curtis, C. E. Morton, S. Bacon, S. Davy, A. Y. S. Wong, A. Mehrkar, L. Tomlinson, I. J. Douglas, R. Mathur, B. MacKenna, P. Ingelsby, R. Croker, J. Parry, F. Hester, S. Harper, N. J. DeVito, W. Hulme, J. Tazare, L. Smeeth, B. Goldacre, and R. M. Eggo. Severity of SARS-CoV-2 alpha variant (B.1.1.7) in england. *Clinical Infectious Diseases*, 2021. ISSN 1058-4838. doi: 10.1093/cid/ciab754. URL <https://doi.org/10.1093/cid/ciab754>.
- J. Guérin, K. Delmas, R. S. Ferreira, and J. Guiochet. Out-Of-Distribution Detection Is Not All You Need. 2022. doi: 10.48550/ARXIV.2211.16158. URL <https://arxiv.org/abs/2211.16158>.
- M. Hall, T. Golubchik, D. Bonsall, L. Abeler-Dörner, M. Limbada, B. Kosloff, A. Schaap, M. De Cesare, G. MacIntyre-Cockett, N. Otecko, W. Probert, O. Ratmann, A. Bulas Cruz, E. Piwowar-Manning, D. N. Burns, M. S. Cohen, D. J. Donnell, S. H. Eshleman, M. Simwinga, S. Fidler, R. Hayes, H. Ayles, C. Fraser, Y. Agyei, N. Beyers, P. Bock, V. Bond, J. Bwalya, A. Cori, A. Deventer, R. Dunbar, W. El-Sadr, L. Emel, S. Floyd, S. Griffith, J. Hargreaves, K. Hauck, T. Headen, G. Hoddinott, A. James, K. Jennings, S. Kanema, B. Kosloff, J. Kruger, R. Kumar, D. Macleod, N. Makola, N. Mandla, E. Miller, A. Moore, L. Mwenge, H. Noble, M. Phiri, M. Pickles, K. Sabapathy, E. Sakala, R. Sauter, K. Shanaube, A. Silumesi, N. Sista, T. Skalland, P. Smith, R. Thomas, S. Vermund, R. White, E. Wilson, B. Yang, K. Yuhas, R. Bowden, V. Calvez, M. Essex, K. Grabowski, R. Gupta, J. Herbeck, J. Kagaayi, P. Kaleebu, J. Lingappa, S. Moyo, V. Novitsky, T. Ndung’u, D. Pillay, T. Quinn, A. Rambaut, J. Seeley, D. Ssemwanga,

- F. Tanser, and M. Wawer. Demographics of sources of HIV-1 transmission in Zambia: A molecular epidemiology analysis in the HPTN 071 PopART study. *The Lancet Microbe*, 5(1):e62–e71, Jan. 2024. ISSN 26665247. doi: 10.1016/S2666-5247(23)00220-3. URL <https://linkinghub.elsevier.com/retrieve/pii/S2666524723002203>.
- A. J. Herbst, G. S. Cooke, T. Bärnighausen, A. KanyKany, F. Tanser, and M.-L. Newell. Adult mortality and antiretroviral treatment roll-out in rural KwaZulu-Natal, South Africa. *Bulletin of the World Health Organization*, 87(10):754–762, Oct. 2009. ISSN 00429686. doi: 10.2471/BLT.08.058982. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2755311/pdf/08-058982.pdf>.
- J. Hermans, V. Begy, and G. Louppe. Likelihood-free MCMC with Amortized Approximate Ratio Estimators, 2019. URL <https://arxiv.org/abs/1903.04057>.
- H. W. Hethcote. The Mathematics of Infectious Diseases. *SIAM Review*, 42(4):599–653, Jan. 2000. ISSN 0036-1445, 1095-7200. doi: 10.1137/S0036144500371907. URL <http://epubs.siam.org/doi/10.1137/S0036144500371907>.
- B. Hitchcock. Clampdown on AIDS information in E. Africa. *New African*, pages 9–10, Jan. 1986. ISSN 0142-9345.
- J. P. Hobert. The data augmentation algorithm : Theory and methodology. 2009. URL <https://api.semanticscholar.org/CorpusID:16162925>.
- M. D. Homan and A. Gelman. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, Jan. 2014a. ISSN 1532-4435.
- M. D. Homan and A. Gelman. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, 15(1):1593–1623, Jan. 2014b. ISSN 1532-4435.
- C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, and B. Cao. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223):497–506, Feb. 2020. ISSN 0140-6736. doi: 10.1016/S0140-6736(20)30183-5. URL <https://www.sciencedirect.com/science/article/pii/S0140673620301835>.
- R. A. Husein, H. Aburajouh, and C. Catal. Large language models for code completion: A systematic literature review. *Computer Standards & Interfaces*, 92:103917, Mar. 2025. ISSN 09205489. doi: 10.1016/j.csi.2024.103917. URL <https://linkinghub.elsevier.com/retrieve/pii/S0920548924000862>.

- D. S. Jones. The Persistence of American Indian Health Disparities. *American Journal of Public Health*, 96(12):2122–2134, Dec. 2006. ISSN 0090-0036. doi: 10.2105/AJPH.2004.054262. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1698152/>.
- K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, and P. Daszak. Global trends in emerging infectious diseases. *Nature*, 451(7181):990–993, Feb. 2008. ISSN 1476-4687. doi: 10.1038/nature06536. URL <https://www.nature.com/articles/nature06536>.
- J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, Aug. 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03819-2. URL <https://www.nature.com/articles/s41586-021-03819-2>.
- J. Kagaayi and D. Serwadda. The History of the HIV/AIDS Epidemic in Africa. *Current HIV/AIDS Reports*, 13(4):187–193, Aug. 2016. ISSN 1548-3576. doi: 10.1007/s11904-016-0318-8. URL <https://doi.org/10.1007/s11904-016-0318-8>.
- D. Kahle and H. Wickham. Ggmap: Spatial Visualization with ggplot2. *The R Journal*, 5(1):144–161, 2013. URL <https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf>.
- K. Katoh. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, July 2002. ISSN 13624962. doi: 10.1093/nar/gkf436. URL <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkf436>.
- R. Ke, E. Romero-Severson, S. Sanche, and N. Hengartner. Estimating the reproductive number R_0 of SARS-CoV-2 in the United States and eight European countries and implications for vaccination. *Journal of Theoretical Biology*, 517:110621, May 2021. ISSN 0022-5193. doi: 10.1016/j.jtbi.2021.110621. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7880839/>.
- W. O. Kermack, A. G. McKendrick, and G. T. Walker. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, 1927. doi: 10.1098/rspa.1927.0118. URL <https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1927.0118>.

- A. B. M. Kharsany, C. Cawood, L. Lewis, N. Yende-Zuma, D. Khanyile, A. Puren, S. Madurai, C. Baxter, G. George, K. Govender, S. Beckett, N. Samsunder, C. Toledo, K. A. Ayalew, K. Diallo, M. Glenshaw, A. Herman-Roloff, E. Wilkinson, T. de Oliveira, S. S. Abdool Karim, and Q. Abdool Karim. Trends in HIV Prevention, Treatment, and Incidence in a Hyperendemic Area of KwaZulu-Natal, South Africa. *JAMA Network Open*, 2(11): e1914378, Nov. 2019. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2019.14378. URL <https://doi.org/10.1001/jamanetworkopen.2019.14378>.
- D. S. Khoury, D. Cromer, A. Reynaldi, T. E. Schlub, A. K. Wheatley, J. A. Juno, K. Subbarao, S. J. Kent, J. A. Triccas, and M. P. Davenport. Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection. *Nature Medicine*, 27:1205—1211, 2021. doi: 10.1038/s41591-021-01377-8.
- D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes, 2013. URL <https://arxiv.org/abs/1312.6114>.
- A. Koch, C. Brierley, M. M. Maslin, and S. L. Lewis. Earth system impacts of the European arrival and Great Dying in the Americas after 1492. *Quaternary Science Reviews*, 207:13–36, Mar. 2019. ISSN 0277-3791. doi: 10.1016/j.quascirev.2018.12.004. URL <https://www.sciencedirect.com/science/article/pii/S0277379118307261>.
- B. Korber, M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya. Timing the Ancestor of the HIV-1 Pandemic Strains. *Science*, 288(5472):1789–1796, June 2000. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.288.5472.1789. URL <https://www.science.org/doi/10.1126/science.288.5472.1789>.
- R. D. Kouyos, V. von Wyl, S. Yerly, J. Böni, P. Rieder, B. Joos, P. Taffé, C. Shah, P. Bürgisser, T. Klimkait, R. Weber, B. Hirschel, M. Cavassini, A. Rauch, M. Battegay, P. L. Vernazza, E. Bernasconi, B. Ledergerber, S. Bonhoeffer, H. F. Günthard, and the Swiss HIV Cohort Study. Ambiguous Nucleotide Calls From Population-based Sequencing of HIV-1 are a Marker for Viral Diversity and the Age of Infection. *Clinical Infectious Diseases*, 52(4):532–539, Feb. 2011. ISSN 1058-4838. doi: 10.1093/cid/ciq164. URL <https://doi.org/10.1093/cid/ciq164>.
- M. E. Kretzschmar, B. Ashby, E. Fearon, C. E. Overton, J. Panovska-Griffiths, L. Pellis, M. Quaife, G. Rozhnova, F. Scarabel, H. B. Stage, B. Swallow, R. N. Thompson, M. J. Tildesley, and D. Villela. Challenges for modelling interventions for future pandemics. *Epidemics*, 38:100546, Mar. 2022. ISSN 1755-4365. doi: 10.1016/j.epidem.2022.100546. URL <https://www.sciencedirect.com/science/article/pii/S1755436522000081>.

- A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(1):430–474, Jan. 2017. ISSN 1532-4435.
- L. A. Kurpiers, B. Schulte-Herbrüggen, I. Ejotre, and D. M. Reeder. Bushmeat and Emerging Infectious Diseases: Lessons from Africa. *Problematic Wildlife*, pages 507–551, Sept. 2015. doi: 10.1007/978-3-319-22246-2_24. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7123567/>.
- J. Larmarange, P. Bachanas, T. Skalland, L. B. Balzer, C. Iwuji, S. Floyd, L. A. Mills, D. Pillay, D. Havlir, M. R. Kamya, H. Ayles, K. Wirth, F. Dabis, R. Hayes, M. Petersen, and for the UT³C consortium. Population-level viremia predicts HIV incidence at the community level across the Universal Testing and Treatment Trials in eastern and southern Africa. *PLOS Global Public Health*, 3(7):e0002157, July 2023. ISSN 2767-3375. doi: 10.1371/journal.pgph.0002157. URL <https://dx.plos.org/10.1371/journal.pgph.0002157>.
- R. Lebelonyane, P. Bachanas, L. Block, F. Ussery, M. G. Alwano, T. Marukutira, S. El Halabi, M. Roland, W. Abrams, G. Ussery, J. A. Miller, S. Lockman, T. Gaolathe, M. P. Holme, S. Hader, L. A. Mills, K. Wirth, N. Bock, and J. Moore. To achieve 95-95-95 targets we must reach men and youth: High level of knowledge of HIV status, ART coverage, and viral suppression in the Botswana Combination Prevention Project through universal test and treat approach. *PLOS ONE*, 16(8):e0255227, Aug. 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0255227. URL <https://dx.plos.org/10.1371/journal.pone.0255227>.
- Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541.
- S. N. Leite, M. Finkler, J. G. Martini, I. Heidemann, M. Verdi, F. Hellmann, and M. F. Vásquez. Management of the health workforce in facing COVID-19: Disinformation and absences in Brazil’s Public Policies. *Ciência & Saúde Coletiva*, 26:1873–1884, 2021.
- Lesotho Ministry of Health LB of S. Lesotho Summary Sheet 2020. Technical report, 2020. URL https://phia.icap.columbia.edu/wp-content/uploads/2021/08/53059_14_LePHIA_Summary-sheet_with-coat-of-arms_WEB_v2.pdf.
- Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T. T. Lam, J. T. Wu, G. F. Gao, B. J. Cowling, B. Yang, G. M. Leung, and Z. Feng. Early Transmission

- Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *New England Journal of Medicine*, 382(13):1199–1207, Mar. 2020. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa2001316. URL <http://www.nejm.org/doi/10.1056/NEJMoa2001316>.
- S. L. Li, R. H. M. Pereira, C. A. Prete Jr, A. E. Zarebski, L. Emanuel, P. J. H. Alves, P. S. Peixoto, C. K. V. Braga, A. A. de Souza Santos, W. M. de Souza, R. J. Barbosa, L. F. Buss, A. Mendrone, C. de Almeida-Neto, S. C. Ferreira, N. A. Salles, I. Marcilio, C.-H. Wu, N. Gouveia, V. H. Nascimento, E. C. Sabino, N. R. Faria, and J. P. Messina. Higher risk of death from COVID-19 in low-income and non-White populations of São Paulo, Brazil. *BMJ Global Health*, 6, 2021. doi: 10.1136/bmjgh-2021-004959. URL <https://gh.bmj.com/content/6/4/e004959>.
- X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdass, C. Kern, J. R. Ledsam, M. K. Schmid, K. Balaskas, E. J. Topol, L. M. Bachmann, P. A. Keane, and A. K. Denniston. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet Digital Health*, 1(6): e271–e297, Oct. 2019. ISSN 25897500. doi: 10.1016/S2589-7500(19)30123-2. URL <https://linkinghub.elsevier.com/retrieve/pii/S2589750019301232>.
- J. Ma. Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling*, 5:129–141, Jan. 2020. ISSN 2468-0427. doi: 10.1016/j.idm.2019.12.009. URL <https://www.sciencedirect.com/science/article/pii/S2468042719300491>.
- F. R. Machado, A. B. Cavalcanti, F. A. Bozza, E. M. Ferreira, F. S. A. Carrara, J. L. Sousa, N. Caixeta, R. Salomao, D. C. Angus, L. C. P. Azevedo, et al. The epidemiology of sepsis in Brazilian intensive care units (the Sepsis PREvalence Assessment Database, SPREAD): An observational study. *The Lancet Infectious Diseases*, 17:1180–1189, 2017.
- S. A. Madhi, V. Baillie, C. L. Cutland, M. Voysey, A. L. Koen, L. Fairlie, S. D. Padayachee, K. Dheda, S. L. Barnabas, Q. E. Bhorat, et al. Efficacy of the ChAdOx1 nCoV-19 COVID-19 vaccine against the B. 1.351 variant. *New England Journal of Medicine*, 384: 1885–1898, 2021. doi: 10.1056/NEJMoa2102214.
- L. E. Magosi, Y. Zhang, T. Golubchik, V. DeGruttola, E. Tchetgen Tchetgen, V. Novitsky, J. Moore, P. Bachanas, T. Segolodi, R. Lebelonyane, M. Pretorius Holme, S. Moyo, J. Makhema, S. Lockman, C. Fraser, M. M. Essex, M. Lipsitch, and On behalf of The Botswana Combination Prevention Project and PANGEA consortium. Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial – BCPP/Ya Tsie trial. *eLife*, 11:e72657, Mar. 2022. ISSN 2050-084X. doi: 10.7554/eLife.72657. URL <https://elifesciences.org/articles/72657>.

- M. Maheu-Giroux and S. Mishra. Evidence with 95-95-95 that ambitious is feasible. *The Lancet HIV*, 0(0), Mar. 2024. ISSN 2352-3018. doi: 10.1016/S2352-3018(24)00028-6. URL [https://www.thelancet.com/journals/lanhiv/article/PIIS2352-3018\(24\)00028-6/fulltext](https://www.thelancet.com/journals/lanhiv/article/PIIS2352-3018(24)00028-6/fulltext).
- M. I. Mahy, K. M. Sabin, A. Feizzadeh, and I. Wanyeki. Progress towards 2020 global HIV impact and treatment targets. *Journal of the International AIDS Society*, 24 (S5):e25779, Sept. 2021. ISSN 1758-2652, 1758-2652. doi: 10.1002/jia2.25779. URL <https://onlinelibrary.wiley.com/doi/10.1002/jia2.25779>.
- Malawi Ministry of Health. Malawi Summary Sheet 2020-2021. Technical report, Mar. 2022. URL https://phia.icap.columbia.edu/wp-content/uploads/2022/03/110322_MPHIA_Summary-sheet-English.pdf.
- M. Malta, S. A. Strathdee, and P. J. Garcia. The Brazilian tragedy: Where patients living at the ‘Earth’s lungs’ die of asphyxia, and the fallacy of herd immunity is killing people. *EClinicalMedicine*, 32, 2021. doi: 10.1016/j.eclinm.2021.100757.
- M. Marston, B. Zaba, and J. W. Eaton. Relative patterns of sexual activity and fertility among HIV positive and negative women-Evidence from 46 DHS. *PloS One*, 13(10): e0204584, 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0204584.
- A. F. Martins, A. P. Zavascki, P. L. Wink, F. C. Z. Volpato, F. L. Monteiro, C. Rosset, F. De-Paris, Á. K. Ramos, and A. L. Barth. Detection of SARS-CoV-2 lineage P.1 in patients from a region with exponentially increasing hospitalisation rate, February 2021 Rio Grande do Sul, Southern Brazil. *Eurosurveillance*, 26:210–276, 2021.
- J. R. Mascola, B. S. Graham, and A. S. Fauci. SARS-CoV-2 Viral Variants—Tackling a Moving Target. *JAMA*, 325(13):1261–1262, Apr. 2021. ISSN 0098-7484. doi: 10.1001/jama.2021.2088. URL <https://doi.org/10.1001/jama.2021.2088>.
- E. Mathieu, H. Ritchie, L. Rodés-Guirao, C. Appel, C. Giattino, J. Hasell, B. Macdonald, S. Dattani, D. Beltekian, E. Ortiz-Ospina, and M. Roser. Coronavirus Pandemic (COVID-19). *Our World in Data*, Mar. 2020. URL <https://ourworldindata.org/covid-vaccinations>.
- M. E. McGovern, T. Bärnighausen, J. A. Salomon, and D. Canning. Using interviewer random effects to remove selection bias from HIV prevalence estimates. *BMC Medical Research Methodology*, 15(1):8, Feb. 2015. ISSN 1471-2288. doi: 10.1186/1471-2288-15-8. URL <https://doi.org/10.1186/1471-2288-15-8>.
- M. McGowan, K. Bärnighausen, A. Berner-Rodoreda, S. A. McMahon, C. Mtaita, J. Mauti, and F. Neuhaan. “The targets...are driving the agenda and that probably needs

- to change”: Stakeholder perspectives on HIV partner notification in sub-Saharan Africa. *BMC Public Health*, 24(1):521, Feb. 2024. ISSN 1471-2458. doi: 10.1186/s12889-023-17422-9. URL <https://doi.org/10.1186/s12889-023-17422-9>.
- N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7 (35):983–999, 2006. URL <http://jmlr.org/papers/v7/meinshausen06a.html>.
- Á. Mendes, L. Carnut, and L. D. d. S. Guerra. Primary Health Care federal funding in the Unified Health System: Old and new dilemmas. *Saúde em Debate*, 42(SPE1):224–243, 2018.
- G. N. Mercer, K. Glass, and N. G. Becker. Effective reproduction numbers are commonly overestimated early in a disease outbreak. *Statistics in Medicine*, 30(9):984–994, Apr. 2011. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.4174. URL <https://onlinelibrary.wiley.com/doi/10.1002/sim.4174>.
- M. J. Miller and B. J. Reich. Bayesian spatial modeling using random Fourier frequencies. *Spatial Statistics*, 48:100598, Apr. 2022. ISSN 2211-6753. doi: 10.1016/j.spasta.2022.100598. URL <https://www.sciencedirect.com/science/article/pii/S2211675322000069>.
- W. C. Miller, K. A. Powers, M. K. Smith, and M. S. Cohen. Community viral load as a measure for assessment of HIV treatment as prevention. *The Lancet Infectious Diseases*, 13(5):459–464, May 2013. ISSN 14733099. doi: 10.1016/S1473-3099(12)70314-6. URL <https://linkinghub.elsevier.com/retrieve/pii/S1473309912703146>.
- E. J. Mills, C. Bakanda, J. Birungi, K. Chan, N. Ford, C. L. Cooper, J. B. Nachega, M. Dybul, and R. S. Hogg. Life Expectancy of Persons Receiving Combination Antiretroviral Therapy in Low-Income Countries: A Cohort Analysis From Uganda. *Annals of Internal Medicine*, 155(4):209, Aug. 2011. ISSN 0003-4819. doi: 10.7326/0003-4819-155-4-201108160-00358. URL <http://annals.org/article.aspx?doi=10.7326/0003-4819-155-4-201108160-00358>.
- M. Mine, K. A. Stafford, R. L. Laws, R. Marima, P. Lekone, D. Ramaabya, K. Makhaola, H. K. Patel, P. Mapondera, F. Wray-Gordon, C. Agbakwuru, L. Okui, S. Matroos, E. Onyadile, J. Ngidi, A. Abimiku, K. Bagapi, B. Nkomo, S. M. Bodika, K. J. Kim, M. Moloney, A. Mitchell, A. Ehoche, F. L. Ussery, S. Y. Hong, S. Keipeile, M. Matlhaga, R. Mathumo, R. Selato, M. E. Charurat, and A. C. Voetsch. Progress towards the UNAIDS 95-95-95 targets in the Fifth Botswana AIDS Impact Survey (BAIS V 2021): A nationally representative survey. *The Lancet HIV*, 0(0), Mar. 2024. ISSN 2352-3018. doi: 10.1016/S2352-3018(24)00003-1. URL [https://www.thelancet.com/journals/lanhiv/article/PIIS2352-3018\(24\)00003-1/fulltext](https://www.thelancet.com/journals/lanhiv/article/PIIS2352-3018(24)00003-1/fulltext).

- J. F. Molez. The historical question of acquired immunodeficiency syndrome in the 1960s in the Congo River basin area in relation to cryptococcal meningitis. *The American Journal of Tropical Medicine and Hygiene*, 58(3):273–276, Mar. 1998. ISSN 0002-9637. doi: 10.4269/ajtmh.1998.58.273.
- M. Monod, A. Blenkinsop, A. Brizzi, Y. Chen, C. Cardoso Correia Perello, V. Jogarah, Y. Wang, S. Flaxman, S. Bhatt, and O. Ratmann. Regularised B-splines Projected Gaussian Process Priors to Estimate Time-trends in Age-specific COVID-19 Deaths. *Bayesian Analysis*, 18(3), Sept. 2023a. ISSN 1936-0975. doi: 10.1214/22-BA1334. URL <https://projecteuclid.org/journals/bayesian-analysis/volume-18/issue-3/Regularised-B-splines-Projected-Gaussian-Process-Priors-to-Estimate-Time/10.1214/22-BA1334.full>.
- M. Monod, A. Brizzi, R. M. Galiwango, R. Ssekubugu, Y. Chen, X. Xi, E. N. Kankaka, V. Ssempijja, L. Abeler-Dörner, A. Akullian, A. Blenkinsop, D. Bonsall, L. W. Chang, S. Dan, C. Fraser, T. Golubchik, R. H. Gray, M. Hall, J. C. Jackson, G. Kigozi, O. Laeyendecker, L. A. Mills, T. C. Quinn, S. J. Reynolds, J. Santelli, N. K. Sewankambo, S. E. F. Spencer, J. Ssekasanvu, L. Thomson, M. J. Wawer, D. Serwadda, P. Godfrey-Faussett, J. Kagaayi, M. K. Grabowski, O. Ratmann, Rakai Health Sciences Program, and PANGEA-HIV consortium. Longitudinal population-level HIV epidemiologic and genomic surveillance highlights growing gender disparity of HIV transmission in Uganda. *Nature Microbiology*, 9(1):35–54, Dec. 2023b. ISSN 2058-5276. doi: 10.1038/s41564-023-01530-8. URL <https://www.nature.com/articles/s41564-023-01530-8>.
- S. S. Morse. Factors in the emergence of infectious diseases. *Emerging Infectious Diseases*, 1(1):7–15, 1995. ISSN 1080-6040. doi: 10.3201/eid0101.950102.
- Mozambique Ministry of Health (MISAU). Mozambique Summary Sheet 2021. Technical report, Dec. 2022. URL https://phia.icap.columbia.edu/wp-content/uploads/2022/12/53059_14_INSIDA_Summary-sheet-Web.pdf.
- R. K. Nash, S. Bhatt, A. Cori, and P. Nouvellet. Estimating the epidemic reproduction number from temporally aggregated incidence data: A statistical modelling approach and software tool. *PLOS Computational Biology*, 19(8):e1011439, Aug. 2023. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1011439. URL <https://dx.plos.org/10.1371/journal.pcbi.1011439>.
- F. G. Naveca, V. Nascimento, V. C. de Souza, A. d. L. Corado, F. Nascimento, G. Silva, Á. Costa, D. Duarte, K. Pessoa, M. Mejía, M. J. Brandão, M. Jesus, L. Gonçalves, C. F. da Costa, V. Sampaio, D. Barros, M. Silva, T. Mattos, G. Pontes, L. Abdalla, J. H. Santos, I. Arantes, F. Z. Dezordi, M. M. Siqueira, G. L. Wallau, P. C. Resende, E. Delatorre,

- T. Gräf, and G. Bello. COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence. *Nature Medicine*, 27:1230—1238, May 2021.
- L.-T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, Jan. 2015. ISSN 1537-1719, 0737-4038. doi: 10.1093/molbev/msu300. URL <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msu300>.
- NHS England. Landmark moment as first NHS patient receives COVID-19 vaccination, Dec. 2020. URL <https://www.england.nhs.uk/2020/12/landmark-moment-as-first-nhs-patient-receives-covid-19-vaccination/>.
- M. A. Nicolelis, R. L. Raimundo, P. S. Peixoto, and C. S. Andreazzi. The impact of super-spreader cities, highways, and intensive care availability in the early stages of the COVID-19 epidemic in Brazil. *Scientific Reports*, 11:1–12, 2021. doi: 10.1038/s41598-021-92263-3.
- K. V. M. d. S. Noronha, G. R. Guedes, C. M. Turra, M. V. Andrade, L. Botega, D. Nogueira, J. A. Calazans, L. Carvalho, L. Servo, and M. F. Ferreira. Pandemia por COVID-19 no Brasil: Análise da demanda e da oferta de leitos hospitalares e equipamentos de ventilação assistida segundo diferentes cenários. *Cadernos de Saúde Pública*, 36, 2020. doi: 10.1590/0102-311X00115320.
- P. Nouvellet, S. Bhatia, A. Cori, K. E. C. Ainslie, M. Baguelin, S. Bhatt, A. Boonyasiri, N. F. Brazeau, L. Cattarino, L. V. Cooper, H. Coupland, Z. M. Cucunuba, G. Cuomo-Dannenburg, A. Dighe, B. A. Djaafara, I. Dorigatti, O. D. Eales, S. L. van Elsland, F. F. Nascimento, R. G. FitzJohn, K. A. M. Gaythorpe, L. Geidelberg, W. D. Green, A. Hamlet, K. Hauck, W. Hinsley, N. Imai, B. Jeffrey, E. Knock, D. J. Laydon, J. A. Lees, T. Mangal, T. A. Mellan, G. Nedjati-Gilani, K. V. Parag, M. Pons-Salort, M. Ragonnet-Cronin, S. Riley, H. J. T. Unwin, R. Verity, M. A. C. Vollmer, E. Volz, P. G. T. Walker, C. E. Walters, H. Wang, O. J. Watson, C. Whittaker, L. K. Whittles, X. Xi, N. M. Ferguson, and C. A. Donnelly. Reduction in mobility and COVID-19 transmission. *Nature Communications*, 12(1):1090, Feb. 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21358-2. URL <https://www.nature.com/articles/s41467-021-21358-2>.
- R. Nyaboke, H. O. Ramadhani, T. Lascko, P. Awuor, E. Kirui, E. Koech, I. Mutisya, C. Ngunu, and R. Wangusi. Factors associated with adherence and viral suppression among patients on second-line antiretroviral therapy in an urban HIV program in Kenya. *SAGE Open Medicine*, 11:205031212311623, Jan. 2023. ISSN 2050-3121, 2050-3121. doi: 10.1177/20503121231162354. URL <http://journals.sagepub.com/doi/10.1177/20503121231162354>.

- T. Nyberg, K. A. Twohig, R. J. Harris, S. R. Seaman, J. Flannagan, H. Allen, A. Charlett, D. De Angelis, G. Dabrera, and A. M. Presanis. Risk of hospital admission for patients with SARS-CoV-2 variant B.1.1.7: Cohort analysis. *BMJ*, page n1412, June 2021. ISSN 1756-1833. doi: 10.1136/bmj.n1412. URL <https://www.bmj.com/lookup/doi/10.1136/bmj.n1412>.
- T. Nyberg, N. M. Ferguson, S. G. Nash, H. H. Webster, S. Flaxman, N. Andrews, W. Hinsley, J. L. Bernal, M. Kall, S. Bhatt, P. Blomquist, A. Zaidi, E. Volz, N. A. Aziz, K. Harman, S. Funk, S. Abbott, T. Nyberg, N. M. Ferguson, S. G. Nash, H. H. Webster, S. Flaxman, N. Andrews, W. Hinsley, J. Lopez Bernal, M. Kall, S. Bhatt, P. Blomquist, A. Zaidi, E. Volz, N. Abdul Aziz, K. Harman, S. Funk, S. Abbott, R. Hope, A. Charlett, M. Chand, A. C. Ghani, S. R. Seaman, G. Dabrera, D. De Angelis, A. M. Presanis, S. Thelwall, R. Hope, A. Charlett, M. Chand, A. C. Ghani, S. R. Seaman, G. Dabrera, D. De Angelis, A. M. Presanis, and S. Thelwall. Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 omicron (B.1.1.529) and delta (B.1.617.2) variants in England: A cohort study. *The Lancet*, 399(10332):1303–1312, Apr. 2022. ISSN 0140-6736. doi: 10.1016/S0140-6736(22)00462-7. URL <https://www.sciencedirect.com/science/article/pii/S0140673622004627>.
- T. Nyberg, P. Bager, I. B. Svalgaard, D. Bejko, N. Bundle, J. Evans, T. G. Krause, J. McMenamin, J. Mossong, H. Mutch, A. Omokanye, A. Peralta-Santos, P. Pinto-Leite, J. Starrfelt, S. Thelwall, L. Veneti, R. Whittaker, J. Wood, R. Pebody, and A. M. Presanis. A standardised protocol for relative SARS-CoV-2 variant severity assessment, applied to Omicron BA.1 and Delta in six European countries, October 2021 to February 2022. *Eurosurveillance*, 28(36):2300048, Sept. 2023. ISSN 1560-7917. doi: 10.2807/1560-7917.ES.2023.28.36.2300048. URL <https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2023.28.36.2300048>.
- M. O’Driscoll, C. Harry, C. A. Donnelly, A. Cori, and I. Dorigatti. A Comparative Analysis of Statistical Methods to Estimate the Reproduction Number in Emerging Epidemics, With Implications for the Current Coronavirus Disease 2019 (COVID-19) Pandemic. *Clinical Infectious Diseases*, 73(1):e215–e223, July 2021. ISSN 1058-4838. doi: 10.1093/cid/ciaa1599. URL <https://doi.org/10.1093/cid/ciaa1599>.
- I. Olier, Y. Zhan, X. Liang, and V. Volovici. Causal inference and observational data. *BMC Medical Research Methodology*, 23(1):227, Oct. 2023. ISSN 1471-2288. doi: 10.1186/s12874-023-02058-5. URL <https://doi.org/10.1186/s12874-023-02058-5>.
- Á. O’Toole, E. Scher, A. Underwood, B. Jackson, V. Hill, J. T. McCrone, R. Colquhoun, C. Ruis, K. Abu-Dahab, B. Taylor, C. Yeats, L. du Plessis, D. Maloney, N. Medd, S. W. Attwood, D. M. Aanensen, E. C. Holmes, O. G. Pybus, and A. Rambaut. Assignment

- of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution*, 7(2):veab064, Dec. 2021. ISSN 2057-1577. doi: 10.1093/ve/veab064. URL <https://doi.org/10.1093/ve/veab064>.
- M. S. Pakkanen, X. Miscouridou, M. J. Penn, C. Whittaker, T. Berah, S. Mishra, T. A. Mellan, and S. Bhatt. Unifying incidence and prevalence under a time-varying general branching process. *Journal of Mathematical Biology*, 87(2):35, Aug. 2023. ISSN 1432-1416. doi: 10.1007/s00285-023-01958-w. URL <https://doi.org/10.1007/s00285-023-01958-w>.
- A. M. Palma, G. Marra, R. Bray, S. Saito, A. C. Awor, M. F. Jalloh, A. Kailembo, W. Kirungi, G. S. Mgomella, P. Njau, A. C. Voetsch, J. A. Ward, T. Bärnighausen, and G. Harling. Correcting for selection bias in HIV prevalence estimates: An application of sample selection models using data from population-based HIV surveys in seven sub-Saharan African countries. *Journal of the International AIDS Society*, 25(8):e25954, Aug. 2022. ISSN 1758-2652. doi: 10.1002/jia2.25954. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9353488/>.
- N. Pantazis, C. Thomadakis, J. Del Amo, D. Alvarez-del Arco, F. M. Burns, I. Fakoya, and G. Touloumi. Determining the likely place of HIV acquisition for migrants in Europe combining subject-specific information and biomarkers data. *Statistical Methods in Medical Research*, 28(7):1979–1997, July 2019. ISSN 0962-2802, 1477-0334. doi: 10.1177/0962280217746437. URL <http://journals.sagepub.com/doi/10.1177/0962280217746437>.
- N. Pantazis, M. Rosinska, A. van Sighem, C. Quinten, T. Noori, F. Burns, H. Cortes Martins, P. D. Kirwan, K. O'Donnell, D. Paraskevis, C. Sommen, D. Zenner, and A. Pharris. Discriminating Between Premigration and Postmigration HIV Acquisition Using Surveillance Data. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 88(2):117, Oct. 2021. ISSN 1525-4135. doi: 10.1097/QAI.0000000000002745. URL https://journals.lww.com/jaids/fulltext/2021/10010/discriminating_between_premigration_and.1.aspx.
- E. U. Patel, S. S. Solomon, G. M. Lucas, A. M. McFall, A. K. Srikrishnan, M. S. Kumar, S. H. Iqbal, S. Saravanan, N. Paneerselvam, P. Balakrishnan, O. Laeyendecker, D. D. Celentano, and S. H. Mehta. Temporal change in population-level prevalence of detectable HIV viraemia and its association with HIV incidence in key populations in India: A serial cross-sectional study. *The Lancet HIV*, 8(9):e544–e553, Sept. 2021. ISSN 2352-3018. doi: 10.1016/S2352-3018(21)00098-9. URL <https://linkinghub.elsevier.com/retrieve/pii/S2352301821000989>.
- I. Pathmanathan, R. Nelson, A. De Louvado, R. Thompson, S. Pals, I. Casavant, M. J. A. Cardoso, D. Ujamaa, J. Bonzela, S. Mikusova, V. Chivurre, S. Tamele, K. Sleeman,

- G. Zhang, C. Zeh, T. Dobbs, A. Vubil, A. Auld, M. Briggs-Hagen, A. Vergara, A. Couto, and D. MacKellar. High Coverage of Antiretroviral Treatment With Annual Home-Based HIV Testing, Follow-up Linkage Services, and Implementation of Test and Start: Findings From the Chókwè Health Demographic Surveillance System, Mozambique, 2014–2019. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 86(4):e97–e105, Apr. 2021. ISSN 1525-4135. doi: 10.1097/QAI.0000000000002583. URL <https://journals.lww.com/10.1097/QAI.0000000000002583>.
- R. H. M. Pereira, C. K. V. Braga, L. M. Servo, B. Serra, P. Amaral, N. Gouveia, and A. Paez. Geographic access to COVID-19 healthcare in Brazil using a balanced float catchment area approach. *Social Science & Medicine*, 273, 2021. ISSN 0277-9536. doi: 10.1016/j.socscimed.2021.113773. URL <https://www.sciencedirect.com/science/article/pii/S0277953621001052>.
- P. N. Perez-Guzman, E. Knock, N. Imai, T. Rawson, Y. Elmaci, J. Alcada, L. K. Whittles, D. Thekke Kanapram, R. Sonabend, K. A. M. Gaythorpe, W. Hinsley, R. G. FitzJohn, E. Volz, R. Verity, N. M. Ferguson, A. Cori, and M. Baguelin. Author Correction: Epidemiological drivers of transmissibility and severity of SARS-CoV-2 in England. *Nature Communications*, 14(1):8099, Dec. 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-44062-9. URL <https://www.nature.com/articles/s41467-023-44062-9>.
- P. N. Perez-Guzman, S. L. Chanda, A. Schaap, K. Shanaube, M. Baguelin, S. T. Nyangu, M. K. Kanyanga, P. Walker, H. Ayles, R. Chilengi, R. Verity, K. Hauck, E. S. Knock, and A. Cori. Pandemic burden in low-income settings and impact of limited and delayed interventions: A granular modelling analysis of COVID-19 in Kabwe, Zambia. *International Journal of Infectious Diseases*, 147:107182, Oct. 2024. ISSN 12019712. doi: 10.1016/j.ijid.2024.107182. URL <https://linkinghub.elsevier.com/retrieve/pii/S1201971224002534>.
- M. Peroni, M. Kurban, S. Y. Yang, Y. S. Kim, H. Y. Kang, and J. H. Song. Extending the Neural Additive Model for Survival Analysis with EHR Data, 2022. URL <https://arxiv.org/abs/2211.07814>.
- N. Perra. Non-pharmaceutical interventions during the COVID-19 pandemic: A review. *Physics Reports*, 913:1–52, May 2021. ISSN 0370-1573. doi: 10.1016/j.physrep.2021.02.001. URL <https://www.sciencedirect.com/science/article/pii/S0370157321000624>.
- M. Pickles, A. Cori, W. J. M. Probert, R. Sauter, R. Hinch, S. Fidler, H. Ayles, P. Bock, D. Donnell, E. Wilson, E. Piwowar-Manning, S. Floyd, R. J. Hayes, C. Fraser, and HPTN 071 (PopART) Study Team. PopART-IBM, a highly efficient stochastic individual-based simulation model of generalised HIV epidemics developed in the context of the HPTN

- 071 (PopART) trial. *PLOS Computational Biology*, 17(9):e1009301, Sept. 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1009301. URL <https://dx.plos.org/10.1371/journal.pcbi.1009301>.
- D. Pillay, J. Herbeck, M. S. Cohen, T. de Oliveira, C. Fraser, O. Ratmann, A. L. Brown, and P. Kellam. PANGEA-HIV: Phylogenetics for generalised epidemics in Africa. *The Lancet Infectious Diseases*, 15(3):259–261, Mar. 2015. ISSN 1473-3099. doi: 10.1016/S1473-3099(15)70036-8. URL <https://www.sciencedirect.com/science/article/pii/S1473309915700368>.
- T. Pillay, M. Cornell, M. P. Fox, J. Euvrard, G. Fatti, K.-G. Technau, N. Sipambo, H. Prozesky, B. Eley, F. Tanser, and L. F. Johnson. Recording of HIV Viral Loads and Viral Suppression in South African Patients Receiving Antiretroviral Treatment: A Multicentre Cohort Study. *Antiviral Therapy*, 25(5):257–266, July 2020. ISSN 1359-6535, 2040-2058. doi: 10.3851/IMP3371. URL <http://journals.sagepub.com/doi/10.3851/IMP3371>.
- C. Pley, M. Evans, R. Lowe, H. Montgomery, and S. Yacoub. Digital and technological innovation in vector-borne disease surveillance to predict, detect, and control climate-driven outbreaks. *The Lancet Planetary Health*, 5(10):e739–e745, Oct. 2021. ISSN 25425196. doi: 10.1016/S2542-5196(21)00141-8. URL <https://linkinghub.elsevier.com/retrieve/pii/S2542519621001418>.
- M. Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. 2003. URL <https://api.semanticscholar.org/CorpusID:265976949>.
- J. A. Polonsky, A. Baidjoe, Z. N. Kamvar, A. Cori, K. Durski, W. J. Edmunds, R. M. Eggo, S. Funk, L. Kaiser, P. Keating, O. L. P. De Waroux, M. Marks, P. Moraga, O. Morgan, P. Nouvellet, R. Ratnayake, C. H. Roberts, J. Whitworth, and T. Jombart. Outbreak analytics: A developing data science for informing the response to emerging pathogens. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1776):20180276, July 2019. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2018.0276. URL <https://royalsocietypublishing.org/doi/10.1098/rstb.2018.0276>.
- E. A. Preble. Impact of HIV/AIDS on African children. *Social Science & Medicine*, 31(6):671–680, Jan. 1990. ISSN 0277-9536. doi: 10.1016/0277-9536(90)90249-R. URL <https://www.sciencedirect.com/science/article/pii/027795369090249R>.
- Public Health England. Investigation of SARS-CoV-2 variants of concern: Technical briefings, Sept. 2021. URL <https://www.gov.uk/government/publications/investigation-of-novel-sars-cov-2-variant-variant-of-concern-20201201>.

- L. Qin, Q. Sun, Y. Wang, K.-F. Wu, M. Chen, B.-C. Shia, and S.-Y. Wu. Prediction of Number of Cases of 2019 Novel Coronavirus (COVID-19) Using Social Media Search Index. *International Journal of Environmental Research and Public Health*, 17(7):2365, Mar. 2020. ISSN 1660-4601. doi: 10.3390/ijerph17072365. URL <https://www.mdpi.com/1660-4601/17/7/2365>.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. URL <https://www.R-project.org/>.
- B. Rai, A. Shukla, and L. K. Dwivedi. Estimates of serial interval for COVID-19: A systematic review and meta-analysis. *Clinical Epidemiology and Global Health*, 9: 157–161, Jan. 2021. ISSN 22133984. doi: 10.1016/j.cegh.2020.08.007. URL <https://linkinghub.elsevier.com/retrieve/pii/S2213398420301895>.
- R. Rajpoot, M. Gour, S. Jain, and V. B. Semwal. Integrated ensemble CNN and explainable AI for COVID-19 diagnosis from CT scan and X-ray images. *Scientific Reports*, 14(1): 24985, Oct. 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-75915-y. URL <https://www.nature.com/articles/s41598-024-75915-y>.
- A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, and O. G. Pybus. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, 5:1403–1407, 2020. doi: 10.1038/s41564-020-0770-5.
- O. T. Ranzani, L. S. Bastos, J. G. M. Gelli, J. F. Marchesi, F. Baião, S. Hamacher, and F. A. Bozza. Characterisation of the first 250 000 hospital admissions for COVID-19 in Brazil: A retrospective analysis of nationwide data. *The Lancet Respiratory Medicine*, 9:407–418, 2021a. doi: 10.1016/S2213-2600(20)30560-9.
- O. T. Ranzani, M. D. T. Hitchings, M. Dorion, T. L. D'Agostini, R. C. de Paula, O. F. P. de Paula, E. F. d. M. Villela, M. S. S. Torres, S. B. de Oliveira, W. Schulz, M. Almiron, R. Said, R. D. de Oliveira, P. Vieira da Silva, W. N. de Araújo, J. C. Gorinchteyn, J. R. Andrews, D. A. T. Cummings, A. I. Ko, and J. Croda. Effectiveness of the CoronaVac vaccine in older adults during a gamma variant associated epidemic of COVID-19 in Brazil: Test negative case-control study. *BMJ (Clinical research ed.)*, 374, 2021b. doi: 10.1136/bmj.n2015. URL <https://www.bmj.com/content/374/bmj.n2015>.
- D. A. Rasmussen, E. M. Volz, and K. Koelle. Phylodynamic Inference for Structured Epidemiological Models. *PLoS Computational Biology*, 10(4):e1003570, Apr. 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003570. URL <https://dx.plos.org/10.1371/journal.pcbi.1003570>.

- O. Ratmann, M. K. Grabowski, M. Hall, T. Golubchik, C. Wymant, L. Abeler-Dörner, D. Bonsall, A. Hoppe, A. L. Brown, T. de Oliveira, A. Gall, P. Kellam, D. Pillay, J. Kagaayi, G. Kigozi, T. C. Quinn, M. J. Wawer, O. Laeyendecker, D. Serwadda, R. H. Gray, and C. Fraser. Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. *Nature Communications*, 10(1):1411, Mar. 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09139-4. URL <https://www.nature.com/articles/s41467-019-09139-4>.
- O. Ratmann, J. Kagaayi, M. Hall, T. Golubchick, G. Kigozi, X. Xi, C. Wymant, G. Nakigozi, L. Abeler-Dörner, D. Bonsall, A. Gall, A. Hoppe, P. Kellam, J. Bazaale, S. Kalibbala, O. Laeyendecker, J. Lessler, F. Nalugoda, L. W. Chang, T. de Oliveira, D. Pillay, T. C. Quinn, S. J. Reynolds, S. E. F. Spencer, R. Ssekubugu, D. Serwadda, M. J. Wawer, R. H. Gray, C. Fraser, M. K. Grabowski, H. Ayles, R. Bowden, V. Calvez, M. Cohen, A. Dennis, M. Essex, S. Fidler, D. Frampton, R. Hayes, J. Herbeck, P. Kaleebu, C. Kityo, J. Lingappa, V. Novitsky, N. Paton, A. Rambaut, J. Seeley, D. Ssemwanga, F. Tanser, T. Lutalo, R. Galiwango, F. Makumbi, N. K. Sewankambo, D. Nabukalu, A. Ndyababo, J. Ssekasanvu, H. Nakawooya, J. Nakukumba, G. N. Kigozi, B. S. Nantume, N. Resty, J. Kambasu, M. Nalugemwa, R. Nakabuye, L. Ssebanobe, J. Nankinga, A. Kayiira, G. Nanfuka, R. Ahimbisibwe, S. Tomusange, R. M. Galiwango, M. Nakalanzi, J. O. Otobi, D. Ankunda, J. L. Ssembatya, J. B. Ssemanda, E. Kato, R. Kairania, A. Kisakyé, J. Batte, J. Ludigo, A. Nampijja, S. Watya, K. Nehemia, S. M. Anyokot, J. Mwinike, G. Kibumba, P. Ssebowa, G. Mondo, F. Wasswa, A. Nantongo, R. Kakembo, J. Galiwango, G. Ssemango, A. D. Redd, J. Santelli, C. E. Kennedy, J. Wagman, and A. Tobian. Quantifying HIV transmission flow between high-prevalence hotspots and surrounding communities: A population-based study in Rakai, Uganda. *The Lancet HIV*, 7(3): e173–e183, Mar. 2020. ISSN 2352-3018. doi: 10.1016/S2352-3018(19)30378-9. URL <https://www.sciencedirect.com/science/article/pii/S2352301819303789>.
- G. Reniers, E. Slaymaker, J. Nakiyingi-Miiro, C. Nyamukapa, A. C. Crampin, K. Herbst, M. Urassa, F. Otieno, S. Gregson, M. Sewe, D. Michael, T. Lutalo, V. Hosegood, I. Kasamba, A. Price, D. Nabukalu, E. Mclean, and B. Zaba. Mortality trends in the era of antiretroviral therapy: Evidence from the Network for Analysing Longitudinal Population based HIV/AIDS data on Africa (ALPHA). *AIDS*, 28(Supplement 4): S533–S542, Nov. 2014. ISSN 0269-9370. doi: 10.1097/QAD.0000000000000496. URL <https://journals.lww.com/00002030-201411004-00015>.
- G. Riutort-Mayol, P.-C. Bürkner, M. R. Andersen, A. Solin, and A. Vehtari. Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. *Statistics and Computing*, 33(1):17, Dec. 2022. ISSN 1573-1375. doi: 10.1007/s11222-022-10167-2. URL <https://doi.org/10.1007/s11222-022-10167-2>.

- Christian. Robert and George. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York, New York, NY, 2nd ed. 2004. edition, 2004. ISBN 1-4757-4145-6.
- R. Rocha, R. Atun, A. Massuda, B. Rache, P. Spinola, L. Nunes, M. Lago, and M. C. Castro. Effect of socioeconomic inequalities and vulnerabilities on health-system preparedness and response to COVID-19 in Brazil: A comprehensive analysis. *The Lancet Global Health*, 9:782–792, 2021. ISSN 2214-109X. doi: 10.1016/S2214-109X(21)00081-4. URL <https://www.sciencedirect.com/science/article/pii/S2214109X21000814>.
- T. A. H. Rocha, N. C. da Silva, A. C. Q. Barbosa, P. V. Amaral, E. Thumé, J. V. Rocha, V. Alvares, and L. A. Facchini. National registry of health facilities: Data reliability evidence. *Ciencia & Saude Coletiva*, 23:229–240, 2018. doi: 10.1590/1413-81232018231.16672015.
- V. Romano, M. Ancillotti, D. Mascalzoni, and R. Biasiotto. Italians locked down: People's responses to early COVID-19 pandemic public health measures. *Humanities and Social Sciences Communications*, 9(1):1–9, Sept. 2022. ISSN 2662-9992. doi: 10.1057/s41599-022-01358-3. URL <https://www.nature.com/articles/s41599-022-01358-3>.
- E. Romero-Severson, H. Skar, I. Bulla, J. Albert, and T. Leitner. Timing and Order of Transmission Events Is Not Directly Reflected in a Pathogen Phylogeny. *Molecular Biology and Evolution*, 31(9):2472–2482, Sept. 2014. ISSN 0737-4038. doi: 10.1093/molbev/msu179. URL <https://doi.org/10.1093/molbev/msu179>.
- E. O. Romero-Severson, I. Bulla, and T. Leitner. Phylogenetically resolving epidemiologic linkage. *Proceedings of the National Academy of Sciences*, 113(10):2690–2695, Mar. 2016. doi: 10.1073/pnas.1522930113. URL <https://www.pnas.org/doi/full/10.1073/pnas.1522930113>.
- E. Rösli, S. Bozkurt, and T. Hernandez-Boussard. Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model. *Scientific Data*, 9(1):24, Jan. 2022. ISSN 2052-4463. doi: 10.1038/s41597-021-01110-7. URL <https://www.nature.com/articles/s41597-021-01110-7>.
- Rothe Camilla, Schunk Mirjam, Sothmann Peter, Bretzel Gisela, Froeschl Guenter, Wallrauch Claudia, Zimmer Thorbjörn, Thiel Verena, Janke Christian, Guggemos Wolfgang, Seilmaier Michael, Drosten Christian, Vollmar Patrick, Zwirgmaier Katrin, Zange Sabine, Wölfel Roman, and Hoelscher Michael. Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany. *New England Journal of Medicine*, 382(10):970–971, Mar. 2020. doi: 10.1056/NEJMc2001468. URL <https://www.nejm.org/doi/full/10.1056/NEJMc2001468>.

- M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Computation*, 14(1):21–41, Jan. 2002. ISSN 0899-7667. doi: 10.1162/089976602753284446. URL <https://doi.org/10.1162/089976602753284446>.
- D. Salnikov, K. Michalewicz, and D. Leonte. Lifted Coefficient of Determination: Fast model-free prediction intervals and likelihood-free model comparison, 2024. URL <https://arxiv.org/abs/2410.08958>.
- F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, Dec. 1977. doi: 10.1073/pnas.74.12.5463. URL <https://www.pnas.org/doi/abs/10.1073/pnas.74.12.5463>.
- J. A. Scott, A. Gandy, S. Mishra, J. Unwin, S. Flaxman, and S. Bhatt. Epidemia: Modeling of epidemics using hierarchical bayesian models, 2020. URL <https://imperialcollegelondon.github.io/epidemia/>.
- E. Semenova, Y. Xu, A. Howes, T. Rashid, S. Bhatt, S. Mishra, and S. Flaxman. Prior-VAE: Encoding spatial priors with variational autoencoders for small-area estimation. *Journal of The Royal Society Interface*, 19(191):20220094, June 2022. ISSN 1742-5662. doi: 10.1098/rsif.2022.0094. URL <https://royalsocietypublishing.org/doi/10.1098/rsif.2022.0094>.
- L. Shami and T. Lazebnik. Economic aspects of the detection of new strains in a multi-strain epidemiological–mathematical model. *Chaos, Solitons & Fractals*, 165:112823, Dec. 2022. ISSN 09600779. doi: 10.1016/j.chaos.2022.112823. URL <https://linkinghub.elsevier.com/retrieve/pii/S0960077922010025>.
- P. M. Sharp and B. H. Hahn. Origins of HIV and the AIDS Pandemic. *Cold Spring Harbor Perspectives in Medicine*, 1(1):a006841, Jan. 2011. ISSN , 2157-1422. doi: 10.1101/cshperspect.a006841. URL <http://perspectivesinmedicine.cshlp.org/content/1/1/a006841>.
- A. Sheikh, J. McMenamin, B. Taylor, and C. Robertson. SARS-CoV-2 Delta VOC in Scotland: Demographics, risk of hospital admission, and vaccine effectiveness. *The Lancet*, 397(10293):2461–2462, 2021. doi: 10.1016/S0140-6736(21)01358-1.
- Y. Shu and J. McCauley. GISAID: Global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance*, 22, 2017. doi: 10.2807/1560-7917.ES.2017.22.13.30494.

- N. Simelela, W. D. F. Venter, Y. Pillay, and P. Barron. A Political and Social History of HIV in South Africa. *Current HIV/AIDS Reports*, 12(2):256–261, June 2015. ISSN 1548-3576. doi: 10.1007/s11904-015-0259-7. URL <https://doi.org/10.1007/s11904-015-0259-7>.
- H. Skar, J. Albert, and T. Leitner. Time from HIV-1 seroconversion to AIDS and death before widespread use of highly-active antiretroviral therapy: A collaborative re-analysis. *The Lancet*, 355(9210):1131–1137, Apr. 2000. ISSN 0140-6736. doi: 10.1016/S0140-6736(00)02061-4. URL <https://www.sciencedirect.com/science/article/pii/S0140673600020614>.
- M. L. Smith and M. W. Hahn. Phylogenetic inference using generative adversarial networks. *Bioinformatics*, 39(9):btad543, Sept. 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad543. URL <https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btad543/7260504>.
- S. S. Solomon, S. H. Mehta, A. M. McFall, A. K. Srikrishnan, S. Saravanan, O. Laeyendecker, P. Balakrishnan, D. D. Celentano, S. Solomon, and G. M. Lucas. Community viral load, antiretroviral therapy coverage, and HIV incidence in India: A cross-sectional, comparative study. *The Lancet HIV*, 3(4):e183–e190, Apr. 2016. ISSN 23523018. doi: 10.1016/S2352-3018(16)00019-9. URL <https://linkinghub.elsevier.com/retrieve/pii/S2352301816000199>.
- J. L. Spouge. A comprehensive estimation of country-level basic reproduction numbers R_0 for COVID-19: Regime regression can automatically estimate the end of the exponential phase in epidemic data. *PLOS ONE*, 16(7):e0254145, July 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0254145. URL <https://dx.plos.org/10.1371/journal.pone.0254145>.
- V. Ssempijja, L. W. Chang, G. Nakigozi, A. Ndyababo, T. C. Quinn, F. Cobelens, M. Wawer, R. Gray, D. Serwadda, and S. J. Reynolds. Results of Early Virologic Monitoring May Facilitate Differentiated Care Monitoring Strategies for Clients on ART, Rakai, Uganda. *Open Forum Infectious Diseases*, 5(10):ofy212, Oct. 2018. ISSN 2328-8957. doi: 10.1093/ofid/ofy212. URL <https://academic.oup.com/ofid/article/doi/10.1093/ofid/ofy212/5123901>.
- V. Ssempijja, M. Nason, G. Nakigozi, A. Ndyababo, R. Gray, M. Wawer, L. W. Chang, E. Gabriel, T. C. Quinn, D. Serwadda, and S. J. Reynolds. Adaptive Viral Load Monitoring Frequency to Facilitate Differentiated Care: A Modeling Study From Rakai, Uganda. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 71(4):1017–1021, Aug. 2020. ISSN 1537-6591. doi: 10.1093/cid/ciz880.

- V. Ssempijja, G. Nakigozi, R. Ssekubugu, J. Kagaayi, G. Kigozi, F. Nalugoda, B. Nantume, J. Batte, G. Kigozi, P. T. Yeh, H. Nakawooya, D. Serwadda, T. C. Quinn, R. H. Gray, M. J. Wawer, K. M. Grabowski, L. W. Chang, A. Van't Hoog, F. Cobelens, and S. J. Reynolds. High Rates of Pre-exposure Prophylaxis Eligibility and Associated HIV Incidence in a Population With a Generalized HIV Epidemic in Rakai, Uganda. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 90(3):291–299, July 2022. ISSN 1525-4135. doi: 10.1097/QAI.0000000000002946. URL <https://journals.lww.com/10.1097/QAI.0000000000002946>.
- Stan Development Team. RStan: The R interface to Stan, 2024. URL <https://mc-stan.org/>.
- O. T. Stirrup and D. T. Dunn. Estimation of delay to diagnosis and incidence in HIV using indirect evidence of infection dates. *BMC Medical Research Methodology*, 18(1):65, Dec. 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0522-x. URL <https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0522-x>.
- E. Štrumbelj, A. Bouchard-Côté, J. Corander, A. Gelman, H. Rue, L. Murray, H. Pesonen, M. Plummer, and A. Vehtari. Past, Present and Future of Software for Bayesian Inference. *Statistical Science*, 39(1), Feb. 2024. ISSN 0883-4237. doi: 10.1214/23-STS907. URL <https://projecteuclid.org/journals/statistical-science/volume-39/issue-1/Past-Present-and-Future-of-Software-for-Bayesian-Inference/10.1214/23-STS907.full>.
- F. Tanser, A. Vandormael, D. Cuadros, A. N. Phillips, T. De Oliveira, A. Tomita, T. Bärnighausen, and D. Pillay. Effect of population viral load on prospective HIV incidence in a hyperendemic rural African community. *Science Translational Medicine*, 9(420):eaam8012, Dec. 2017. ISSN 1946-6234, 1946-6242. doi: 10.1126/scitranslmed.aam8012. URL <https://www.science.org/doi/10.1126/scitranslmed.aam8012>.
- H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, D. Doolabh, S. Pillay, E. J. San, N. Msomi, K. Mlisana, A. von Gottberg, S. Walaza, M. Allam, A. Ismail, T. Mohale, A. J. Glass, S. Engelbrecht, G. Van Zyl, W. Preiser, F. Petruccione, A. Sigal, D. Hardie, G. Marais, N.-y. Hsiao, S. Korsman, M.-A. Davies, L. Tyers, I. Mudau, D. York, C. Maslo, D. Goedhals, S. Abrahams, O. Laguda-Akingba, A. Alisoltani-Dehkordi, A. Godzik, C. K. Wibmer, B. T. Sewell, J. Lourenço, L. C. J. Alcantara, S. L. Kosakovsky Pond, S. Weaver, D. Martin, R. J. Lessells, J. N. Bhiman, C. Williamson, and T. de Oliveira. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 592(7854):438–443, Apr. 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03402-9. URL <https://www.nature.com/articles/s41586-021-03402-9>.

- R. N. Thompson, J. E. Stockwin, R. D. van Gaalen, J. A. Polonsky, Z. N. Kamvar, P. A. Demarsh, E. Dahlqwist, S. Li, E. Miguel, T. Jombart, J. Lessler, S. Cauchemez, and A. Cori. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics*, 29:100356, Dec. 2019. ISSN 1755-4365. doi: 10.1016/j.epidem.2019.100356. URL <https://www.sciencedirect.com/science/article/pii/S1755436519300350>.
- R. N. Thompson, T. D. Hollingsworth, V. Isham, D. Arribas-Bel, B. Ashby, T. Britton, P. Challenor, L. H. K. Chappell, H. Clapham, N. J. Cunniffe, A. P. Dawid, C. A. Donnelly, R. M. Eggo, S. Funk, N. Gilbert, P. Glendinning, J. R. Gog, W. S. Hart, H. Heesterbeek, T. House, M. Keeling, I. Z. Kiss, M. E. Kretzschmar, A. L. Lloyd, E. S. McBryde, J. M. McCaw, T. J. McKinley, J. C. Miller, M. Morris, P. D. O'Neill, K. V. Parag, C. A. B. Pearson, L. Pellis, J. R. C. Pulliam, J. V. Ross, G. S. Tomba, B. W. Silverman, C. J. Struchiner, M. J. Tildesley, P. Trapman, C. R. Webb, D. Mollison, and O. Restif. Key questions for modelling COVID-19 exit strategies. *Proceedings of the Royal Society B: Biological Sciences*, 287(1932):20201405, Aug. 2020. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.2020.1405. URL <https://royalsocietypublishing.org/doi/10.1098/rspb.2020.1405>.
- K. A. Twohig, T. Nyberg, A. Zaidi, S. Thelwall, M. A. Sinnathamby, S. Aliabadi, S. R. Seaman, R. J. Harris, R. Hope, J. Lopez-Bernal, E. Gallagher, A. Charlett, D. De Angelis, A. M. Presanis, G. Dabrera, C. Koshy, A. Ash, E. Wise, N. Moore, M. Mori, N. Cortes, J. Lynch, S. Kidd, D. Fairley, T. Curran, J. McKenna, H. Adams, C. Fraser, T. Golubchik, D. Bonsall, M. Hassan-Ibrahim, C. Malone, B. Cogger, M. Wantoch, N. Reynolds, B. Warne, J. Maksimovic, K. Spellman, K. McCluggage, M. John, R. Beer, S. Afifi, S. Morgan, A. Marchbank, A. Price, C. Kitchen, H. Gulliver, I. Merrick, J. Southgate, M. Guest, R. Munn, T. Workman, T. Connor, W. Fuller, C. Bresner, L. Snell, A. Patel, T. Charalampous, G. Nebbia, R. Batra, J. Edgeworth, S. Robson, A. Beckett, D. Aanensen, A. Underwood, C. Yeats, K. Abudahab, B. Taylor, M. Menegazzo, G. Clark, W. Smith, M. Khakh, V. Fleming, M. Lister, H. Howson-Wells, L. Berry, T. Boswell, A. Joseph, I. Willingham, C. Jones, C. Holmes, P. Bird, T. Helmer, K. Fallon, J. Tang, V. Raviprakash, S. Campbell, N. Sheriff, V. Blakey, L.-A. Williams, M. Loose, N. Holmes, C. Moore, M. Carlile, V. Wright, F. Sang, J. Debebe, F. Coll, A. Signell, G. Betancor, H. Wilson, S. Eldirdiri, A. Kenyon, T. Davis, O. Pybus, L. du Plessis, A. Zarebski, J. Raghwan, M. Kraemer, S. Francois, S. Attwood, T. Vasylyeva, M. Escalera Zamudio, B. Gutierrez, M. E. Torok, W. Hamilton, I. Goodfellow, G. Hall, A. Jahun, Y. Chaudhry, M. Hosmillo, M. Pinckert, I. Georgana, S. Moses, H. Lowe, L. Bedford, J. Moore, S. Stonehouse, C. Fisher, A. Awan, J. BoYes, J. Breuer, K. Harris, J. Brown, D. Shah, L. Atkinson, J. Lee, N. Storey, F. Flaviani, A. Alcolea-Medina, R. Williams, G. Vernet, M. Chapman, L. Levett, J. Heaney, W. Chatterton, M. Pusok, L. Xu-McCrae, D. Smith, M. Bashton, G. Young, A. Holmes, P. Randell, A. Cox, P. Madona, F. Bolt, J. Price, S. Mookerjee, M. Ragonnet-

Cronin, F. F. Nascimento, D. Jorgensen, I. Siveroni, R. Johnson, O. Boyd, L. Geidelberg, E. Volz, A. Rowan, G. Taylor, K. Smollett, N. Loman, J. Quick, C. McMurray, J. Stockton, S. Nicholls, W. Rowe, R. Poplawski, A. McNally, R. Martinez Nunez, J. Mason, T. Robinson, E. O'Toole, J. Watts, C. Breen, A. Cowell, G. Sluga, N. Machin, S. Ahmad, R. George, F. Halstead, V. Sivaprakasam, W. Hogsden, C. Illingworth, C. Jackson, E. Thomson, J. Shepherd, P. Asamaphan, M. Niebel, K. Li, R. Shah, N. Jesudason, L. Tong, A. Broos, D. Mair, J. Nichols, S. Carmichael, K. Nomikou, E. Aranday-Cortes, N. Johnson, I. Starinskij, A. da Silva Filipe, D. Robertson, R. Orton, J. Hughes, S. Vattipally, J. Singer, S. Nickbakhsh, A. Hale, L. Macfarlane-Smith, K. Harper, H. Carden, Y. Taha, B. Payne, S. Burton-Fanning, S. Waugh, J. Collins, G. Eltringham, S. Rushton, S. O'Brien, A. Bradley, A. Maclean, G. Mollett, R. Blacow, K. Templeton, M. McHugh, R. Dewar, E. Wastenge, S. Dervisevic, R. Stanley, E. Meader, L. Coupland, L. Smith, C. Graham, E. Barton, D. Padgett, G. Scott, E. Swindells, J. Greenaway, A. Nelson, C. McCann, W. Yew, M. Andersson, T. Peto, A. Justice, D. Eyre, D. Crook, T. Sloan, N. Duckworth, S. Walsh, A. Chauhan, S. Glaysher, K. Bicknell, S. Wyllie, S. Elliott, A. Lloyd, R. Impey, N. Levene, L. Monaghan, D. Bradley, T. Wyatt, E. Allara, C. Pearson, H. Osman, A. Bosworth, E. Robinson, P. Muir, I. Vipond, R. Hopes, H. Pymont, S. Hutchings, M. Curran, S. Parmar, A. Lackenby, T. Mbisa, S. Platt, S. Miah, D. Bibby, C. Manso, J. Hubb, M. Chand, G. Dabrera, M. Ramsay, D. Bradshaw, A. Thornton, R. Myers, U. Schaefer, N. Groves, E. Gallagher, D. Lee, D. Williams, N. Ellaby, I. Harrison, H. Hartman, N. Manesis, V. Patel, C. Bishop, V. Chalker, J. Ledesma, K. Twohig, M. Holden, S. Shaaban, A. Birchley, A. Adams, A. Davies, A. Gaskin, A. Plimmer, B. Gatica-Wilcox, C. McKerr, C. Moore, C. Williams, D. Heyburn, E. De Lacy, E. Hilvers, F. Downing, G. Shankar, H. Jones, H. Asad, J. Coombes, J. Watkins, J. Evans, L. Fina, L. Gifford, L. Gilbert, L. Graham, M. Perry, M. Morgan, M. Bull, M. Cronin, N. Pacchiarini, N. Craine, R. Jones, R. Howe, S. Corden, S. Rey, S. Kumziene-SummerhaYes, S. Taylor, S. Cottrell, S. Jones, S. Edwards, J. O'Grady, A. Page, A. Mather, D. Baker, S. Rudder, A. Aydin, G. Kay, A. Trotter, N.-F. Alikhan, L. de Oliveira Martins, T. Le-Viet, L. Meadows, A. Casey, L. Ratcliffe, D. Simpson, Z. Molnar, T. Thompson, E. Acheson, J. Masoli, B. Knight, S. Ellard, C. Auckland, C. Jones, T. Mahungu, D. Irish-Tavares, T. Haque, J. Hart, E. Witele, M. Fenton, A. Dadrah, A. Symmonds, T. Saluja, Y. Bourgeois, G. Scarlett, K. Loveson, S. Goudarzi, C. Fearn, K. Cook, H. Dent, H. Paul, D. Partridge, M. Raza, C. Evans, K. Johnson, S. Liggett, P. Baker, S. Bonner, S. Essex, R. Lyons, K. Saeed, A. Mahanama, B. Samaraweera, S. Silveira, E. Pelosi, E. Wilson-Davies, R. Williams, M. Kristiansen, S. Roy, C. Williams, M. Cotic, N. Bayzid, A. Westhorpe, J. Hartley, R. Jannoo, H. Lowe, A. Karamani, L. Ensell, J. Prieto, S. Jeremiah, D. Grammatopoulos, S. Pandey, L. Berry, K. Jones, A. Richter, A. Beggs, A. Best, B. Percival, J. Mirza, O. Megram, M. Mayhew, L. Crawford, F. Ashcroft, E. Moles-Garcia, N. Cumley, C. Smith, G. Bucca, A. Hesketh, B. Blane, S. Girgis, D. Leek, S. Sridhar, S. Forrest,

C. Cormie, H. Gill, J. Dias, E. Higginson, M. Maes, J. Young, L. Kermack, R. Gupta, C. Ludden, S. Peacock, S. Palmer, C. Churcher, N. Hadjirin, A. Carabelli, E. Brooks, K. Smith, K. Galai, G. McManus, C. Ruis, R. Davidson, A. Rambaut, T. Williams, C. Balcazar, M. Gallagher, Á. O'Toole, S. Rooke, V. Hill, K. Williamson, T. Stanton, S. Michell, C. Bewshea, B. Temperton, M. Michelsen, J. Warwick-Dugdale, R. Manley, A. Farbos, J. Harrison, C. Sambles, D. Studholme, A. Jeffries, L. Jackson, A. Darby, J. Hiscox, S. Paterson, M. Iturriza-Gomara, K. Jackson, A. Lucaci, E. Vamos, M. Hughes, L. Rainbow, R. Eccles, C. Nelson, M. Whitehead, L. Turtle, S. Haldenby, R. Gregory, M. Gemmell, C. Wierzbicki, H. Webster, T. de Silva, N. Smith, A. Angyal, B. Lindsey, D. Groves, L. Green, D. Wang, T. Freeman, M. Parker, A. Keeley, P. Parsons, R. Tucker, R. Brown, M. Wyles, M. Whiteley, P. Zhang, M. Gallis, S. Louka, C. Constantinidou, M. Unnikrishnan, S. Ott, J. Cheng, H. Bridgewater, L. Frost, G. Taylor-Joyce, R. Stark, L. Baxter, M. Alam, P. Brown, D. Aggarwal, A. Cerda, T. Merrill, R. Wilson, P. McClure, J. Chappell, T. Tsoleridis, J. Ball, D. Buck, J. Todd, A. Green, A. Trebes, G. MacIntyre-Cockett, M. de Cesare, A. Alderton, R. Amato, C. Ariani, M. Beale, C. Beaver, K. Bellis, E. Betteridge, J. Bonfield, J. Danesh, M. Dorman, E. Drury, B. Farr, L. Foulser, S. Goncalves, S. Goodwin, M. Gourtovaia, E. Harrison, D. Jackson, D. Jamroz, I. Johnston, L. Kane, S. Kay, J.-P. Keatley, D. Kwiatkowski, C. Langford, M. Lawniczak, L. Letchford, R. Livett, S. Lo, I. Martincorena, S. McGuigan, R. Nelson, S. Palmer, N. Park, M. Patel, L. Prestwood, C. Puethe, M. Quail, S. Rajatileka, C. Scott, L. Shirley, J. Sillitoe, M. Spencer Chapman, S. Thurston, G. Tonkin-Hill, D. Weldon, D. Rajan, I. Bronner, L. Aigrain, N. Redshaw, S. Lensing, R. Davies, A. Whitwham, J. Liddle, K. Lewis, J. Tovar-Corona, S. Leonard, J. Durham, A. Bassett, S. McCarthy, R. Moll, K. James, K. Oliver, A. Makunin, J. Barrett, and R. Gunson. Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: A cohort study. *The Lancet Infectious Diseases*, 22 (1):35–42, Jan. 2022. ISSN 1473-3099. doi: 10.1016/S1473-3099(21)00475-8. URL <https://www.sciencedirect.com/science/article/pii/S1473309921004758>.

Uganda Ministry of Health. Uganda Summary Sheet 2020-2021. Technical report, Aug. 2022. URL <https://phia.icap.columbia.edu/wp-content/uploads/2022/08/UPHIA-Summary-Sheet-2020.pdf>.

UNAIDS. Understanding Fast-Track: Accelerating action to end the AIDS epidemic by 2030. Technical report, 2015. URL https://www.unaids.org/sites/default/files/media_asset/201506_JC2743_Understanding_FastTrack_en.pdf.

UNAIDS. Global AIDS Strategy 2021–2026., 2020. URL https://www.unaids.org/sites/default/files/media_asset/global-AIDS-strategy-2021-2026_en.pdf.

- United Nations. Sustainable development goals. *United Nations, S-1018, New York, NY, 10017*, 2015.
- Gabriel. Valiente. Algorithms on trees and graphs with python code. In *Algorithms on Trees and Graphs with Python Code*, Texts in Computer Science. Springer International Publishing, Cham, 2nd ed. 2021. edition, 2021. ISBN 978-3-030-81885-2.
- A. Van Sighem, F. Nakagawa, D. De Angelis, C. Quinten, D. Bezemer, E. O. De Coul, M. Egger, F. De Wolf, C. Fraser, and A. Phillips. Estimating HIV Incidence, Time to Diagnosis, and the Undiagnosed HIV Epidemic Using Routine Surveillance Data:. *Epidemiology*, 26(5):653–660, Sept. 2015. ISSN 1044-3983. doi: 10.1097/EDE.0000000000000324. URL <http://journals.lww.com/00001648-201509000-00005>.
- A. Vehtari, A. Gelman, and J. Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, Sept. 2017. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-016-9696-4. URL <http://link.springer.com/10.1007/s11222-016-9696-4>.
- A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. Rank-Normalization, Folding, and Localization: An Improved $R^{\hat{}}$ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2):667–718, June 2021. ISSN 1936-0975, 1931-6690. doi: 10.1214/20-BA1221. URL <https://projecteuclid.org/journals/bayesian-analysis/volume-16/issue-2/Rank-Normalization-Folding-and-Localization--An-Improved-R-%cb%86-for/10.1214/20-BA1221.full>.
- L. Veiga E Silva, M. D. P. de Andrade Abi Harb, A. M. Teixeira Barbosa Dos Santos, C. A. de Mattos Teixeira, V. H. Macedo Gomes, E. H. Silva Cardoso, M. S da Silva, N. L. Vijaykumar, S. Venâncio Carvalho, A. Ponce de Leon Ferreira de Carvalho, and C. R. Lisboa Frances. COVID-19 Mortality underreporting in Brazil: Analysis of data from government internet portals. *Journal of Medical Internet Research*, 22, Aug. 2020.
- R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. T. Walker, H. Fu, A. Dighe, J. T. Griffin, M. Baguelin, S. Bhatia, A. Boonyasiri, A. Cori, Z. Cucunubá, R. FitzJohn, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, D. Laydon, G. Nedjati-Gilani, S. Riley, S. van Elsland, E. Volz, H. Wang, Y. Wang, X. Xi, C. A. Donnelly, A. C. Ghani, and N. M. Ferguson. Estimates of the severity of coronavirus disease 2019: A model-based analysis. *The Lancet Infectious Diseases*, 20(6):669–677, June 2020. ISSN 1473-3099, 1474-4457. doi: 10.1016/S1473-3099(20)30243-7. URL [https://www.thelancet.com/article/S1473-3099\(20\)30243-7/fulltext](https://www.thelancet.com/article/S1473-3099(20)30243-7/fulltext).

- R. Viana, S. Moyo, D. G. Amoako, H. Tegally, C. Scheepers, C. L. Althaus, U. J. Anyaneji, P. A. Bester, M. F. Boni, M. Chand, W. T. Choga, R. Colquhoun, M. Davids, K. De-
forche, D. Doolabh, L. du Plessis, S. Engelbrecht, J. Everatt, J. Giandhari, M. Gio-
vanetti, D. Hardie, V. Hill, N.-Y. Hsiao, A. Iranzadeh, A. Ismail, C. Joseph, R. Joseph,
L. Koopile, S. L. Kosakovsky Pond, M. U. G. Kraemer, L. Kuate-Lere, O. Laguda-
Akingba, O. Lesetedi-Mafoko, R. J. Lessells, S. Lockman, A. G. Lucaci, A. Maharaj,
B. Mahlangu, T. Maponga, K. Mahlakwane, Z. Makatini, G. Marais, D. Maruapula,
K. Masupu, M. Matshaba, S. Mayaphi, N. Mbhele, M. B. Mbulawa, A. Mendes,
K. Mlisana, A. Mnguni, T. Mohale, M. Moir, K. Moruisi, M. Mosepele, G. Motsatsi,
M. S. Motswaledi, T. Mphoyakgosi, N. Msomi, P. N. Mwangi, Y. Naidoo, N. Ntuli,
M. Nyaga, L. Olubayo, S. Pillay, B. Radibe, Y. Ramphal, U. Ramphal, J. E. San,
L. Scott, R. Shapiro, L. Singh, P. Smith-Lawrence, W. Stevens, A. Strydom, K. Sub-
ramoney, N. Tebeila, D. Tshiabuila, J. Tsui, S. van Wyk, S. Weaver, C. K. Wibmer,
E. Wilkinson, N. Wolter, A. E. Zarebski, B. Zuze, D. Goedhals, W. Preiser, F. Treurnicht,
M. Venter, C. Williamson, O. G. Pybus, J. Bhiman, A. Glass, D. P. Martin, A. Ram-
baut, S. Gaseitsiwe, A. von Gottberg, and T. de Oliveira. Rapid epidemic expan-
sion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature*, 603(7902):
679–686, Mar. 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04411-y. URL
<https://www.nature.com/articles/s41586-022-04411-y>.
- E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J.
Laydon, G. Dabrera, Á. O’Toole, R. Amato, M. Ragonnet-Cronin, I. Harrison, B. Jackson,
C. V. Ariani, O. Boyd, N. J. Loman, J. T. McCrone, S. Gonçalves, D. Jorgensen, R. Myers,
V. Hill, D. K. Jackson, K. Gaythorpe, N. Groves, J. Sillitoe, D. P. Kwiatkowski, COVID-19
Genomics UK (COG-UK) consortium, S. Flaxman, O. Ratmann, S. Bhatt, S. Hopkins,
A. Gandy, A. Rambaut, and N. M. Ferguson. Assessing transmissibility of SARS-CoV-2
lineage B.1.1.7 in england. *Nature*, 593:266–269, May 2021.
- E. M. Volz. Complex Population Dynamics and the Coalescent Under Neutrality. *Ge-
netics*, 190(1):187–201, Jan. 2012. ISSN 1943-2631. doi: 10.1534/genetics.111.134627.
URL <https://doi.org/10.1534/genetics.111.134627>.
- E. M. Volz and S. D. W. Frost. Inferring the Source of Transmission with Phylogenetic
Data. *PLOS Computational Biology*, 9(12):e1003397, Dec. 2013. ISSN 1553-7358. doi:
10.1371/journal.pcbi.1003397. URL [https://journals.plos.org/ploscompbiol/article?
id=10.1371/journal.pcbi.1003397](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003397).
- E. M. Volz, S. L. Kosakovsky Pond, M. J. Ward, A. J. Leigh Brown, and S. D. W. Frost.
Phylogenetics of Infectious Disease Epidemics. *Genetics*, 183(4):1421–1430, Dec.
2009. ISSN 1943-2631. doi: 10.1534/genetics.109.106021. URL [https://doi.org/10.
1534/genetics.109.106021](https://doi.org/10.1534/genetics.109.106021).

- M. Voysey, S. A. C. Clemens, S. A. Madhi, L. Y. Weckx, P. M. Folegatti, P. K. Aley, B. Angus, V. L. Baillie, S. L. Barnabas, Q. E. Bhorat, et al. Single-dose administration and the influence of the timing of the booster dose on immunogenicity and efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine: A pooled analysis of four randomised trials. *The Lancet*, 397:881–891, 2021. doi: 10.1016/S0140-6736(21)00432-3.
- J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604, Feb. 2007. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.2006.3754. URL <https://royalsocietypublishing.org/doi/10.1098/rspb.2006.3754>.
- J. Wallinga and P. Teunis. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. *American Journal of Epidemiology*, 160(6):509–516, Sept. 2004. ISSN 0002-9262. doi: 10.1093/aje/kwh255. URL <https://doi.org/10.1093/aje/kwh255>.
- G. T. Wassie, A. G. Azene, G. M. Bantie, G. Dessie, and A. M. Aragaw. Incubation Period of Severe Acute Respiratory Syndrome Novel Coronavirus 2 that Causes Coronavirus Disease 2019: A Systematic Review and Meta-Analysis. *Current Therapeutic Research*, 93:100607, 2020. ISSN 0011393X. doi: 10.1016/j.curtheres.2020.100607. URL <https://linkinghub.elsevier.com/retrieve/pii/S0011393X20300333>.
- M. J. Wawer, N. K. Sewankambo, D. Serwadda, T. C. Quinn, N. Kiwanuka, C. Li, T. Lutalo, F. Nalugoda, C. A. Gaydos, L. H. Moulton, S. Ahmed, R. H. Gray, L. A. Paxton, F. Wabwire-Mangen, and M. O. Meehan. Control of sexually transmitted diseases for AIDS prevention in Uganda: A randomised community trial. *The Lancet*, 353(9152):525–535, Feb. 1999. ISSN 01406736. doi: 10.1016/S0140-6736(98)06439-3. URL <https://linkinghub.elsevier.com/retrieve/pii/S0140673698064393>.
- F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, and Y.-Z. Zhang. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269, Mar. 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2008-3. URL <https://www.nature.com/articles/s41586-020-2008-3>.
- J. T. Wu, S. Mei, S. Luo, K. Leung, D. Liu, Q. Lv, J. Liu, Y. Li, K. Prem, M. Jit, J. Weng, T. Feng, X. Zheng, and G. M. Leung. A global assessment of the impact of school closure in reducing COVID-19 spread. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 380(2214):20210124, Jan. 2022. ISSN 1364-503X, 1471-2962. doi: 10.1098/rsta.2021.0124. URL <https://royalsocietypublishing.org/doi/10.1098/rsta.2021.0124>.

- C. Wymant, F. Blanquart, T. Golubchik, A. Gall, M. Bakker, D. Bezemer, N. J. Croucher, M. Hall, M. Hillebregt, S. H. Ong, O. Ratmann, J. Albert, N. Bannert, J. Fellay, K. Fransen, A. Gourlay, M. K. Grabowski, B. Gunsenheimer-Bartmeyer, H. F. Günthard, P. Kivelä, R. Kouyos, O. Laeyendecker, K. Liitsola, L. Meyer, K. Porter, M. Ristola, A. van Sighem, B. Berkhout, M. Cornelissen, P. Kellam, P. Reiss, C. Fraser, and BEEHIVE Collaboration. Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. *Virus Evolution*, 4(1):vey007, Jan. 2018a. ISSN 2057-1577. doi: 10.1093/ve/vey007. URL <https://doi.org/10.1093/ve/vey007>.
- C. Wymant, M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. De Cesare, A. Gall, M. Cornelissen, C. Fraser, and STOP-HCV Consortium, The Maela Pneumococcal Collaboration, and The BEEHIVE Collaboration. PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. *Molecular Biology and Evolution*, 35(3):719–733, Mar. 2018b. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msx304. URL <https://academic.oup.com/mbe/article/35/3/719/4653772>.
- X. Xi. *Bayesian Methods for Source Attribution Using HIV Deep Sequence Data*. PhD thesis, Imperial College London, 2022.
- X. Xi, S. E. F. Spencer, M. Hall, M. K. Grabowski, J. Kagaayi, and O. Ratmann. Inferring the Sources of HIV Infection in Africa from Deep-Sequence Data with Semi-Parametric Bayesian Poisson Flow Models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(3):517–540, June 2022. ISSN 0035-9254, 1467-9876. doi: 10.1111/rssc.12544. URL <https://academic.oup.com/jrssc/article/71/3/517/7067604>.
- J. Xiao, J. Hu, G. He, T. Liu, M. Kang, Z. Rong, L. Lin, H. Zhong, Q. Huang, A. Deng, W. Zeng, X. Tan, S. Zeng, Z. Zhu, J. Li, D. Gong, D. Wan, S. Chen, L. Guo, Y. Li, Y. Li, L. Sun, W. Liang, T. Song, J. He, and W. Ma. The time-varying transmission dynamics of COVID-19 and synchronous public health interventions in China. *International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases*, 103:617–623, Feb. 2021. ISSN 1878-3511. doi: 10.1016/j.ijid.2020.11.005.
- W. Yamaka, S. Lomwanawong, D. Magel, and P. Maneejuk. Analysis of the Lockdown Effects on the Economy, Environment, and COVID-19 Spread: Lesson Learnt from a Global Pandemic in 2020. *International Journal of Environmental Research and Public Health*, 19(19):12868, Oct. 2022. ISSN 1660-4601. doi: 10.3390/ijerph191912868. URL <https://www.mdpi.com/1660-4601/19/19/12868>.
- Y. Ye, A. Pandey, C. Bawden, D. M. Sumsuzzman, R. Rajput, A. Shoukat, B. H. Singer, S. M. Moghadas, and A. P. Galvani. Integrating artificial intelligence with mechanistic epidemiological modeling: A scoping review of opportunities and challenges.

Nature Communications, 16(1):581, Jan. 2025. ISSN 2041-1723. doi: 10.1038/s41467-024-55461-x. URL <https://www.nature.com/articles/s41467-024-55461-x>.

C.-J. Yu, Z.-X. Wang, Y. Xu, M.-X. Hu, K. Chen, and G. Qin. Assessment of basic reproductive number for COVID-19 at global level: A meta-analysis. *Medicine*, 100(18):e25837, May 2021. ISSN 0025-7974, 1536-5964. doi: 10.1097/MD.00000000000025837. URL <https://journals.lww.com/10.1097/MD.00000000000025837>.

Y. Zhang, C. Wymant, O. Laeyendecker, M. K. Grabowski, M. Hall, S. Hudelson, E. Piwowar-Manning, M. McCauley, T. Gamble, M. C. Hosseinipour, N. Kumarasamy, J. G. Hakim, J. Kumwenda, L. A. Mills, B. R. Santos, B. Grinsztejn, J. H. Pilotto, S. Chariyalertsak, J. Makhema, Y. Q. Chen, M. S. Cohen, C. Fraser, and S. H. Eshleman. Evaluation of Phylogenetic Methods for Inferring the Direction of Human Immunodeficiency Virus (HIV) Transmission: HIV Prevention Trials Network (HPTN) 052. *Clinical Infectious Diseases*, 72(1):30–37, Jan. 2021. ISSN 1058-4838. doi: 10.1093/cid/ciz1247. URL <https://doi.org/10.1093/cid/ciz1247>.

Zhu Na, Zhang Dingyu, Wang Wenling, Li Xingwang, Yang Bo, Song Jingdong, Zhao Xiang, Huang Baoying, Shi Weifeng, Lu Roujian, Niu Peihua, Zhan Faxian, Ma Xuejun, Wang Dayan, Xu Wenbo, Wu Guizhen, Gao George F, and Tan Wenjie. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine*, 382(8):727–733, Feb. 2020. doi: 10.1056/NEJMoa2001017. URL <https://www.nejm.org/doi/full/10.1056/NEJMoa2001017>.

SITOGRAPHY

PNAD COVID-19 - pesquisa nacional por amostra de domicílios COVID-19. 2020.

SRAG 2020 - banco de dados de síndrome respiratória aguda grave, (2020). URL <https://opendatasus.saude.gov.br/dataset/srag-2020>.

SRAG 2021 e 2022- banco de dados de síndrome respiratória aguda grave, (2021). URL <https://opendatasus.saude.gov.br/dataset/srag-2021-e-2022>.

Cadastro nacional de estabelecimentos de saúde, 2021. URL <https://datasus.saude.gov.br/transferencia-de-arquivos/>.

I. Africa. We partner with scientific leaders in Africa for relevant and sustainable vaccine discovery and development, 2024. URL <https://www.iavi.org/africa/iavi-africa/>.

artic-network. Fieldbioinformatics, 2021. URL <https://artic.network/>.

M. Betancourt. Robust Gaussian Process Modeling, Sept. 2020. URL https://betanalpha.github.io/assets/case_studies/gaussian_processes.html.

Brazilian Ministry of Health. National Vaccination Campaign against COVID-19, 2022. URL <https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao>.

Brazilian Ministry of Health. Campanha nacional de vacinação contra COVID-19. 2021. URL <https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao>.

Brazilian Ministry of Labor and Employment. Brazilian Classification of Occupations, 2021. URL <https://www.ocupacoes.com.br/>.

Britannica. World War I - Casualties, Armistice, Legacy | Britannica, July 2024. URL <https://www.britannica.com/event/World-War-I/Killed-wounded-and-missing>.

A. Brizzi. Dealing with missed generations of infections with EpiEstim, 2024. URL https://mrc-ide.github.io/EpiEstim/articles/EpiEstim_backimputation.html.

CADDE-CENTRE. Covid19_brazil_hfr, 2022. URL https://github.com/CADDE-CENTRE/covid19_brazil_hfr.

CDC. Uganda Reduces Rates of Mother-to-child Transmission of HIV, June 2024. URL <https://www.cdc.gov/global-hiv-tb/php/success-stories/uganda-pmtct.html>.

Center for Systems Science and Engineering (CSSE). COVID-19 Data Repository by the CSSEGI at Johns Hopkins University., Jan. 2020. URL <https://github.com/CSSEGISandData/COVID-19>.

Centers for Disease Control (CDC). A Cluster of Kaposi's Sarcoma and Pneumocystis carinii Pneumonia among Homosexual Male Residents of Los Angeles and range Counties, California, 1981. URL <https://www.cdc.gov/mmwr/preview/mmwrhtml/00001114.htm>.

cov-lineages. Pangolin, 2020. URL <https://github.com/cov-lineages/pangolin>.

Department of Health and Social Care. Daily tests processed and testing capacity (UK): 20 March to 22 September 2020, Sept. 2020. URL <https://www.gov.uk/government/publications/daily-tests-processed-and-testing-capacity-uk-20-march-to-22-september-2020>.

European Centre for Disease Prevention and Control (ECDC). The European Surveillance System (TESSy), May 2017. URL <https://www.ecdc.europa.eu/en/publications-data/european-surveillance-system-tessy>.

GISAID. GISAID website, 2021. URL <https://www.gisaid.org>.

National Civil Registry. Transparency Portal, 2021. URL <https://transparencia.registrocivil.org.br/inicio>.

E. Ortiz-Ospina and M. Roser. Global health. *Our World in Data*, 2016.

Our World in Data. Data Page: Number of people infected with HIV by region, 2016. URL <https://ourworldindata.org/grapher/number-of-people-infected-with-hiv-by-region>.

P. Pyramid. Population Pyramids of the World from 1950 to 2100, 2023. URL <https://www.populationpyramid.net/uganda/2020/>.

Rede Genomica FioCruz. Dashboard, 2021. URL <https://www.genomahcov.fiocruz.br/dashboard/>.

RHSP. Status Report 2018, 2018. URL https://www.rhsp.org/bulletins/RHSP_Annual_Report_2018.pdf.

Swiss Agency for Development and Cooperation SDC. The Global Fund to Fight AIDS, Tuberculosis and Malaria – Global Fund, 2002. URL <https://www.dfae.admin.ch/deza/en/home/partnerships-mandates/partnerships-multilateral-organisations/weitere-organisationen-netzwerke/gfatm.html>.

- UNAIDS. Global AIDS Strategy 2021–2026., 2020. URL https://www.unaids.org/sites/default/files/media_asset/global-AIDS-strategy-2021-2026_en.pdf.
- UNAIDS. Countries, 2022. URL <https://www.unaids.org/en/regionscountries/countries>.
- UNAIDS. The path that ends AIDS: UNAIDS Global AIDS Update 2023, 2023. URL <https://www.unaids.org/en/resources/documents/2023/global-aids-update-2023>.
- United Nations. Goal 3: Ensure healthy lives and promote well-being for all at all ages, 2016. URL <https://www.un.org/sustainabledevelopment/health/>.
- United States Department of State. PEPFAR's Five-year Strategy. Fulfilling America's Promise to End the HIV/AIDS Pandemic by 2030, 2022. URL https://www.state.gov/wp-content/uploads/2022/11/PEPFARs-5-Year-Strategy_WAD2022_FINAL_COMPLIANT_3.0.pdf.
- WHO. Constitution of the World Health Organization, 1948. URL <https://www.who.int/about/accountability/governance/constitution>.
- WHO. International Health Regulations (2005) – Third edition, 2005. URL <https://www.who.int/publications-detail-redirect/9789241580496>.
- WHO. Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection (2013), 2013. URL <https://www.who.int/publications-detail-redirect/9789241505727>.
- WHO. *Consolidated Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection: Recommendations for a Public Health Approach*. WHO Guidelines Approved by the Guidelines Review Committee. World Health Organization, Geneva, 2nd edition, 2016. ISBN 978-92-4-154968-4. URL <http://www.ncbi.nlm.nih.gov/books/NBK374294/>.
- WHO. COVID-19 Public Health Emergency of International Concern (PHEIC) Global research and innovation forum, 2020a. URL [https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-\(pheic\)-global-research-and-innovation-forum](https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-(pheic)-global-research-and-innovation-forum).
- WHO. Situation report - 39, Feb. 2020b. URL <https://www.who.int/publications/m/item/situation-report---39>.
- WHO. Genomic sequencing of SARS-CoV-2, a guide to implementation for maximum impact on public health, 2021. URL <https://www.who.int/publications/i/item/9789240018440>.

WHO. Historical working definitions and primary actions for SARS-CoV-2 variants, Feb. 2024. URL <https://www.who.int/publications/m/item/historical-working-definitions-and-primary-actions-for-sars-cov-2-variants>.

APPENDIX A



SUPPLEMENTARY

FIGURES AND TABLES FOR

“ESTIMATING FLUCTUATING HOSPITAL FATALITY RATES IN LMIC SETTINGS”

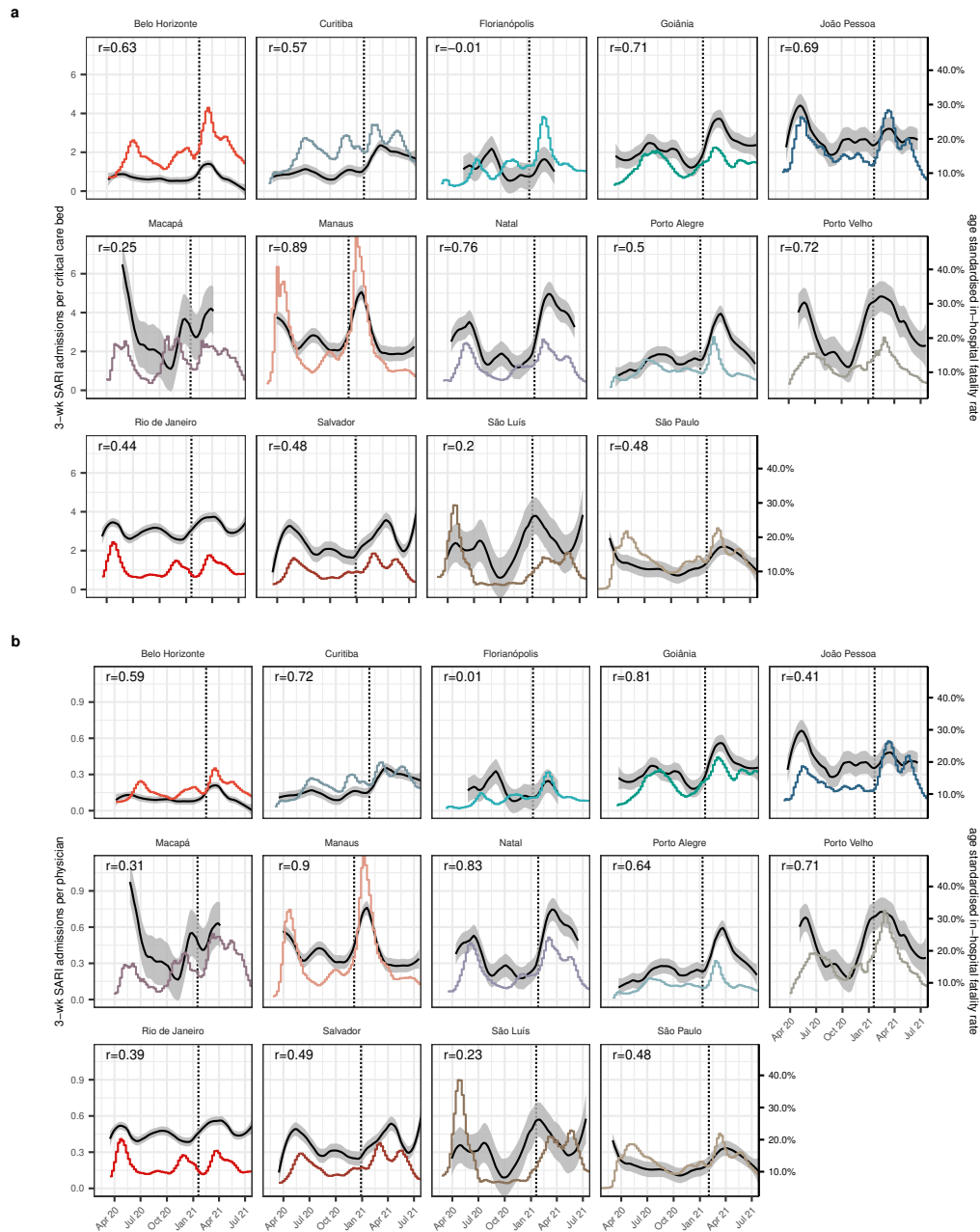


Figure A.1: Time evolution of pandemic healthcare pressure indices, part 1. SARI admissions in this and the following two weeks per hospital resource are shown in colour, with y-axis on the left. In (a), demand per critical care bed is shown, and in (b) demand per physician. Non-parametric estimates of age-standardised COVID-19 in-hospital fatality rates are shown in black, with 95% confidence intervals as grey ribbons, and y-axis on the right. Pearson correlation coefficients (r) are shown in the upper left corner, and dates of Gamma's first detection as vertical black lines. This figure is included in the accepted manuscript provided to *Nature Medicine* for further editing.

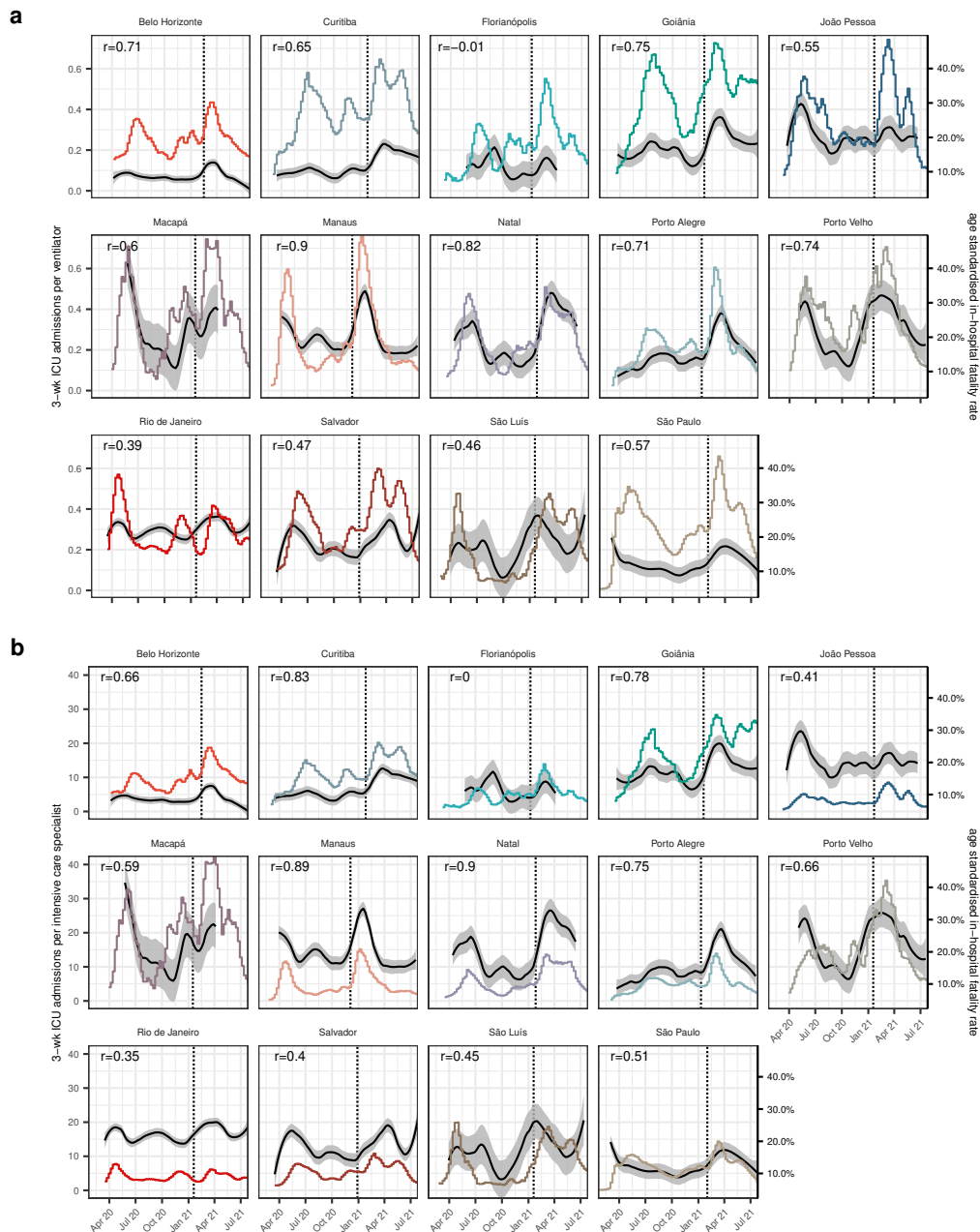


Figure A.2: Time evolution of pandemic healthcare pressure indices, part 2. ICU admissions in this and the following two weeks per hospital resource are shown in colour, with y-axis on the left. In (a), demand per ventilator is shown, and in (b) demand per intensive care specialist. Non-parametric estimates of age-standardised COVID-19 in-hospital fatality rates are shown in black, with 95% confidence intervals as grey ribbons, and y-axis on the right. Pearson correlation coefficients (r) are shown in the upper left corner, and dates of Gamma's first detection as vertical lines. This figure is included in the accepted manuscript provided to *Nature Medicine* for further editing.

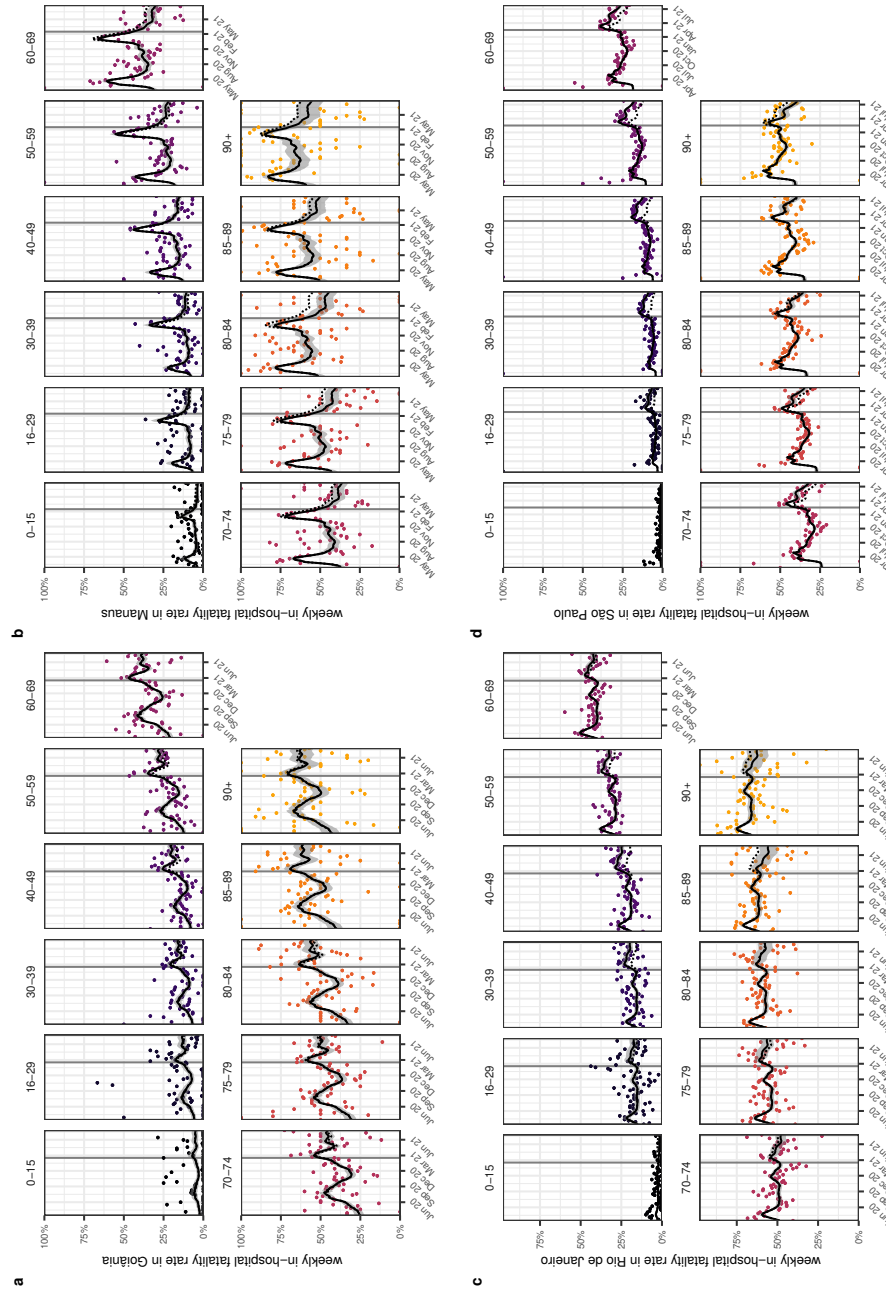


Figure A.3: Model fits to age-specific COVID-19 in-hospital fatality rates. Weekly, age-specific in-hospital fatality rates are shown as dots. Posterior median estimates of the expected in-hospital fatality rates across variants from the Bayesian multi-strain fatality model, Eq. (4.9b), are shown on the y-axis (black line) along with 95% credible intervals (grey ribbon). The expected in-hospital fatality rates of non-Gamma variants, Eq. (4.7a), are shown as dotted line. The date of Gamma's first detection is indicated as a vertical line. Data are shown for Goiânia (a), Manaus (b), Rio de Janeiro (c), and São Paulo (d). This figure is included in the accepted manuscript provided to *Nature Medicine* for further editing.

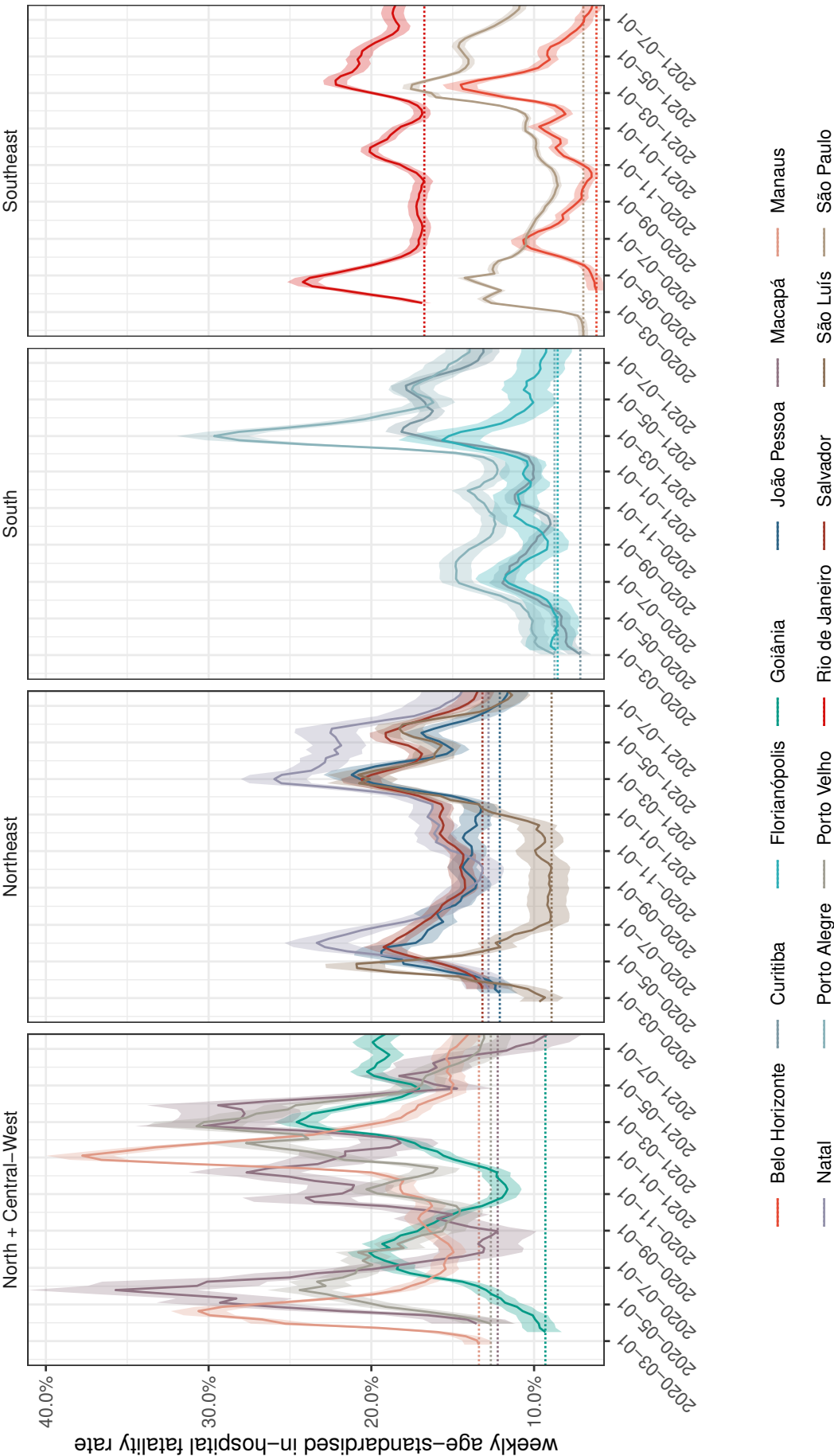


Figure A.4: Estimated, weekly age-standardised COVID-19 in-hospital fatality rates assuming that all hospitalisations with unreported outcome were discharged. Posterior median estimates (line) are shown with 95%CrIs (ribbon), and the lowest estimated fatality rates during the observation period in each state capital (dotted horizontal line).

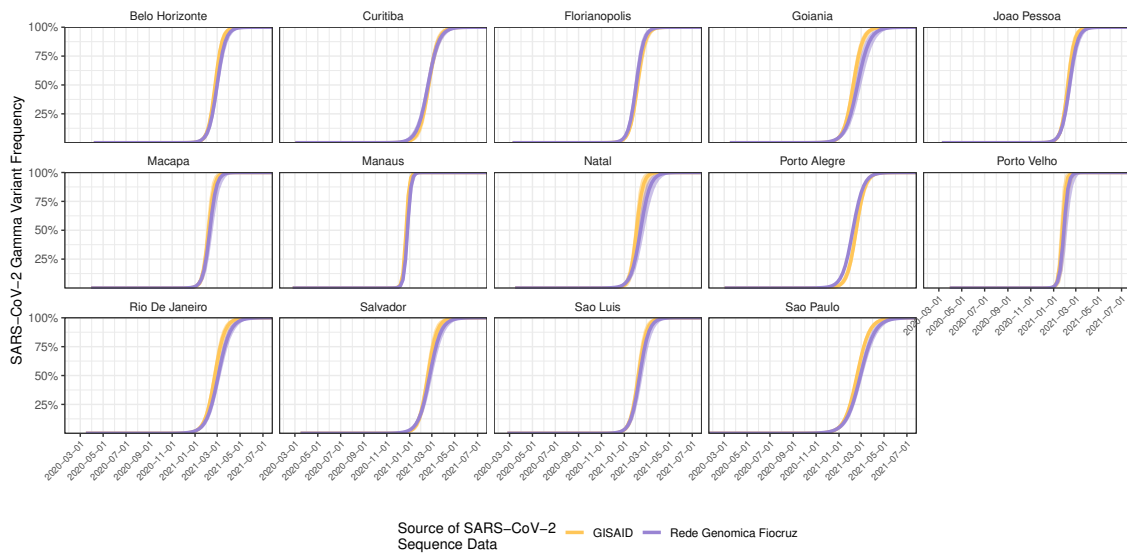


Figure A.5: Estimated temporal dynamics of Gamma frequency across the 14 cities when using sequence data from Rede Genômica FioCruz. Posterior median estimates (dots) are shown along with 95% credible intervals (linrange) for the central analysis (yellow, n=7,221 samples) and sensitivity analysis (purple, n=17,070 samples).

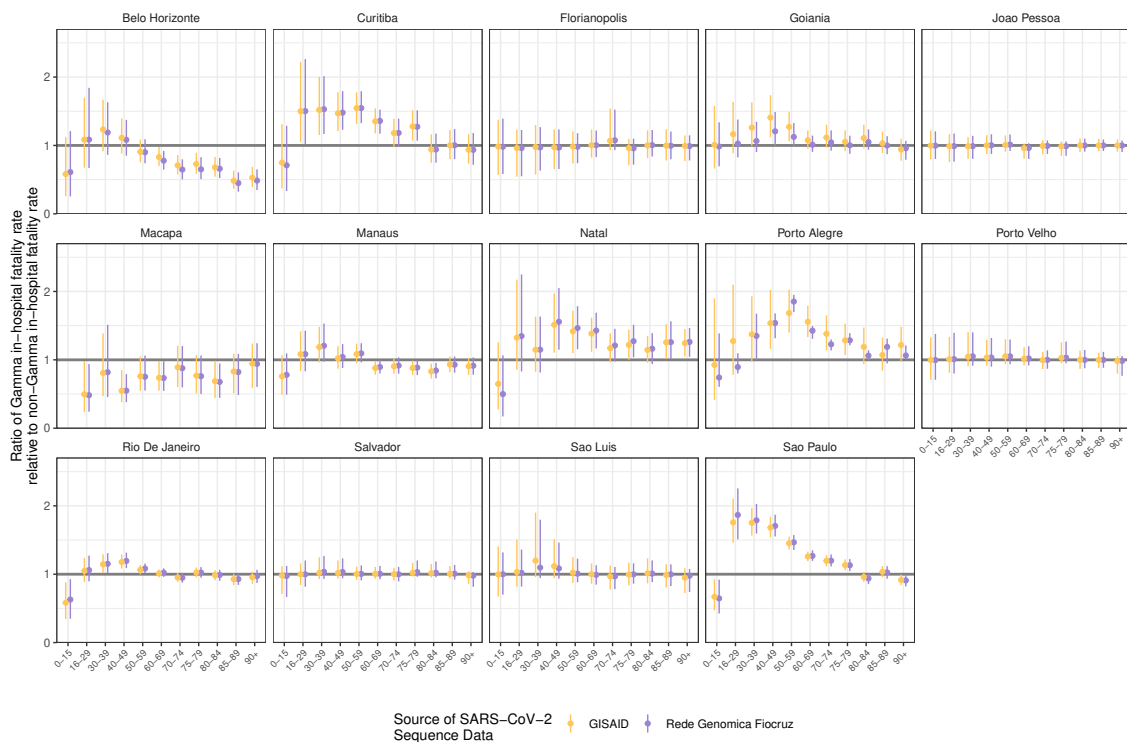


Figure A.6: Estimated effect of healthcare pressure on in-hospital fatality rates when using sequence data from Rede Genômica FioCruz. Estimated effect of Gamma infection on in-hospital fatality rates when using sequence data from Rede Genômica FioCruz. Posterior median estimates (dots) are shown along with 95% credible intervals (linrange) for the central analysis (yellow, n=7,221 samples) and sensitivity analysis (purple, n=17,070 samples).

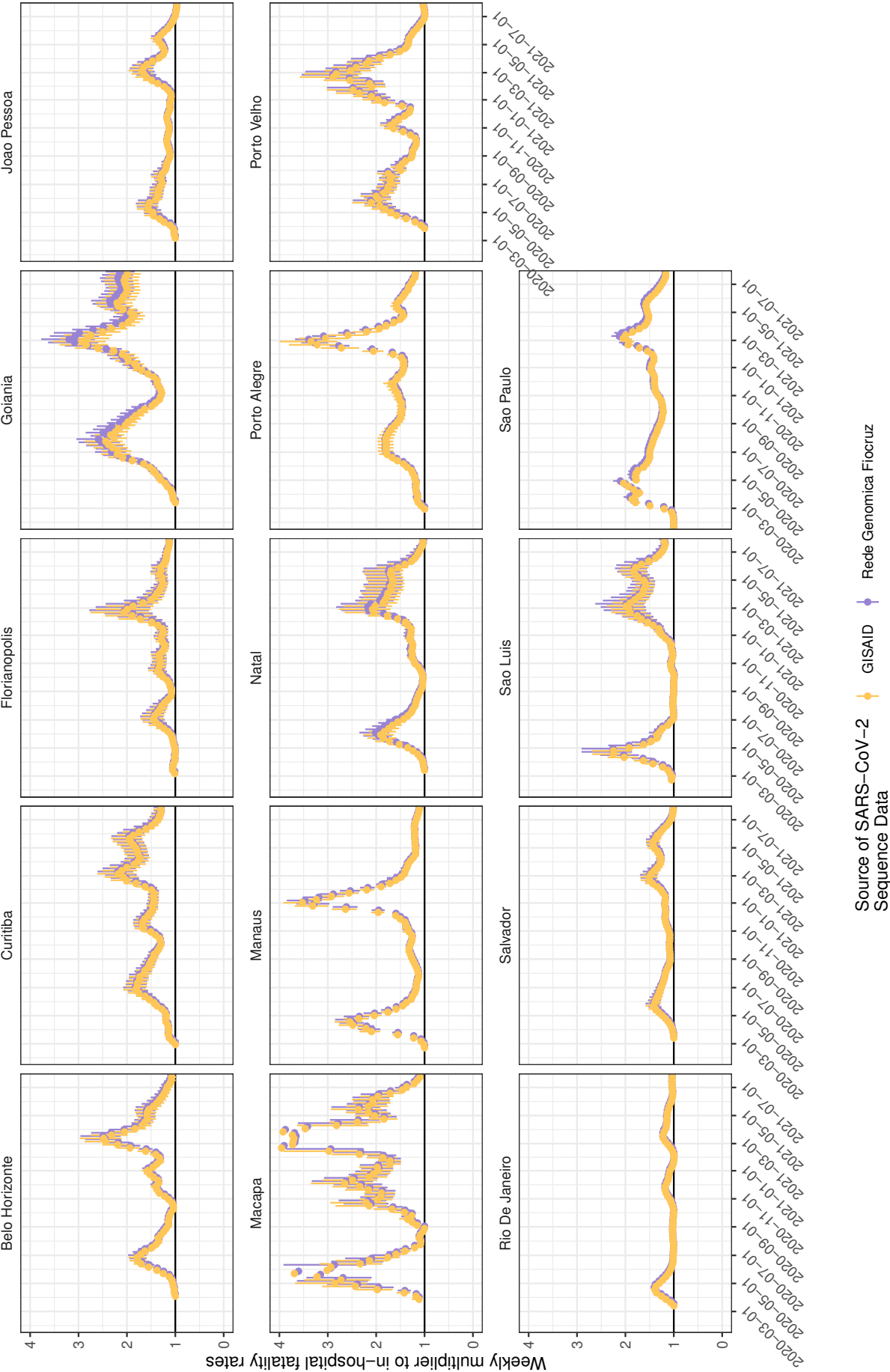


Figure A.7: Estimated effect of healthcare pressure on in-hospital fatality rates when using sequence data from Rede Genômica FioCruz. Posterior median estimates (dots) are shown along with 95% credible intervals (linorange) for the central analysis (yellow, n=3,683 samples) and sensitivity analysis (purple, n=301 samples).

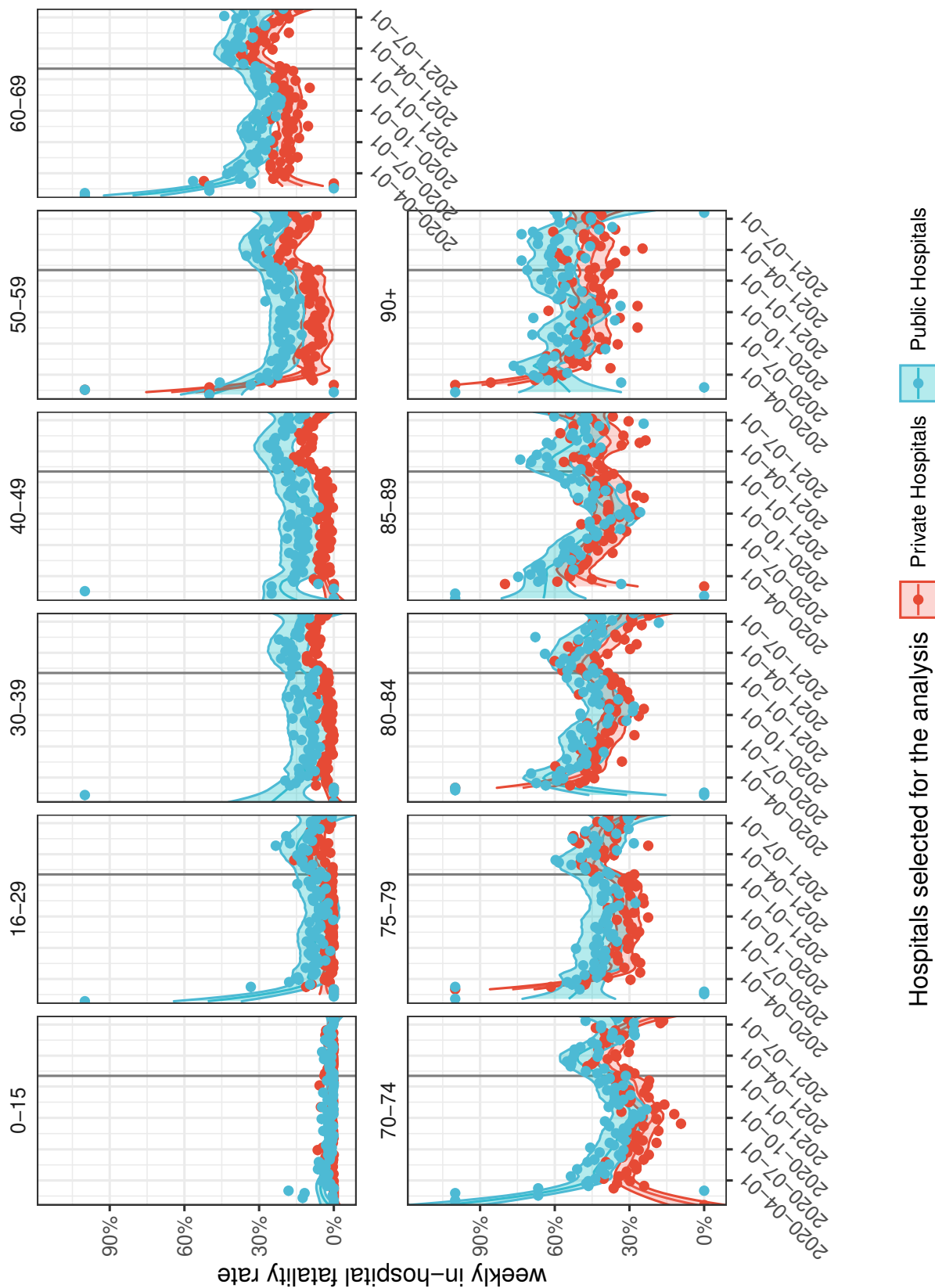


Figure A.8: Non-parametric estimates of COVID-19 in-hospital fatality rates in public and private hospitals by age group in São Paulo. Nearly all of São Paulo's hospitals could be categorised into public or private healthcare facilities, and COVID-19 in-hospital fatality rates were calculated among unvaccinated patients in public hospitals (blue), and private hospitals (red). Non-parametric, smoothed loess mean estimates (line) are shown with 95% confidence intervals (ribbon). Weekly data are shown as dots, and the date of Gamma's first detection is indicated as a vertical dotted black line.



Figure A.9: Time to death in COVID-19 attributable hospital admissions. Fatalities in weekly COVID-19 attributable hospital admissions were stratified by time to death (1 week or less, 1-2 weeks, 2-3 weeks or more than 3 weeks since admission). We show the proportion of deaths in each time to death stratum (colours). Data for each city are in facets. For reference, the time evolution of weekly hospital admissions is shown with the black line.

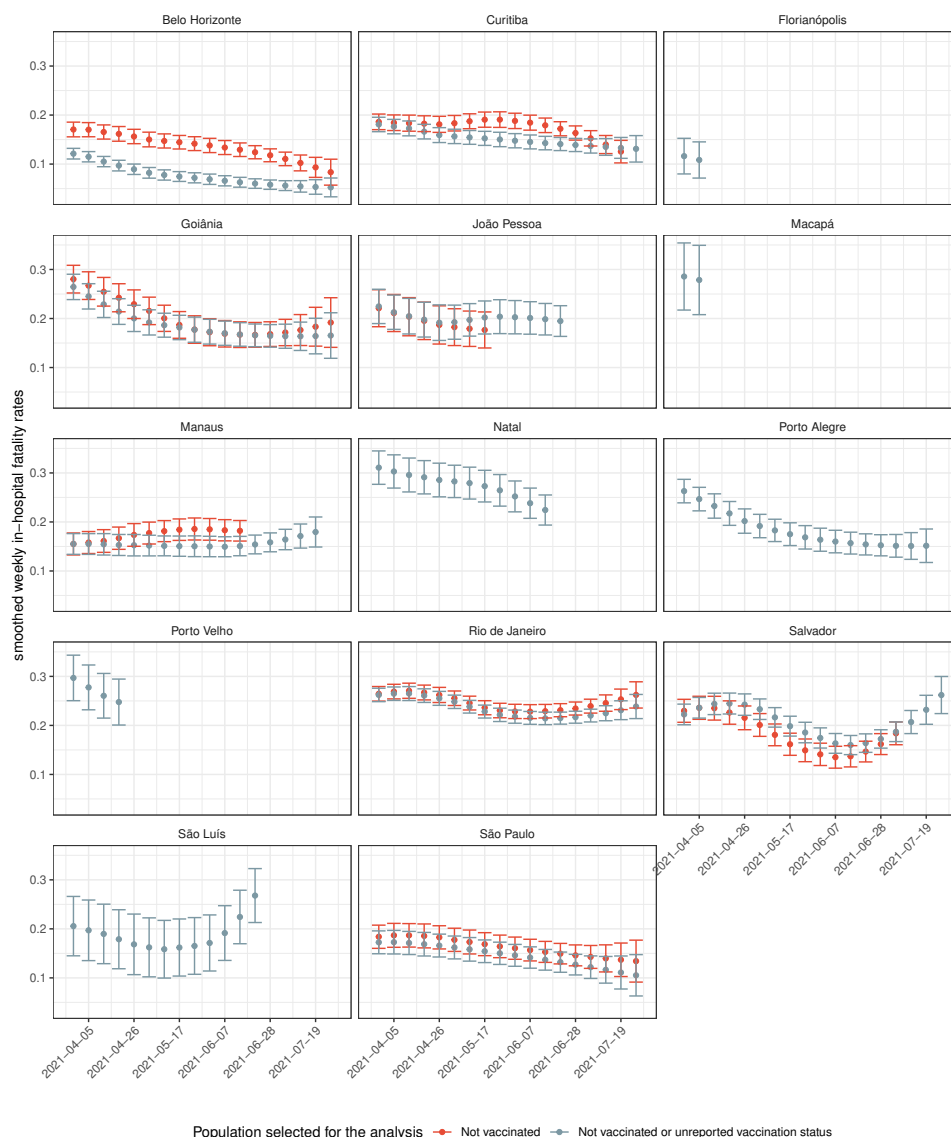


Figure A.10: Age-standardised COVID-19 fatality rates in reported unvaccinated patients only. From the denominator of COVID-19 attributable hospitalised residents, we further excluded patients with unreported vaccination status on or after March 29, 2021, and re-calculated the non-parametric, smoothed estimates of age-standardised COVID-19 fatality rates. To avoid extrapolation, we only included estimates obtained prior to the last date for which there was at least one hospital admission per age group. Mean estimates (dots) are shown along with 95% confidence intervals (errorbars) for two patient groups depending on reported vaccination status (colour): patients without evidence of vaccination (n=100,746) and confirmed unvaccinated patients (n=45,422)

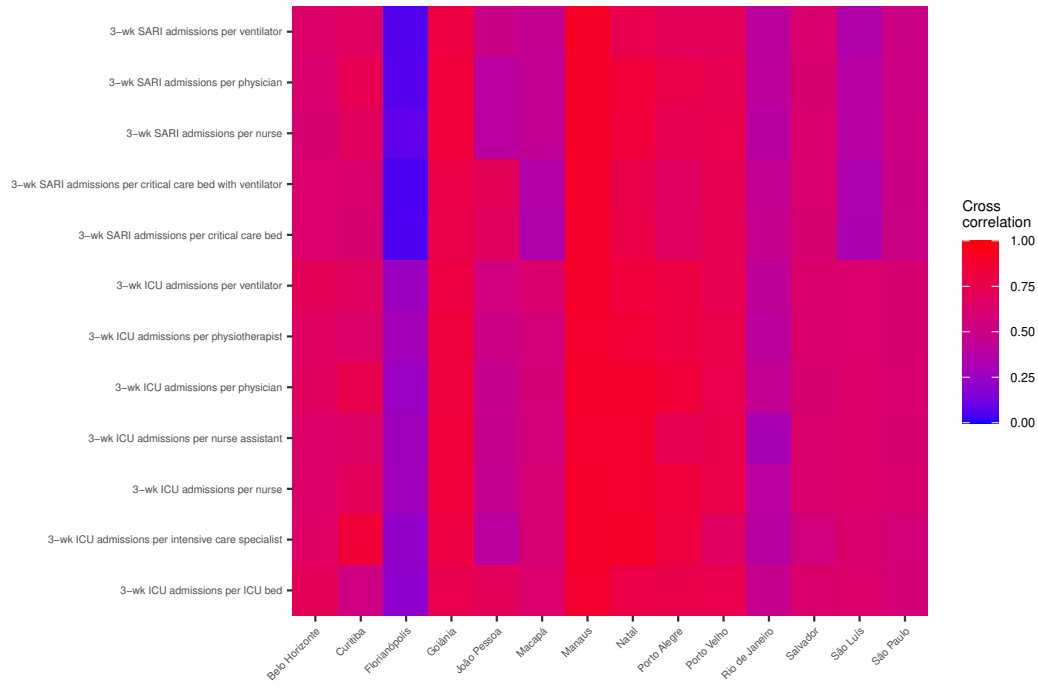


Figure A.11: Measure of cross correlations between age-standardised in-hospital fatality rates and each pandemic healthcare pressure index. For each city and healthcare pressure index, we report the maximum cross-correlation value calculated over lags of up to 5 weeks.

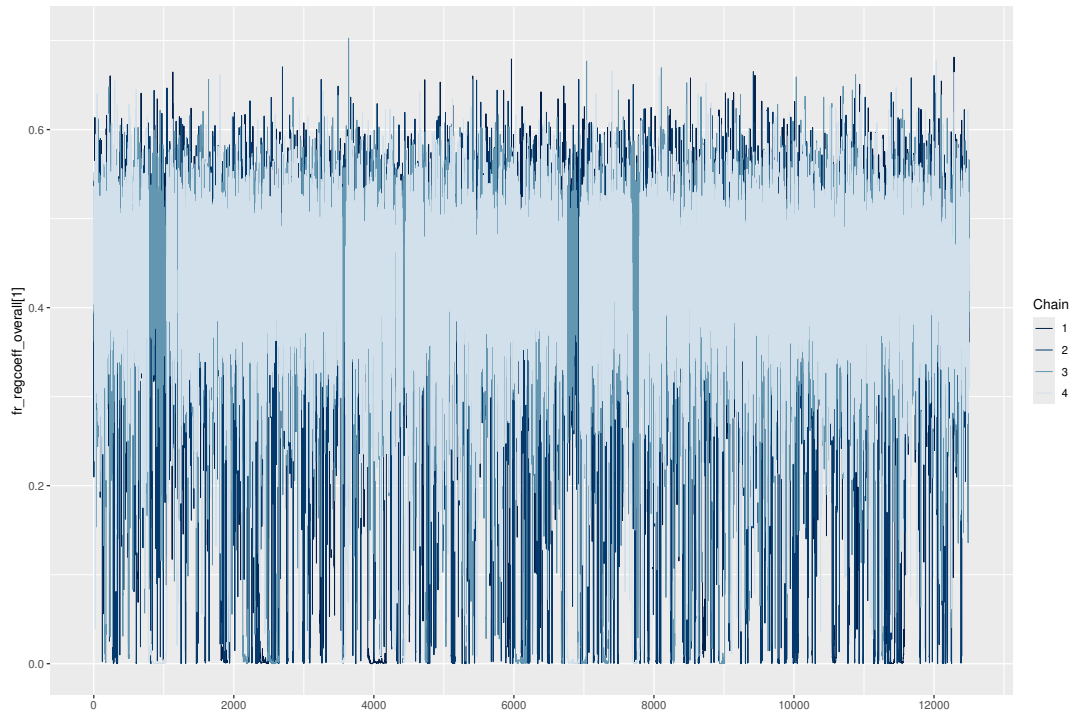


Figure A.12: Traceplot of parameter with lowest effective sample size. This corresponded to one of the β_l for the city of Macapá, see Eq. (4.7)

State capital	State	Observation period		SARS-CoV-2 Gamma variant	
		Start	End	Estimated emergence (posterior median)	First detection (observed)
Belo Horizonte	Minas Gerais	06/04/2020	26/07/2021	15/02/2021	15/12/2020
Curitiba	Paraná	02/03/2020	26/07/2021	26/01/2021	22/10/2020
Florianópolis	Santa Catarina	09/03/2020	26/07/2021	08/01/2021	11/12/2020
Goiânia	Goiás	16/03/2020	26/07/2021	19/01/2021	28/11/2020
João Pessoa	Paraíba	09/03/2020	26/07/2021	21/01/2021	20/12/2020
Macapá	Amapá	30/03/2020	26/07/2021	16/01/2021	14/12/2020
Manaus	Amazonas	24/02/2020	26/07/2021	04/12/2020	04/12/2020
Natal	Rio Grande do Norte	16/03/2020	26/07/2021	26/01/2021	06/12/2020
Porto Alegre	Rio Grande do Sul	02/03/2020	26/07/2021	10/01/2021	21/11/2020
Porto Velho	Rondônia	30/03/2020	26/07/2021	18/01/2021	25/12/2020
Rio de Janeiro	Rio de Janeiro	16/03/2020	26/07/2021	19/01/2021	29/11/2020
Salvador	Bahia	16/03/2020	26/07/2021	28/12/2020	02/12/2020
São Luís	Maranhão	24/02/2020	26/07/2021	19/01/2021	04/12/2020
São Paulo	São Paulo	20/01/2020	26/07/2021	01/02/2021	01/11/2020

Table A.1: **Sampling frame of study.**

Healthcare pressure index	Definition [†]	CBO family
3-wk SARI admissions per physician	$x_{l,w}^{\text{SARI-adm-per-physician-02}} = \left(\sum_{i=0}^2 h_{l,w+i}^{\text{SARI}} \right) / n_{l,w}^{\text{physicians}}$	2231, 2251, 2252, 2253
3-wk ICU admissions per physiotherapist	$x_{l,w}^{\text{ICU-adm-per-physiotherapist-02}} = \left(\sum_{i=0}^2 h_{l,w+i}^{\text{ICU}} \right) / n_{l,w}^{\text{physiotherapists}}$	2236
3-wk ICU admissions per physician	$x_{l,w}^{\text{ICU-adm-per-physician-02}} = \left(\sum_{i=0}^2 h_{l,w+i}^{\text{ICU}} \right) / n_{l,w}^{\text{physicians}}$	2231, 2251, 2252, 2253
3-wk ICU admissions per nurse assistant	$x_{l,w}^{\text{ICU-adm-per-nurse-assist-02}} = \left(\sum_{i=0}^2 h_{l,w+i}^{\text{ICU}} \right) / n_{l,w}^{\text{nurse-assist}}$	3222
3-wk ICU admissions per nurse	$x_{l,w}^{\text{ICU-adm-per-nurse-02}} = \left(\sum_{i=0}^2 h_{l,w+i}^{\text{ICU}} \right) / n_{l,w}^{\text{nurses}}$	2235
3-wk ICU admissions per intensive care specialist	$x_{l,w}^{\text{ICU-adm-per-intensivist-02}} = \left(\sum_{i=0}^2 h_{l,w+i}^{\text{ICU}} \right) / n_{l,w}^{\text{intensivist}}$	225150
3-wk SARI admissions per ventilator	$x_{l,w}^{\text{SARI-adm-per-ventilator-02}} = \left(\sum_{i=0}^2 h_{l,w+i}^{\text{SARI}} \right) / n_{l,w}^{\text{ventilators}}$	64
3-wk SARI admissions per critical care bed with ventilator	$x_{l,w}^{\text{SARI-adm-per-crit-care-bed-vent-02}} = \left(\sum_{i=0}^2 h_{l,w+i}^{\text{SARI}} \right) / n_{l,w}^{\text{crit-care-beds-vent}}$	64, 95
3-wk SARI admissions per critical care bed	$x_{l,w}^{\text{SARI-adm-per-crit-care-bed-02}} = \left(\sum_{i=0}^2 h_{l,w+i}^{\text{SARI}} \right) / n_{l,w}^{\text{crit-care-beds}}$	51, 75, 76, 95
3-wk ICU admissions per ventilator	$x_{l,w}^{\text{ICU-adm-per-ventilator-02}} = \left(\sum_{i=0}^2 h_{l,w+i}^{\text{ICU}} \right) / n_{l,w}^{\text{ventilators}}$	64
3-wk ICU admissions per ICU bed	$x_{l,w}^{\text{ICU-adm-per-ICU-bed-02}} = \left(\sum_{i=0}^2 h_{l,w+i}^{\text{ICU}} \right) / r_{l,w}^{\text{ICU-beds}}$	51, 75, 76

[†] Data on hospital resources were obtained from CNES (cne [2021]), and were aggregated by month and location. Data on hospital admissions were retrieved from the SIVEP-Gripe platform. Hospital personnel data were defined according to the 2002 Brazilian Classification of Occupations [Classificação Brasileira de Ocupações (CBO)] (Brazilian Ministry of Labor and Employment [2021]). Hospital equipment data were defined according to the CNES code classification

Table A.2: Definitions of the healthcare pressure indices.

	Assumed efficacy against severe infection requiring hospitalisation		Notes and references
	1 st dose	2 nd dose	
Sinovac	31.6%	71.4%	From Ranzani et al. [2021b] , assuming vaccine effectiveness against hospitalisation is similar to effectiveness against death.
Covishield	40%	91.4%	From Voysey et al. [2021] , Madhi et al. [2021] , Clemens et al. [2021] , and calibrating effectiveness against symptomatic infection against severe infection using Fig. 3A in Khoury et al. [2021] .
Janssen and Pfizer	40%	91.4%	No effectiveness studies found in Brazil. Janssen and Pfizer contributed to 15.07% of administered by the end of the study period. Hence, used same values as for Covishield.

Table A.3: **COVID-19** vaccine efficacy estimates in populations in which **SARS-CoV-2** Gamma is circulating.

Location	Reported COVID-19 attributable in-hospital deaths [†]				Underreporting-adjusted COVID-19 attributable in-hospital deaths [‡]			
	Total	Ages			Total	Ages		
		0-49	50-74	75+		0-49	50-74	75+
Belo Horizonte	7640	808	3680	3152	7691	817	3705	3169
Curitiba	7535	1162	3941	2432	7535	1162	3941	2432
Florianópolis	882	85	441	356	892	86	446	360
Goiânia	6103	1006	3110	1987	6245	1036	3181	2028
João Pessoa	2810	508	1326	976	3635	674	1730	1231
Macapá	992	240	507	245	1011	243	520	248
Manaus	10006	1922	5390	2694	10168	1961	5483	2724
Natal	3059	467	1448	1144	3512	561	1688	1263
Porto Alegre	5133	517	2625	1991	5265	541	2701	2023
Porto Velho	2156	468	1200	488	2472	561	1361	550
Rio de Janeiro	23398	3116	11537	8745	28052	3947	13988	10117
Salvador	7667	1206	3754	2707	8508	1402	4206	2900
São Luís	2163	372	1049	742	2546	428	1256	862
São Paulo	40744	5483	20301	14960	42768	5795	21302	15671

[†] COVID-19 attributable in-hospital deaths were defined as deaths in hospitalised unvaccinated residents with PCR confirmed or clinically diagnosed COVID-19 infection, or severe respiratory infection with no other confirmed cause. Data capture admissions and deaths until 26 July 2021.

[‡] To reported counts were added expected fatal outcomes in COVID-19 attributable hospital admissions in unvaccinated residents with unreported outcomes.

Table A.4: Reported and underreporting-adjusted COVID-19 attributable in-hospital deaths.

City	Longest observed duration of COVID-19 in-hospital fatality rates above 50% (in weeks, by age group shown in years)										
	0-15	16-29	30-39	40-49	50-59	60-69	70-74	75-79	80-84	85-89	90+
Belo Horizonte	0	0	0	0	0	0	0	0	0	0	0
Curitiba	0	0	0	0	0	0	3	0	2	0	6
Florianópolis	0	0	0	0	0	0	0	8	7	5	14
Goiânia	0	0	0	0	0	0	7	9	12	22	32
João Pessoa	0	0	0	0	0	2	8	18	44	24	19
Macapá	0	7	0	8	9	11	14	18	23	53	31
Manaus	0	0	0	1	4	9	10	14	51	26	51
Natal	0	0	0	0	1	24	16	25	30	48	47
Porto Alegre	0	0	0	0	0	6	9	10	12	56	62
Porto Velho	0	0	0	0	0	8	25	30	16	48	64
Rio de Janeiro	0	0	0	0	0	0	12	52	72	72	70
Salvador	0	0	0	0	1	0	9	10	13	28	29
São Luís	1	0	0	0	0	1	2	14	11	17	12
São Paulo	0	1	0	0	0	4	7	2	8	20	20

Table A.5: Longest observed duration of COVID-19 in-hospital fatality rates above 50%.

	Date	Healthcare resources per 100,000 population [†]						
		Belo Horizonte	Curitiba	Florianópolis	Goiânia	João Pessoa	Macapá	Manaus
Intensive Care Specialists	March 2020	2	3	3	2	7	0	2
	July 2021	2 (0%)	3 (0%)	4 (33%)	2 (0%)	6 (-14%)	0	2 (0%)
Nurses	March 2020	249	185	251	151	212	129	114
	July 2021	306 (23%)	228 (23%)	299 (19%)	180 (19%)	238 (12%)	164 (27%)	152 (33%)
Physicians	March 2020	631	482	615	368	309	124	155
	July 2021	674 (7%)	520 (8%)	686 (12%)	394 (7%)	345 (12%)	139 (12%)	166 (7%)
Physiotherapists	March 2020	74	65	86	50	76	47	19
	July 2021	92 (24%)	84 (29%)	117 (36%)	59 (18%)	90 (18%)	56 (19%)	29 (53%)
Nurse assistants	March 2020	584	356	479	371	309	318	290
	July 2021	723 (24%)	491 (38%)	601 (25%)	427 (15%)	367 (19%)	350 (10%)	328 (13%)
CC Beds	March 2020	41	41	35	51	27	10	11
	July 2021	56 (37%)	63 (54%)	54 (54%)	86 (69%)	59 (119%)	28 (180%)	28 (155%)
CC Beds with ventilator	March 2020	40	41	30	36	23	9	11
	July 2021	55 (38%)	62 (51%)	51 (70%)	70 (94%)	54 (135%)	25 (178%)	26 (136%)
ICU Beds	March 2020	34	33	24	32	22	6	9
	July 2021	48 (41%)	56 (70%)	46 (92%)	66 (106%)	51 (132%)	22 (267%)	24 (167%)
Ventilators	March 2020	83	79	86	77	57	21	38
	July 2021	101 (22%)	112 (42%)	108 (26%)	113 (47%)	87 (53%)	41 (95%)	58 (53%)
		Natal	Porto Alegre	Porto Velho	Rio de Janeiro	Salvador	São Luís	São Paulo
Intensive Care Specialists	March 2020	4	6	2	5	6	1	3
	July 2021	4 (0%)	6 (0%)	2 (0%)	6 (20%)	7 (17%)	2 (100%)	4 (33%)
Nurses	March 2020	162	290	168	185	229	223	183
	July 2021	201 (24%)	334 (15%)	250 (49%)	228 (23%)	265 (16%)	258 (16%)	219 (20%)
Physicians	March 2020	313	626	237	299	329	207	324
	July 2021	348 (11%)	654 (4%)	293 (24%)	326 (9%)	344 (5%)	229 (11%)	358 (10%)
Physiotherapists	March 2020	43	67	44	35	73	53	41
	July 2021	55 (28%)	79 (18%)	76 (73%)	41 (17%)	80 (10%)	71 (34%)	55 (34%)
Nurse assistants	March 2020	378	624	461	291	374	497	214
	July 2021	451 (19%)	868 (39%)	608 (32%)	422 (45%)	432 (16%)	569 (14%)	271 (27%)
CC Beds	March 2020	41	55	37	44	55	37	37
	July 2021	68 (66%)	88 (60%)	86 (132%)	56 (27%)	67 (22%)	58 (57%)	52 (41%)
CC Beds with ventilator	March 2020	36	53	35	38	46	33	30
	July 2021	63 (75%)	86 (62%)	82 (134%)	50 (32%)	63 (37%)	48 (45%)	48 (60%)
ICU Beds	March 2020	27	45	31	22	42	24	18
	July 2021	53 (96%)	82 (82%)	78 (152%)	31 (41%)	60 (43%)	37 (54%)	37 (106%)
Ventilators	March 2020	75	102	68	72	83	65	66
	July 2021	123 (64%)	156 (53%)	147 (116%)	94 (31%)	109 (31%)	87 (34%)	86 (30%)

[†] Numbers summarise healthcare facility level microdata on personnel (nurses, nurse assistants, physiotherapists, physicians and intensive care specialists), and reported equipment (critical care beds, ICU beds, ventilators) from Brazil's National Register of Health Facilities (Cadastro Nacional de Estabelecimentos de Saúde - CNES). Values in brackets give the percent increase from March 2020 to July 2021. CC stands for critical care.

Table A.6: Healthcare resources, March 2020 to July 2021.

City	Maximum $\hat{R}$	Mininum ESS
Belo Horizonte	1.0003	27857
Curitiba	1.0003	12923
Florianópolis	1.0003	8700
Goiânia	1.0003	21118
João Pessoa	1.0003	8121
Macapá	1.0009	1687
Manaus	1.0002	31551
Natal	1.0003	13478
Porto Alegre	1.0002	24679
Porto Velho	1.0008	7936
Rio de Janeiro	1.0004	34369
Salvador	1.0002	8222
São Luis	1.0005	8610
São Paulo	1.0001	34373

Table A.7: **Model maximum $\hat{R}$ and minimum effective sample sizes for each city.**

APPENDIX B



SUPPLEMENTARY

FIGURES AND TABLES FOR

“AGE AND GENDER PROFILES OF HIV INFECTION BURDEN AND VIRAEMIA”

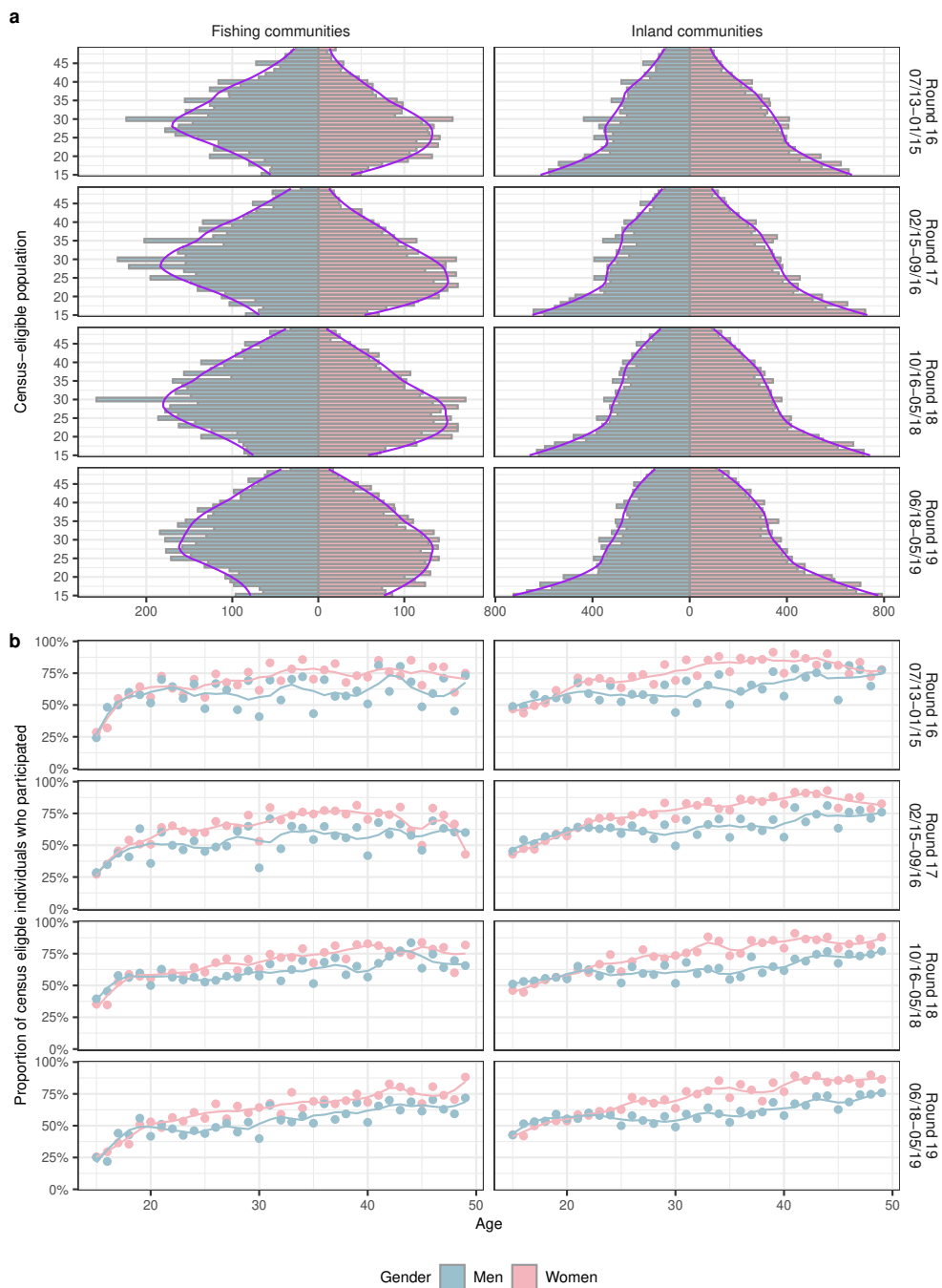


Figure B.1: RCCS population size estimates and participation rates, 2013-2019 (a) Age- and gender-specific population sizes reported by households heads during the census preceding each survey round (bars), overlaid with smooth population size estimates obtained with **LOESS** (purple line). (b) Age- and gender-specific empirical survey participation rates in the census-eligible populations (dots), overlaid with smooth estimates obtained with **LOESS** (lines).

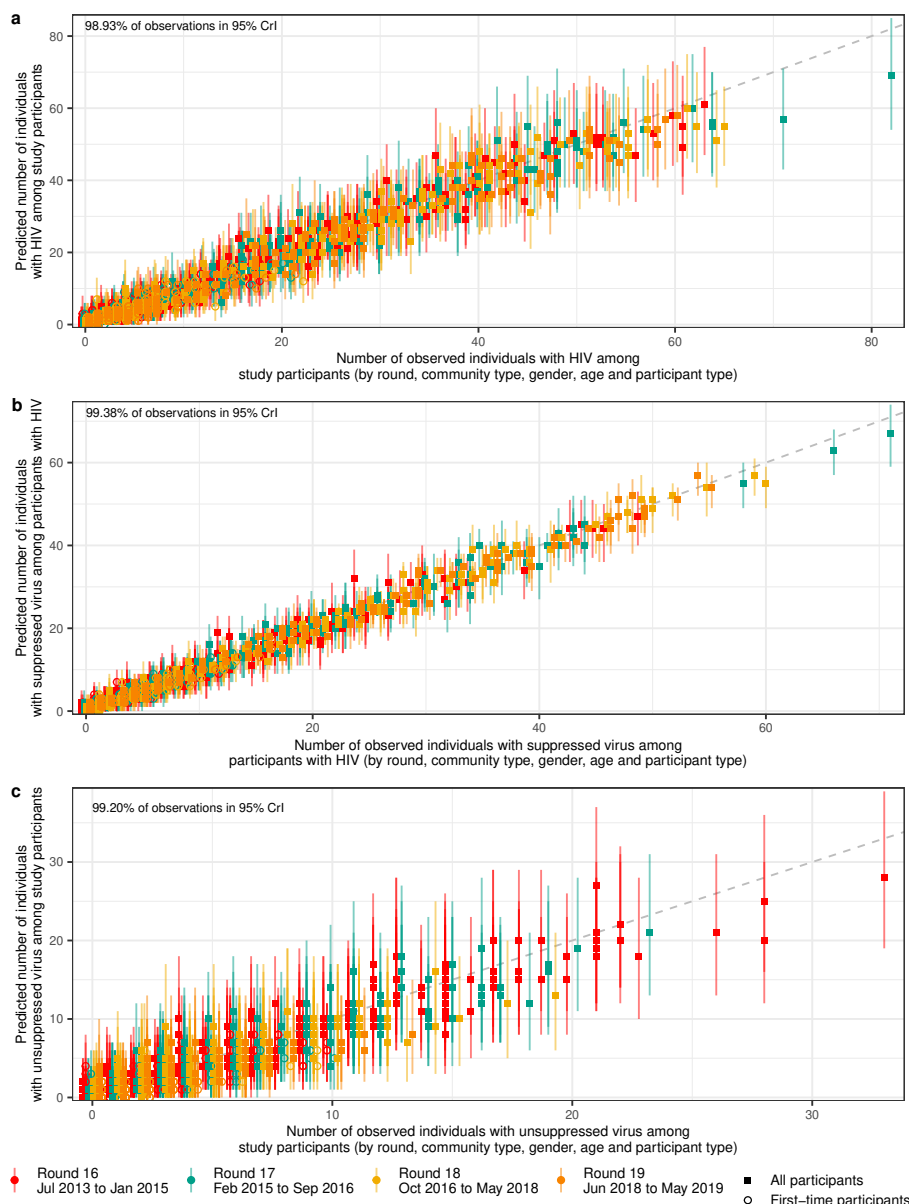


Figure B.2: Consistency of fitted models with available data. (a) Posterior predictive median estimates of the number of individuals with **HIV** among participants (dot) are plotted on the y-axis versus the observed number in each survey round-community type-gender-age-participant type stratum, along with 95% posterior predictive intervals (errorbars). (b) Posterior predictive median estimates of the number of individuals suppressed virus among participants with **HIV** (dot) are plotted on the y-axis versus the observed number in each survey round-community type-gender-age-participant type stratum, along with 95% posterior predictive intervals (errorbars). (c) Posterior predictive median estimates of the number of individuals suppressed virus among participants (dot) are plotted on the y-axis versus the observed number in each survey round-community type-gender-age-participant type stratum, along with 95% posterior predictive intervals (errorbars).

Location	Gender	Survey round	Census eligible individuals (n)	Participants (n)	Participation rates
Fishing communities	Women	Round 16 07/13-01/15	2,757	1,867	67.7%
		Round 17 02/15-09/16	3,224	2,077	64.4%
		Round 18 10/16-05/18	3,359	2,225	66.2%
		Round 19 06/18-05/19	3,236	1,968	60.8%
	Men	Round 16 07/13-01/15	3,546	2,067	58.3%
		Round 17 02/15-09/16	4,068	2,163	53.2%
		Round 18 10/16-05/18	4,224	2,535	60%
		Round 19 06/18-05/19	4,002	2,123	53%
	Women	Round 16 07/13-01/15	11,346	7,816	68.9%
		Round 17 02/15-09/16	11,990	8,377	69.9%
		Round 18 10/16-05/18	12,193	8,331	68.3%
		Round 19 06/18-05/19	12,893	8,495	65.9%
	Men	Round 16 07/13-01/15	10,541	6,256	59.3%
		Round 17 02/15-09/16	10,939	6,716	61.4%
		Round 18 10/16-05/18	11,076	6,722	60.7%
		Round 19 06/18-05/19	11,736	6,765	57.6%

Table B.1: Census eligible population size and participation rates by community type, gender and survey round.

Location	Gender	Survey round	First-time participants (n)	First-time participants with HIV (n)	First-time participants exhibiting viraemia (n)	Ratio of first-time participants with HIV relative to round 16	Ratio of first-time participants exhibiting viraemia relative to round 16
Fishing	Women	Round 16 (07/13-01/15)	492	198	100	1.00	1.00
		Round 17 (02/15-09/16)	670	241	87	1.22	0.87
		Round 18 (10/16-05/18)	688	231	57	1.17	0.57
		Round 19 (06/18-05/19)	540	176	29	0.89	0.29
	Men	Round 16 (07/13-01/15)	529	122	88	1.00	1.00
		Round 17 (02/15-09/16)	577	134	77	1.1	0.88
		Round 18 (10/16-05/18)	766	159	57	1.3	0.65
		Round 19 (06/18-05/19)	505	103	21	0.84	0.24
Inland	Women	Round 16 (07/13-01/15)	2,553	310	131	1.00	1.00
		Round 17 (02/15-09/16)	2,421	324	102	1.05	0.78
		Round 18 (10/16-05/18)	2,318	235	76	0.76	0.58
		Round 19 (06/18-05/19)	2,447	238	43	0.77	0.33
	Men	Round 16 (07/13-01/15)	2,028	118	67	1.00	1.00
		Round 17 (02/15-09/16)	1,898	108	58	0.92	0.87
		Round 18 (10/16-05/18)	1,846	92	43	0.78	0.64
		Round 19 (06/18-05/19)	1,937	91	30	0.77	0.45

Table B.2: Characteristics of RCCS first-time participants by community type, gender and survey round, 2013-2019.

Location	Gender	Survey Round	Age profile of individuals with HIV			Age profile of individuals with unsuppressed virus		
			25% quantile	Median	75% quantile	25% quantile	Median	75% quantile
Fishing	Female	Round 16 (07/13-01/15)	24.4 (24.2-24.5)	29.3 (29.2-29.5)	34.9 (34.7-35.0)	22.2 (22.0-22.4)	26.7 (26.5-27.0)	32.2 (31.9-32.6)
		Round 17 (02/15-09/16)	24.5 (24.4-24.6)	29.3 (29.2-29.5)	35.0 (34.8-35.1)	23.0 (22.8-23.3)	27.5 (27.2-27.8)	32.8 (32.4-33.2)
		Round 18 (10/16-05/18)	25.6 (25.5-25.8)	30.4 (30.2-30.5)	36.0 (35.8-36.1)	22.5 (22.1-22.9)	27.1 (26.8-27.4)	32.0 (31.5-32.4)
		Round 19 (06/18-05/19)	27.1 (26.9-27.3)	32.2 (32.0-32.3)	37.7 (37.6-37.9)	22.2 (21.7-22.8)	28.3 (27.7-29.0)	34.5 (33.9-35.0)
	Male	Round 16 (07/13-01/15)	27.6 (27.5-27.8)	32.3 (32.1-32.5)	37.8 (37.5-38.0)	26.3 (26.1-26.5)	30.4 (30.2-30.7)	36.2 (35.8-36.6)
		Round 17 (02/15-09/16)	28.8 (28.6-29.0)	33.0 (32.8-33.2)	37.9 (37.7-38.1)	27.5 (27.3-27.8)	31.8 (31.5-32.1)	36.9 (36.5-37.2)
		Round 18 (10/16-05/18)	29.5 (29.3-29.6)	34.1 (33.9-34.2)	39.4 (39.2-39.5)	27.1 (26.7-27.5)	31.8 (31.5-32.2)	36.8 (36.5-37.2)
		Round 19 (06/18-05/19)	30.6 (30.4-30.9)	35.4 (35.2-35.6)	40.7 (40.5-40.9)	28.3 (27.9-28.7)	31.8 (31.4-32.3)	36.0 (35.4-36.7)
Inland	Female	Round 16 (07/13-01/15)	25.7 (25.6-25.9)	31.5 (31.4-31.7)	37.7 (37.5-37.9)	23.4 (23.2-23.7)	28.4 (28.2-28.7)	34.8 (34.4-35.3)
		Round 17 (02/15-09/16)	25.5 (25.4-25.7)	31.8 (31.6-31.9)	38.2 (38.0-38.4)	22.5 (22.2-22.8)	27.3 (27.0-27.6)	32.9 (32.5-33.2)
		Round 18 (10/16-05/18)	26.8 (26.6-26.9)	33.1 (32.9-33.2)	39.3 (39.2-39.5)	22.3 (22.1-22.6)	27.1 (26.7-27.5)	32.4 (32.0-32.8)
		Round 19 (06/18-05/19)	27.7 (27.5-27.9)	33.8 (33.6-34.0)	40.4 (40.2-40.6)	23.5 (23.0-24.0)	29.0 (28.5-29.5)	34.9 (34.4-35.5)
	Male	Round 16 (07/13-01/15)	28.9 (28.7-29.2)	34.4 (34.1-34.6)	39.9 (39.6-40.2)	27.1 (26.8-27.4)	31.9 (31.5-32.3)	37.7 (37.3-38.2)
		Round 17 (02/15-09/16)	29.9 (29.6-30.1)	35.5 (35.2-35.7)	40.9 (40.6-41.1)	27.6 (27.2-27.9)	32.6 (32.1-33.1)	39.1 (38.5-39.6)
		Round 18 (10/16-05/18)	30.4 (30.1-30.6)	35.6 (35.3-35.8)	40.7 (40.5-41.0)	26.5 (25.9-27.0)	32.0 (31.6-32.5)	37.2 (36.7-37.8)
		Round 19 (06/18-05/19)	30.5 (30.1-30.8)	36.7 (36.4-37.0)	42.1 (41.8-42.3)	26.2 (25.6-26.7)	31.4 (30.8-32.0)	37.7 (36.9-38.5)

Table B.3: Age profile of individuals with HIV and individuals with unsuppressed virus in the population by community type, gender and survey round, 2013-2019. Age profiles are reported as posterior median estimates with 95% credible intervals in parentheses.

Model	Round	PDI	EBFMI	ESS
HIV prevalence in first-time participants and all participants	16	0.00 %	0.96	6360
	17	0.00 %	0.98	7598
	18	0.00 %	1.00	7442
	19	0.00 %	0.99	12589
Viral suppression rates among people with HIV	16	0.00 %	0.99	3531
	17	0.00 %	0.98	2234
	18	0.00 %	0.99	2172
	19	0.00 %	1.01	1647
Prevalence of unsuppressed viraemia in first-time participants and all participants	16	0.01 %	0.96	3032
	17	0.00 %	0.97	3277
	18	0.06 %	0.98	5128
	19	0.04 %	0.96	3044

Table B.4: **Convergence diagnostics for all models.** PDI=Proportion of Divergent Iterations, EBFMI= Effective Bayesian Fraction of Missing Information, ESS= Minimum effective sample size for the model parameters.

APPENDIX C

PROOFS

C.1 Proof of Algorithm 1 working.

Let $\mathcal{N} = (\mathcal{V}, E)$ be a connected transmission network (a directed tree). Let $(m_i^{\text{hist}})_{i \in \mathcal{V}}$ be the minimum plausible infection date for each individual in the network. The aim of the proof is to show that the output $(m_i^*)_{i \in \mathcal{V}}$ satisfies the following properties.

A Constraints Eqs. (5.2) and (5.4). These can be jointly written as:

$$m_i^* \geq \min\left(m_i^{\text{hist}}, m_{\text{source}(i)}^*\right) \quad \forall i \in \mathcal{V}. \quad (\text{C.1})$$

B For each $i \in \mathcal{V}$, m_i^* is “minimal”, meaning it is the smallest number satisfying Eq. (C.1).

As the output is initialized with $(m_i^{\text{hist}})_{i \in \mathcal{V}}$ and every update corresponds with an increase in value, property **A** is trivially satisfied by the algorithm. We are left to prove statement **B**.

The proof proceeds by induction, exploiting the uniqueness of the source for each node, except for the root node.

Base case: The output of the algorithm for the root node i_0 equals its initialisation value:

$m_{i_0}^* = m_{i_0}^{\text{hist}}$. Any value $m_{i_0}^* < m_{i_0}^{\text{hist}}$ would violate the constraint Eq. (5.4) and so $m_{i_0}^*$ is “minimal”.

Inductive step: Assume that the statement holds for all nodes i at distance d from the root node i_0 . Consider a node j at distance $d + 1$ from the root node i_0 . By construction of the algorithm, we have:

$$m_j^* = \min\left(m_j^{\text{hist}}, m_{\text{source}(j)}^*\right). \quad (\text{C.2})$$

If $m_{\text{source}(j)}^* = m_{\text{source}(j)}^{\text{hist}}$, then m_j^* is “minimal” by an argument analogous to the one for the base case. Otherwise, $m_j^* = m_{\text{source}(j)}^*$. By the inductive hypothesis, $m_{\text{source}(j)}^*$ is “minimal”, and so m_j^* is “minimal”.

An analogous proof works for the M-algorithm (Algorithm 2), with the main differences being that we now want to prove that each component of $(M_i^*)_{i \in \mathcal{V}}$ is maximal, and where we have to account for the fact that each node may have multiple recipients. Both adjustments are trivial.