Quantitative analysis of the mode and tempo of virus molecular evolution



Mahan Ghafari

University College

University of Oxford

A thesis presented for the degree of

Doctor of Philosophy

Trinity 2022

"The Revelations of Devout and Learn'd

Who rose before us, and as Prophet burn'd,

Are all but Stories, which, awoke from Sleep

They told their comrades and to Sleep return'd"

– Omar Khayyam

Abstract

Quantitative analysis of the mode and tempo of virus molecular evolution

Mahan Ghafari, University College, University of Oxford

Estimating the rate of nucleotide and amino acid substitutions is fundamental to our understanding of how fast species evolve. It also enables us to reconstruct their natural history. The development of sophisticated statistical methods and software in molecular sequencing analysis has made it easier for biologists to test and compare various evolutionary hypotheses and infer evolutionary parameters. However, the use of these software packages as a 'blackbox' without paying close attention to the underlying assumptions can lead to spurious rate estimations across different timescales which often result in erroneous inferences of virus evolutionary histories. In this thesis, I develop quantitative methods to understand biological mechanisms that can alter the molecular evolutionary rate of viruses over time and investigate the implications of having a time-dependent rate of evolution for reconstructing the phylogenetic history of viruses. I also explore alternative methods to estimating the time of origin and reconstructing the evolutionary trajectory of viruses using mobility data, national vital statistics, and serology surveys. I show how three evolutionary processes of rapid within-host evolution during chronic infections, purifying selection, and site saturation can determine the changes in evolutionary rate of viruses over different timescales, from a few weeks to millions of years. I also leverage excess mortality and seroprevalence data to correct for under-reporting of COVID-19-related deaths, reconstruct early transmission dynamics of SARS-CoV-2, and estimate the start of the Iranian epidemic as well its growth rate in the absence of high-quality genome sequence and province-level epidemiological data. This work will build towards the goal of making more accurate estimations of the timescale of phylogenetic histories and the rate at which species evolve. It also further strengthens the conceptual link between phylogenetics, population genetics, and molecular evolution.

Acknowledgments

I thank my fellow graduate students and postdocs in the Katzourakis and Pybus groups, past and present, for their continued friendship and support. I am extremely grateful to my examiners Christopher Dye and Erik Volz for their thorough and thoughtful comments and my co-authors on the papers making up this thesis. I would also like to thank my supervisors Aris and Oliver who have always been supportive and have given me the freedom to do the research that I enjoy. I would like to particularly thank Aris who was primarily responsible for supervising my thesis work. I would also like to express my gratitude to the members of staff at the Doctoral Training Centre, University College, Department of Zoology, and Peter Medawar Building for Pathogen Research who have always been patient with me and supported me during my DPhil. Finally, I would like to thank my family for their love and support. This thesis is dedicated to my mother, Nazli, who always went above and beyond to help me be where I am today but never got to finally see me in the graduation gown. After I lost her during the second year of my DPhil, there were countless days when everything felt too unbearable. Had it not been for the continued encouragement from my wife, I would have never been able to complete this work. So, I am most grateful for her love and support.

Statement of contribution and associated publications

Chapter 2. A mechanistic evolutionary model explains the time-Dependent pattern of substitution rates in viruses

Published as: **Ghafari M**, Pybus OG, Simmonds P, and Katzourakis A. A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses. *Current Biology* 31: 4689-4696e5, 2021.

MG wrote the chapter and conducted all the analysis under the supervision from AK. PS prepared the hepatitis C virus alignments. All authors provided editorial assistance.

Chapter 3. Reconstructing the deep evolutionary history of simian immunodeficiency virus

MG wrote the chapter and conducted all the analysis under the supervision from AK. PS prepared the HIV and SIV alignments. All authors provided editorial assistance.

Chapter 4. Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1

Published as: **Ghafari M**, du Plessis L, Raghwani J, Bhatt S, Xu B, Pybus O, and Katzourakis A. Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza. *Molecular Biology and Evolution* 39: msac009, 2022.

MG wrote the chapter and conducted all the analysis under the supervision from AK. All authors provided editorial assistance.

Chapter 5. Investigating the evolutionary origins of the first three SARS-CoV-2 variants of concern

MG wrote the chapter and conducted all the analysis under the supervision from AK and DBW. All authors provided editorial assistance.

Chapter 6. Lessons for preparedness and reasons for concern from the early COVID-19 epidemic in Iran

Published as: **Ghafari M**, Hejazi B, Karshenas A, Dascalu S, Kadvidar A, Khosravi MA, Abbasalipour M, Heydari M, Zeinali S, Ferretti L, Ledda A, and Katzourakis A. Lessons for preparedness and reasons for concern from the early COVID-19 epidemic in Iran. *Epidemics* 36: 100472, 2021.

MG wrote the chapter and conducted all the analysis under the supervision from AK. MAK, MA, and SZ prepared the genomic samples from Iran. All authors provided editorial assistance.

Chapter 7. Inferring SARS-CoV-2 transmission dynamics using excess mortality data

MG wrote the chapter and conducted all the analysis under the supervision from AK.

Part a, published as: **Ghafari M**, Kadivar A, and Katzourakis A. Excess deaths associated with the Iranian COVID-19 epidemic: A province-level analysis. *International Journal of Infectious Diseases* 107: 101-115, 2021.

MG wrote the manuscript and conducted all the analysis under the supervision from AK. All authors provided editorial assistance.

Part b, accepted for publication in *Nature Communications*: **Ghafari M**, Watson OJ, Karlinsky A, and Katzourakis A. A framework for reconstructing SARS-CoV-2 transmission dynamics using excess mortality data.

MG wrote the manuscript under the supervision from AK. MG and OJW contributed equally to the analysis. AKar and MG performed the calculations for weekly excess mortality. OJW and AKar provided assistance for generating the figures.

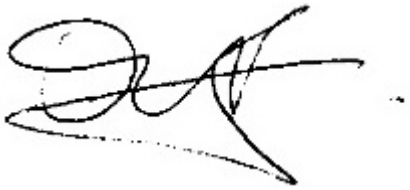
Other manuscripts published during my DPhil that are not related to this thesis:

- **Ghafari M**, Lumby CK, Weissman DB, and Illingworth CJR. Inferring transmission bottleneck size from viral sequence data using a novel haplotype reconstruction method. *International Journal of Virology* 94: e00014-20, 2020.
- Wikramaratna PS, Paton RS, **Ghafari M**, and Lourenco J. Estimating the false-negative test probability of SARS-CoV-2 by RT-PCR. *Eurosurveillance* 25: 2000568, 2020.
- Dascalu S, Flammer PG, **Ghafari M**, Henson SC, Nascimento R, Bonsall MB. Engaging religious institutions and faith-based communities in public health initiatives: a case study of the Romanian Orthodox Church during the COVID-19 pandemic. *Frontiers in public health* 9: 768091, 2021.
- de Oliveira T, Lutucuta S, Nkengasong J, Morais J, Paixco JP, Neto Z, Afonso P, Miranda J, David K, Inglas L, Katzourakis A, **Ghafari M**, Singh L, Tshiabuila D, Martin D, Lessells RJ. A novel variant of interest of SARS-CoV-2 with multiple spike mutations detected through travel surveillance in Africa. *medRxiv* 21254323, 2021.
- Hinch R, Panovska-Griffiths J, Probert WJM, Ferretti L, Wymant C, Di Lauro F, Baya Nikolas, **Ghafari M**, Abeler-Dorner L, The COVID-19 Genomics UK (COG-UK) consortium, Fraser C. Estimating SARS-CoV-2 Variant Fitness and the Impact of Interventions in England using Statistical and Geo-Spatial Agent-Based Model. Accepted: *Philosophical Transactions of the Royal Society A*, 2022.

As Mahan's supervisors, we confirm that the information presented here is representative of his contributions to the work listed in this thesis.



Professor Aris Katzourakis
Professor of Evolution and Genomics
Department of Zoology
University of Oxford



Professor Oliver G Pybus
Professor of Evolution and Infectious Disease
Department of Zoology
University of Oxford

Table of Contents

Chapter 1: Introduction 1
1.1 Abstract 1
1.2 Molecular clock and neutral theory 2
1.3 Molecular dating and time dependency of evolutionary rate estimates 3
1.3.1 Calibration methods 4
1.3.2 Time-dependency of evolutionary rate estimates 9
1.4 Sources of variation in molecular rate estimates 12
1.4.1 Synonymous vs. nonsynonymous substitution rate 12
1.4.2 Site saturation 13
1.4.3 Transient mutations 14
1.4.4 Complex adaptation 16
1.4.5 Other factors 18
1.5 Empirical applications of molecular dating methods 19
1.6 Alternative methods to estimating viral dynamics 20
1.7 Thesis outline 24
1.8 References 27
Chapter 2: A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses 34
2.1 Abstract 34
2.2 Introduction 35
2.3 Results 37
2.3.1 Power-law rate decay can emerge due to site saturation 34
2.3.2 Conundrum of rate calibrations 35
2.3.3 The Prisoner of War model of virus evolution 37
2.4 Discussion 49
2.5 Methods 52
2.5.1 Resource availability 52
2.5.2 Data sets and simulation codes 52
2.5.3 Mathematical formulation 56
2.6 References 65
Chapter 3: Deep evolutionary history of simian and prosimian immunodeficiency viruses 69

3.1 Abstract 69
3.2 Introduction 70
3.3 Results 73
3.4 Discussion 78
3.5 Methods 81
3.5.1 Data sets 81
3.5.2 Molecular clock analyses and phylogeny reconstruction 82
3.6 References 84
Chapter 4: Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza 87
4.1 Abstract 87
4.2 Introduction 88
4.3 Results 91
4.4 Discussion 98
4.5 Materials and Methods 100
4.5.1 Phylogenetic analyses 101
4.5.2 Estimating the site frequency and number of non-neutral sites 102
4.5.3 Resource availability 103
4.6 References 104
Chapter 5: Investigating the evolutionary origins of the first three SARS-CoV-2 variants of concern 108
5.1 Abstract 108
5.2 Introduction 108
5.3 Results 110
5.3.1 Emergence of VOCs: an evolutionary dilemma 111
5.3.2 Between-host model of VOC emergence 111
5.3.3 Within-host model of VOC emergence 112
5.3.4 Fitness landscapes 114
5.3.5 Landscape 1: single mutations 116
5.3.6 Landscape 2: additive mutations 120
5.3.7 Landscape 3: fitness plateau crossing 124
5.4 Discussion 128
5.4.1 Cryptic transmission of VOCs in humans 129

5.4.2 Possibility of reverse zoonosis 130
5.4.3 Role of recombination 131
5.4.4 Shifting landscape 131
5.4.5 Future VOCs 132
5.5 Methods 134
5.5.1 Between-host model of VOC emergence 134
5.5.2 Within-host model of VOC emergence 137
5.5.3 Simulation setup 137
5.6 References 139
Chapter 6: Lessons for preparedness and reasons for concern from the early COVID-19 epidemic in Iran 143
6.1 Abstract 143
6.2 Introduction 144
6.3 Results 147
6.3.1 The situation in Iran 147
6.3.2 Origins of the epidemic in Iran: a phylogenetic analysis 149
6.3.3 Estimating the under-reporting of incidence using air-travel data 150
6.3.4 Estimating the under-reporting of cumulative deaths using excess mortality data 153
6.3.5 Early-stage importation risk from China 155
6.3.6 Evaluating effectiveness of intervention measures and the dynamics of the outbreak 155
6.4 Discussion 157
6.5 Methods 162
6.5.1 Data on aggregate number of cases and deaths 162
6.5.2 Phylogenetic analysis 163
6.5.3 Estimating point-prevalence using air travel data 164
6.5.4 Estimating seasonal excess mortality 165
6.5.5 Evaluating the importation risk from China in January 166
6.5.6 SEIR model 167
6.5.7 Resource availability 169
6.6 References 170
Chapter 7: Inferring SARS-CoV-2 transmission dynamics using excess mortality data 176

Part a: Excess deaths associated with the Iranian COVID-19 epidemic: A province-level analysis 176
7a.1 Abstract 176
7a.2 Introduction 177
7a.2.1 Historical trends of all-cause mortality in Iran 177
7a.3 Results 180
7a.3.1 Did the COVID-19 epidemic in Iran start during fall 2019? 180
7a.3.2 The geographical spread of COVID-19 in Iran 181
7a.3.3 Estimating the level of exposure across the population 184
7a.4 Discussion 188
7a.4.1 Limitations 191
7a.5 Methods 192
7a.6 References 201
Part b: A framework for reconstructing SARS-CoV-2 transmission dynamics using excess mortality data 204
7b.1 Abstract 204
7b.2 Introduction 204
7b.3 Results 206
7b.4 Discussion 216
7b.5 Methods 218
7b.5.1 Excess mortality 218
7b.5.2 Transmission model 219
7b.5.3 Data availability 222
7b.6 References 224
Chapter 8: Discussion 230
8.1 Conclusions and final thoughts 230
8.2 Future directions 236
8.3 Concluding remarks 240
8.4 References 241
Appendices 244
Appendix 1: Supplementary material for chapter 2 244
Appendix 2: Supplementary material for chapter 4 251
Appendix 3: Supplementary material for chapter 5 255

Appendix 4: Supplementary material for chapter 6 262
Appendix 5: Estimates of anti-SARS-CoV-2 antibody seroprevalence in Iran 276
Appendix 6: Supplementary material for chapter 7 part b 281

Chapter 1: Introduction

1.1 Abstract

The observation of a constant rate of protein evolution in the 1960s by Zuckerkland and Pauling gave rise to the notion of the existence of a so-called 'molecular evolutionary clock' which formed the basis for using genetic data to estimate the timescale of major evolutionary events in various species. It was also an instrumental piece of evidence in support of the neutral theory of molecular evolution developed by Kimura in 1968. Over the last two decades, the availability of high throughput sequencing data has expedited the widespread applications of molecular evolutionary clocks across a range of fields. However, genomic data has revealed that estimated rates of molecular evolution in many biological systems are dependent on the time frame of measurement, as elucidated by Ho and many others. This has caused some confusion and ambiguity in the literature about the underlying causes of time-dependent rate phenomenon and how to validate and compare rates that are estimated across different timescales. One of the major implications of having a time-dependent rate of molecular evolution is that the estimated rates over a given time window may not be transferable to analyses of other time windows. In this chapter, I describe the concept of the molecular clock, some of the common calibration methods used in molecular dating, and the underlying biological mechanisms that can influence the estimated rate of molecular evolution. Finally, I describe alternative methods to inferring evolutionary and demographic history of pathogens over time using mobility, all-cause mortality, and serological data.

1.2 Molecular clock and neutral theory

The first evidence of the molecular evolutionary clock emerged in the early 1960s when Zuckerkandl and Pauling were investigating haemoglobin protein similarities in various species including cow, pig, and bony fish among other animals [1-3]. What they found was that the divergence in the amino acid sequence of haemoglobin increased with the evolutionary distance between species. Based on this observation, they made the assumption that the rate of amino acid changes is constant over time. To calibrate the molecular clock, they used the paleontological estimate of 100-160 million years for the horse-human divergence and, given the observation of 18 amino acid differences separating human and horse, inferred an evolutionary rate of 1 amino acid substitution per ~15 million years. They then applied their clock estimate to the globin genes of other animals and suggested that the gorilla-human split may have occurred nearly 11 million years ago. Studies on other proteins such as cytochrome *c* and fibrinopeptide later showed similar patterns of amino acid change over time, further strengthening the evidence for clocklike evolution of proteins [4-6].

These early studies laid the foundations for the neutral theory of evolution, proposed by Kimura in 1968, which states that the rate at which new mutations reach fixation only depends on the rate at which they first appear in a population [7]. In other words, the fate of a neutral mutation in a population is solely determined by genetic drift. To illustrate this, let us consider a haploid population of size N which produces new mutations at rate u per generation. We know from population genetics theory that the probability of fixation of a de novo neutral mutation (starting at size 1) is $1/N$ [8]. Thus, with Nu new mutations being produced every generation, their fixation rate is $Nu \times 1/N = u$. This implies that the rate of substitution of a neutral

mutation is the same as the mutation rate and is independent of the population size. Although in Kimura's original proposal he defined neutral theory based on the constancy of evolutionary rate *per unit time*, he later recognised that the neutral theory should predict a constant substitution rate *per generation* [9].

The concept of neutral theory of molecular evolution was highly contested among biologists who, at the time, considered selection to be the dominant evolutionary force behind amino acid substitutions. This led to heated debates about the neutralist vs. selectionist view of evolution which largely revolved around the relative percentages of polymorphic and fixed alleles that are neutral or under selection (for a recent review, see [1] and references therein). There is now a growing body of evidence suggesting that some features of molecular evolution clearly point towards evidence in favour of selection which cannot be explained by the neutral theory [11]. Nevertheless, neutral theory had a profound impact on the field of molecular evolutionary biology and led to the development of several scientific frameworks to contrast the impact of natural selection and genetic drift in the study of molecular data [12]. In particular, several statistical approaches were developed to explicitly measure the amount of adaptive evolution within a species and to test whether any observed patterns of within-species polymorphism and between-species divergence are consistent with the null model of neutral evolution [13-15].

1.3 Molecular dating and time dependency of evolutionary rate estimates

The molecular clock gradually moved away from its role in the selectionist-neutralist debate and became increasingly appreciated for its applications in evolutionary biology as a tool for inferring evolutionary rates and timescales. Like any other clock,

the molecular clock also requires a source of information for its calibration. Here, I explain some of the commonly used molecular clock calibration methods.

1.3.1 Calibration methods

Zuckerkandl and Pauling transformed our understanding of evolutionary timescales by using genetic data to estimate divergence time between different species for the first time. In its simplest form, the molecular clock assumes a single clock that is homogeneous across all taxa and takes a linear relationship between the amount of divergence (i.e., Hamming distance), d, and time since the common ancestor, t, such that for any pair of sequences evolving at rate u, the amount of divergence is given by $d=ut$, as long as every de novo mutation occurs at a unique site (infinite site model). Unless we have a priori information about the evolutionary rate, u, molecular dating methods require the clock to be calibrated in order to estimate dates in absolute time. There are multiple methods to calibrate the molecular clock.

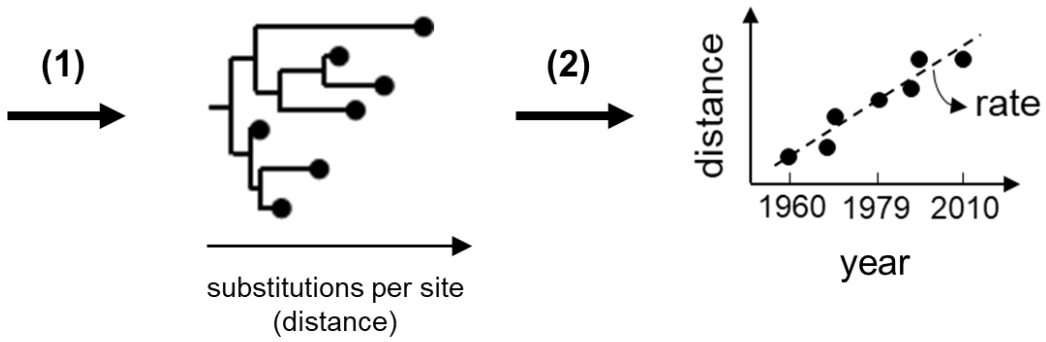
(i) Heterochronous sampling: One of the main sources of information for calibrating the molecular clock in rapidly evolving viruses and bacteria is the age of the molecular samples -- such genomic data sets with known sampling times (also known as heterochronous or time stamped data sets) [16, 17]. Crucially, for the samples to be useful for molecular clock analysis, they must cover a sufficiently broad window of observation such that the population has measurably evolved within that period [18-20].

The root-to-tip regression method was one of the earliest and most common molecular clock approaches using heterochronous data to infer the time to the most recent common ancestor of HIV pandemic strains [21]. Figure 1.1A shows a generic root-to-tip regression analysis which consists of two steps: (1) inferring a

phylogenetic tree using maximum likelihood or Bayesian approaches and (2) assuming a constant rate of evolution over time to infer divergence times. One can then fit a line of least squares to the distance from the root of the tree to each of the tips as a function of their respective sampling times to find the evolutionary rate (slope of the line), and the age of the most recent common ancestor (x-intercept). The root-to-tip regression method has several shortcomings including the use of a single clock rate across all branches of the tree which does not take into account rate variations among lineages (e.g., the impact of selection; see Figure 1.1B). It also does not correctly capture the uncertainty in rate estimations as it discards information regarding the branching order of the tree.

More sophisticated approaches based on maximum likelihood or Bayesian molecular clock incorporate the uncertainty in estimates of tree topology, branch lengths, demographic dynamics, and rate variations over time. In particular, challenges to strict clock models that use a single homogenous rate across all taxa have driven the development of relaxed clock models which, as the name suggests, are able to relax the assumption of rate homogeneity among lineages [22, 23]. There is now a wide range of relaxed-clock models, implemented in various Bayesian phylogenetic programs such as BEAST [24, 25].

A



B

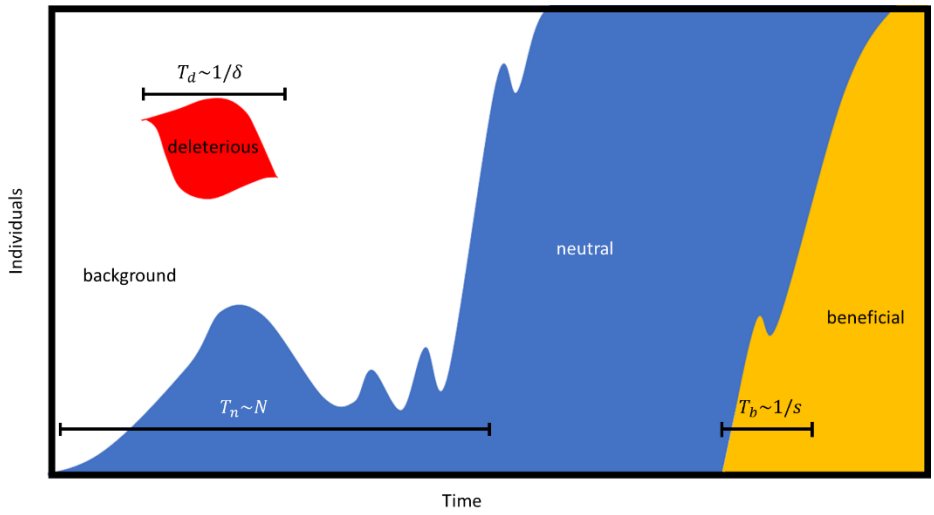


Figure 1.1**: Molecular evolutionary rate estimates based on a root-to-tip regression method and population genetic factors influencing molecular clock rate**. **(A)** Estimating rates of molecular evolution using a regression-based approach. (1) A nucleotide substitution model is applied to sequences sampled at different time points and used to estimate a phylogeny whose branch lengths represent estimated genetic distances. (2) The estimated tree is rooted, and linear regression is used to fit the root-to-tip genetic distances to the sampling date of each sequence. The gradient of the regression line is an estimate of the molecular evolutionary rate. **(B)** The classic interpretation of molecular evolutionary rates. New mutations are produced every generation and may reach high frequencies or die out over time depending on their selective coefficient. In the case of a constant population of size *N*, the probabilities and rates of fixation depend on whether a de novo mutation is deleterious (red) with fitness disadvantage δ, beneficial (yellow) with fitness advantage *s*, or neutral (blue) relative to the wild type (white) population. Horizontal black lines show the typical timescales for persistence of a segregating deleterious mutation, $T_d \sim 1/\delta$ generations, fixation of a neutral mutation, $T_d \sim N$ generations, and fixation of a strongly beneficial mutation, $T_d \sim 1/s$ generations. Transient mutations can elevate the rate over short timescales before they are purged from the population while adaptive substitutions can result in much quicker fixation events and increase the inferred substitution rates relative to predictions based on neutral theory.

Ancient DNA data is another source of data for heterochronous sampling methods that can be used for clock calibrations based on samples that are hundreds or even thousands of years old. Newly developed next-generation sequencing methods have allowed us to sequence ancient human populations and perform direct analysis of genetic relationships between contemporary humans, Neanderthals, and Denisovan hominins among other archaic human population groups in the last hundred thousand years [26-28]. Some ancient DNA samples have known ages, for instance as documented dates of collection, but others are too old for their ages to be known exactly. The latter can be dated using radiocarbon dating techniques [29] which adds a degree of uncertainty in the age of the sample that can also be incorporated in phylogenetic analysis of molecular clock measurements [30]. Ancient DNA studies have also contributed to investigations of evolutionary history of viruses, including smallpox and HBV in 1600s, 1918 H1N1 influenza A pandemic, and the hepatitis B virus samples collected from the Bronze age to the Medieval period [31-33].

(ii) Biogeographic calibration: Geological events can also be used for calibration provided that they can be tied to evolutionary divergence events among the sampled taxa [34]. These divergences can be a result of the formation of connections between previously isolated habitats or dispersal across barriers and establishment of new populations such as the diversification associated with continental fragmentation (i.e., pieces of continents that have broken off from main continental masses to form distinct islands) or the colonisation of newly formed islands. One of the examples of using biogeographic calibrations in the analysis of virus evolutionary history is the study of simian immunodeficiency virus (SIV) sampled in Drills on the island of Bioko in Equatorial Guinea, which were isolated

from mainland Africa as sea levels rose nearly 10,000 years ago [35]. This geological event enabled researchers to investigate evolutionary history of SIVs over longer timescales.

(iii) Endogenous viral elements: The discovery of copies of DNA and RNA viruses integrated in the genomes of plants and animals, also known as endogenous viral elements (EVEs), has led to the use of EVE integration events as a calibration method to study evolution of viruses over longer periods [36-38]. By vertically passing from parents to offspring, they become part of the host genomes and create a genomic fossil record that stretches back to millions of years. Recent characterisation of genomes of mammals and birds revealed that genetic material derived from all known viral genome types and replication strategies can enter the animal germ line, greatly broadening the scope of paleovirological studies [39].

(iv) Host co-divergence: In some cases, the timescale of pathogenic organisms can be inferred by assuming a strong degree of co-divergence with the evolution of the host organism. For instance, in bacterial populations, co-divergence of endosymbionts and their eukaryotic hosts can be as a calibration point for estimating their long-term evolutionary rates [40]. In viruses, predictions on the longevity of virus lineages from EVE fossil records are further supported by observations of the co-speciation of viruses and hosts [41]. For example, foamy viruses have a long history of co-speciation with their hosts stretching back to more than a hundred million years [42-44]. Other examples include co-evolution of papillomaviruses and spumaviruses with their hosts, consistent with virus-host co-speciation over 400 million years [45, 46]. However, there is also evidence of multiple host-switching events in pathogens which questions the general reliability of co-evolution assumptions for all pathogens [47-49].

(v) Other calibration methods: Perhaps the most commonly used methods of calibration, particularly in vertebrates, are based on palaeontological evidence where a minimum constraint on the age of a clade is based on the timing of its earliest fossil representative [50, 51]. There are various other calibration methods (see [52] for a recent review on calibration methods) including those based on archaeological and anthropological evidence. For instance, diagnoses of human infection from archaeological sites and historical documentations have been used to provide a calibration point for the analysis of evolutionary history of smallpox in China in the 4th century AD [53].

1.3.2 Time-dependency of evolutionary rate estimates

From large-scale meta-analyses based on the implementation of molecular dating methods for various organisms across different timescales, it became increasingly clear that the evolutionary rate estimates are negatively correlated with the timescale of measurement [52]. This relationship was also observed in both RNA and DNA viruses [58] and led to the notion of time-dependent rate phenomenon (TDRP) in the field of molecular evolution. The immediate implication of TDRP is that a rate estimate over one time window may not be transferable to analyses of other time windows. Therefore, a naïve extrapolation of short-term rate estimates can lead to significant underestimation of the timing of ancient divergence events. Similarly, using calibrations based on long-term rate estimates causes the timing of more recent evolutionary events to be overestimated (Figure 1.2). The earliest evidence in support of TDRP was from palaeontological studies where disparities between date estimates based on molecular data and fossil records were observed and also genetic studies of carnivores, humans, and birds [54-57]. Subsequent

studies showed evidence of TDRP across a broad range of taxa including worms, insects, fish, primates, and various multicellular organisms (see [57] for a review).

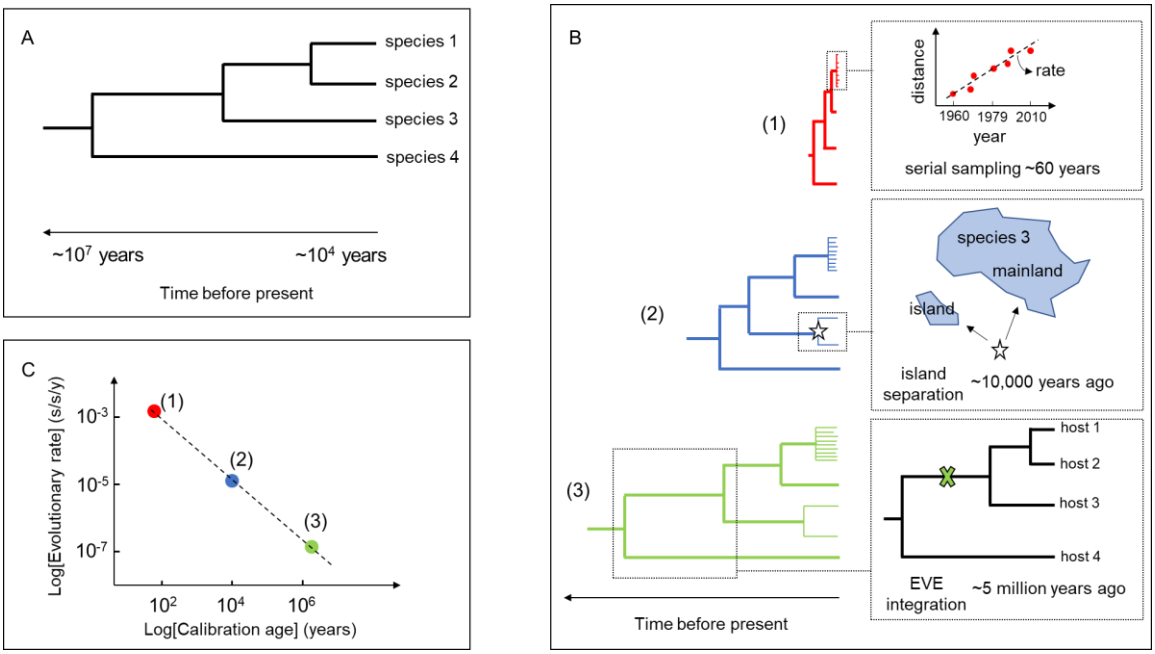


Figure 1.2: **Time dependency of evolutionary rate estimates using different calibration methods**. (**A**) A hypothetical phylogenetic tree of four virus species with known divergence times. (**B**) An attempt to reconstruct the time tree in (A) based on (1) short-term rate calibrations using heterochronous data from sequences within a subpopulation of species 1, (2) biogeographical rate calibrations of island separation for subspecies 3 found in mainland and island, and (3) integration of an endogenous viral element into the common ancestor of host species of viruses 1, 2, and 3 (marked with an "X" on the tree) which took place approximately 5 million years ago. (**C**) Evolutionary rate estimates of the 4 virus species based on the three rate calibration methods shown in (B). The dashed line represents a power-law dependency between the rate estimates and calibration age as demonstrated by Aiewsakun and Katzourakis [58].

Perhaps one of the most well-documented pieces of evidence in support of TDRP in viruses includes the evolutionary rate estimates for primate lentiviruses, a group of pathogens that include SIV. Estimates of the age of primate lentiviruses based on the use of heterochronous sequences are up to hundred times younger [60] than those based on biogeographic calibrations [35]. Even estimates based on the biogeographic calibration (separation of the island of Bioko from mainland which took place nearly 10,000 years ago) are roughly one hundred times younger than estimates based on paleovirological estimates that show the presence of endogenous lentiviruses in lemurs indicating that these viruses have circulated in some primates for at least hundreds of thousands of years [61]. Further evidence in support of ancient history of lentiviruses come from the presence of a rabbit endogenous lentivirus which suggests that lentiviruses may have been present in mammals for at least 12 million years [62, 63].

There is also evidence in support of TDRP in bacterial populations, although there have been fewer studies on the evolutionary rates of bacteria across long timescales. It has been shown that the ratio of nonsynonymous to synonymous substitutions are strongly correlated with the timespan of observation in several related bacterial taxa [64]. Using calibration methods based on host co-speciation shows that endosymbiotic bacteria evolve at an apparent rate that is much lower than the standard rate. However, rate estimates for *Vibrio cholera* based on heterochronous sequences during the cholera pandemics are hundred folds higher than standard rates used for calibrations [65].

1.4 Sources of variation in molecular rate estimates

Many factors can influence the evolutionary rate estimates across different timescales which can lead to TDRP. Here, I summarise some that are particularly relevant for viruses and other rapidly evolving pathogens.

1.4.1 Synonymous vs. nonsynonymous substitution rate

From the 1970s to 1990s, there was a shift in focus from measuring molecular clock rates in protein sequences to DNA sequences [66]. This led to a growing body of evidence on the discrepancy between the rate of synonymous and nonsynonymous substitutions in DNA. Many studies showed that while in accordance with the standard principles of neutral theory the rate of synonymous substitutions per generation was constant (i.e., organisms with shorter generations had more synonymous changes in a given time window), nonsynonymous changes appeared to grow at a constant rate per year [67, 68]. This discrepancy between the rate of synonymous and nonsynonymous substitutions was partly reconciled in the nearly neutral theory of evolution by Ohta [69, 70] which proposes that, conditional on the existence of an inverse relationship between effective population size and generation time of a given organism, if most amino acid substitutions are mildly deleterious, it would disproportionately increase the rate of effectively neutral substitutions in smaller population. This can, in turn, offset the effect of longer generation times in smaller population sizes resulting in a roughly constant rate of nonsynonymous substitutions per unit time. On the other hand, because noncoding (synonymous) changes tend to be more neutral, irrespective of population size, their rate of evolution only depends on generation time.

In quantitative terms, the nearly neutral theory proposes that the rate of substitutions, ρ, in a haploid population is given by $\rho = N_e u P_{fix}$ where N_e is the effective population size, u is the mutation rate per generation, and $P_{fix} = \frac{1-e^{-s}}{1-e^{-sN_e}}$ is the probability of fixation of de novo mutations with selective coefficient s. Under strict neutrality (i.e., $|s| \ll 1/N_e$), $P_{fix}=1/N_e$. Therefore, we can recover the results from neutral theory (i.e., $\rho = u$). For mutations that are highly deleterious (i.e., $-s \gg 1/N_e$), P_{fix} declines exponentially with increasing N_e. Thus, there is no contribution from these mutations to increasing substitution rate, ρ. However, mutations that are 'nearly neutral' (i.e., $|N_e s| \sim 1$) can fix in smaller populations through genetic drift while being purged by selection in larger population sizes. For example, in a haploid population with effective size N_e=1,000 the probability of fixation of a deleterious mutation with selective disadvantage s=-0.001 is nearly 58% of the fixation probability of neutral mutations, whereas for a population with an effective size N_e=10,000, there is only a 0.05% relative chance for deleterious mutations to reach fixation.

1.4.2 Site saturation

Site saturation (also known as substitution saturation or mutational saturation) is the result of repeated substitutions occurring at the same site over time which erodes historical evolutionary signal in sequence data and results in the underestimation of amount of genetic change that has happened in a sequence, particularly over longer evolutionary timescales (Figure 1.3A). Site saturation also results in long branch attraction, a form of systematic bias in phylogenetic analysis where distant lineages have misleadingly short branches [71].

To illustrate the impact of site saturation on molecular clock estimates, let us consider a pair of sequences that have diverged from their most recent common ancestor $t/2$ generations ago and have since been evolving under a constant and homogenous substitution rate, u. Therefore, the proportion of pairwise differences, $p(t)$, initially grows linearly with time (i.e., $p(t)=ut$). However, given that there is a finite number of evolving sites, there comes a time, $t^* \sim 1/u$, when every new substitution occurs on the background of a site that has already changed at least once (i.e., those sites have reached the saturation point). If the pair diverge for timescales $t \gg t^*$ without having any intermediate samples taken from them, any estimation of molecular clock rate based on the amount of observed number of substitutions is going to significantly underestimate the true rate, u. In chapter 2, I investigate how substitution saturation can lead to a power-law decay in the estimated rate of molecular evolution over time.

1.4.3 Transient mutations

Particularly over short timescales, evolutionary rate estimates can be elevated by the inclusion of deleterious mutations, which tend to be removed from the population by purifying selection over longer periods (Figure 1.3B). To provide a quantitative explanation for why deleterious mutations only have a transient impact on evolutionary rate estimates, let us consider a simple evolutionary scenario where mutations are sufficiently rare in a population such that they do not influence each other's chance of reaching fixation (i.e., isolated mutations). Under such conditions, the dynamics of a de novo mutation with fitness disadvantage δ can be described by a continuous-time branching process such that at every generation each mutant individuals can go through a binary fission at rate $1-\delta$ or die at rate 1. Therefore, the probability that a

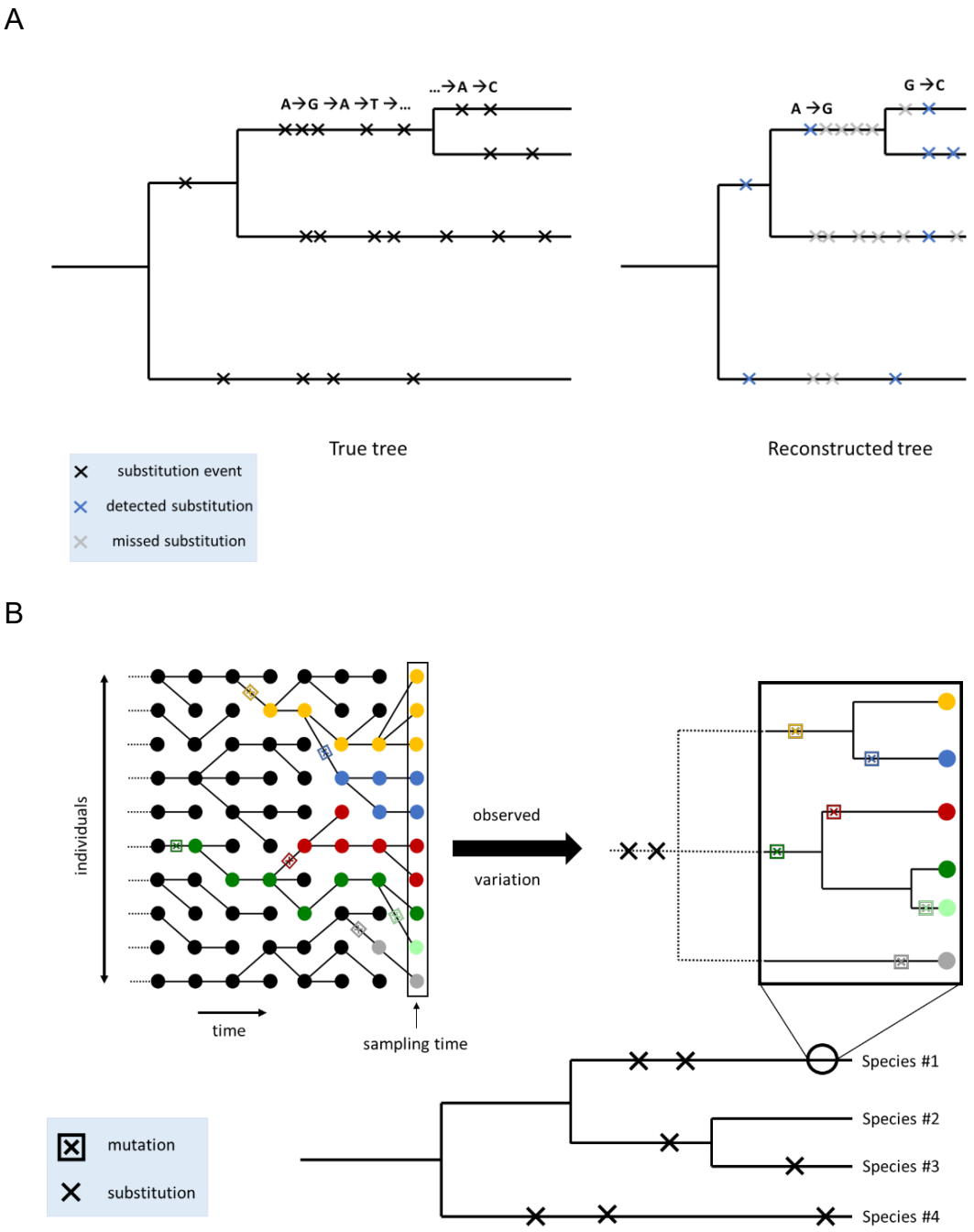


Figure 1.3: **Examples of the underlying factors influencing evolutionary rate estimates. (A)** Over long timescales, large divergence times between collected sequences may result in multiple hits and numerous substitutions going undetected (gray). As a result, fewer substitutions are observed (blue) which gives an apparent drop in inferred substitution rates. (**B**) Over short timescales, many individuals carry mutations (coloured circles) that represent transient polymorphisms that are removed from the population by fixation or elimination over longer timescales (species tree). Many of these are expected to be slightly deleterious mutations. Thus, younger branches carry an excess of sequence changes, thereby elevating the apparent estimated evolutionary rate. Sequence changes on older internal branches better represent true substitution events (balancing and frequency-dependent selection notwithstanding). Over the shortest of timescales the estimated rate across the whole population should approach (but never exceed) the rate at which mutations are introduced into the population.

mutant lineage persists for at least *t* generations in the population is given by [72]

$$P(T>t)=\left(1+\frac{e^{\delta t}-1}{\delta}\right)^{-1}\approx\begin{cases}1/t & \text{for } 1\ll t\ll 1/\delta\\ \delta e^{-\delta t} & \text{for } t\gg 1/\delta\end{cases}.$$

This suggests that the dynamics of a deleterious mutant lineage resembles an effectively neutral population for a transient period as long as $t\ll 1/\delta$ -- note that under neutrality, the probability that a mutant lineage persists for *t* generations is approximately equal to 1/*t* [8]. However, as the lineage persists for a longer period of time $t\gg 1/\delta$ and, thereby, reach a larger size, its 'deleteriousness' would be felt and the probability that it goes extinct grows exponentially (i.e., it will almost never grow to a size much larger than $1/\delta$). Therefore, during a transient period where the mutant is effectively neutral (i.e., $1\ll t\ll 1/\delta$), the molecular rate estimates from the population will be elevated with respect to a purely neutral background. Nicolaisen and Desai showed [73] that even under more complex evolutionary dynamics where one can relax the assumption about mutations being rare, purifying selection still creates similar dynamics whereby the effective substitution rate is elevated relative to a neutral rate over a period of time that is inversely proportional to the selective coefficient of the deleterious mutants. In chapter 3, I show how purifying selection impacts the evolutionary rate estimates in pH1N1 influenza and SARS-CoV-2 over the span of one year.

1.4.4 Complex adaptation

Another key evolutionary process that can influence the molecular rate estimates is complex adaptation. While standard evolutionary adaptation proceeds by the sequential fixation of beneficial mutations, some adaptations may have non-adaptive origins where multiple individually neutral or deleterious mutations need to

interact together (i.e., epistasis) in order for the adaptation to take place [74]. Under such conditions, at least some individuals must first acquire one or more neutral or deleterious intermediate mutations and cross a metaphoric fitness plateau or valley before reaching the higher-fitness state. Such evolutionary paths are inherently complex in the sense that none of the intermediate mutations have a selective advantage relative to the background population. Classic population genetics theory suggests such adaptive changes are very rare and only occur at very slow rates in small populations where the effect of genetic drift is strong [75]. However, several recent studies have demonstrated that complex adaptations can occur much faster than what we would expect in larger populations [76-78].

There are two general ways a population can cross a fitness valley (if intermediate mutations are all deleterious) or plateau (if intermediate mutations are neutral). Either (i) each of the intermediates sequentially fix over time via genetic drift as predicted by the shifting balance theory [75] or (ii) intermediates continue to circulate at relatively low frequencies for an extended period of time, through a process known as stochastic tunnelling [79], before producing the necessary combination of mutations that provide a selective advantage. The former evolutionary path has the molecular clock signature that resembles the dynamics proposed by the nearly neutral theory, whereas the latter has a very different impact on the molecular clock. Stochastic tunnelling creates an evolutionary dynamic where a group of individually neutral or deleterious mutations can reach fixation together at the same time, thereby creating a transient elevation in the rate of molecular evolution. In chapter 5, I investigate whether the clustered emergence of three SARS-CoV-2 variants of concern in late 2020 with many extra mutations

relative to the background population from which they first emerged could have been a result of complex adaption or not.

1.4.5 Other factors

Besides the biological processes mentioned above, several other factors may cause the rate of molecular evolution to vary with time. One such factor is sequencing error which can create an apparent increase in sequence diversity and in turn elevate the amount of observed evolutionary changes over short timescales [96, 97]. In viruses, adaptive changes resulting from positive selection due to recent host-virus evolutionary arm races can also lead to lineage-specific elevated rates [98, 99]. Misspecification of calibration points used for rate estimation can result in overestimation (underestimation) of true rates if those points are incorrectly estimated to be younger (older) than their true age [100, 101].

Over longer timescales, apart from substitution saturation which I explained in the previous section, choosing an incorrect substitution model that underestimates the amount of genetic distance between samples or does not account for rate heterogeneity across sites can lead to an apparent drop in the evolutionary rate estimates [102]. Substitution model misspecification is perhaps the single most important factor biasing rate estimations over longer timescales. This is because the two main assumptions behind all conventional substitution models is that (1) the process of nucleotide or amino acid changes is purely Markovian (memoryless) and (2) that every site evolves independently over time. Such models do not account for complex interactions between different parts of the genome such as epistasis or gene overlapping which are likely to alter the pattern of substitutions across the

genome in very complex and non-Markovian ways (i.e., future changes in the genome are dependent on the changes that came before them) [103, 104].

1.5 Empirical applications of molecular dating methods

Accurate estimation of rates of nucleotide or amino acid substitutions is one of the fundamental aspects of molecular evolutionary biology. In population genetics, estimates of substitution rates allow us to evaluate the effects of genetic drift and natural selection, and to understand the processes of population divergence and speciation. At the macro-evolutionary scale, substitution rates can be used for molecular clock dating, which enables us to infer phylogenetic histories on a natural timescale of years or millennia. This has led to the widespread popularity across a range of disciplines of statistical methods that infer substitution rates from molecular sequence data.

There is a large body of literature on nucleotide and amino acid substitution rate estimates for both RNA and DNA viruses [105-108]. One of the earliest examples of molecular dating methods in viruses was the analysis of enterovirus evolutionary history where Takeda and colleagues used a substitution rate of ~$5x10^{-3}$ substitutions per site per year to date the start of the outbreak using a linear relationship between the amount of genetic distance from the tree root of sequences of enterovirus and their respective collection dates [109]. In another example, a nucleotide substitution rate of ~$5x10^{-4}$ substitutions per site per year to calculate the dates of emergence of various genotype 2 subtypes of hepatitis C virus back to more than 500 hundred years ago. This finding has major implications for understanding the association of genotype 2 infections in areas where the slave trades operated several hundred years ago [110].

One of the more recent examples of using molecular clock rates was during the 2014-2016 West Africa Ebolavirus epidemic where there was controversy and concern regarding the virus' "mutation rate", because early estimates of the substitution rate (~$1.9x10^{-3}$ substitutions per site per year) from the epidemic appeared to be approximately twice the average rate across multiple outbreaks (~$0.9x10^{-3}$ substitutions per site per year) [111, 112]. This led to a debate among scientists who, at the time, believed the elevated rates were a sign that the virus is becoming more transmissible [113]. However, it was later showed that this was merely a sign of time-dependent rate effects due to purifying selection whereby an excess of mildly deleterious variants caused the inferred rate to be higher over shorter timescales [112].

Most recently, during the current COVID-19 pandemic, phylogenetic analysis of SARS-CoV-2 allowed researchers to determine the evolutionary rate of ~$1.1x10^{-3}$ substitutions per site per year for the virus and estimate the starting date of the COVID-19 pandemic to late November - early December 2019 [114, 115].

1.6 Alternative methods to estimating viral dynamics

Viruses, particularly RNA viruses, are unique because of their ability to grow rapidly in size and accumulate significant amount of genetic variation over a short period of time in a host population. Therefore, their evolutionary dynamics are highly influenced by the mode of transmission as well as spatial and temporal dynamics of spread. Viral phylodynamics provide a powerful framework to study how various epidemiological, evolutionary, and immunodynamic processes come into play to shape viral dynamics [80-82]. However, in the absence of genomic data, other methods that leverage mobility, all-cause mortality, and seroprevalence data can

provide insight into viral transmission dynamics. These approaches may not only help us find, say, the starting date of an epidemic or its growth rate and prevalence over time but also allow us to better understand viral phenotypes such as transmissibility and virulence. Furthermore, nonrepresentative genomic samples can lead to biased inference of epidemiological parameters when important spatiotemporal gaps in the samples are present [83]. Under such circumstances, independent analyses based on epidemiological methods can improve the inferential power and accuracy of phylodynamic methods.

Real-time mobility and detailed case data including travel history can play a crucial role in understanding the pattern of disease spread. In particular, we can use mobility data to understand the role of case importation and non-pharmaceutical interventions in altering the mode of transmission dynamics [84, 85]. For example, characterising the initial phase of the SARS-CoV-2 pandemic in Wuhan, China, and its dissemination into other regions was primarily determined by human mobility data [86, 87]. By further incorporating data on the global aerial transportation networks and high-resolution geolocation collected from devices such as smartphones, we can access an unprecedented amount of information to finely detail human movements at local, national, and global scales to help refine interventions in near real-time [88, 89]. In chapter 6, I develop a method to use aerial transportation data to assess the importation risk of COVID-19 cases from China to Iran at the early stages of the pandemic. I also show how to use air travel data to estimate the outbreak size in Iran during February and March 2020.

Seroprevalence surveys also play an important role in addressing gaps in knowledge of the cumulative disease incidence. These surveys use antibody tests to estimate the percentage of people in a population who have antibodies against a

particular pathogen which allows us to quantify what percentage of the population have already been infected with the pathogen. Seroprevalence measurements have also been a critical component of the response to SARS-CoV-2 pandemic [90, 91], particularly in areas where there is inadequate viral diagnostic testing and incomplete understanding of the rates of mild and asymptomatic infections [92].

All-cause mortality can also act as an objective indicator of the burden of epidemics [93]. By comparing the total number of registered deaths during a period of public health emergency to some 'background' level in a period in the past with no major crisis, we can estimate the level of excess mortality that is attributable to that specific public health crisis (Figure 1.4). Historical trends have long been used to estimate death toll from extreme events including the current COVID-19 pandemic [94]. Understanding the true extent of mortality associated with infectious diseases using excess mortality is particularly importance in resource-poor settings where limited diagnostic capacity and access to healthcare leads to significant under-reporting of true cases and deaths associated with a disease [95]. In chapter 7 part a, I use excess mortality data to estimate the level of under-reporting in SARS-CoV-2 related deaths in Iran. I also use age-stratified infection fatality ratio estimates for COVID-19 to assess the level of exposure to the disease across the country. I then examine the degree of correlation between the estimated exposure level with seroprevalence data for different provinces in chapter 7 part b.

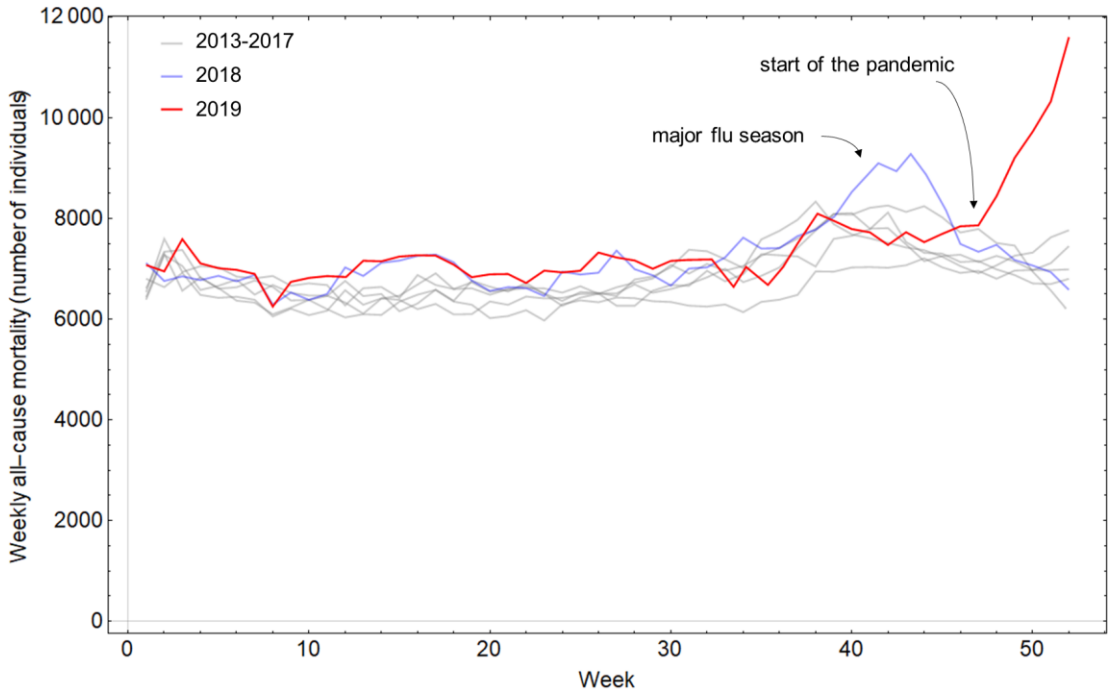


Figure 1.4: **Weekly all-cause mortality trend over time**. Trajectory of all-cause mortality from 2013 to 2019 generated using hypothetical data for illustrative purposes only. Arrows indicate points in time where there are substantial levels of excess mortality either due to a major flu season (blue) or the start of the COVID-19 pandemic (red).

1.7 Thesis outline

The term “tempo and mode in evolution”, which inspired the title of this thesis, was first coined by George Simpson in 1944 [116]. He argued that various population genetic parameters such as generation time, population size, mutation rate, and strength of selection which determine the microevolutionary processes can also collectively explain the macroevolutionary processes based on paleontological evidence. In his words, “tempo” has to do with “evolutionary rates…, their acceleration and deceleration [over time]” while the word “mode” involves “the study of the way, manner, or pattern of evolution, a study in which tempo is a basic factor, but which embraces considerably more than tempo” (pp. xvii-xviii). This captures the essence of what I will be exploring in the next 7 chapters of my thesis.

This thesis is divided into two main parts: in chapters 2 to 5, I explore various biological mechanisms that alter the rate of molecular evolutionary clock over time and investigate the ways I can reconstruct the evolutionary history of viruses over short (months to years) and long (millions of years) timescales despite the presence of TDRP. In chapters 6 and 7, I evaluate the SARS-COV-2 pandemic response in Iran and use various phylodynamic and epidemiological methods to reconstruct the early transmission dynamics across the country and estimate the prevalence of disease over time using mobility data, serological surveys, and all-cause mortality data.

In chapter 2, I develop a predictive mechanistic model of virus evolution that explains the empirical observation of a universal power-law rate decay across different virus groups over time. I show how the rapid genetic saturation of some sites, together with the host constraint on other sites, can create a time-dependent

rate dynamic whereby sites can partially or fully saturate according to how fast they accumulate substitutions over time. This model reproduces empirically observed TDRP patterns, and the inflection points where time-dependent rate changes become manifest due to site saturation.

In chapter 3, I reconstruct the evolutionary history of SIV and its divergence from HIV using the mechanistic evolutionary model I developed in chapter 2. I show that our model not only reproduces the known divergence times between different HIV-1 and HIV-2 groups (date back to nearly one hundred years ago), it also accurately estimates the divergence time between the two SIV drill subspecies found in the island of Bioko and mainland without the need for any biogeographic calibrations. I also find that the most recent common ancestor of all SIV species lived nearly one million years ago and that they share a common ancestor with grey mouse lemur prosimian immunodeficiency virus (pSIV) nearly 17 million years ago. This result matches the estimate from paleovirological analyses, thus bridging the gap between ancient and extant viral evolutionary studies on SIV.

In chapter 4, I characterise the molecular evolutionary dynamics of SARS-CoV-2 and 2009 pandemic H1N1 influenza during the first 12 months of their respective pandemics and identify the source of TDRP during that period. I show that time dependency in estimates of the mean substitution rate is dominated by elevated rates at terminal branches, whereas the I rate of substitution at internal branches does not exhibit a time-dependent decay. I find that the elevated rate at terminal branches is strongly correlated with an increasing number of segregating nonneutral sites, demonstrating the role of purifying selection in generating the time dependency of evolutionary parameters during pandemics.

In chapter 5, I systematically investigate the likelihood of various evolutionary pathways that would lead to the emergence of the first three SARS-CoV-2 variants of concern. I analytically find a range of parameter values for mutation rate, population size, and selective advantage that are consistent with the data on the timing of the emergence of Alpha, Beta, and Gamma variants of SARS-CoV-2. After finding the consistent range of parameter values, I further investigate whether they correspond to biologically reasonable scenarios in terms of the frequencies of the intermediate mutations prior to the emergence of VOCs, total number of mutations required to produce VOCs, total number of successful VOC lineages produced over time, and the timing between the emergence of different VOC lineages.

In chapter 6, I provide a detailed analysis of the SARS-CoV-2 epidemic in Iran. I use epidemiological and genetic data from air travellers with a travel history to Iran along with the first few whole genomes samples from inside the country to determine the start date and early growth rate of the epidemic. Next, I provide a province-level analysis to find the pattern of spread during winter and spring, highlighting the degree of under-reporting in both point prevalence and deaths. Finally, I evaluate the effect of non-pharmaceutical interventions on growth rate of the epidemic and use an SEIR model to reconstruct the early transmission dynamics across the country.

Chapter 7 is about estimating attributable deaths to COVID-19 and developing a framework for reconstructing transmission dynamics of SARS-CoV-2. It is divided into two parts that are conceptually and methodologically linked. In part a, I use seasonal all-cause mortality data provided by the National Organisation for Civil Registration of Iran to estimate excess mortality in all 31 provinces of Iran from January to September 2020. I then apply a population-weighted infection fatality

ratio to calculate the level of exposure to SARS-CoV-2 over time in each province. Finally, I compare the estimated levels of exposure to seroprevalence measurements collected during this period.

In part b, I provide a new framework for reconstructing SARS-CoV-2 transmission dynamics using excess mortality data in regions of the world where there is significant under-reporting of COVID-19-related deaths. By taking the Iranian COVID-19 epidemic as an example, I show excellent agreement between our model estimates and representative, province-level seroprevalence measurements across the country as well as the pattern of daily hospital admissions in all the 31 provinces of Iran throughout the pandemic. Despite a relatively young age structure in Iran, this analysis reveals that the infection fatality ratio in most provinces is comparable to high-income countries with a larger percentage of older adults, suggesting that limited access to medical services, coupled with undercounting of COVID-19-related deaths, can have a significant impact on accurate estimation of COVID-19 fatalities.

1.8 References

1. E. Zuckerkandl and L. Pauling, *Molecular disease, evolution, and genic heterogeneity.* In: Kasha M, Pullman B (eds) Horizons in biochemistry; Academic, New York, 1962: p. 189-225.
2. E. Zuckerkandl, R.T. Jones, and L. Pauling, *A Comparison of Animal Hemoglobins by Tryptic Peptide Pattern Analysis.* Proceedings of the National Academy of Sciences of the United States of America, 1960. **46**(10): p. 1349-1360.
3. E. Zuckerkandl and L. Pauling, *Evolutionary Divergence and Convergence, in Proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins; Academic, New York.* 1965: p. 97-166.
4. E. Margoliash, *Primary Structure and Evolution of Cytochrome C.* Proceedings of the National Academy of Sciences of the United States of America, 1963. **50**(4): p. 672-703.
5. R. F. Doolittle and B. Blomback, *Amino-Acid Sequence Investigations of Fibrinopeptides from Various Mammals - Evolutionary Implications.* Nature, 1964. **202**(492): p. 147-152.

6. V. M. Sarich and A.C. Wilson, *Immunological Time Scale for Hominid Evolution.* Science, 1967. **158**(3805): p. 1200-1203.
7. M. Kimura, *Evolutionary rate at the molecular level.* Nature, 1968. **217**(5129): p. 624-6.
8. D. S. Fisher, *Evolutionary dynamics. In: Bouchaud JP, Mezard M, Dalibard J (eds) Complex Systems, Lecture Notes of the Les Houches Summer School; Springer-Verlag, Amsterdam.* 2007.
9. M. Kimura, *The neutral theory of molecular evolution. Cambridge University Press, Cambridge.* 1983.
10. S. Y. W. Ho, *The Molecular Evolutionary Clock: Theory and Practice. Springer, Cham, Switzerland.* 2021.
11. A. D. Kern and M.W. Hahn, *The Neutral Theory in Light of Natural Selection.* Molecular Biology and Evolution, 2018. **35**(6): p. 1366-1371.
12. M. Nei, *Selectionism and neutralism in molecular evolution.* Molecular Biology and Evolution, 2005. **22**(12): p. 2318-2342.
13. R. R. Hudson, M. Kreitman, and M. Aguade, *A Test of Neutral Molecular Evolution Based on Nucleotide Data.* Genetics, 1987. **116**(1): p. 153-159.
14. F. Tajima, *Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.* Genetics, 1989. **123**(3): p. 585-595.
15. J. H. Mcdonald and M. Kreitman, *Adaptive Protein Evolution at the Adh Locus in Drosophila.* Nature, 1991. **351**(6328): p. 652-654.
16. A. Rambaut, *Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies.* Bioinformatics, 2000. **16**(4): p. 395-399.
17. A. J. Drummond, et al., *Genealogies from time-stamped sequence data.* In: Buck CE, Maillard AR (eds) Tools for constructing chronologies: crossing disciplinary boundaries. London: Springer-Verlag, 2004: p. 149–171.
18. A. J. Drummond, et al., *Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data.* Genetics, 2002. **161**(3): p. 1307-1320.
19. A. J. Drummond, et al., *Measurably evolving populations.* Trends in Ecology & Evolution, 2003. **18**(9): p. 481-488.
20. A. Rambaut, et al., *Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen).* Virus Evolution, 2016. **2**(1).
21. B. Korber, et al., *Timing the ancestor of the HIV-1 pandemic strains.* Science, 2000. **288**(5472): p. 1789-1796.
22. M. J. Sanderson, *A nonparametric approach to estimating divergence times in the absence of rate constancy.* Molecular Biology and Evolution, 1997. **14**(12): p. 1218-1231.
23. J. L. Thorne, H. Kishino, and I.S. Painter, *Estimating the rate of evolution of the rate of molecular evolution.* Molecular Biology and Evolution, 1998. **15**(12): p. 1647-1657.
24. A. J. Drummond, et al., *Relaxed phylogenetics and dating with confidence.* Plos Biology, 2006. **4**(5): p. 699-710.
25. A. J. Drummond and M.A. Suchard, *Bayesian random local clocks, or one rate to rule them all.* BMC Biology, 2010. **8**: p. 114.
26. J. P. Noonan, et al., *Sequencing and analysis of Neanderthal genomic DNA.* Science, 2006. **314**(5802): p. 1113-1118.

27. M. Meyer, et al., *A High-Coverage Genome Sequence from an Archaic Denisovan Individual.* Science, 2012. **338**(6104): p. 222-226.
28. I. Olalde, et al., *The Beaker phenomenon and the genomic transformation of northwest Europe.* Nature, 2018. **555**(7695): p. 190-198.
29. T. P. Guilderson, P.J. Reimer, and T.A. Brown, *The boon and bane of radiocarbon dating.* Science, 2005. **307**(5708): p. 362-364.
30. M. Molak, et al., *Phylogenetic Estimation of Timescales Using Ancient DNA: The Effects of Temporal Sampling Scheme and Uncertainty in Sample Ages.* Molecular Biology and Evolution, 2013. **30**(2): p. 253-262.
31. A. T. Duggan, et al., *17th Century Variola Virus Reveals the Recent History of Smallpox.* Current Biology, 2016. **26**(24): p. 3407-3412.
32. Z. P. Ross, et al., *The paradox of HBV evolution as revealed from a 16th century mummy.* Plos Pathogens, 2018. **14**(1): p. e1006750
33. M. Toppinen, et al., *Bones hold the key to DNA virus history and epidemiology.* Scientific Reports, 2015. **5**: p. 17226
34. S. Y. W. Ho and S. Duchene, *Molecular-clock methods for estimating evolutionary rates and timescales.* Molecular Ecology, 2014. **23**(24): p. 5947-5965.
35. M. Worobey, et al., *Island Biogeography Reveals the Deep History of SIV.* Science, 2010. **329**(5998): p. 1487-1487.
36. P. Aiewsakun and A. Katzourakis, *Endogenous viruses: Connecting recent and ancient viral evolution.* Virology, 2015. **479**: p. 26-37.
37. V. A. Belyi, A.J. Levine, and A.M. Skalka, *Sequences from Ancestral Single-Stranded DNA Viruses in Vertebrate Genomes: the Parvoviridae and Circoviridae Are More than 40 to 50 Million Years Old.* Journal of Virology, 2010. **84**(23): p. 12458-12462.
38. M. Horie, et al., *Endogenous non-retroviral RNA virus elements in mammalian genomes.* Nature, 2010. **463**(7277): p. 84-U90.
39. A. Katzourakis and R.J. Gifford, *Endogenous Viral Elements in Animal Genomes.* Plos Genetics, 2010. **6**(11): p. e1001191.
40. C. H. Kuo and H. Ochman, *Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria.* Biology Direct, 2009. **4**(4): p. 35.
41. P. M. Sharp and P. Simmonds, *Evaluating the evidence for virus/host co-evolution.* Current Opinion in Virology, 2011. **1**(5): p. 436-441.
42. W. M. Switzer, et al., *Ancient co-speciation of simian foamy viruses and primates.* Nature, 2005. **434**(7031): p. 376-380.
43. A. Katzourakis, et al., *Macroevolution of Complex Retroviruses.* Science, 2009. **325**(5947): p. 1512-1512.
44. P. Aiewsakun and A. Katzourakis, *Time dependency of foamy virus evolutionary rate estimates.* BMC Evolutionary Biology, 2015. **15**: p. 119.
45. K. Van Doorslaer, et al., *Unique genome organization of non-mammalian papillomaviruses provides insights into the evolution of viral early proteins.* Virus Evolution, 2017. **3**(2): p. vex027
46. P. Aiewsakun and A. Katzourakis, *Marine origin of retroviruses in the early Palaeozoic Era.* Nature Communications, 2017. **8**: p. 13954.
47. R. E. Ricklefs and S.M. Fallon, *Diversification and host switching in avian malaria parasites.* Proceedings of the Royal Society B-Biological Sciences, 2002. **269**(1494): p. 885-892.

48. M. A. Charleston and D.L. Robertson, *Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny.* Systematic Biology, 2002. **51**(3): p. 528-535.
49. G. W. Harkins, et al., *Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts.* Virology Journal, 2009. **6**: p. 104.
50. M. A. Gandolfo, K.C. Nixon, and W.L. Crepet, *Selection of fossils for calibration of molecular dating models.* Annals of the Missouri Botanical Garden, 2008. **95**(1): p. 34-42.
51. J. F. Parham, et al., *Best Practices for Justifying Fossil Calibrations.* Systematic Biology, 2012. **61**(2): p. 346-359.
52. S. Y. W. Ho, et al., *Time-dependent rates of molecular evolution.* Molecular Ecology, 2011. **20**(15): p. 3087-3101.
53. Y. Li, et al., *On the origin of smallpox: Correlating variola phylogenics with historical smallpox records.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(40): p. 15787-15792.
54. R. K. Wayne, B. Vanvalkenburgh, and S.J. Obrien, *Molecular Distance and Divergence Time in Carnivores and Primates.* Molecular Biology and Evolution, 1991. **8**(3): p. 297-319.
55. N. Howell, et al., *The pedigree rate of sequence divergence in the human mitochondrial genome: There is a difference between phylogenetic and pedigree rates.* American Journal of Human Genetics, 2003. **72**(3): p. 659-670.
56. J. Garcia-Moreno, *Is there a universal mtDNA clock for birds?* Journal of Avian Biology, 2004. **35**(6): p. 465-468.
57. S. Y. W. Ho, et al., *Time-dependent estimates of molecular evolutionary rates: evidence and causes.* Molecular Ecology, 2015. **24**(24): p. 6007-6012.
58. P. Aiewsakun and A. Katzourakis, *Time-Dependent Rate Phenomenon in Viruses.* Journal of Virology, 2016. **90**(16): p. 7184-7195.
59. S. Duchene, E.C. Holmes, and S.Y.W. Ho, *Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates.* Proceedings of the Royal Society B-Biological Sciences, 2014. **281**(1786): p. 20140732.
60. P. M. Sharp, et al., *The origins of acquired immune deficiency syndrome viruses: where and when?* Philosophical Transactions of the Royal Society B-Biological Sciences, 2001. **356**(1410): p. 867-876.
61. R. J. Gifford, et al., *A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(51): p. 20362-20367.
62. A. Katzourakis, et al., *Discovery and analysis of the first endogenous lentivirus.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(15): p. 6261-6265.
63. Z. Keckesova, et al., *Identification of a RELIK orthologue in the European hare (Lepus europaeus) reveals a minimum age of 12 million years for the lagomorph lentiviruses.* Virology, 2009. **384**(1): p. 7-11.
64. E. P. C. Rocha, et al., *Comparisons of dN/dS are time dependent for closely related bacterial genomes.* Journal of Theoretical Biology, 2006. **239**(2): p. 226-235.
65. L. Feng, et al., *A Recalibrated Molecular Clock and Independent Origins for the Cholera Pandemic Clones.* Plos One, 2008. **3**(12): p. e4053.

66. M. Nei, Y. Suzuki, and M. Nozawa, *The Neutral Theory of Molecular Evolution in the Genomic Era.* Annual Review of Genomics and Human Genetics, 2010. **11**: p. 265-289.
67. C. D. Laird, B.L. Mcconaughy, and B.J. Mccarthy, *Rate of Fixation of Nucleotide Substitutions in Evolution.* Nature, 1969. **224**(5215): p. 149-154.
68. D. E. Kohne, *Evolution of higher-organism DNA.* Q Rev Biophys, 1970. **3**(3): p. 327-75.
69. T. Ohta, *Population size and rate of evolution.* J Mol Evol, 1972. **1**(4): p. 305-14.
70. T. Ohta, *Slightly Deleterious Mutant Substitutions in Evolution.* Nature, 1973. **246**(5428): p. 96-98.
71. J. Bergsten, *A review of long-branch attraction.* Cladistics, 2005. **21**(2): p. 163-193.
72. D. G. Kendall, *On the Generalized Birth-and-Death Process.* Annals of Mathematical Statistics, 1948. **19**(1): p. 1-15.
73. L. E. Nicolaisen and M.M. Desai, *Distortions in Genealogies due to Purifying Selection and Recombination.* Genetics, 2012. **195**(11): p. 3589–3600.
74. C. Pal and B. Papp, *Evolution of complex adaptations in molecular systems.* Nature Ecology & Evolution, 2017. **1**(8): p. 1084-1092.
75. S. Wright, *The Shifting Balance Theory and Macroevolution.* Annual Review of Genetics, 1982. **16**: p. 1-19.
76. D. B. Weissman, et al., *The rate at which asexual populations cross fitness valleys.* Theoretical Population Biology, 2009. **75**(4): p. 286-300.
77. D. B. Weissman, M.W. Feldman, and D.S. Fisher, *The Rate of Fitness-Valley Crossing in Sexual Populations.* Genetics, 2010. **186**(4): p. 1389-1410.
78. M. Ghafari and D.B. Weissman, *The expected time to cross extended fitness plateaus.* Theoretical Population Biology, 2019. **129**: p. 54-67.
79. Y. Iwasa, F. Michor, and M.A. Nowak, *Stochastic tunnels in evolutionary dynamics.* Genetics, 2004. **166**(3): p. 1571-1579.
80. B. T. Grenfell, et al., *Unifying the epidemiological and evolutionary dynamics of pathogens.* Science, 2004. **303**(5656): p. 327-332.
81. O. G. Pybus and A. Rambaut, *Evolutionary analysis of the dynamics of viral infectious disease.* Nature Reviews Genetics, 2009. **10**(8): p. 540-550.
82. E. M. Volz, K. Koelle, and T. Bedford, *Viral Phylodynamics.* Plos Computational Biology, 2013. **9**(3): p. e1002947.
83. A. Kalkauskas, et al., *Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk.* Plos Computational Biology, 2021. **17**: p. 1–27.
84. C. O. Buckee, et al., *Aggregated mobility data could help fight COVID-19.* Science, 2020. **368**(6487): p. 145-146.
85. B. Rader, et al., *Crowding and the shape of COVID-19 epidemics.* Nature Medicine, 2020. **26**(12): p. 1829–1834.
86. M. U. G. Kraemer, et al., *The effect of human mobility and control measures on the COVID-19 epidemic in China.* Science, 2020. **368**(6490): p. 493-497
87. S. Zhao, et al., *Quantifying the association between domestic travel and the exportation of novel coronavirus (2019-nCoV) cases from Wuhan, China in 2020: a correlational analysis.* Journal of Travel Medicine, 2020. **27**(2): p. 1-3.
88. L. Ferretti, et al., *Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing.* Science, 2020. **368**(6491): p. 619-628.

89. C. Wymant, et al., *The epidemiological impact of the NHS COVID-19 app.* Nature, 2021. **594**(7863): p. 408-412.
90. D. B. Larremore, et al., *Estimating SARS-CoV-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys.* Elife, 2021. **10**: p. e64206.
91. L. F. Buss, et al., *Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic.* Science, 2021. **371**(6526): p. 288-292.
92. D. Sutton, et al., *Universal Screening for SARS-CoV-2 in Women Admitted for Delivery.* New England Journal of Medicine, 2020. **382**(22): p. 2163-2164.
93. C. Whittaker, et al., *Under-reporting of deaths limits our understanding of true burden of covid-19.* Bmj-British Medical Journal, 2021. **375**: p. n2239.
94. A. Karlinsky and D. Kobak, *Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset.* Elife, 2021. **10**: p. e69336.
95. P. G. T. Walker, et al., *The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries.* Science, 2020. **369**(6502): p. 413-422.
96. A. G. Clark and T. S. Whittam, *Sequencing errors and molecular evolutionary analysis.* Molecular Biology and Evolution, 1992. **9**: p. 744-752.
97. G. M. Jenkins, et al., *The Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis.* Journal of Molecular Evolution, 2002. **54**: p. 156-165.
98. E. Ruiz-Pesini, et al., *Effects of purifying and adaptive selection on regional variation in human mtDNA.* Science, 2004. **303**: p. 223-226.
99. D. Mishmar, et al., *Natural selection shaped regional mtDNA variation in humans.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**: p. 171-176.
100. J. Sullivan and P. Joyce, *Model selection in phylogenetics.* Annual Reviews of Ecology, Evolution, and Systematisc, 2005. **36**: p. 445-466.
101. J. Gatesy, *A tenth crucial question regarding model use in phylogenetics.* Trends in Ecology and Evolution, 2007. **22**: p. 509-510.
102. J. Soubrier, et al., *The influence of rate heterogeneity among sites on the time dependence of molecular rates.* Molecular Biology and Evolution, 2012. **29**: p. 3345-3358.
103. C. Kosiol and N. Goldman, *Markovian and Non-Markovian Protein Sequence Evolution: Aggregated Markov Process Models.* Journal of Molecular Biology, 2011. **411**: p. 910-923.
104. F. Rizzato, A. Rodriguez, A. Liao, *Non-Markovian effects on protein sequence evolution due to site dependent substitution rates.* BMC Bioinformatics, 2016. **17**: p. 1-9.
105. G. M. Jenkins, A. Rambaut, O. G. Pybus, *Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis.* Journal of Molecular Evolution, 2002. **54**: p. 156-165.

106. R. Sanjuan, *From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses.* Plos Pathogens, 2012. **8**: e1002685.

107. L. A. Shackelton, et al., *High rate of viral evolution associated with the emergence of carnivore parvovirus.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**: p. 379-384.

108. S. Duffy, L. A. Shackelton, and E. C. Holmes, *Rates of evolutionary change in viruses: patterns and determinants.* Nature Review Genetics, 2008. **9**: p. 267-276.

109. N. Takeda, M. Tanimura, and K. Miyamura, *Molecular evolution of the major capsid protein VP1 of enterovirus 70.* Journal of Virology, 1994. **68**: p. 854-862.

110. P. V. Markov, et al., *Phylogeography and molecular epidemiology of hepatitis C virus genotype 2 in Africa.* Journal of General Virology, 2009. **90**: p. 2086-2096.

111. S. K. Gire, et al., *Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak.* Science, 2014. **345**: p. 1369-1372.

112. E. C. Holmes, et al., *The evolution of Ebola virus: Insights from the 2013–2016 epidemic.* Science, 2016. **538**: p. 163-200.

113. D. J. Blackley, et al., *Reduced evolutionary rate in reemerged Ebola virus transmission chains.* Science Advances, 2016. **2**: e1600378.

114. J. Lu, et al., *Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China.* Cell, 2020. **181**: p. 997-1003.e9.

115. S. Duchene, et al., *Temporal signal and the phylodynamic threshold of SARS-CoV-2.* Virus Evolution, 2020. **6**: p. veaa061.

116. G. G. Simpsoon, *Tempo and Mode in Evolution.* New York: Columbia University Press.

Chapter 2: A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses

2.1 Abstract

Estimating viral timescales is fundamental to understanding the evolutionary biology of viruses. Molecular clocks are widely used to reveal the recent evolutionary histories of viruses but may severely underestimate their longer-term origins because of the inverse correlation between inferred rates of evolution and the timescale of their measurement. Here we provide a predictive mechanistic model that readily explains the rate decay phenomenon over a wide range of timescales, recapitulates the ubiquitous power-law rate decay with a slope of -0.65. We show that standard substitution models fail to correctly estimate divergence times once the most rapidly-evolving sites saturate, typically after hundreds of years in RNA viruses and thousands of years in DNA viruses. Our model successfully re-creates the observed pattern of decay and explains the evolutionary processes behind the time-dependent rate phenomenon. We then apply our model to re-date the diversification of genotypes of hepatitis C virus to 423,000 (95% HPD: 394,000 - 454,000) years before present, a time preceding the dispersal of modern humans out of Africa, and also showed that the most recent common ancestor of sarbecoviruses dates back to 21,000 (95% HPD: 19,000 - 22,000) years ago, nearly thirty times older than previous estimates. This creates a new perspective for our understanding the origins of these viruses and also suggests that a substantial revision of evolutionary timescales of other viruses can be similarly achieved.

2.2 Introduction

The timescale over which viruses evolve and how this process is connected to host adaptation has been an area of considerable research and methodological progress in recent decades. Mammalian RNA viruses, in particular, exhibit extraordinarily rapid genomic change [1-3] and analyses of their genetic variation has enabled detailed reconstruction of the emergence of viruses such as HIV-1[4], hepatitis C virus [5] and influenza A virus [6]. RNA viruses display evolutionary change over short-timescales (weeks to months) and can alter a substantial part of their genomes following a host switch [7-10]. Well characterised examples for both RNA and DNA viruses include the emergence of HIV-1 in humans from a chimpanzee reservoir [4, 11], and the adaptation of myxomatosis in rabbits [12].

These rapid rates of virus sequence change stand in striking contrast with evidence for extreme conservation of virus genome sequences over long periods of evolution and at higher taxonomic levels. Inferred short term rates of virus sequence change should create completely unrecognisable genome sequences if they were naively extrapolated over thousands or even hundreds of years, yet endogenous viral elements (EVEs) that integrated into host genomes throughout mammalian evolution are recognisably similar to contemporary genera and families of *Bornaviridae*, *Parvoviridae* and *Circoviridae* among many examples [13-15]. This observation is complemented by evidence from studies of virus / host co-evolution [16-18], and more recently from analysis of viruses recovered from ancient DNA and RNA in archaeological remains [19-22], that indicate a remarkable degree of conservation in viral genome sequences and their inter-relationships at genus and family levels. This dichotomy has been attributed to the time-dependent rate

phenomenon (TDRP), which is the observation that apparent rates of evolution are dependent on timescales of measurement [23, 24].

The TDRP has been explained by processes such as sequence site saturation, purifying selection, short-term changes in selection pressure, and potential errors in estimation of short-term substitution rates [23-26]. Empirically, substitution rates across RNA and DNA viruses show a striking linear relationship between log-transformed rates and timescale of measurement despite the large variation among viruses in their initial short term substitution rates [25]. The gradients of regressions of observation time against estimated evolutionary rates are consistently around -0.65, for all virus groups for which long term substitution rates can be calculated or inferred. Crucially, the observation of a universal power-law rate decay in nearly all viruses suggests that there is a common underlying evolutionary process. However, we do not have a systematic biological explanation for this observation. Various factors have been invoked to account for these patterns, including purifying selection, site saturation, and sequencing errors though none of these alone have been shown to generate a power law rate decay [24, 26].

One proposal is that the primary driver of virus evolution over long evolutionary time-scales is host adaptation, in which virus sequence change is severely curtailed by stringent fitness constraints [27]. Viruses exist within a tightly-constraining host niche to which they rapidly adapt; paradoxically, their high mutation rates, large population sizes and consequent ability to adapt rapidly serve to restrict their long-term diversification and sustained sequence change, rendering them evolutionary “Prisoners of War” (PoW). This idea posits that, over longer timescales, rates of viral evolution will be bounded by the rate of evolution of their hosts [27]. However, the exact timescales over which these various evolutionary events occur and the extent

to which they contribute to changing virus sequences over time is still largely unknown.

Here we develop a new model of the longer-term evolutionary rate dynamics of viruses that explains the empirical observation of a universal power-law rate decay across different virus groups. Our model is biologically motivated and based on a minimal number of realistic assumptions. We show how the rapid genetic saturation of some sites, together with the host constraint on other sites, can create a time-dependent rate dynamic whereby sites can partially or fully saturate according to how fast they accumulate substitutions over time. This process occurs chronologically from sites evolving the fastest to those that evolve epistatically, to those that evolve at the hosts' substitution rate. This model precisely reproduces empirically observed TDRP patterns and the inflection points where time-dependent rate changes become manifest due to site saturation. We demonstrate that the model predictions are robust to intrinsic and marked differences in substitution rates among different virus groups or to assumptions about the relative proportion of sites evolving at different rates.

2.3 Results

2.3.1 Power-law rate decay can emerge due to site saturation

First, we show how a time-dependent rate effect emerges when estimating the rate of sequence divergence using a standard evolutionary model. Suppose that a sequence has diverged from its ancestor for t years under a constant and uniform substitution rate per site per year (SSY), μ. The proportion of pairwise differences between the derived sequence and its ancestor, $p(t)$, initially accumulates linearly, i.e. $p(t) \approx \mu t$, until it reaches a point where every new substitution occurs on the

background of a site that has already changed at least once; this limit, hereafter called the saturation frequency, α, occurs at time $t^* \approx \alpha / \mu$ (see Equation S1). As the derived sequence continues to diverge beyond the saturation point, t^*, the observed proportion of pairwise differences, $\hat{p}$, remains effectively unchanged. If there are no intermediate samples from the derived sequence before it reaches the saturation point, any conventional substitution model is not able to correctly count the number substitutions and, therefore, underestimates the true rate. Thus, without enough temporal information from the derived sequence before reaching the saturation point, using any conventional substitution model, the inferred genetic distance (i.e., the expected number of substitutions per site), $\hat{d}$, remains constant and the estimated substitution rate, $\hat{\mu}$, declines as the divergence time increases, $\hat{\mu} \propto 1/t$, which manifests itself with a power-law rate decay with slope -1 on a log-transformed plot.

2.3.2 Conundrum of rate calibrations

As an illustrative example, we consider how time-dependent rate effects pose a challenge to estimation of the substitution rate of foamy viruses (FVs) over time. FVs are a group of retroviruses in the subfamily of *Spumaretrovirinae* that have been isolated from a broad range of mammals [28]. By tracking the evolution of FVs in a population of African Green Monkeys over short timescales (i.e., less than a decade), it was estimated that their evolutionary rate is approximately 3.8×10^{-4} substitutions per site per year (SSY) [29]. By contrast, using the very long and stable co-speciation history of FVs with their hosts, which goes back to more than a hundred million years [30], their long-term rate of evolution has been estimated to be nearly 1.7×10^{-8} SSY [31], almost four orders of magnitude slower than their

short-term rates. If we calibrate the amount of divergence (i.e., the expected number of substitutions per site) between FV sequences over time using their short-term evolutionary rate estimates, then nearly all sites in the virus genome should have acquired a substitution (i.e., saturation point) after ~2,500 years. Beyond this time, any standard model of sequence evolution (substitution model) underestimates the amount of sequence divergence and, thus, the inferred substitution rates, $\hat{\mu}$, drop sharply as the timespan of observation, t, increases such that the slope of the rate decay on a log-transformed plot is -1 (Figure 2.1A).

Even though there is a sharp decline in the inferred substitution rate of FVs over time due to site saturation, it still overestimates their true long-term evolutionary rate based on co-speciation history after a hundred million years (see purple line in Figure 2.1A). One way to resolve this is by dividing the sites in the FV genome into two "rate groups" such that a fraction of sites, m_1, evolve at a fast rate (say, $\mu_1 = 1 \times 10^{-3}$ SSY), and the remaining sites, $m_2 = 1 - m_1$, evolve at a slower rate (say, $\mu_2 = 1.7 \times 10^{-8}$ SSY). The allocated fraction, m_1, can be specified such that the mean substitution rate, $\langle \mu \rangle = m_1 \mu_1 + m_2 \mu_2$, is equal to the observed short-term evolutionary rate of FV, $\hat{\mu} = 3.8 \times 10^{-4}$ SSY. Therefore, a standard substitution model can reliably estimate short-term rates up to and before the saturation of rapidly-evolving sites, m_1, beyond which time the number of substitutions at those sites is underestimated and a time-dependent rate decay emerges. However, over longer timescales, new substitutions at the slowly-evolving sites, m_2, accumulate such that the inferred rate, $\hat{\mu}$, gradually plateaus at the rate which corresponds to the long-term substitution rate of FV, μ_2. This new plateau will only last until the proportion of slow-evolving

sites also saturate, which takes more than 100 million years to occur, beyond which time the entire genome is saturated and a power-law rate decay with slope -1 emerges (see cyan line in Figure 2.1A).

2.3.3 The Prisoner of War model of virus evolution

Even though an evolutionary model for FV substitution rate based on two rate groups can recover its inferred short-term and long-term rates, it still fails to accurately predict the observed substitution rates over intermediate timescales. Aiewsakun and Katzourakis [31] made the empirical observation that the time-dependent rate of FVs follows a power-law decay with slope -0.65 [32]. More importantly, they showed that not only FVs but all other virus groups for which a correlation could be performed follow a similar universal power-law rate decay with the same slope [25], suggesting a common underlying process involved in rate decay.

The Prisoner of War (PoW) model of evolution is based on the principle that a virus sequence is divided into M substitution rate groups ranging from sites evolving very rapidly at rate, $\mu_{\max}$, similar to virus mutation rates, to sites evolving more slowly due to epistatic and compensatory substitutions, all the way to sites evolving at the host substitution rate, $\mu_{\min}$. The fraction of sites, m_i, allocated to each rate group, i, is an exponentially distributed number, $m_i = Ce^{\lambda i}$, where C is the normalisation factor and λ is the exponent coefficient that determines whether the majority of sites are slowly- ($\lambda < 0$) or rapidly- ($\lambda > 0$) evolving. As the virus sequence evolves, sites belonging to the fastest rate group saturate first, typically after $t \sim 1/\mu_{\max}$ years, followed by the saturation of sites in the next fastest rate group, which takes longer to occur, and so on. Depending on how divergent a virus

sequence is with respect to its ancestor, sites that belong to more slowly-evolving rate groups may also reach partial or complete saturation (Equation S4). Thus, at any given timespan, the virus explores only a subset of its sites (i.e., is trapped inside a 'prison cell' in sequence space) and does not have access to explore substitutions at every position in its genome. This chronological saturation of sites gives rise to a power-law decay in the inferred substitution rate over time with slope -0.65 on a log-transformed graph, supporting the empirical observation by Aiewsakun and Katzourakis [25] (Figure 2.1B).

Building upon a collection of virus evolutionary rate estimates from >130 publications [25], we use 396 nucleotide substitution rate estimates across six major viral groups to find the line of best fit between the PoW model of time-dependent substitution rates and the evolutionary rate estimates for each viral group (i.e., the data that is collected from the literature) using the geometric least squares method[33]. This enables us to estimate a mean and maximum substitution rate (i.e., $\langle\mu\rangle$ and $\mu_{\max}$) for each virus group.

Our results show that upon the saturation of the fastest-evolving sites a power-law rate decay emerges with the slope -0.65 (95% HPD: -0.72, -0.52) across all viral groups (Figure 2.2). The inflection point in the rate curve, which signals the saturation of rapidly-evolving sites, occurs typically after 100 to 1,000 years in most RNA and DNA viruses. We further find that in double-stranded DNA viruses (dsDNA), the short-term substitution rate (i.e., the flat part of the time-dependent rate curve, $\langle\mu\rangle = \sum_{i=1}^{M} m_i \mu_i$) and the fastest-evolving rate group, $\mu_{\max}$, have the lowest rates compared to all other virus groups. Together with reverse-transcribing DNA

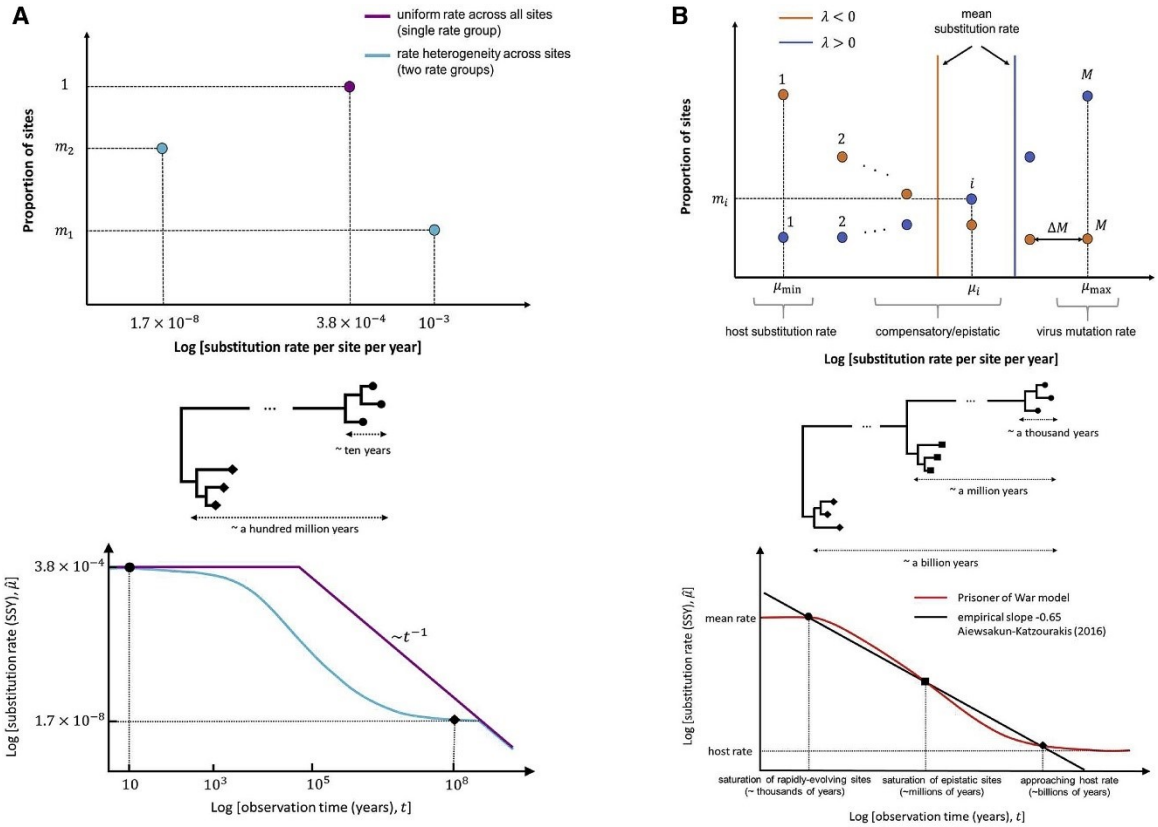


Figure 2.1: **Time-dependent rate phenomenon in foamy viruses (FVs) under different evolutionary models.** The figure shows how time-dependent rate phenomenon emerges (**A**) using a standard evolutionary model with a single or double rate group and (**B**) using the PoW model with several rate groups. Distribution of the fraction of sites per rate group when (**A**, top) there is a single rate group (purple), i.e. all sites evolve at the same substitution rate, or there are two rate groups (blue) such that a fraction, $m_1 = 0.38$, are evolving at a faster rate (10^{-3} SSY) compared to the remaining sites, $m_2 = 1 - m_1$, which are evolving more slowly (10^{-6} SSY). (**B**, top) There are M rate groups that are equally spaced on a log-scale, with a common ratio, ΔM, from the slowest rate (group 1), $\mu_{\min} = 10^{-9}$ SSY, to the fastest (group M), $\mu_{\max}$. According to the PoW model, a fraction of sites, m_i, belonging to rate group $i \in \{1,2,\ldots,M\}$, evolving at rate μ_i, is an exponentially distributed number with exponent parameter λ such that λ is less (greater) than zero when the majority (minority) of sites evolve at the host substitution rate. (**A**, bottom) Schematic plot of the evolutionary rate trajectory of FVs over time assuming that their rate is inferred using a standard substitution model with a single rate group that is calibrated based on their short-term substitution rates (circular nodes on the tree) and two rate groups based on their mean and long-term substitution rates (diamond-shaped nodes on the tree). (**B**, bottom) Evolutionary rate trajectory of viruses over time under the PoW model whereby the short-term rates can be inferred until when the fraction of sites belonging to the fastest rate group reach saturation (the inflection point in the curve) beyond which point a time-dependent rate decay emerges. The virus substitution rates can be reliably inferred across all timescales without a need for any rate calibrations and the pattern of rate decay is aligned with the empirical slope of -0.65 [25].

(RT-DNA) and single-stranded DNA viruses (ssDNA), dsDNA viruses have typically 1-2 orders of magnitude slower short-term substitution rates and fastest-evolving rate groups compared to RNA viruses (Table 2.1). Conversely, these rates are very similar among the positive-strand RNA viruses (+ssRNA), negative-strand RNA viruses (-ssRNA), and reverse-transcribing RNA viruses (RT-RNA). We find that the estimated rate at the fastest-evolving sites in RNA viruses is $\mu_{\max} \approx 4\times10^{-2}$ SSY which is very close to their estimated mutation rates. Therefore, these sites can begin to saturate after only a few decades or hundreds of years. On the other hand, we find that a large proportion of slow-evolving sites are not saturated over the span of more than 1 billion years and that the rate curve in none of the virus groups has completely plateaued at the host substitution rate (Figure 2.2).

Table 2.1: **Estimated short-term and maximum substitution rate SSY according to the PoW model.** The table shows the inferred short-term substitution rate and the rate of substitution at the fastest-evolving rate group across 6 virus groups. Parentheses show to the 95% confidence intervals.

Viral group	Type of virus	Short-term substitution rate, $\langle\mu\rangle$	Fastest rate group, $\mu_{\max}$
Group I	dsDNA virus	$2(0.3-16)\times10^{-5}$	$3(0.6-10)\times10^{-3}$
Group II	ssDNA virus	$3(1-6)\times10^{-4}$	$2(1-3)\times10^{-2}$
Group IV	(+)ssRNA virus	$2(1-4)\times10^{-3}$	$4(3-6)\times10^{-2}$
Group V	(-)ssRNA virus	$1(0.7-3)\times10^{-3}$	$4(3-6)\times10^{-2}$
Group VI	RT-RNA virus	$1(0.7-3)\times10^{-3}$	$4(3-6)\times10^{-2}$
Group VII	RT-DNA virus	$5(1-20)\times10^{-5}$	$4(2-10)\times10^{-3}$

To ensure that our model predictions are not biased towards a particular virus family with more evolutionary rate estimates (i.e., more data points to fit to the PoW model in Equation S4), we remove all the short-term rate estimates (i.e., any rate that is measured over a timespan of less than 100 years) within each viral group except for one virus family or genus to re-calibrate the mean substitution rates (Table S1). We find that despite the broad variation in short-term rates across all

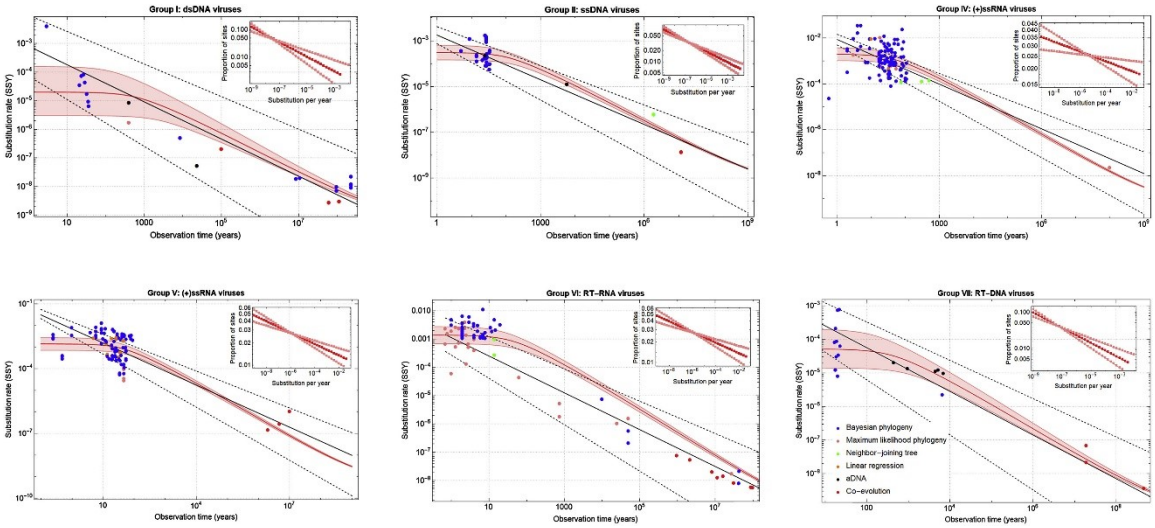


Figure 2.2: **Estimated time-dependent rate curves for each virus group according to the PoW model.** A total of 389 viral rate estimates (coloured circles representing various phylogenetic methods used for rate estimation) was collected from more than 130 publications, 23 estimates for Baltimore group I, 32 for group II, 123 for group IV, 106 for group V, 85 for group VI, and 20 for group VII. The insets show the proportion of sites in each rate group. Every rate group, including the ones with the smallest proportion, are well-represented in the genome, i.e. $m_{\text{minimum}} \gg 10^{-4} > 1/L$, where L is the typical genome size of an RNA virus. The red lines show the best fit and shaded area the 95% confidence intervals for each virus group ($\Delta M = 1.58$ and $\alpha_M = \alpha = 3/4$).

viral groups, the shape of the sigmoid curve, exponent coefficient λ, and $\mu_{\max}$ are robust to such changes and are not an artifact of systematic biases in selecting rate estimates from a particular virus family. We note that, in group VI, the large difference in evolutionary rates between *Lentivirus* and *Deltaretrovirus* families results in a noticeably different pattern of rate decay over short timescales (Figure S2) and the long-term rates are more aligned with the predictions based on the *Deltaretrovirus* re-calibration. The short-term substitution rate of *Lentivirus* families are 1-2 orders of magnitude higher than the *Deltaretrovirus* family. The latter evolves at rates similar to RT-DNA viruses. We also see that a larger fraction of sites in DNA viruses tend to evolve at rates closer to the host substitution rates (i.e., majority of sites in the virus sequence are slow-evolving) compared to RNA viruses, which largely have an equal proportion of sites in every rate group (i.e., the exponent coefficient λ is close to zero). We also carried out a similar sensitivity analysis at the level of virus genera which further confirms that the rate curves predicted by the PoW model remain accurate at this level and are not an artefact of measured rates at the level of Baltimore groups (Figure S2).

The formulation of the PoW model allows for a one-to-one map between the relative genetic distance of any given pair of sequences and divergence time. Therefore, by estimating the genetic distance between sequences using some distance metric, we can convert them into divergence time (Equation S5 and S6). Given that the shape of the rate curve is robust to changes in the short-term substitution rates, $\langle\mu\rangle$, across virus families and genera, the transformation from relative genetic distance to time can be applied to estimate the divergence times among any given virus sequences. To test the validity of this approach in recovering known divergence times, we re-estimate the time to the most recent common

ancestor of various species of FV and compare the results with estimates based on host calibrations [32]. Because of the long history of co-speciation, the virus phylogenies have the same topology as the host. Therefore, there can be a one-to-one comparison between estimated divergence times based on their phylogenies. First, by fixing the rate of substitution at the fastest-evolving sites to those inferred for RNA viruses (Table 2.1) and using known long-term substitution rate estimates of FV from the literature to find the line of best fit for the PoW rate curve, we estimate the FV short-term substitution rate to be 6.1 x 10^{-6} (95% HPD: 5.4 x 10^{-6} - 7.0 x 10^{-6}) which is in agreement with previous estimates [34]. We then construct a distance tree using the Jukes-Cantor (JC69) or Hasegawa, Kishino and Yano (HKY85) model and convert branch lengths into divergence time using the PoW transformation (see Methods). The result confirms that we can reliably recover true divergence times between most samples (some are different by up to a factor of 2) without calibrating the dates of any nodes on the tree (Figures 2.3A and S3A).

To further illustrate the radical effect of applying the PoW model to virus evolutionary timescales, we analyse a heterochronous dataset of complete hepatitis C virus (HCV) genome sequences that represent its component genotypes and subtypes (Figure 2.3B). First, using a standard HKY+ Γ_4 substitution model, we find that the mean substitution rate of HCV is 8.3 x 10^{-4} (95%HPD: 7.3 x 10^{-4} - 9.5 x 10^{-4}) SSY. Then, by using the predicted value of $\mu_{\max}$ for viruses that belong to group IV, $\mu_{\max} = 3.65 \times 10^{-2}$ (Table 2.1), and the inferred median short-term substitution rate $\langle \mu \rangle = 8.3 \times 10^{-4}$, we can construct a PoW-transformed time tree for HCV (see Methods). We find that there is a clear separation of timescales for the diversification of HCV variants within genotypes ($\sim$50 – 2,000 years), among subtypes ($\sim$1,000 –

80,000 years), and among genotypes (~80,000 – 200,000 years) with an estimated time to the most recent common ancestor (TMRCA) of 423,000 (95% HPD: 394,000 - 454,000) years before present (BP) for HCV (Figure 2.3B). While the predicted divergence times for some of the within-genotype variants using the PoW model is similar to those obtained using a standard HKY+ Γ_4 substitution model, the latter estimates the TMRCA of HCV to be only 940 (95% HPD: 820 – 1100) years BP with no clear separation of timescales for among-genotype diversifications. These results contrast with estimates of 500 – 2,000 years of genotype diversification by simple extrapolation from short term rates, while among-subtype divergence times of 1,000 – 80,000 years are up to 50 times higher than the 300 – 500 years estimated in previous molecular epidemiological analysis [35-37]. The revised, very early evolutionary origin of HCV genotypes (394,000 years, 454,000 years 95% HPD) predicted by our model is striking. While these early dates still fit proposed hypotheses for multiple and potentially relatively recent zoonotic sources of HCV in humans, associated with different genotypes [38, 39], the existence of a common ancestor of HCV before human migration of Africa (150,000 BP) support an alternative scenario where HCV diversified within anatomically modern humans. HCV genotypes may have arisen from geographical separation in Africa (genotypes 1, 2, 4, 5, 7) and migrational separation of human populations migrating out of Africa into Asia (genotypes 3, 6 and 8).

We also carried out a similar analysis to investigate the origins of the SARS-CoV-2 sarbecovirus lineage (Figure 2.3C). By finding the mean substitution rate of the sarbecovirus lineage to be 5.6×10^{-4} (95%HPD: 3.5×10^{-4} - 7.6×10^{-4}) using a standard HKY+ Γ_4 substitution model (see Methods), we find that while the PoW-transformed phylogeny recovers the previous estimates for SARS-CoV and SARS-

CoV-2 diversification from their most closely related bat virus over short timescales (i.e., less than hundreds of years BP), it extends the TMRCA back to 21,000 (95% HPD: 19,000 - 22,000) years BP, nearly 30 times older than previous estimates[40]. The 95% HPD represents the uncertainty that may arise from the choice of substitution model, inferred genetic distance between each pair of sequence, and the inferred tree topology. Our results indicate that humanity may have been exposed to these viruses since the Paleolithic if they had come into contact with their natural hosts. Also, our date estimates of the origin of the sarbecovirus lineage are in remarkable concordance with signatures of selection on human genomic datasets that indicate an arms race with corona-like viruses dating back to 25,000 years BP [41], providing an external comparator for our methodology. We also note that even if sarbecovirus origin were estimated to be more recent, the pattern of selection could still be attributed to deeper coronavirus ancestry.

To test the impact of changing the substitution model on the estimated divergence times, we compared our analyses on FV, HCV, and sarbecovirus datasets using both the JC69 and HKY substitution models and find minimal difference between the estimates (see Figures 2.3 and S3). To further assess the impact of uncertainty in the clock model on estimated divergence times, we allow the short-term rates and $\mu_{\max}$ for each virus dataset to vary in accordance with its inferred posterior mean rate distriubtion and the geometric least square cost function, respectively (see Methods). Our results show that while the median TMRCA estimates for the three datasets are very similar to results in Figure 2.3, the confidence intervals are much wider. We find that the TMRCA for the HCV dataset is 427,000 (95% HPD: 153,000 - 826,000) years BP and for the sarbecovirus dataset is 25,000 (95% HPD: 5,000 - 73,000) years BP (Figure S3). We note that

the higher level of uncertainty in the estimated divergence times due to the clock model is somewhat artificial and can vary widely depending on the choice of the clock model (e.g., strict/relaxed clock) and rate prior. For instance, in the sarbecovirus dataset, because the alignments are from diverse viral populations with deep evolutionary histories, time-dependent rate effects become manifest over the span of 10-50 years of rate measurement. As a result, using an uncorrelated relaxed clock which allows each branch of the phylgeny to have its own evolutionary rate creates a very wide variation in the inferred mean substitution rate. The sigmoid shape of the time-dependent rate decay in the PoW model makes a very specific assumption about how the substitution rate varies as the time-span of rate measurement increases. Therefore, by allowing the short-term rate to vary in accordance with the posterior rate distribution, we can generate unwanted uncertainty in the PoW-transformed estimated divergence times.

2.4 Discussion

The PoW model creates an over-arching evolutionary framework that can reconcile and incorporate timescales derived from co-evolutionary and ancient DNA studies. Further substantive re-evaluations of timescales of other RNA and DNA viruses using this approach may provide new insights into their origins and evolutionary dynamics. We show how these can alter paradigms about how we think that particular viruses have evolved. Application of the PoW will place ancestors of divergent virus sequences far further back into the past than conventional reconstructions. We obtained a good fit between the pattern of modelled and observed substitution rate decay gradients over time using only a minimal number of assumptions about mutational fitness effects and proportion of sites evolving at a particular rate. We showed that our method is robust to substantial differences in

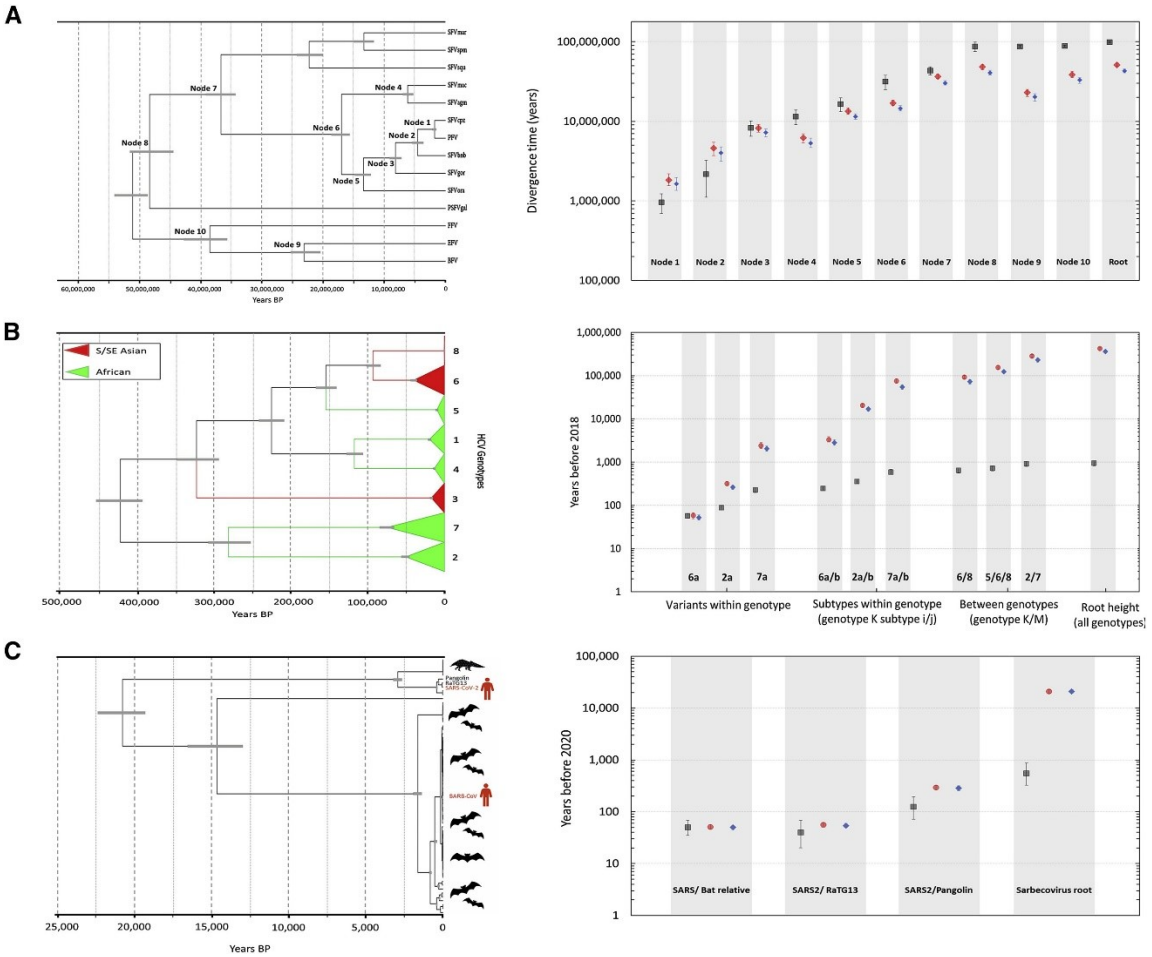


Figure 2.3: **The PoW-transformed time-calibrated phylogenies and estimated divergence times for FV, HCV, and sarbecovirus datasets.** (**A**) FV phylogeny and estimated divergence times between labelled internal nodes using the PoW model (red) and host calibrations [32] (black). (**B**) HCV (including all 8 genotypes) phylogeny and estimated divergence times for variants within genotypes, subtypes, and between genotypes using the PoW model (red) and a strict clock with HKY+ Γ_4 substitution model (black). (**C**) SARS-CoV-2 sarbecovirus phylogeny based on the non-recombinant alignment 3 (NRA3) [40] constructed using the PoW model (red) and a standard HKY+ Γ_4 substitution model with uncorrelated relaxed clock (black).

substitution rates among viral groups. By finding the short-term substitution rate (the flat part of the modelled rate decay) and the value of the fastest-evolving rate group (which sets the inflection point of the curve), the PoW model can reconstruct corrected substitution rates for virus genotypes with increasingly divergent nucleotide sequences.

While the empirical observation of the power-law rate decay has enabled the reconstruction of the timescales of association between some viruses and their hosts [32], these approaches are based on using a top-down description of rate decay that lacks an underlying biological basis. Furthermore, they require the use of multiple internal calibration points in order to estimate timescales. The PoW model does not require such calibrations and it does not exhibit substantial rate decay over short timescales (i.e., the flat part of the rate curve) before the fastest-evolving sites have saturated). This enables reliable inference of divergence time over shallow timescales; over such timescales a naive extrapolation of substitution rates using the empirical power-law can produce inaccurate divergence date estimations.

Our mechanistic model allows for a fraction of sites to evolve at different rates due to epistasis or nucleotide biases, requires a minimal number of assumptions, and no additional calibration information. This provides for the first time, a bottom-up model that can account for the empirical observation of the TDRP. The model is compatible with the notion of host-driven constraints on virus evolution [27], which represents a special case of the PoW model, but it does not require the constraints to be host driven and is generalisable. Furthermore, we find the substitution rate at the fastest-evolving rate groups in RNA viruses to be 1-2 orders of magnitude faster than DNA viruses. This provides RNA viruses with an access to a wider range of

sites that can evolve at intermediate substitution rates which, in turn, provides them with more possibilities for epistatic substitutions. We know that this is indeed the case for most RNA viruses due to their compact genomes, overlapping reading frames, and secondary structures.

2.5 Methods

2.5.1 Resource availability

Lead contact

Further information and requests for resources and datasets should be directed to and will be fulfilled by the Lead Contact, Aris Katzourakis aris.katzourakis@zoo.ox.ac.uk.

Data and code availability

All datasets and codes required to reproduce the analyses are available at https://github.com/mg878/PoW_model.

2.5.2 Data sets and simulation codes

Sarbecovirus data set

To minimise the effect of recombination when inferring time-tree phylogenies of Sarbecovirus, we use the putative recombination-free alignment from Boni et al. [40] with 66 sequences (also called the non-recombinant alignment 3, NRA3).

Hepatitis C virus data set

All coding complete genome sequences of HCV were downloaded from GenBank in May 2019. Those with annotated sample dates were then quality tested (completeness, lack of internal gaps, stop codons, ambiguous bases), and then

filtered for sequence similarity to each other. A threshold of 0.2 nucleotide sequence divergence (over the whole genome) was used to extract single examples of each subtype that were dated and supplemented these with 17 further sequences that were the most divergent examples of the same subtype (threshold of 17%). The set was further supplemented with examples of genotypes 1a, 1b and 3a, and references sequences of all subtypes to produce a final alignment of 120 sequences.

Foamy virus data set

We use the manually-curated *pol* nucleotide (3,351 nucleotides) alignments of 14 extant FVs from Aiewsakun and Katzourakis [25]. The dataset was examined for potential recombination by Aiewsakun and Katzourakis and no evidence for significant recombination was found. Because the viral tree topology is closely aligned with the host phylogeny, we assumed a long history of cospeciation for these viruses and matched the viral tree topology to that of the hosts (enforced certain taxon sets to be monophyletic in the BEAST analysis).

R package for the construction of PoW-transformed phylogenies

For time-stamped datasets (i.e., genomic samples collected at different time points), we first infer the mean rate using standard substitution models. We then construct an ultrametric distance tree using either the JC69 or HKY85 substitution models. To convert the distance trees into a PoW-transformed time tree, we first subsample distance trees that are produced by BEAST and convert each of them to time trees using the PoW model (see Equation S5 and Equation S6) assuming that the short-term rate (i.e., flat part of the sigmoid curve) is equal to the median substitution rate inferred from the previous step and that the fastest-evolving rate group (which sets

the inflection point) matches with the inferred μ_{max} from Table 2.1 based on the Baltimore group that the virus belongs to. Finally, we use TreeAnnotator to build a consensus tree and find the estimated median and 95% HPD node heights. This approach fully captures the uncertainty that may arise from the substitution model and tree topology. We note that while in the first step we may use substitution models with rate heterogeneity (such as a gamma distribution) and various clock models to infer the short-term rate, in the second step (i.e., constructing ultrametric distance trees) we must use a strict clock (i.e., rate = 1 across all branches) to infer the genetic distance between every pair of sequences.

To further capture the effect of variation in the inferred mean rate (from BEAST) and μ_{max} (from geometric least square fit) on the clock model and, ultimately, the PoW-transformed ultrametric distance trees, we can either randomly draw numbers from the posterior rate distribution or any other appropriate statistical distriubtions to find a range of parameter values for the short-term substitution rate, $\langle\mu\rangle$, and μ_{max}. This enables us to convert each sampled distance tree into a time tree with a unique pair of values for $\langle\mu\rangle$ and μ_{max}.

In summary, the following steps are required for the implementation of the PoW model:

Step 1: Estimating the short-term substitution rate. Here, we infer the short-term rate, $\langle\mu\rangle$, using time-stamped sequences by following a conventional BEAST analysis. If there are not enough time-stamped sequences in a given dataset to reliably infer the substitution rate, the rate can be given as an input to the code either in the form of a rate distribution or a fixed value based on known rate estimates from other studies.

Step 2: Estimating divergence. In order for the PoW model to generate a rescaled time tree, we need to construct a rooted distance tree to infer the amount of genetic distance between samples. We do so by constructing an ultrametric distance tree in BEAST using a standard JC69 or HKY substitution model (with no rate heterogeneity).

Step 3: Constructing a PoW-transformed time tree. Following the inference of the short-term substitution rate and construction of the distance tree in Steps 1 and 2, we can use the PoW model to convert the distances into time using Equation S5 (for distance trees constructed using the JC69 substitution model) and Equation S6 (for distance trees constructed using the HKY85 substitution model) such that the value of $\mu_{\max}$ is determined from Table 2.1 based on the Baltimore group that the virus dataset belongs to. We subsample from the posterior rate distribution and converged distance trees in Steps 1 and 2 by taking 1,000 random samples (ignoring the first 10% of the total Markov Chain Monte Carlo as burn-in) and converting each sampled tree using the PoW model.

Step 4: Constructing a consensus time tree. Finally, we construct a maximum clade credibility tree from the 1,000 PoW-transformed time trees in Step 3 using TreeAnnotator.

Substitution rate inference of simulated data sets

We simulate a neutral haploid Wright-Fisher population of size N_e with L evolving sites under a constant mutation rate μ per site such that every nucleotide (A, C, G, and T) can mutate to any other nucleotide at the same rate $\mu/3$ – mutation rate is equal to substitution rate under neutrality. We then sample from the entire population at two time points with an increasingly wider time gap, t^*. Initially, we

allow the population to evolve for $10N_e$ generations before taking the first sample to ensure that neutral coalescent events reach their steady state distribution and that the population, on average, coalesces every $2N_e$ generations. We then take the second sample t^* generations later and repeat this process 100 times to generate replicate sequences at both time points and run each set of simulations in BEAST 1.10 to estimate the substitution rate (Figure S1). We load the simulated sequences (along with their sampling times) on BEAST and use a strict molecular clock with a continuous-time Markov chain reference prior on substitution rates, a constant population coalescent prior, and a Jukes-Cantor substitution model. For every simulated set, the Markov chain Monte Carlo was run for 10,000,000 steps and parameter convergence was inspected visually.

2.5.3 Mathematical formulation

Power-law rate decay due to site saturation

For a sequence that has diverged from its ancestor *t* generations ago under a constant and uniform substitution rate μ per site per year, the proportion of pairwise differences, $p(t)$, is given by [42]

$$p(t) = \alpha(1 - e^{-\mu t/\alpha}) \quad \text{(S1)}$$

such that α is the maximum proportion of pairwise differences and is given by $\alpha = 1 - \sum_i \pi_i^2$ where π_i is the base frequency of the *i*th nucleotide or amino acid. Assuming that d is the 'true' genetic distance between a pair of homologous sequences (i.e., $d = \mu t$), we can estimate the observed genetic distance, $\hat{d}$, with an observed proportion of pairwise differences, $\hat{p}$, using the Felseinstein's 1981 substitution model [43]

$$\hat{d} = \mu t = -\alpha_M Ln\{1 - \hat{p} / \alpha_M\} \quad \text{(S2)}$$

where α_M is the expected saturation frequency set by the substitution model. If the model correctly identifies the saturation frequency, i.e. $\alpha_M = \alpha$, Equation S2 accurately predicts the true genetic distance, i.e. $\hat{d} = d$, as long as the divergence time $t \ll \alpha / \mu$. As $p(t)$ approaches saturation frequency at $t^* \approx \alpha / \mu$, the observed proportion of pairwise differences will be bound by the number of evolving sites. For instance, if the saturation frequency is $\alpha = 3/4$, i.e. a standard Jukes-Cantor substitution model, to distinguish between an observed pairwise difference of $\hat{p}^* = 0.74$ and $\hat{p}^* = 0.741$ requires approximately one thousand evolving sites, all evolving at rate μ. Thus, beyond t^*, the estimated distance, $\hat{d}$, will remain effectively unchanged. In other words, if the pair diverge beyond the saturation point, the inferred rate follows a power-law rate drop with a slope -1 on a log-log plot (see grey curve in Figure S1A).

In the presence of purifying selection and/or amino acid and nucleotide biases, the substitution model in Equation S2 may overestimate the maximum proportion of pairwise differences, i.e. $\alpha_M > \alpha$. For instance, if we apply a Jukes-Cantor measure of nucleotide distance to a pair of sequences with a particular site preference that equally favours only two (out of the four) nucleotides, i.e. $\alpha = 1/2$, the proportion of pairwise differences reaches saturation much earlier than what the chosen substitution model would predict. As a result, similar to the previous example ($\alpha_M = \alpha$), the estimated rate drops as a power-law with slope -1 (orange curve in Figure S1A) after reaching the saturation point, i.e. $\hat{\mu} \approx \alpha_M Ln\{\alpha_M / (\alpha_M - \alpha)\} / t$.

In the opposite extreme, i.e., when $\alpha_M < \alpha$, the substitution model underestimates the true saturation frequency. Thus, the observed proportion of pairwise differences surpasses the level predicted by the substitution model, i.e. $\hat{p}/\alpha_M \gtrsim 1$, at which point the estimated substitution rate, $\hat{\mu} \approx \alpha Ln\{\alpha/(\alpha-\alpha_M)\}/t$, goes to infinity (blue curve in Figure S1A).

Saturation of sites in the presence of rate heterogeneity

In the presence of rate heterogeneity, a fraction of sites may evolve at a rate that is much slower (or faster) than some other sites. In principle, this process may involve *M* different rate groups such that each group *i* evolves at rate μ_i and occupies a fraction of sites m_i. Thus, the proportion of pairwise differences would be given by

$$p(t) = \sum_{i=1}^{M} m_i \alpha_i (1 - e^{-\mu_i t/\alpha_i}) \quad \text{(S3)}$$

such that $\sum_i m_i = 1$ with the mean substitution rate $\langle \mu \rangle = \sum_i m_i \mu_i$. If we apply a measure of distance based on Equation S2 to an evolutionary process with rate heterogeneity, Equation S3, it can reliably infer the mean substitution rate up to when the sites belonging to the fastest substitution rate, $\mu_{\max}$, reach saturation after $t \sim \mu_{\max}^{-1}$ at which point the pattern of time-dependent rate decay emerges and the mean substitution rate gradually plateaus at a value corresponding to the substitution rate of the slowest-evolving sites, $\mu_{\min}$. Once all the rate groups reach saturation at time $t \sim \mu_{\min}^{-1}$, the power-law rate decay with slope -1 emerges.

For instance, if the sequence evolution involves two rate groups, i.e. *M*=*2*, a fraction of sites m_1 may evolve neutrally at rate $\mu_1 = \mu$ and the remaining sites $(1 - m_1)$ evolve epistatically such that a pair of sites need to mutate simultaneously to

recover the wild-type fitness, i.e. $\mu_2 = \mu^2$. Assuming the saturation frequency across all sites is equal and that the model correctly identifies their frequency, i.e. $\alpha_i = \alpha_M = \alpha$, we can use Equation S2 to recover the expected substitution rate $\langle \mu \rangle = m_1\mu + (1-m_1)\mu^2$. As the fast-evolving sites approach the saturation point at $t_1 \approx \alpha/\mu$, the rate decay emerges and a sharp decline in estimated substitution rate follows while the remaining fraction of sites, $(1-m_1)$, keep accumulating new substitutions at rate μ^2, slowing down the speed of the rate decay until those sites also reach saturation at $t_2 \approx \alpha/\mu^2$ beyond which point the entire genome reaches saturation and the power-law rate decay with slope -1 emerges. Figure S1B shows that as the proportion of slow-evolving sites increases, the mean substitution rate goes down and the slope of the time-dependent rate decay becomes less steep.

Although our focus so far has been on the saturation of pairwise differences and how it can create a time-dependent rate effect, the same holds true when tracking the evolutionary changes of a large number of sequences through time. Using a standard Jukes-Cantor substitution model on a set of simulated sequences, both in the absence and presence of rate heterogeneity, we can recreate similar patterns of time-dependent rate decay and show that, over longer timescales, i.e. when the divergence time between two populations is much longer than the typical coalescent times, the variation in inferred substitution rates is dominated by the saturation along the longest (internal) branch connecting the two populations (Figure S1C-H). We also find that, over short timescales, systematic under-estimation of the Time to the Most Recent Common Ancestor (TMRCA) results in inflated substitution rate estimates (Equation S6).

Saturation of sites under the PoW model

Under the PoW model, the virus substitution rate is categorised into M discrete rate classes such that there is a fixed incremental difference between any two consecutive rate groups, $\mu_{i+1} = \Delta M \mu_i$, with a common ratio ΔM. Rate groups range from those evolving the fastest, at rate $\mu_{\max}$, to the ones evolving at the host substitution rate, $\mu_{\min}$. The fraction of sites, m_i, in each rate group i, evolving at rate, μ_i, is an exponentially distributed number, $m_i = Ce^{\lambda i}$, where C is the normalisation factor, $C = 1/\sum_{j=1}^{M} e^{\lambda j}$, and the exponent coefficient, λ, sets the tendency of sites to be either mostly slowly- ($\lambda < 0$) or rapidly- ($\lambda > 0$) evolving. Given a fixed common ratio between consecutive rate groups, the substitution rate at the fastest-evolving sites can be determined by finding the total number of rate groups, M, which, in turn, sets the inflection point for when the time-dependent rate decay emerges. Once the fastest-evolving sites reach the saturation point, other rate groups that evolve more slowly (e.g. via epistatic and compensatory substitutions), saturate chronologically as the timespan of rate measurement, t, increases. This chronological saturation effect continues until the inferred rate decays to the host substitution rate, $\mu_{\min}$. Therefore, the time-dependent rate curve, according to the PoW model is given by

$$\hat{\mu}(t) = -\alpha_M Ln\left(1 - \frac{1}{\alpha_M}\sum_{i=1}^{M} \alpha m_i \left(1 - e^{-\mu_i t/\alpha}\right)\right)/t \qquad \text{(S4)}$$

where the observed genetic distance between the derived and ancestral sequences, $\hat{d} = \hat{\mu} t$, is proportional to the timespan of rate measurement, t. While over short timescales, i.e. $t \ll 1/\mu_{\max}$, several methodological (e.g. internal node calibration errors) and biological (e.g. purifying selection) artefacts may inflate the

substitution rate estimates in viruses (i.e., such that $\hat{\mu}(t)$ underestimates the inferred rates), over longer time-scales (i.e., typically after a few years) the rate estimates are expected to be closer to the mean (short-term) substitution rate, $\langle \mu \rangle = \sum_{i=1}^{M} m_i \mu_i$. Over such timescales, there is no site saturation, i.e., the rate curve is flat, and the inferred short-term substitution rates can be reliably used to estimate divergence times between samples without significant influence from site saturation. We also note that the finer the gap between consecutive rate groups, ΔM , gets, the more accurate the estimated rate curve becomes. However, typically, less than 50 rate groups (i.e., $M < 50$) is sufficient for all predictions. For a fixed ΔM , the exponent coefficient, λ , together with the number of rate groups, M , are the two free parameters of the PoW model which set the short-term, $\langle \mu \rangle$, and maximum, $\mu_{\max}$, substitution rates for any given data set. While Equation S4 assumes that the saturation frequency, α , across all sites is the same, there can be instances where, due to site preferences, some mutations do not appear at certain positions in the virus genome. This can result in a reduction in saturation frequency at those sites (i.e., $\alpha < 3/4$). However, to avoid overparametrising the model, in the absence of sufficient data, we assume identical saturation frequencies across all sites.

Distance tree transformation using the PoW model

Equation S4 allows for a one-to-one map between the inferred genetic distance and divergence time. Therefore, by estimating the genetic distance (in units of substitutions), $\hat{d}$, between any pair of sequences under a JC69 substitution model (i.e., $\alpha_M = \alpha = 3/4$) we can solve Equation S5 to find the divergence time, *t*, since the common ancestor of each pair using the PoW model.

$$\hat{d} = -3/4 Ln\left(1 - \sum_{i=1}^{M} m_i \left(1 - e^{-4\mu_i t/3}\right)\right) \quad \text{(S5)}$$

More generally, we can apply other, more complex, substitution models to infer the genetic distance between pairs of sequence. For instance, under the Tamura-Nei substitution model (TN93) where there is an analytically tractable formula for distance [44], the PoW-transformed equation to find divergence time, *t*, is given by

$$\hat{d} = \frac{2\pi_T \pi_C}{\pi_Y}(a_1 - \pi_R b) + \frac{2\pi_A \pi_G}{\pi_R}(a_2 - \pi_Y b) + 2\pi_Y \pi_R b, \quad \text{(S6)}$$

$$\widehat{k_1} = \frac{a_1 - \pi_R b}{\pi_Y b},$$

$$\widehat{k_2} = \frac{a_2 - \pi_Y b}{\pi_R b},$$

where

$$a_1 = -Ln\left(1 - \frac{\pi_Y S_1}{2\pi_T \pi_C} - \frac{V}{2\pi_Y}\right),$$

$$a_2 = -Ln\left(1 - \frac{\pi_R S_2}{2\pi_A \pi_G} - \frac{V}{2\pi_R}\right),$$

$$b = -Ln\left(1 - \frac{V}{2\pi_Y \pi_R}\right)$$

$$S_1 = 2\pi_T p_{TC}(t),$$

$$S_2 = 2\pi_A p_{AG}(t),$$

$$V = 2\pi_T p_{TA}(t) + 2\pi_T p_{TG}(t) + 2\pi_C p_{CA}(t) + 2\pi_C p_{CG}(t),$$

such that k_1 and k_2 are the two different types of transition rates (transversions are all assumed to occur at the same rate) according to the TN93 model, π_i is the nucleotide equilibrium base frequency, and $p_{ij}(t)$ is the transition probability (not to be confused with transition rate) to go from nucleotide *i* to *j* according to the PoW model, $(i, j) = \{A, C, G, T\}$. While the equilibrium base frequencies and transition rates can be found numerically from the phylogenetic analysis using BEAST, the transition probabilities are found from the eigenvalues of the transition matrix $P(t) = \{p_{ij}(t)\} = e^{Qt}$ (see ref [44] for more details on the calculations). For example, according to the PoW model, $p_{TA}(t) = \sum_{i=1}^{M} m_i \pi_A (1 - e^{-\beta_i t})$, where β_i is the transversion rate from rate group *i*. We can also find a relationship between the average substitution rate per rate group, μ_i, and β_i which is given by

$$\mu_i = 2\beta_i (k_1 \pi_T \pi_C + k_2 \pi_A \pi_G + \pi_Y \pi_R).$$

Similarly, Equation S6 can be used for the PoW transformation to be applied to a wider range of substitution models such as the HKY85 substitution model (i.e., $k_1 = k_2$).

Saturation of sites for simulated datasets

In Figure S1C-H, we recreate the time-dependent pattern of rate decay both in the absence and presence of rate heterogeneity across sites, using a standard substitution model on simulated data. We find that while the inferred substitution rates exhibit a power-law rate decay with slope -1 over longer time intervals (see Figure S1C,D), the inferred TMRCAs tend to be overestimated with a similar (inverse) power-law trend, i.e. $\hat{t} \sim 1/\hat{\mu}$ (see Figure S2E,F). We also find an

unexpected time-dependent rate effect over short timescales. This occurs when the observation gap, t^*, is much shorter than the expected coalescent time of the population, i.e. $t^* \ll 2N_e$. This also results in the underestimation of true TMRCAs which systematically makes worse predictions for higher substitution rates. The expected rate curves (dashed lines shown in Figure S2C,D) can be approximated by replacing $p(t)$ from Equation S1 into $\hat{p}$ from Equation S2 which is given by

$$\hat{\mu}(t^*) \approx -\alpha_M Ln\left\{1-\frac{1}{L\alpha_M}\left\lfloor L\sum_{i=1}^{M} m_i\alpha(1-e^{-\mu_i(t^*+2N_e)/\alpha})\right\rfloor\right\}/(t^*+T) \quad \text{(S6)}$$

such $\lfloor . \rfloor$ is the floor function which represents the finite size effect of having L evolving sites on saturation frequency. The mean divergence time between the two populations is approximately $t \approx t^* + 2N_e$ and the inferred divergence time is $\hat{t} \approx t^* + T$ – this resembles the mis-calibration effects reported elsewhere (see Equation 2 in ref [45]). The reason why Equation S4 only works as an approximate is that the median inferred TMRCA from simulation results, T, also varies with respect to observation gap t^* (see Figure S1E,F). However, for $t^* \gg 2N_e$, the variation in T becomes negligible compared to $t^*$and only has second-order effects on inferred substitution rates. Figure S2G,H show the tree topology under the two extremes, $t^* \ll 2N_e$ and $t^* \gg 2N_e$, respectively. It indicates that, over long timescales, the time-dependent rate effects are dominated by the very long (and saturated) branch connecting the two populations that are t^* generations apart. As a result, the decay dynamics looks very similar to the analytical results in Figure S1A,B where we estimate the substitution rate between a pair of sequences separated by a very long branch.

2.6 References

1. Duffy, S., L.A. Shackelton, and E.C. Holmes, *Rates of evolutionary change in viruses: patterns and determinants.* Nat Rev Genet, 2008. **9**(4): p. 267-76.
2. Pybus, O.G. and A. Rambaut, *Evolutionary analysis of the dynamics of viral infectious disease.* Nat Rev Genet, 2009. **10**(8): p. 540-50.
3. Holland, J.J., *Transitions in understanding of RNA viruses: a historical perspective.* Curr Top Microbiol Immunol, 2006. **299**: p. 371-401.
4. Sharp, P.M., et al., *Origins and evolution of AIDS viruses: estimating the time-scale.* Biochem.Soc.Trans., 2000. **28**(2): p. 275-282.
5. Nakano, T., et al., *Viral gene sequences reveal the variable history of hepatitis C virus infection among countries.* Journal of Infectious Diseases, 2004. **190**(6): p. 1098-1108.
6. Nelson, M.I. and E.C. Holmes, *The evolution of epidemic influenza.* Nat Rev Genet, 2007. **8**(3): p. 196-205.
7. Sawyer, S.L., M. Emerman, and H.S. Malik, *Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G.* PLoS Biol, 2004. **2**(9): p. E275.
8. Li, W., et al., *Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2.* Embo j, 2005. **24**(8): p. 1634-43.
9. Allison, A.B., et al., *Host-specific parvovirus evolution in nature is recapitulated by in vitro adaptation to different carnivore species.* PLoS Pathog, 2014. **10**(11): p. e1004475.
10. Bhatt, S., et al., *The evolutionary dynamics of influenza A virus adaptation to mammalian hosts.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1614): p. 20120382.
11. Wain, L.V., et al., *Adaptation of HIV-1 to its human host.* Mol Biol Evol, 2007. **24**(8): p. 1853-60.
12. Alves, J.M., et al., *Parallel adaptation of rabbit populations to myxoma virus.* Science, 2019. **363**(6433): p. 1319-1326.
13. Katzourakis, A. and R.J. Gifford, *Endogenous viral elements in animal genomes.* PLoS Genet, 2010. **6**(11): p. e1001191.
14. Katzourakis, A., et al., *Discovery and analysis of the first endogenous lentivirus.* Proc Natl Acad Sci U S A, 2007. **104**(15): p. 6261-5.
15. Gifford, R.J., et al., *A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution.* Proc Natl Acad Sci U S A, 2008. **105**(51): p. 20362-7.
16. Sharp, P.M. and P. Simmonds, *Evaluating the evidence for virus/host co-evolution.* Curr Opin Virol, 2011. **1**(5): p. 436-41.
17. Katzourakis, A., et al., *Macroevolution of complex retroviruses.* Science, 2009. **325**(5947): p. 1512.
18. Aiewsakun, P. and A. Katzourakis, *Marine origin of retroviruses in the early Palaeozoic Era.* Nat Commun, 2017. **8**: p. 13954.
19. Muhlemann, B., et al., *Ancient human parvovirus B19 in Eurasia reveals its long-term association with humans.* Proc Natl Acad Sci U S A, 2018. **115**(29): p. 7557-7562.
20. Muhlemann, B., et al., *Ancient hepatitis B viruses from the Bronze Age to the Medieval period.* Nature, 2018.
21. Duggan, A.T., et al., *17(th) Century Variola Virus Reveals the Recent History of Smallpox.* Curr Biol, 2016. **26**(24): p. 3407-3412.

22. Dux, A., et al., *Measles virus and rinderpest virus divergence dated to the sixth century BCE.* Science, 2020. **368**(6497): p. 1367-1375.
23. Duchene, S., E.C. Holmes, and S.Y. Ho, *Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates.* Proc Biol Sci, 2014. **281**(1786).
24. Ho, S.Y., et al., *Time-dependent rates of molecular evolution.* Mol Ecol, 2011. **20**(15): p. 3087-101.
25. Aiewsakun, P. and A. Katzourakis, *Time-Dependent Rate Phenomenon in Viruses.* Journal of Virology, 2016. **90**(16): p. 7184-7195.
26. Wertheim, J.O. and S.L.K. Pond, *Purifying Selection Can Obscure the Ancient Age of Viral Lineages.* Molecular Biology and Evolution, 2011. **28**(12): p. 3355-3365.
27. Simmonds, P., P. Aiewsakun, and A. Katzourakis, *Prisoners of war - host adaptation and its constraints on virus evolution.* Nat Rev Microbiol, 2019. **17**(5): p. 321-328.
28. Meiering, C.D. and M.L. Linial, *Historical perspective of foamy virus epidemiology and infection.* Clinical Microbiology Reviews, 2001. **14**(1): p. 165-169.
29. Schweizer, M., et al., *Genetic stability of foamy viruses: Long-term study in an African green monkey population.* Journal of Virology, 1999. **73**(11): p. 9256-9265.
30. Katzourakis, A., et al., *Macroevolution of Complex Retroviruses.* Science, 2009. **325**(5947): p. 1512-1512.
31. Switzer, W.M., et al., *Ancient co-speciation of simian foamy viruses and primates.* Nature, 2005. **434**(7031): p. 376-380.
32. Aiewsakun, P. and A. Katzourakis, *Time dependency of foamy virus evolutionary rate estimates.* Bmc Evolutionary Biology, 2015. **15**.
33. Crawford, G. and C. Williams, *A Note on the Analysis of Subjective Judgment Matrices.* Journal of Mathematical Psychology, 1985. **29**(4): p. 387-405.
34. Membrebe, J.V., et al., *Bayesian Inference of Evolutionary Histories under Time-Dependent Substitution Rates* Molecular Biology and Evolution, 2019. **36**(8): p. 1793-1803.
35. Simmonds, P. and D.B. Smith, *Investigation of the pattern of diversity of hepatitis C virus in relation to times of transmission.* Journal of Viral Hepatitis, 1997. **4**: p. 69-74.
36. Markov, P.V., et al., *Phylogeography and molecular epidemiology of hepatitis C virus genotype 2 in Africa.* Journal of General Virology, 2009. **90**: p. 2086-2096.
37. Iles, J.C., et al., *Phylogeography and epidemic history of hepatitis C virus genotype 4 in Africa.* Virology, 2014. **464**: p. 233-243.
38. Pybus, O.G. and R.R. Gray, *VIROLOGY The virus whose family expanded.* Nature, 2013. **498**(7454): p. 310-311.
39. Pybus, O.G. and J. Theze, *Hepacivirus cross-species transmission and the origins of the hepatitis C virus.* Current Opinion in Virology, 2016. **16**: p. 1-7.
40. Boni, M.F., et al., *Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic.* Nature Microbiology, 2020. **5**(11): p. 1408-1412.
41. Souilmi, Y., et al., *An ancient coronavirus-like epidemic drove adaptation in East Asians from 25,000 to 5,000 years ago.* Curr Biol, 2021. **31**: p. 1-11.

42. Tajima, F. and M. Nei, *Estimation of Evolutionary Distance between Nucleotide-Sequences.* Molecular Biology and Evolution, 1984. **1**(3): p. 269-285.
43. Felsenstein, J., *Evolutionary Trees from DNA-Sequences - a Maximum-Likelihood Approach.* Journal of Molecular Evolution, 1981. **17**(6): p. 368-376.
44. Yang, Z., *Molecular Evolution: A Statistical Approach.* Oxford University Press, 2014.
45. Ho, S.Y.W., et al., *Time dependency of molecular rate estimates and systematic overestimation of recent divergence times.* Molecular Biology and Evolution, 2005. **22**(7): p. 1561-1568.

Statement of Authorship for joint/multi-authored papers for PGR thesis.

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses
Publication Status	Published
Publication Details	Ghafari M, Pybus OG, Simmonds P, and Katzourakis A. A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses. Current Biology 31: 4689-4696e5, 2021.

Student confirmation

Student Name	Mahan Ghafari
Contribution to the paper	MG wrote the paper and conducted all the analysis under the supervision from AK. PS prepared the hepatitis C virus alignments. All authors provided editorial assistance.
Signature 	Date 11/05/2022

Supervisor confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title	Professors Aris Katzourakis and Oliver G Pybus	
Supervisor comments		
Signature 		Date 11/05/2022

Chapter 3: Deep evolutionary history of simian and prosimian immunodeficiency viruses

3.1 Abstract

While paleovirological evidence supports the idea that the ancestors of modern simian immunodeficiency viruses (SIVs) have been evolving in nonhuman primates for millions of years, standard molecular clock methods calibrated using contemporary sequences significantly underestimate their time of origin. This discrepancy is attributed to the time-dependent nature of evolutionary rate estimates whereby the rate of virus evolution varies with respect to the time window of its measurement. Here we show that by correcting for such time-dependent rate effects using a recently developed mechanistic evolutionary model, we can successfully reconstruct the evolutionary history of SIV across different timescales and estimate the age of lineages that gave rise to HIV-1 and HIV-2. We also find that the most recent common ancestor of SIVs lived nearly a million years ago and that most SIV lineages are at least tens of thousands of years old. Our phylogenetic estimates also show that the most recent common ancestor of SIV and prosimian immunodeficiency virus (pSIV) lived around 17 million years ago which suggests that the most likely explanation for the origins of primate lentiviruses is a terrestrial transfer of lentiviruses between African and Malagasy primates or transfer mediated by an aerial vector species. These results may offer an explanation as to why lentiviruses are not as pathogenic in nonhuman primates due to their very long history of coevolution with their hosts, as opposed to the very recent evolutionary origins of highly pathogenic HIV strains in humans, which may have resulted in the

development of strong immune defence mechanisms that slowed or blocked retroviral reproduction.

3.2 Introduction

The discovery of the human immunodeficiency viruses types 1 and 2 (HIV-1, HIV-2) in the 1980s [1-3] led to the identification of another primate lentivirus known as simian immunodeficiency virus (SIV) that are close relatives of HIV and found in various non-human primates in sub-Saharan Africa including chimpanzees, gorillas, and sooty mangabeys. Subsequent to their discovery, there were extensive investigations into the origins of HIV and identification of cross-species transmission events from SIVs to humans (see ref. [4] for a recent review). These studies revealed that SIVcpz transmissions from chimpanzee subspecies, *Pan troglodytes troglodytes*, and SIVsmm transmission from sooty mangabeys, *Cercocebus torquatus atys*, were responsible for the emergence of HIV-1 and HIV-2 strains respectively in humans [5-7].

Molecular epidemiological studies based on extant SIV and HIV sequences showed a very high rate of nucleotide sequence change over time that predicted a very recent evolutionary origin for these primate lentiviruses [4, 8]. If we take HIV's molecular clock rate estimate of around 10^{-3} substitutions per nucleotide per year and extrapolate that across the entire primate lentivirus phylogeny, it leads us to estimate that the most recent common ancestor of the entire group lived no earlier than a couple of centuries ago. More advanced molecular clock methods based on relaxed clocks which, to a certain degree, account for rate variation across different taxa, would push the common ancestor of extant SIV sequences in chimpanzee back to a few hundreds to 2,000 years before present [9]. However, all of these

estimates based on molecular clock analysis stand in stark contrast to paleovirological evidence from endogenous viral elements (EVEs) that show lentiviruses have infiltrated their hosts' germlines millions of years ago. These examples include the rabbit endogenous lentivirus type K which became endogenised in European rabbits about 12 million years ago [9] and two prosimian endogenous lentiviruses, fat-tailed dwarf lemur (pSIVdfl) and grey mouse lemur (pSIVgml), that independently integrated into the lemur's genome nearly 4 million years ago [10, 11]. The sheer resemblance of these integrated virus elements to extant viruses suggests that they have evolved at a much slower rate over time and in a host-dependent manner.

Molecular clock predictions of a recent evolutionary origin for HIV-1 and HIV-2 are supported by their initially restricted distribution in Central / West Africa - if either had widely infected humans much earlier than the late nineteenth century, they would have been exported to the Americas and northern Europe via slave trade routes between the sixteenth and nineteenth century and likely persisted in those populations until the present. However, predictions of relatively recent origins for the more divergent strains of SIV and other primate lentiviruses are contradicted by the results of an analysis of strains of SIV infecting four simian species on the island of Bioko. The fauna of this island had been separated from the mainland for around 10,000 years following the rise in sea levels at the end of the last ice age [12]. Therefore, SIV strains infecting these monkeys must have diverged from their mainland relatives for at least this period. However, SIV strains showed much less sequence divergence than would have been expected by extrapolation of short-term substitution rates. Indeed, using the island separation as a biogeographic calibration point, the authors were able to push the estimate of the time to the most

recent common ancestor of SIV strains infecting monkeys and apes in Africa back to nearly 76,000 years before present, with a corresponding clock rate estimate of ~10^{-5} substitutions per nucleotide per year. However, they also noted that this is still likely an underestimate as they found evidence of substitution saturation in their phylogenetic analysis [12].

These independent lines of evidence on the evolutionary history of primate lentiviruses, as outlined above, from the discovery of EVEs in lemur's genome dating back to millions of years ago to recent evolutionary origins of HIV to the discovery of SIV on the Bioko Island, all suggest that lentiviruses evolve at markedly different rates across different timescales and there is currently no unified evolutionary framework to explain this phenomenon. The time-dependency of evolutionary rate estimates, however, is not unique to primate lentiviruses and have also been reported in other retroviruses [13, 14]. One of the major consequences of the time-dependent nature of evolutionary rate estimates is that it results in significant underestimation of the timescale for deeper evolutionary histories of a virus if a shallow calibration point is chosen for the analysis. Similarly, a relatively slow paeleovirological rate estimate (such as the rate inferred from comparison of Bioko and mainland SIV strains) is incompatible with rates of sequence change of HIV-1 and other lentiviruses estimated from recent sampling.

Through developing a unified framework for understanding the variation of evolutionary rates in primate lentiviruses over time, we can shed light on their evolutionary history and better understand widespread lentivirus-primate interactions which is particularly relevant for clinically important pathogens such as HIV-1. In this study, we use a mechanistic evolutionary model that takes into account the time-dependent rate effects in viruses [15] to reconstruct the

evolutionary history of SIV and estimate the age of SIV lineages that gave rise to HIV-1 and HIV-2. We also estimate the time to the most recent common ancestor of SIV and the pSIV lentiviruses infecting lemurs and discuss how it may explain the origins of primate lentiviruses.

3.3 Results

We infer phylogenies for 25 species of SIV and every human-to-human transmissible lineage of HIV (HIV-1 groups M/N/O and HIV-2 groups A/B) and compare them against their host phylogenies (see Methods) to identify cross-species transmission events and assess the degree of lentivirus-host association. In agreement with previous studies [16, 17], we documented a substantial degree of cross-host species transmission SIV lineages. While there was a tendency for more closely related species to exchange SIV strains, there was also evidence of long codivergence between some SIV lineages and their hosts [18] that stretches back to hundreds of thousands of years (Figure 3.1).

We also compare the estimated time to the most recent common ancestor (TMRCA) of HIV and SIV strains using the prisoner of war (PoW) model [15] and a standard relaxed clock model. We use extant HIV and SIV sequences from the *pol* locus (Figure 3.2A) and SIV sequences that only cover the *integrase* region from the Bioko Island and other regions of the world (Figure 3.2B). We find that while the TMRCA estimates for HIV-1 and HIV-2 groups based on the PoW model and a standard molecular clock method are largely similar over short timescales, they differ significantly for SIV lineages across longer evolutionary periods. Indeed, molecular clock calibrations using short-term rate estimates of SIV cannot accurately account for saturation effects and fail to reconstruct their deep

evolutionary history (see estimated TMRCA using a relaxed clock model in ref. [9] shown in Figure 3.2A).

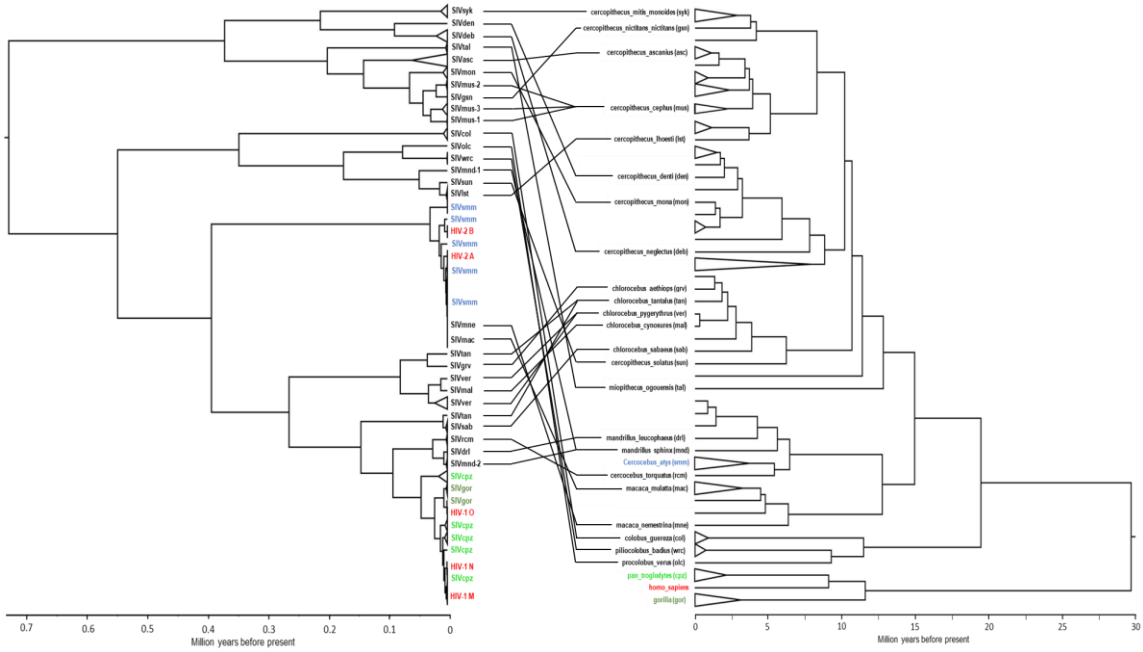
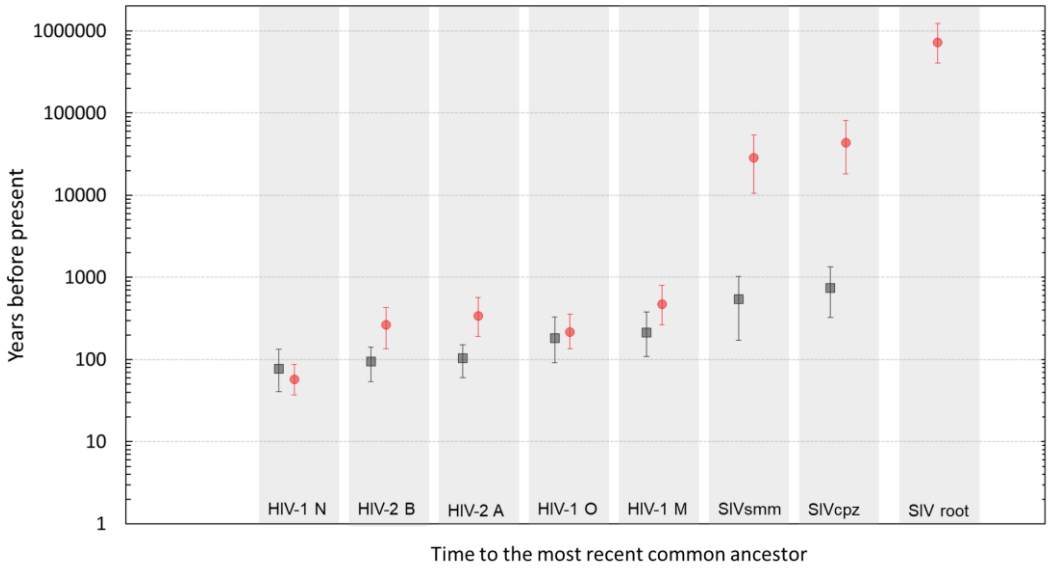


Figure 3.1: **Co-evolutionary history of SIVs (left) and their simian host (right).** SIV PoW-transformed time tree (left) is compared to that of their hosts (right, previously published in [33]). Solid lines between the two trees indicate SIV-host associations. Species abbreviations for the SIV lineages are cpz (chimpanzee), gor (gorilla), mnd (mandrill), drl (drill), rcm (red-capped mangabey), sab (sabaeus monkey), tan (tantalus monkey), ver (vervet monkey), mal (malbrouck monkey), grv (grivet monkey), mac (*rhesus macaque*), mne (pig-tailed macaque), smm (sootey mangabey), lst (L'hoest's monkey), sun (sun-tailed monkey), wrc (western red colobus), olc (olive colobus), col (*colobus guereza*), mus (moustached guenon), gsn (greater spot-nosed monkey), mon (mona monkey), asc (red-tailed guenon), tal (northern talapoin), deb (de brazza's monkey), den (dent's mona monkey), and syk (sykes's monkey).

A



B

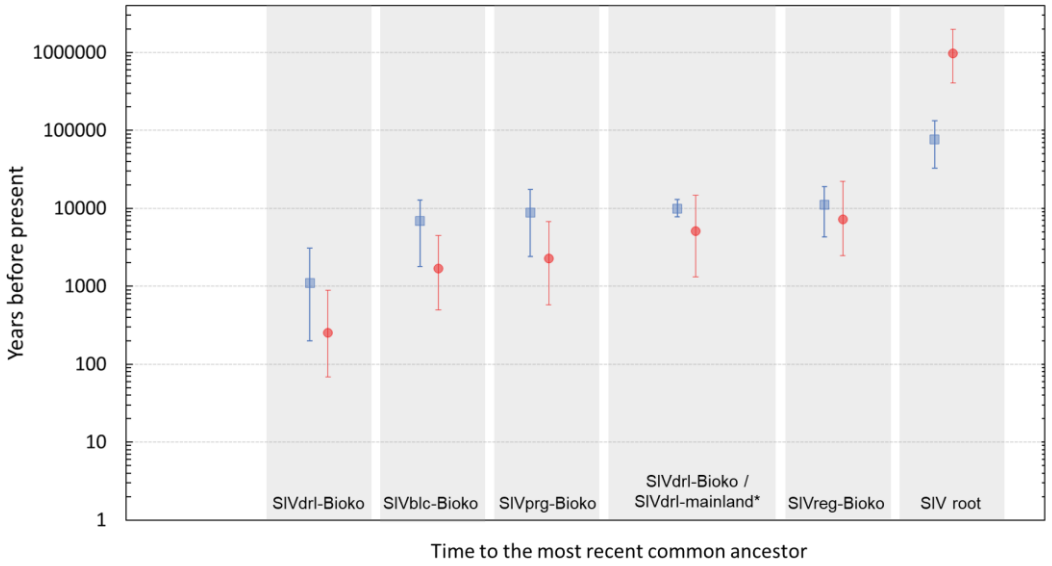


Figure 3.2: **Estimated divergence times for HIV and SIV strains. (A)** Estimated divergence times for HIV-1 and HIV-2 strains, SIV root, and SIV in chimpanzee (SIVcpz) and sootey mangabeys (SIVsmm) using a standard relaxed clock model (black) [9] and PoW-transformed time tree based on the *pol* dataset (red). **(B)** Estimated divergence times for the SIV root and various SIV lineages in the island of Bioko using a biogeographic calibration point and a relaxed clock model (blue) [12] and PoW-transformed time tree based on the integrase gene dataset (red). Vertical lines around each point in both graphs represents the 95% HPD. *The ancestor of SIV drill (SIVdrl) and SIV drill Bioko (SIVdrl-Bioko) is dated back to ~10,000 years ago during the isolation of Bioko island from mainland [12]. Species abbreviations for the SIV lineages are drill (drl), blc (black colobus), preuss's guenon (prg), and red-eared guenon (reg).

The problem with using standard substitution models for estimating TMRCA of SIV is twofold in the case of the analysis of SIV strains discovered on the Bioko Island (see estimated TMRCA using a relaxed clock model in ref. [12] shown in Figure 3.2B). This is because using a biogeographic calibration point (island separation around 10,000 years before present) would imply that any TMRCA estimate that is significantly younger than this calibration point is overestimated (e.g., most recent common ancestor of the SIV drill lineage in Bioko Island). Similarly, any TMRCA estimate that is significantly older than this biogeographic calibration point is underestimated (e.g., SIV root). For both the *pol* locus and *integrase* gene data sets, we find that the TMRCA of all SIV lineages is nearly 1 million years before present with predictions based on the *integrase* gene data set being slightly older as they only include SIV samples and have a slightly slower estimated molecular clock rate compared to the *pol* data set which includes both HIV and SIV samples (see Methods).

The discovery of lentiviruses in strepsirrhine primates on the island of Madagascar raises questions about the evolutionary history of primate lentiviruses, particularly the time and direction of transfer of pSIV between Africa and Madagascar [11]. Given that Madagascar has been isolated from Africa for nearly 160 million years, the presence of a lentivirus in lemurs on the island could be an indication of either (1) host-virus cospeciation dating back to 85 million years ago when the strepsirrhine-haplorrhine split occurred or (2) transfer of lentiviruses between Malagasy strepsirrhines and African haplorrhines during the last episode of colonisation of Madagascar via a land bridge by a group of mammals or a nonprimate vector nearly 14 million years ago or (3) transfer of lentiviruses via an aerial vector species (see ref. [11] for a detailed discussion on these three

scenarios). This implies that, at a minimum, there has been an association between haplorrhine and strepsirrhine lentivirus lineages with their hosts for 14 million years when the most recent terrestrial mammal invasion occurred [19]. To evaluate the likelihood of these three evolutionary scenarios, we used the PoW model to estimate the TMRCA of pSIV and SIV. Our analysis based on the *pol* locus suggests that the most recent common ancestor of SIV and pSIV lived 17.1 (95% highest posterior density: 13.2 - 20.7) million years before present (Figure 3.3). We note that in order to estimate this, we need to account for the fact that the pSIV insertion into *M. murinus* germline took place nearly 4.2 million years ago and that the SIV and pSIV samples are therefore not contemporaneous [10] (see Methods). The ~17 million years estimate for the TMRCA of pSIV and SIV is much younger than the ~85 million years prediction based on scenario (1). Therefore, our result does not support the ancestral codivergence as a likely scenario for the evolution of primate lentiviruses. However, the TMRCA estimate based on the PoW model is closely aligned with the timing of the most recent colonisation of Madagascar via a land bridge or a nonprimate vector around 14 million years ago as described in scenario (2). This result is also compatible with scenario (3) whereby an aerial vector such as bats transferred the lentivirus between Madagascar and the mainland, but this scenario has an extra level of complexity as it requires two independent spillover events which is less likely to have taken place.

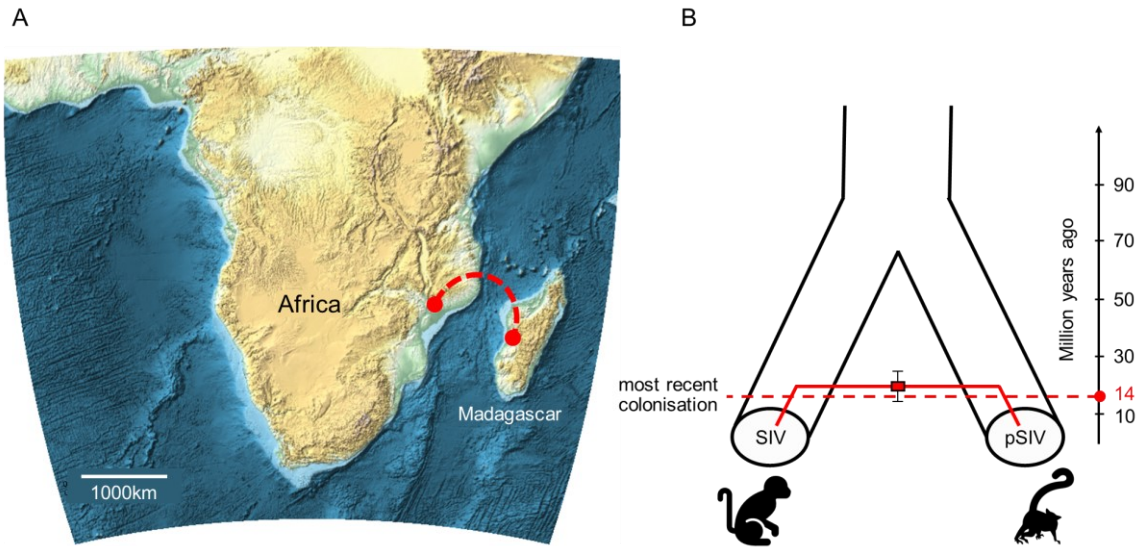


Figure 3.3: **Evolutionary origins of SIV and pSIV. (A)** Map showing the geographical location of Magdagascar relative to Africa and **(B)** the evolutionary timeline of simians and prosimians hosts (black cyndrical tree) along with the SIV/pSIV estimated divergence time according to the PoW model (red solid line). The estimated time to the most recent common ancestor of pSIV and SIV according to the PoW model is 17.1 (95% HPD: 13.2 - 20.7) million years before present which is in agreement with the molecular and fossil data suggesting the most recent colonisation of Madagascar by primates among other terrestrial mammals occurred 14 million years ago (red dashed line).

3.4 Discussion

Our analysis offered a unifying framework for understanding the evolutionary history of human, simian, and prosimian immunodeficiency viruses across vast timescales. The molecular clock analysis based on the PoW model showed, for the first time, that the most recent common ancestor of all SIV strains existed nearly one million years ago. It also successfully reproduced the TMRCA estimates for five human-to-human transmissible strains of HIV and four strains of SIV on the island of Bioko without the need for any external calibration points, giving further credence to its accuracy. Our TMRCA estimates for HIV-1 M, HIV-2 A, and B are slightly older than the estimates based on standard relaxed clock methods. This suggests substitution saturation may have influenced HIV divergence time estimates over a timescale of ~150-200 years. A recent study on HIV-1 M provides evidence for the existence of such saturation effects over a timescale of a few decades [20] which can in turn

result in the underestimation of their evolutionary rates over such timescales. Therefore, further investigations into the possibility of a longer evolutionary history of HIV may be warranted.

The old evolutionary history of SIVs indicates that there exists some degree of codivergence with their hosts for at least tens to hundreds of thousands of years followed by many intraclade host switching events which gave rise to new strains with different primate species hosting the virus at different points in their evolutionary trajectory. This also implies that humans may have occasionally been exposed to SIVs over such periods. The fact that such potential encounters never led to nascent HIVs becoming established much earlier is likely due to changes in human social behaviour, increased mobility, urbanisation, and various other factors that have only emerged around or after late 19th century [7, 21].

Our unified framework also answered one of the longstanding questions on the evolutionary history of primate lentiviruses [10, 11]. By finding the TMRCA of SIV and pSIV to be nearly 17 million years old, we were able to rule out ancestral codivergence of primate lentiviruses into haplorrhine and strepsirrhine lineages as a likely model of primate lentivirus evolution. Instead, we found the land-bridge hypothesis whereby a terrestrial mammal or a nonprimate vector transferred lentiviruses between haplorrhine and strepsirrhine in mainland Africa and Madagascar to be a more plausible scenario given that the timing of the most recent colonisation of Madagascar greatly overlaps with our estimate for the TMRCA of SIV and pSIV [19]. These findings can also offer new insights to the evolution of ancient innate immune defence mechanisms in simian and prosimian hosts against lentivirus infections which could lead to new HIV treatments or vaccines. Given the evidence based on the PoW model about the long history of coevolution between

many SIV strains and their simian hosts for tens to hundreds of thousands of years, it is possible that one of the reasons why lentiviruses are not as pathogenic in non-human primates is because these species have been infected with lentiviruses for a very long period and, as a result, have developed strong immune defence mechanisms that slowed or blocked retroviral reproduction as opposed to the very recent evolutionary origins of HIV strains in humans [22-24].

Our findings also exposed the fundamental flaws of conventional methods for rate estimation based on fixed calibration points and their erroneous inference of the SIV evolutionary history. Such calibration techniques do not fully capture time-dependent rate effects across different timescales which can completely distort the inferred evolutionary timescales of primate lentiviruses [9, 12, 13, 25]. We showed that a calibration based on extant SIVs is not reliable for inferring their longer evolutionary trajectory; nor biogeographic or paleovirological calibrations can be used to infer more recent evolutionary histories of SIVs and HIVs. However, we demonstrated that the PoW model offers a solution to time-dependent rate effects by providing a uniform framework for integrating evolutionary rates across all timescales [15].

The PoW model has previously been applied to reconstruct the phylogeny of foamy viruses which display strong cospeciation with their host and has successfully recovered their deep evolutionary history [15]. The model was also used to reconstruct the evolutionary origins of the sarbecovirus lineage and was found to be in remarkable concordance with signatures of a selection of human genomic datasets that indicate an arms race with corona-like viruses dating back to 25,000 years before present [26], providing an external comparator for this methodology. Given that we have now successfully reconstructed the evolutionary history of

primate lentiviruses using this method, we can look much more widely at other viruses, re-evaluate the timescales of their deeper evolution and gain insights into host relationships that are key to understanding their ability to cause disease.

3.5 Methods

3.5.1 Data sets

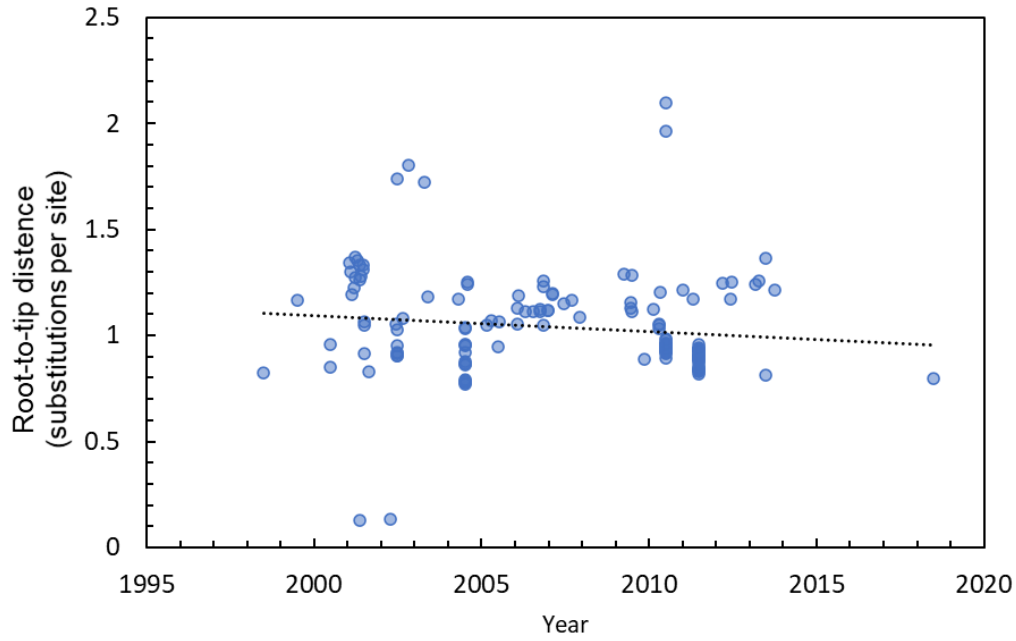
We generate two sets of alignments. One includes 331 SIV and HIV alignments of *pol* locus from 25 lineages of SIV, 3 groups of HIV-1 (M/N/O), and 2 groups of HIV-2 (A/B). The other includes 341 SIV alignments of integrase gene from 18 lineages of SIV including all the alignments taken from the island of Bioko (see ref. [12]). We downloaded all genome sequences from GenBank in 2021 with sampling dates obtained from the Los Alamos National Laboratory HIV sequence database (http://hiv.lanl.gov/content/index) and prepared codon-aligned sequences using MUSCLE [27]. To improve the accuracy of phylogenetic inference, we exclude recombinant sequences using BootScanning in the RDP4 package [28] and remove ambiguously aligned regions using GBlocks [29].

The reason for choosing the *pol* locus is because, unlike *env* and *gag* loci, it tends to be functionally more restricted which makes the alignment of significantly divergent sequences more reliable. Secondly, the sequenced SIV samples from the island of Bioko only cover the integrase region of *pol*. Therefore, we want to investigate whether there is any noticeable difference between the estimated divergence times using the *pol* and integrase gene alignments. The other reason is that the pSIV sequence only covers the *pol* region [11].

3.5.2 Molecular clock analyses and phylogeny reconstruction

To infer the TMRCA for SIV and HIV lineages, we use the PoW model of phylogenetic tree reconstruction [15]. Initially, we examine the temporal signal in the time-stamped samples in each data sets using TempEst [30] by plotting the root-to-tip divergence against sampling time based on a maximum likelihood tree which is constructed using IQTREE [31] under a general time-reversible model with a discrete gamma distribution (GTR+ Γ_4). We find that the integrase gene data set does not have enough temporal signal to reliably infer a clock rate (Figure 3.4A). Therefore, we use the dated SIV samples from the *pol* data set, which has sufficient temporal signal (see Figure 3.4B), to estimate the short-term substitution rate using a standard relaxed clock model under GTR+ Γ_4 in BEAST [32]. Similarly, we estimate the short-term substitution rate for the *pol* data set in BEAST. Next, we construct an ultrametric distance tree for each data set using the HKY95 substitution model and convert distance trees into a PoW-transformed time tree as outlined in ref. [15]. We note that the standard PoW model only allows for the reconstruction of phylogenies from samples that belong to the same time period. However, given that the pSIV sample is from a germline insertion that took place nearly 4.2 million years ago, we need to account for the amount of divergence that pSIV would have accumulated since its insertion until present if it were to evolve at a rate that is similar to that of SIV lineages. This allows us to correctly estimate the total amount of divergence since the most recent common ancestor of pSIV and SIV using the PoW model.

A



B

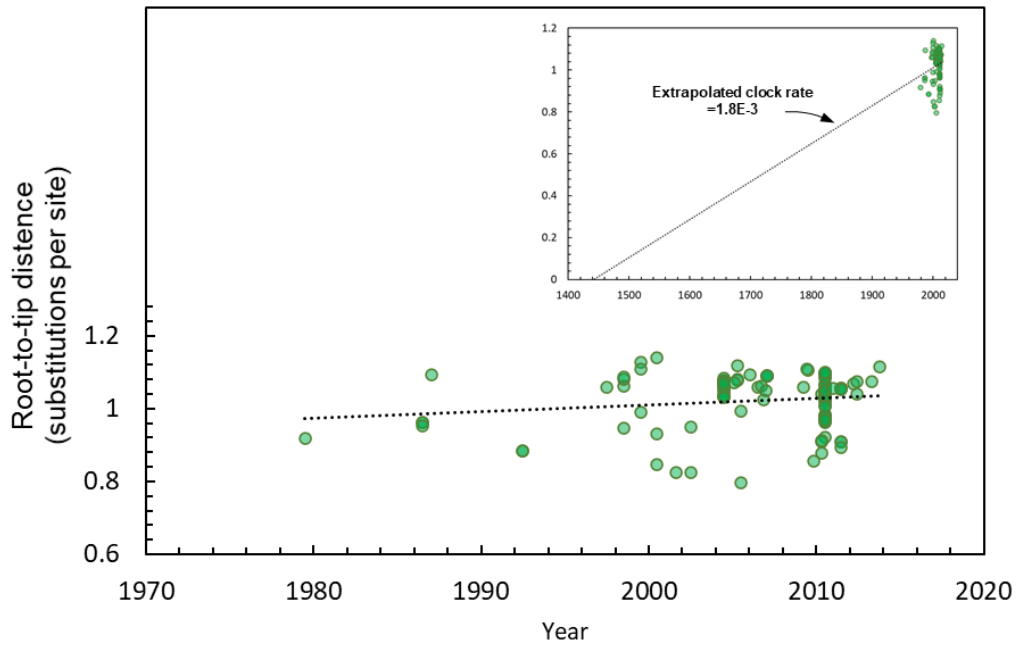


Figure 3.4: **Evaluating temporal signal in the time-stamped data**. Root-to-tip distance as a function of sampling time for dated SIV samples in (A) the integrase gene data set and (B) *pol* data set. The plots are based on maximum likelihood tree reconstructions with a root poition that maximises the residual mean squared for the regression of root-to-tip divergence and sampling time (black dashed line).

3.6 References

1. F. Barresinoussi, et al., *Isolation of a T-Lymphotropic Retrovirus from a Patient at Risk for Acquired Immune-Deficiency Syndrome (AIDS).* Science, 1983. **220**(4599): p. 868-871.
2. R. C. Gallo, et al., *Frequent Detection and Isolation of Cytopathic Retroviruses (Htlv-Iii) from Patients with Aids and at Risk for Aids.* Science, 1984. **224**(4648): p. 500-503.
3. F. Clavel, et al., *Molecular-Cloning and Polymorphism of the Human Immune-Deficiency Virus Type-2.* Nature, 1986. **324**(6098): p. 691-695.
4. P. M. Sharp and B.H. Hahn, *Origins of HIV and the AIDS Pandemic.* Cold Spring Harbor Perspectives in Medicine, 2011. **1**(1): p. a006841.
5. V. M. Hirsch, et al., *An African Primate Lentivirus (SIVsm) Closely Related to HIV-2.* Nature, 1989. **339**(6223): p. 389-392.
6. T. Huet, et al., *Genetic Organization of a Chimpanzee Lentivirus Related to HIV-1.* Nature, 1990. **345**(6273): p. 356-359.
7. B. H. Hahn, et al., *AIDS as a zoonosis: Scientific and public health implications.* Science, 2000. **287**(5453): p. 607-614.
8. B. Korber, et al., *Timing the ancestor of the HIV-1 pandemic strains.* Science, 2000. **288**(5472): p. 1789-1796.
9. J. O. Wertheim and M. Worobey, *Dating the Age of the SIV Lineages That Gave Rise to HIV-1 and HIV-2.* Plos Computational Biology, 2009. **5**(5): p. e1000377.
10. C. Gilbert, et al., *Parallel Germline Infiltration of a Lentivirus in Two Malagasy Lemurs.* Plos Genetics, 2009. **5**(3): p. e1000425.
11. R. J. Gifford, et al., *A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(51): p. 20362-20367.
12. M. Worobey, et al., *Island Biogeography Reveals the Deep History of SIV.* Science, 2010. **329**(5998): p. 1487-1487.
13. P. Aiewsakun and A. Katzourakis, *Time-Dependent Rate Phenomenon in Viruses.* Journal of Virology, 2016. **90**(16): p. 7184-7195.
14. S. Y. W. Ho, et al., *Time dependency of molecular rate estimates and systematic overestimation of recent divergence times.* Molecular Biology and Evolution, 2005. **22**(7): p. 1561-1568.
15. M. Ghafari, et al., *A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses.* Current Biology, 2021. **31**(21): p. 4689-4696.e5.
16. M. A. Charleston and D.L. Robertson, *Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny.* Systematic Biology, 2002. **51**(3): p. 528-535.
17. S. M. Bell and T. Bedford, *Modern-day SIV viral diversity generated by extensive recombination and cross-species transmission.* Plos Pathogens, 2017. **13**(7): p. e1006466.
18. P. M. Sharp, et al., *Origins and evolution of AIDS viruses: estimating the time-scale.* Biochemical Society Transactions, 2000. **28**: p. 275-282.
19. Tattersall, I., *Historical biogeography of the strepsirrhine primates of Madagascar.* Folia Primatol (Basel), 2006. **77**: p. 477-487.

20. V. Druelle and R.A. Neher, *Reversions to consensus are positively selected in HIV-1 and bias substitution rate estimates. .* bioRxiv, 2022. https://doi.org/10.1101/2022.02.13.480259.
21. J. D. de Sousa, et al., *High GUD Incidence in the Early 20(th) Century Created a Particularly Permissive Time Window for the Origin and Initial Spread of Epidemic HIV Strains.* Plos One, 2010. **5**(3): p. e9936.
22. M. Ohainle and H.S. Malik, *A balancing act between primate lentiviruses and their receptor.* Proceedings of the National Academy of Sciences of the United States of America, 2021. **118**(20): p. e2104741118.
23. R. M. Russell, et al., *CD4 receptor diversity represents an ancient protection mechanism against primate lentiviruses.* Proceedings of the National Academy of Sciences of the United States of America, 2021. **118**(13).
24. S. Joas, et al., *Species-specific host factors rather than virus-intrinsic virulence determine primate lentiviral pathogenicity.* Nature Communications, 2018. **9**: p. 1371
25. P. Simmonds, P. Aiewsakun, and A. Katzourakis, *Prisoners of war - host adaptation and its constraints on virus evolution.* Nature Reviews Microbiology, 2019. **17**(5): p. 321-328.
26. Y. Souilmi, et al., *An ancient coronavirus-like epidemic drove adaptation in East Asians from 25,000 to 5,000 years ago.* Curr Biol, 2021. **31**: p. 1-11.
27. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Research, 2004. **32**(5): p. 1792-1797.
28. D. P. Martin, et al., *RDP4: Detection and analysis of recombination patterns in virus genomes.* Virus Evolution, 2015. **1**(1): p. vev003
29. Castresana, J., *Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.* Molecular Biology and Evolution, 2000. **17**(4): p. 540-552.
30. A. Rambaut, et al., *Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen).* Virus Evolution, 2016. **2**(1).
31. L. T. Nguyen, et al., *IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.* Molecular Biology and Evolution, 2015. **32**: p. 268-274.
32. M. A. Suchard, et al., *Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10.* Virus Evolution, 2018. **4**(1).
33. S. Alvarez-Carretero, et al., *A species-level timeline of mammal evolution integrating phylogenomic data.* Nature, 2022. **602**(7896): p. 263-267.

Statement of Authorship for joint/multi-authored papers for PGR thesis.

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	Deep evolutionary history of simian and prosimian immunodeficiency viruses
Publication Status	Unpublished and unsubmitted work written in a manuscript style
Publication Details	Ghafari M, Simmonds P, and Katzourakis A. Deep evolutionary history of simian and prosimian immunodeficiency viruses. (Unpublished)

Student confirmation

Student Name	Mahan Ghafari
Contribution to the paper	MG wrote the paper and conducted all the analysis under the supervision from AK. PS prepared the HIV and SIV alignments. All authors provided editorial assistance.
Signature 	Date 11/05/2022

Supervisor confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title	Professors Aris Katzourakis and Oliver G Pybus	
Supervisor comments		
Signature 		Date 11/05/2022

Chapter 4: Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza

4.1 Abstract

High throughput sequencing enables rapid genome sequencing during infectious disease outbreaks and provides an opportunity to quantify the evolutionary dynamics of pathogens in near real-time. One difficulty of undertaking evolutionary analyses over short timescales is the dependency of the inferred evolutionary parameters on the timespan of observation. Crucially, there is an increasing number of molecular clock analyses using external evolutionary rate priors to infer evolutionary parameters. However, it is not clear which rate prior is appropriate for a given time window of observation due to the time-dependent nature of evolutionary rate estimates. Here, we characterise the molecular evolutionary dynamics of SARS-CoV-2 and 2009 pandemic H1N1 (pH1N1) influenza during the first 12 months of their respective pandemics. We use Bayesian phylogenetic methods to estimate the dates of emergence, evolutionary rates, and growth rates of SARS-CoV-2 and pH1N1 over time and investigate how varying sampling window and dataset sizes affects the accuracy of parameter estimation. We further use a generalised McDonald-Kreitman test to estimate the number of segregating non-neutral sites over time. We find that the inferred evolutionary parameters for both pandemics are time-dependent, and that the inferred rates of SARS-CoV-2 and pH1N1 decline by ~50% and ~100%, respectively, over the course of one year. After at least 4 months since the start of sequence sampling, inferred growth rates and emergence dates remain relatively stable and can be inferred reliably using a

logistic growth coalescent model. We show that the time-dependency of the mean substitution rate is due to elevated substitution rates at terminal branches which are 2-4 times higher than those of internal branches for both viruses. The elevated rate at terminal branches is strongly correlated with an increasing number of segregating non-neutral sites, demonstrating the role of purifying selection in generating the time-dependency of evolutionary parameters during pandemics.

4.2 Introduction

Rapid whole-genome sequencing has become part of pathogen surveillance systems and is important to both infection control and enables a detailed investigation of the epidemiological and evolutionary characteristics of pathogens. Quantifying infectious disease evolution enables the inference of parameters such as times of origin, epidemic growth rates, and evolutionary rates [1-3].

One of the perils of making such inferences over short time periods (i.e., a few months or years) is that the inferred parameters of interest may vary over the timespan of observation [4]. While several factors can lead to the misestimation of substitution rates (such as mis-identification of coalescent population models, changes in natural selection, differing replication rates in various host reservoirs, or sequencing errors), misspecification of clock models and saturation of nucleotide changes can result in rate underestimation [5-8]. This nonstationarity can render findings that are confusing or conflicting such that an estimated substitution rate over one time window is not transferable to the analysis of another [9]. For example, during the 2014-2016 West Africa Ebolavirus epidemic there was controversy and concern regarding the virus' "mutation rate", because early estimates of the substitution rate from the epidemic appeared to be approximately twice the average

rate across multiple outbreaks [10, 11]. Confusion such as this is also common in the SARS-CoV-2 literature and arises in part due to an incomplete understanding of estimated substitution rates by users of phylogenetic analysis software. Given the importance of phylogenetic dating and clock estimation for all SARS-CoV-2 genomic epidemiology worldwide, understanding the pattern of inferred substitution rate over time and investigating the potential underlying mechanisms involved in creating time-dependent rate effects in viruses can shed light on the molecular evolutionary dynamics of viruses.

Another major obstacle is that during an ongoing outbreak the level of sequence diversity may be so low that the amount of accrued evolutionary change is not sufficient to make informative phylogenetic inference possible [12]. Several statistical tests have been developed to ensure enough temporal signal is present in a set of temporally sampled genome sequences to reliably infer evolutionary parameters [13-15].

More specifically, there is extensive evidence that purifying selection in viruses can result in varied estimates for rates of substitution and evolutionary rate ratio, dN/dS, across different timescales [16-18]. It is likely that a greater proportion of polymorphisms observed among genomic sequences sampled early in an epidemic are segregating deleterious mutations, which will only persist for a limited time before being eliminated by purifying selection [19]. The duration of this time-dependent effect may be prolonged due to incomplete purifying selection in populations with very large effective population sizes [20].

While evidence of strong purifying selection has mostly been reported at the species level and over very long timescales (i.e., thousands to millions of years),

using inference methods based on the dN/dS ratio [21-23], there have been few studies at the intra-population level and over short timescales, mainly because there is often no opportunity to collect sufficiently large number of samples through time to track low-frequency variants [4, 24]. Crucially, while purifying selection has been put forward as a possible explanation for the time-dependency of substitution rates over such timescales, there has not been any systematic studies to investigate the role of purifying selection and quantify its contribution to altering the inferred substitution rate of viruses over time. Furthermore, using standard phylogenetic methods to compute the dN/dS ratio for conspecific sequences sampled from a single population over short timescales may be inappropriate as the differences between sequences over such timescales represent segregating polymorphisms as opposed to fixed substitutions along independent lineages. The former has been shown to produce very different estimates of the dN/dS ratio over time [25, 26].

In this study, we aim to quantify and identify the source of time-dependency of virus substitution rate estimates over short time periods and characterise estimates of the molecular clock rate, time of origin, growth rate, and number of non-neutral sites for different datasets that represent different timescales of genomic observation. We show that time-dependency in estimates of the mean substitution rate is dominated by elevated rates at terminal branches, whereas the estimated rate of substitution at internal branches does not exhibit a time-dependent decay. We then use a generalised McDonald-Kreitman test based on nucleotide site frequencies that allows purifying selection to be quantified over short timescales and demonstrate that there is a strong correlation between the elevated rates at terminal branches and the high number of low frequency non-neutral sites in both SARS-CoV-2 and pH1N1 genomes.

4.3 Results

We use Bayesian phylogenetic methods implemented in BEAST v.1.10 to estimate the molecular clock rate, times of origins, and growth rates of SARS-CoV-2 and pH1N1 pandemics. The dataset varied in size and temporal sampling range and were created to reflect the way in which real datasets accumulate in size and diversity during the course of an epidemic. Specifically, we aggregate all available samples up to each respective month and infer the parameters of interest (Figure 4.1; Figure S1). We first compare the results for SARS-CoV-2 using two coalescent growth priors: exponential growth and logistic growth. While this is not a comprehensive comparison between all phylogenetic models that can be used to explain the evolutionary dynamics, it allows us to find the better-fitting model that introduces less error in parameter estimation -- the demographic model is effectively a nuisance variable. Our analysis suggests, except for the month of January, the logistic growth coalescent tree prior is a better fit to the data (Table 4.1).

Table 4.1: **Log-marginal likelihoods of exponential and logistic growth models with increasing temporal ranges of sampling dates.** Taking exponential growth as the null model, we select the logistic growth model for any dataset with a positive Bayes factor.

Month of sampling	Number of samples	Log-marginal likelihood exponential growth model	Log-marginal likelihood logistic growth model	Bayes factor
Jan	41	-41022.77	-41025.56	-2.79
Feb	121	-43062.00	-43047.11	+14.88
Mar	181	-44627.77	-44607.95	+19.78
Apr	241	-46524.75	-46506.25	+18.53
May	301	-48587.39	-48548.15	+39.24
Jun	361	-52310.61	-52248.23	+62.38
Jul	421	-55138.43	-55036.60	+101.83
Aug	481	-58580.94	-58430.09	+150.85
Sep	541	-62101.22	-61805.93	+295.29
Oct	601	-65433.20	-65005.09	+428.11
Nov	661	-69121.03	-68921.98	+199.05
Dec	721	-73094.26	-72389.69	+704.57

As reported in previous studies, we find that there is not enough temporal signal in the data to reliably infer the evolutionary parameters during the first two months of the SARS-CoV-2 pandemic [27]. This results in the underestimation of substitution rates as well as high statistical uncertainty for the parameter estimates. The lack of temporal signal in the SARS-CoV-2 samples is also evident from the number of new singletons (i.e., single nucleotide variations in the dataset compared to the ancestral sequence) added the dataset per month during the first 2-3 months (Figure S2). On the other hand, for pH1N1, the molecular clock rate is up to 2 times higher for the first three months than for the following months (Figure 4.1A,B). While the inferred substitution rate of SARS-CoV-2 tends to decrease as we increase the timespan of measurement, the rate for pH1N1 does not change considerably after the first 3-4 months of measurement, in agreement with previous findings [4].

Both the inferred times of origin and growth rates of SARS-CoV-2 and pH1N1 in early months have wide credible intervals due to the high uncertainty associated with small sample sizes and narrow sampling windows. The precision of these two inferred parameters increases with the addition of data from later months and remain roughly consistent after the first 4-5 months of measurement (Figure 4.1C-F). Using the logistic growth model, we find the estimated times of origin for SARS-CoV-2 and pH1N1 to be 2019-10-28 (95% HPD: 2019-09-30, 2019-11-24) and 2009-01-18 (95% HPD: 2008-12-14, 2009-02-22), respectively, which is also in agreement with previous studies [2, 3, 24]. However, under the exponential growth model, the inferred time of origin of SARS-CoV-2 samples significantly diverge from the expectation after 6 months of sampling and yield unreliable estimates (Figure 4.1C,D). This is likely because after the first few months, the growth rate declines and the population dynamics deviates from the exponential growth model (Figure

4.1E,F). This, in turn, results in the underestimation of growth rate of SARS-CoV-2 and inferring an older time of origin. We also note that, from a molecular epidemiology perspective, using the exponential growth coalescent model would not be appropriate nor realistic particularly over longer timescales (i.e., >6 months) as the pandemic dynamics did not continue to grow exponentially (i.e., the growth rate started to drop after the first few months due to various behavioural changes in the population and the implementation of non-pharmaceutical interventions). Despite the difference in the estimated time of origin, both the exponential and logistic growth models infer very similar clock rates for SARS-CoV-2 over time.

Purifying selection has been often cited as one of the main evolutionary processes contributing to the elevation of estimated molecular evolutionary rates over short timescales [24]. The argument is that low-frequency deleterious mutations can segregate in a population for some time before being purged because of purifying selection. Therefore, the proportion of all changes that are deleterious is high when the phylogenetic tree is short, and lower when the tree is longer as it takes more time for shared neutral or advantageous changes to accrue with respect to the ancestral state. To evaluate this hypothesis, we investigate the correlation between the number of non-neutral polymorphisms and estimated substitution rates over time. Our results show that there is a higher proportion of segregating mutations at low frequencies (<15%) during the first few months of observation, in both the ORF1b gene of SARS-CoV-2 and the HA gene of pH1N1 (Figure 4.2A,B). In particular, the proportion of low frequency non-neutral sites in ORF1ab gene was higher during the first 4-5 months of observation and had a roughly fourfold drop during that period compared to a twofold drop in HA gene. After the first 5 months, the proportion of low frequency non-neutral sites remains

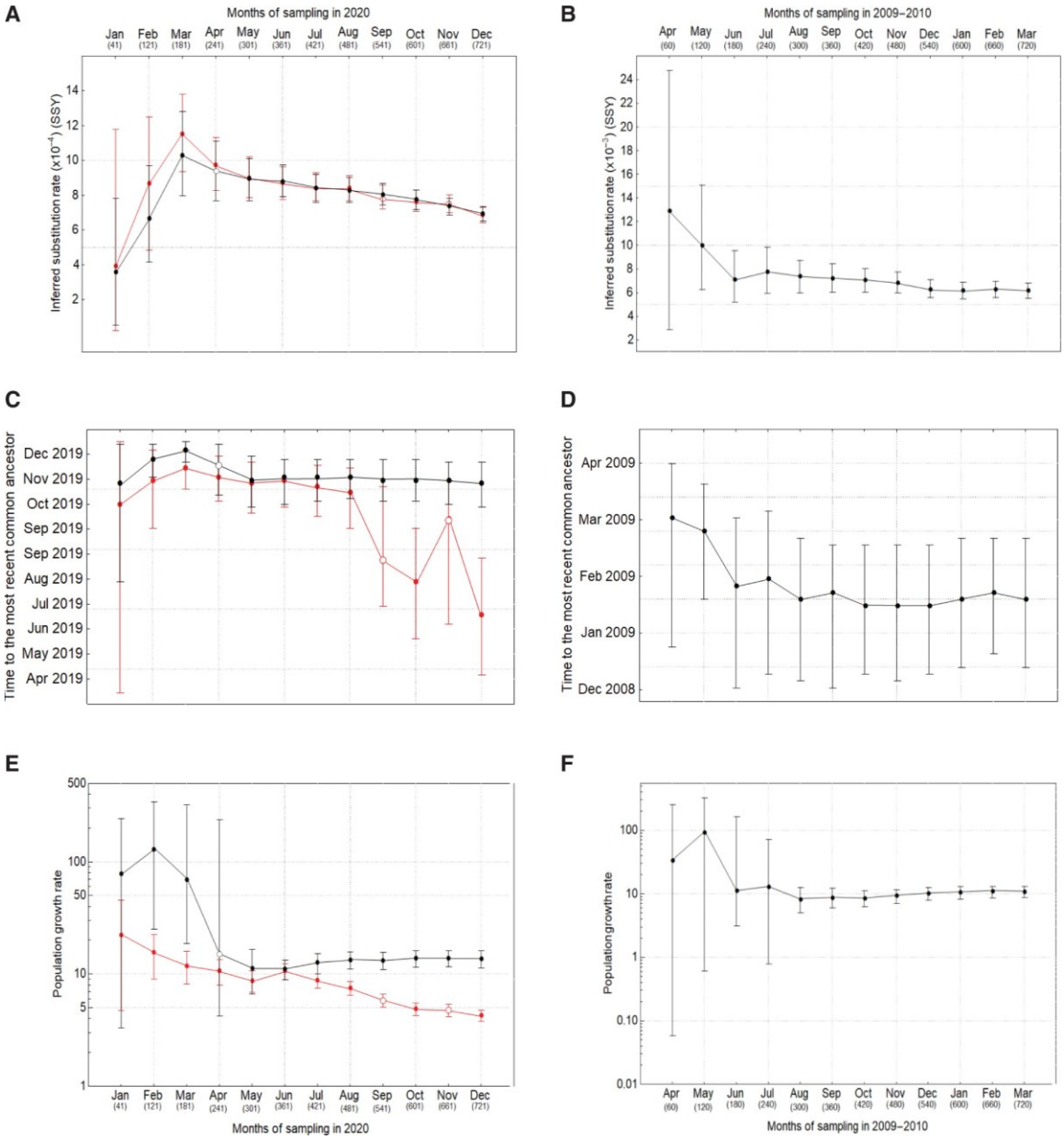


Figure 4.1: **Phylogenetic estimates of substitution rates, time to the most recent common ancestor, and population growth rate for SARS-CoV-2 and pH1N1.** (**A, B**) Inferred rates, (**C, D**) times of origin, and (**E, F**) growth rates of SARS-CoV-2 (left column) and pH1N1 influenza (right column) using an exponential (red) and a logistic (black) growth model. Open circles represent nonconvergence for at least one parameter in the Bayesian analysis. Note that y-axes are not on the same scale for SARS-CoV-2 and pH1N1.

roughly constant at around 15% in ORF1ab, while it shows an uptick from 15% to 17% in HA towards the beginning of 2010. The slight increase in the proportion of non-neutral sites in HA during this period is also in agreement with a rise in the relative genetic diversity of pH1N1 around the world in late 2009/early 2010 which may also have resulted in an increase in the number of segregating deleterious mutations in the population [28]. We note that while it is unlikely for mutations in the low frequency class to have strongly positive fitness effects, they may also contain neutral or adaptive mutations that have not reached sufficiently high frequencies yet. We also find that the number of low frequency replacement sites is always greater than silent sites in ORF1ab while, for HA, silent sites are in majority in most months (Figure 4.2C,D). The total number of non-neutral sites in ORF1ab is higher because it is a much longer gene (~21,000 nt) compared to HA (~1,800 nt).

One of the impacts of purifying selection over short timescales is that the number of replacements on terminal branches should be higher than that on internal branches. To measure this effect during the SARS-CoV-2 and pH1N1 epidemics, we used a branch-specific two parameter molecular clock model in BEAST to infer separate rates of substitution for terminal and internal branches (see Figure 4.2E, F). Our results show that for both viruses, the rate of substitution on terminal branches are 2-5 times higher than on internal branches. The difference is more dramatic in SARS-CoV-2 where the average ratio of substitution rates at terminal branches are 4 times higher than that of internal branches while it is only 2 times higher in HA during the first 12 month of observation. Figures 4.2E,F also show that the mean and credible intervals of the inferred substitution rate at terminal and internal branches are non-overlapping during several months of rate measurement which demonstrates that there is a significant difference between the rate of

evolution at these two branch categories. This also agrees with the observation of a higher proportion of non-neutral sites in SARS-CoV-2 samples compared to pH1N1 because there is going to be more deleterious mutations at terminal branches of SARS-CoV-2. Further, there is a clear decline in the estimated rate at terminal branches as the sampling window increases. In contrast, the estimated rate at internal branches gradually increases through time (with the exception of the first 2-3 months).

We also investigate the role of sequencing error as a potential confounder for the elevated rates at terminal branches. By masking more than 110 sites of SARS-CoV-2, in addition to our standard filtering step (see Methods section), which are suggested to be prone to recurrent sequencing error and appear to be highly homoplastic, we re-calculate the rate of evolution at terminal and internal branches (Figure S3). We find that while the overall inferred rates drop for the alignments with masked sites (due to their lower genetic variation compared to unmasked alignments), the time-dependent rate drop at terminal branches is retained (Figure S3B). Therefore, sequencing error cannot be the underlying source for time-dependency of evolutionary rates. In Figure 4.2 G,H, we further show that there is an overall significant correlation between the rate of substitution at terminal branches for the unmasked alignments and number of non-neutral sites in both SARS-Cov-2 (R^2=0.74 and $p < 0.001$) and pH1N1 (R^2=0.53 and $p < 0.01$) (Figure 4.2 G,H). By examining the site frequency spectrum for both datasets, we can also see that most of the variation comes from the low-frequency regime (<15%) with very limited variation in the mid- to high-frequency regimes (Figure S2).

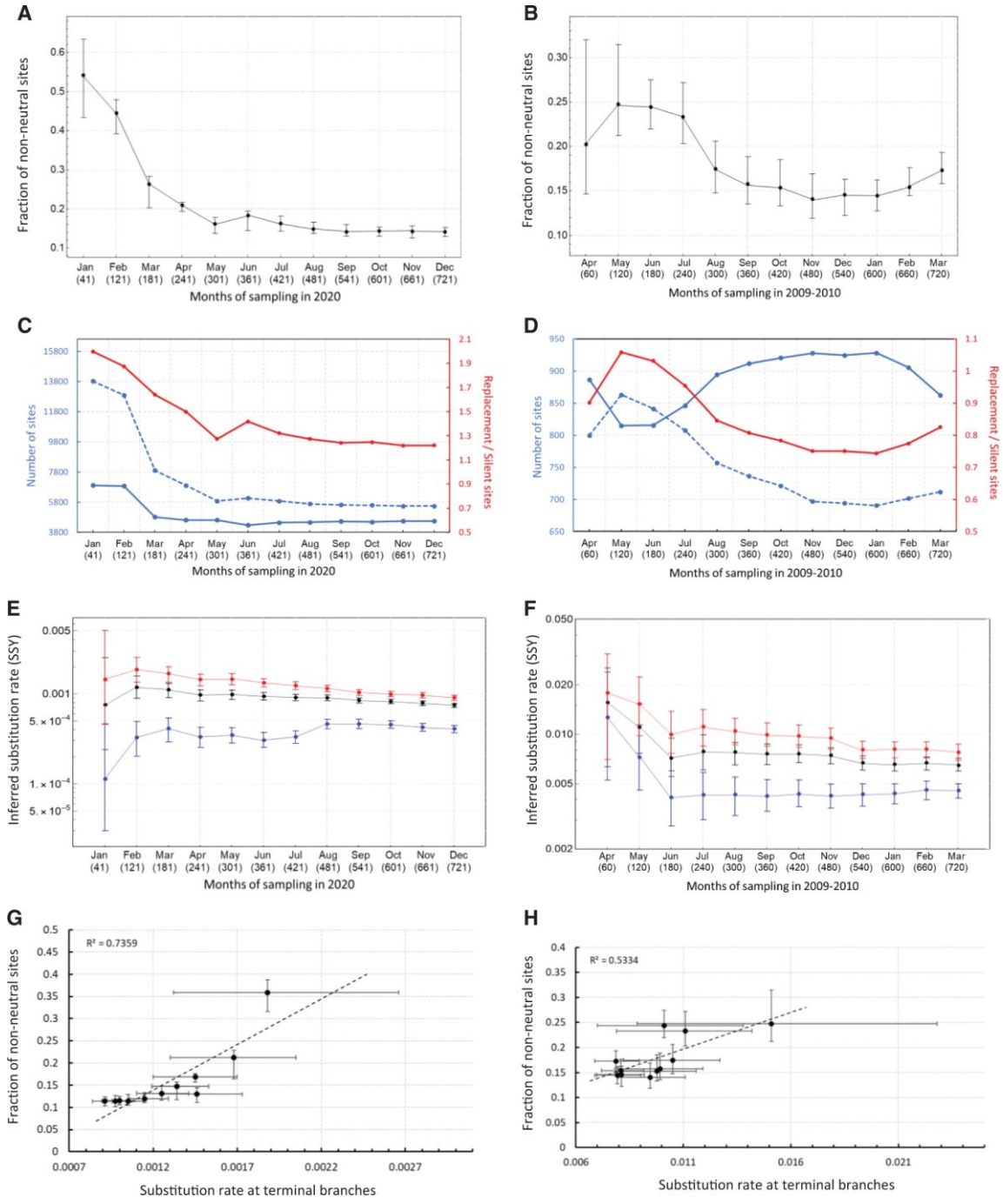


Figure 4.2: **Estimating the fraction of nonneutral sites and substitution rate at terminal and internal branches for SARS-CoV-2 and pH1N1 (A, B)** Number of non-neutral sites over time for SARS-CoV-2 (left column) and pH1N1 (right column). **(C, D)** Number of replacement (dashed blue line), silent (solid blue line) sites, and their ratio (red line) over time. **(E, F)** Mean clock rate (black), and the rates at the terminal (red) and internal (blue) branches. The MCMC chains for the first month of sampling SARS-CoV-2 and the first three months of pH1N1 do not converge using the logistic growth coalescent model. Instead, the exponential growth coalescent model was used. **(G, H)** Correlation coefficient between the substitution rate at terminal branches and number of non-neutral sites -- excluding the estimates from the first month of sampling due to inadequate temporal signal and significant uncertainty in the inferred parameters.

Furthermore, in the pH1N1 dataset, we can see a dip in the inferred rate at terminal branches in December 2009 (Figure 2F). There is also a sudden loss of low frequency genetic diversity in the dataset during the same period which is partially recovered in February and March (see the trajectory of newly added singletons and low frequency variants in Figure S2D,F). On the other hand, the pattern of substitution rate at internal branches is non-monotonic over time. In particular, we see an increase in the rate at internal branches of SARS-CoV-2 from August onwards. Similarly, there is a steady rise in the rate at internal branches of pH1N1 from February onwards.

4.4 Discussion

We found that the over-abundance of deleterious mutations during the early stages of both the SARS-CoV-2 and pH1N1 pandemics strongly correlates with higher substitution rates at terminal branches relative to internal branches of inferred phylogenetic trees. Once there is enough temporal signal in the data to reliably estimate relevant epidemiological and evolutionary parameters, the mean substitution rates decline over the course of one year of rate measurement while the estimated time of origin and growth rate are more stable and remain roughly the same after the first 4-5 months of measurement. We found that this declining pattern in mean substitution rate is caused by the rate decay at terminal branches while the rate at internal branches does not exhibit such a pattern over time. In particular, the gradual increase in the substitution rate at internal branches over the span of several months for both SARS-CoV-2 and pH1N1 influenza could be related to signatures of adaptive evolution. We also found that sites that are prone to recurrent sequencing error and appear to be highly homoplastic do not make a significant contribution to the time-dependent pattern of substitution rates. While both

sequencing error and deleterious mutations appear as singletons or low-frequency variations in the data, deleterious mutations are strongly biased towards changing amino acids whereas sequencing errors are equally likely to be synonymous or non-synonymous. Therefore, they do not make similar contributions to altering the inferred substitution rates over time.

Our results provided further evidence that short-term rate estimates are subject to time-dependent rate effects largely due to transient polymorphisms [4, 24]. In addition, our estimated mean substitution rates, times of origin, and growth rates agree with previous studies of SARS-CoV-2 [3, 12, 29, 30] and pH1N1 [2, 4, 24, 31]. We found that while the time-dependent rate effect for pH1N1 is less pronounced after the first 3 months of sampling, the same effect in SARS-CoV-2 is more pronounced and continues even after 12 months of sampling. This may be the result of a much higher mean substitution rate in pH1N1 compared to SARS-CoV-2. We also found a higher proportion of low frequency non-neutral sites for SARS-CoV-2 dataset during the first 8 months of observations and a higher ratio of replacement to silent sites during the entire one year of observation compared to pH1N1. The same method used in this study to infer the number of low frequency non-neutral sites has also been used to characterise the molecular evolutionary dynamics of other RNA viruses including seasonal coronaviruses and emerging avian influenza viruses [32-34].

By dividing the data into successively longer temporal intervals, we showed that the over-abundance of deleterious mutations at terminal branches are the main reason behind the gradual decay in mean substitution rate over time in SARS-CoV-2 and pH1N1. Further work can be done to quantify the long-term effect of purifying selection on the time-dependency of substitution rates.

4.5 Materials and Methods

We downloaded all SARS-CoV-2 sequences from GISAID and pH1N1 influenza sequences from GenBank and align them using MUSCLE v3.8.425 [35] -- a complete metadata table acknowledging the authors, originating and submitting laboratories of the SARS-CoV-2 sequence data is available in **Table S1**. For the SARS-CoV-2 dataset, we mask the first 54 and last 240 sites of the alignments and only include complete sequences that are more than 29,000 nt long with high coverage as determined by GISAID's default search option (i.e., entries with <1% Ns and <0.05% unique amino acid substitutions). Furthermore, for part of the analysis where we investigate the role of sequencing error in time-dependent rate effects at terminal branches for SARS-CoV-2, we mask an additional 114 sites that appear to be highly homoplastic or prone to recurrent sequencing error [36]. We specifically select pH1N1 sequences from GenBank for the full coding region of the hemagglutinin (HA) gene and ensure that none of the SARS-CoV-2 and pH1N1 samples have undergone serial passaging.

To investigate the effects of the temporal range of sampling dates on the accuracy of parameter estimation, we incrementally increase the size of each dataset by adding 60 genomes, chosen randomly, for every additional month of sampling. Thus, after 12 months of sampling since the first sequence was uploaded on GISAID and/or GenBank, we have 720 samples. We note that due to a lack of temporal signal in the early SARS-CoV-2 samples (i.e., most of the early samples were almost completely identical) and failure of the Markov Chain Monte Carlo (MCMC) chains for the phylogenetic analyses to converge, we only used 41 samples collected between 24 December 2019 to 31 January 2020 (labelled as 'January sequences' in our analysis) and took 20 additional samples during the next

month to match with the 120 samples used for pH1N1 by the end of the second month of sampling (see **Figure S1**).

4.5.1 Phylogenetic analyses

We use BEAST v1.10 [37] for the Bayesian phylogenetic analysis of the entire dataset using an HKY+Γ substitution model with a Laplace prior (mean=0 and scale=100) on the coalescent growth rate, a Lognormal prior (mean=1 and stdev=2) on the coalescent population size, and a continuous time Markov chain prior on the evolutionary clock rate. For the first part of the analysis, we use a strict clock model and exponential and logistic growth coalescent demographic models of SARS-CoV-2 evolution and only the logistic growth model for the HA segment of pH1N1. To quantify the relative fit of the two coalescent models for SARS-CoV-2, we compute their log marginal likelihoods using the generalised stepping-stone sampling method and compare their Bayes factors [38-40]. We note that while the growth coefficient of the two coalescent models is not expected to converge to the same value (as they correspond to intrinsically different population dynamics), the two models are formally nested, and the likelihood function of the exponential model exists as a limit of the likelihood function of the logistic model. Therefore, even though the growth rate is effectively a nuisance variable for our purposes, the likelihoods of the two models can be compared via a likelihood ratio test to select the better-fitting one [41].

For each dataset, we perform MCMC runs for 100 million steps, sample trees every 10 thousand steps, and remove the first 10% of the steps as burin-in. We ensure that the effective sample size for every parameter of interest is >200 using Tracer v1.7 [42]. For the second part of the analysis, we use a two-parameter

molecular clock model with one strict clock rate for terminal branches and one strict clock for internal branches, using the same priors as before. This molecular clock model is a version of a fixed local clock whereby rather than having a single global rate on all branches, the terminal and internal branches are allowed to evolve according to different evolutionary rates while rate constancy is assumed along the respective branches [43, 44]. We use an MCMC chain of length 50 million steps, sampling every 1000 steps and evaluate sampling of the parameters of interest using Tracer v1.7.

4.5.2 Estimating the site frequency and number of non-neutral sites

We use the adapt-a-rate package [45-47], a generalised McDonald-Kreitman test, to estimate the number of non-neutral sites by assuming that deleterious mutations are mostly confined to the low frequency range (0%-15%), neutral mutations to the mid frequency range (15%-75%), and adaptive mutations to the high frequency range (75%-100%). We then estimate the site frequency spectra by comparing the main alignments to an ancestral sequence. For SARS-CoV-2, the ancestral sequence is the earliest sample collected from Wuhan, Wuhan/IPBCAMS-WH-01/2019, and for pH1N1 it is the earliest sample from Mexico, ACQ99614|A/Mexico/4108/2009. The choice of the cut-off frequencies is based on the diffusion approximation of allele frequencies whereby, at equilibrium, most deleterious and adaptive mutations are confined to frequencies <15% and >75%, respectively [46]. It has further been shown that the exact choice of the cut-off frequency for the three frequency classes does not significantly change the estimated number of non-neutral sites [46].

We use the ORF1ab gene of SARS-CoV-2 and HA gene of pH1N1 influenza for this analysis. We also carried out a similar analysis for the other genes of SARS-CoV-2, including the S gene. However, because of the limited genetic diversity present in the sequences and their relatively short size, our generalised McDonald-Kreitman test using adapt-a-rate is unable to estimate the number of non-neutral sites. The reason for choosing to analyse the HA gene of pH1N1 is primarily because of the availability of a large number of sequences from GenBank for the full coding region of HA which allows us to randomly sample 60 alignments per month from April 2009 to March 2010.

4.5.3 Resource availability

Datasets required to reproduce the analyses are available at https://github.com/mg878/twoclock_rate. The adapt-a-rate package is also available at https://github.com/jnarag/teaspoon.

4.6 References

1. C. Fraser, et al., *Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings.* Science, 2009. **324**(5934): p. 1557-1561.
2. G. J. D. Smith, et al., *Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic.* Nature, 2009. **459**(7250): p. 1122-U107.
3. J. Lu, et al., *Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China.* Cell, 2020. **181**(5): p. 997-1001.
4. A. G. Meyer, et al., *Time dependence of evolutionary metrics during the 2009 pandemic influenza virus outbreak.* Virus Evolution, 2015. **1**(1).
5. J. Sullivan and P. Joyce, *Model selection in phylogenetics.* Annual Review of Ecology Evolution and Systematics, 2005. **36**: p. 445-466.
6. Ghafari, M., et al., *A Mechanistic Evolutionary Model Explains the Time-Dependent Pattern of Substitution Rates in Viruses.* Current Biology, 2021. **31**(21): p. 4689-4696.e5.
7. S. Duchêne, E.C. Holmes, and S.Y.W. Ho, *Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates.* P Roy Soc B-Biol Sci, 2014. **281**(1786): p. 20140732.
8. A. G. Clark and T.S. Whittam, *Sequencing Errors and Molecular Evolutionary Analysis.* Molecular Biology and Evolution, 1992. **9**(4): p. 744-752.
9. S. Y. W. Ho, et al., *Time-dependent rates of molecular evolution.* Molecular Ecology, 2011. **20**(15): p. 3087-3101.
10. S. K. Gire, et al., *Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak.* Science, 2014. **345**(6202): p. 1369-1372.
11. E. C. Holmes, et al., *The evolution of Ebola virus: Insights from the 2013-2016 epidemic.* Nature, 2016. **538**(7624): p. 193-200.
12. S. Duchêne, et al., *Temporal signal and the phylodynamic threshold of SARS-CoV-2.* Virus Evolution, 2020. **6**(2).
13. S. Duchêne, et al., *The Performance of the Date-Randomization Test in Phylogenetic Analyses of Time-Structured Virus Data.* Molecular Biology and Evolution, 2015. **32**(7): p. 1895-1906.
14. G. G. R. Murray, et al., *The effect of genetic structure on molecular dating and tests for temporal signal.* Methods in Ecology and Evolution, 2016. **7**(1): p. 80-89.
15. S. Duchêne and D.A. Duchêne, *Estimating evolutionary rates and timescales from time-stamped data*, in *The Molecular Evolutionary Clock* 2020, Springer, Cham. p. 157-174.
16. E. C. Holmes, *Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus.* Journal of Virology, 2003. **77**(20): p. 11296-11298.
17. A. L. Hughes and M.A.K. Hughes, *More effective purifying selection on RNA viruses than in DNA viruses.* Gene, 2007. **404**(1-2): p. 117-125.
18. P. M. Sharp, et al., *The origins of acquired immune deficiency syndrome viruses: where and when?* Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 2001. **356**(1410): p. 867-876.
19. T. T. Y. Lam, et al., *Genomic Analysis of the Emergence, Evolution, and Spread of Human Respiratory RNA Viruses.* Annual Review of Genomics and Human Genetics, Vol 17, 2016. **17**: p. 193-218.

20. M. Woodhams, *Can deleterious mutations explain the time dependency of molecular rate estimates?* Molecular Biology and Evolution, 2006. **23**(12): p. 2271-2273.
21. J. O. Wertheim and S.L.K. Pond, *Purifying Selection Can Obscure the Ancient Age of Viral Lineages.* Molecular Biology and Evolution, 2011. **28**(12): p. 3355-3365.
22. S. Y. W. Ho, et al., *Time dependency of molecular rate estimates and systematic overestimation of recent divergence times.* Molecular Biology and Evolution, 2005. **22**(7): p. 1561-1568.
23. O. G. Pybus, et al., *Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution.* Molecular Biology and Evolution, 2007. **24**(3): p. 845-852.
24. J. Hedge, S.J. Lycett, and A. Rambaut, *Real-time characterization of the molecular epidemiology of an influenza pandemic.* Biology Letters, 2013. **9**(5).
25. S. Kryazhimskiy and J.B. Plotkin, *The Population Genetics of dN/dS.* Plos Genetics, 2008. **4**(12): p. e1000304.
26. E. P. C. Rocha, et al., *Comparisons of dN/dS are time dependent for closely related bacterial genomes.* Journal of Theoretical Biology, 2006. **239**(2): p. 226-235.
27. S. Duchêne, et al., *Temporal signal and the phylodynamic threshold of SARS-CoV-2.* Virus Evolution, 2020. **6**(2): p. veaa061.
28. Y. C. F. Su, et al., *Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection.* Nature Communications, 2015. **6**.
29. L. du Plessis, et al., *Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK.* Science, 2021. **371**(6530): p. 708-712.
30. E. Volz, et al., *Report 5 - Phylogenetic analysis of SARS-CoV-2.* MRC Centre for Global Infectious Disease Analysis, 2020.
31. Rambaut, A. and E. Holmes, *The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic.* PLoS Currents, 2009. **1: RRN1003**.
32. S. Bhatt, et al., *The evolutionary dynamics of influenza A virus adaptation to mammalian hosts.* Philosophical Transactions of the Royal Society B-Biological Sciences, 2013. **368**(1614): p. 20120382.
33. J. Lu, et al., *Molecular Evolution, Diversity, and Adaptation of Influenza A(H7N9) Viruses in China.* Emerging Infectious Diseases, 2018. **24**(10): p. 1795-1805.
34. K. E. Kistler and T. Bedford, *Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses OC43 and 229e.* Elife, 2021. **10**: p. e64509.
35. R. C. Edgar, *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Research, 2004. **32**(5): p. 1792-1797.
36. N. De Maio, et al., *Issues with SARS-CoV-2 sequencing data.* Virological.org, 2020. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473(20 October 2020).
37. M. A. Suchard, et al., *Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10.* Virus Evolution, 2018. **4**(1).

38. G. Baele, P. Lemey, and M.A. Suchard, *Genealogical Working Distributions for Bayesian Model Testing with Phylogenetic Uncertainty.* Systematic Biology, 2016. **65**(2): p. 250-264.
39. Y. Fan, et al., *Choosing among Partition Models in Bayesian Phylogenetics.* Molecular Biology and Evolution, 2011. **28**(1): p. 523-532.
40. Baele, G. and P. Lemey, *Bayesian model selection in phylogenetics and genealogy-based population genetics*, in *Bayesian phylogenetics, methods, algorithms, and applications*. 2014, CPC Press: Boca Raton (Florida). p. 59–93.
41. O. G. Pybus and A. Rambaut, *GENIE: estimating demographic history from molecular phylogenies.* Bioinformatics, 2002. **18**(10): p. 1404-1405.
42. A. Rambaut, et al., *Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7.* Systematic Biology, 2018. **67**(5): p. 901-904.
43. A. J. Drummond and M.A. Suchard, *Bayesian random local clocks, or one rate to rule them all.* BMC Biology, 2010. **8**: p. 114
44. A. D. Yoder and Z.H. Yang, *Estimation of primate speciation dates using local molecular clocks.* Molecular Biology and Evolution, 2000. **17**(7): p. 1081-1090.
45. J. Raghwani, S. Bhatt, and O.G. Pybus, *Faster Adaptation in Smaller Populations: Counterintuitive Evolution of HIV during Childhood Infection.* Plos Computational Biology, 2016. **12**(1).
46. S. Bhatt, E.C. Holmes, and O.G. Pybus, *The Genomic Rate of Molecular Adaptation of the Human Influenza A Virus.* Molecular Biology and Evolution, 2011. **28**(9): p. 2443-2451.
47. S. Bhatt, A. Katzourakis, and O.G. Pybus, *Detecting natural selection in RNA virus populations using sequence summary statistics.* Infection Genetics and Evolution, 2010. **10**(3): p. 421-430.

Statement of Authorship for joint/multi-authored papers for PGR thesis.

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

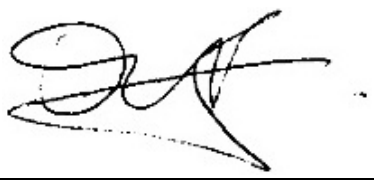
Title of Paper	Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza
Publication Status	Published
Publication Details	Ghafari M, du Plessis L, Raghwani J, Bhatt S, Xu B, Pybus O, and Katzourakis A. Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza. Molecular Biology and Evolution 39: msac009, 2022.

Student confirmation

Student Name	Mahan Ghafari
Contribution to the paper	MG wrote the paper and conducted all the analysis under the supervision from AK. All authors provided editorial assistance.
Signature 	Date 11/05/2022

Supervisor confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title	Professors Aris Katzourakis and Oliver G Pybus	
Supervisor comments		
Signature  		Date 11/05/2022

Chapter 5: Investigating the evolutionary origins of the first three SARS-CoV-2 variants of concern

5.1 Abstract

The emergence of Variants of Concern (VOCs) of SARS-CoV-2 with increased transmissibility, immune evasion properties, and virulence poses a great challenge to public health. Despite unprecedented efforts to increase genomic surveillance, fundamental facts about the evolutionary origins of VOCs remain largely unknown. One major uncertainty is whether the VOCs evolved during transmission chains of many acute infections or during long-term infections within single individuals. We test the consistency of these two possible paths with the observed dynamics, focusing on the clustered emergence of the first three VOCs, Alpha, Beta, and Gamma, in late 2020, following a period of relative evolutionary stasis. We consider a range of possible fitness landscapes, in which the VOC phenotypes could be the result of single mutations, multiple mutations that each contribute additively to increasing viral fitness, or epistatic interactions among multiple mutations that do not individually increase viral fitness—a “fitness plateau”. Our results suggest that the timing and dynamics of the VOC emergence, together with the observed number of mutations in VOC lineages, are in best agreement with the VOCs having evolved within single individuals with long-term infections.

5.2 Introduction

For the first 8 months of the SARS-CoV-2 pandemic, the virus exhibited a very slow pace of adaptation, with D614G being the only persistent adaptive substitution that appears to have resulted in an increased transmissibility of the virus [1-3]. However, during the second half of 2020, three designated variants of concern (VOCs) of

SARS-CoV-2, Alpha, Beta, and Gamma, emerged independently and in quick succession [4-6]. No other VOC emerged until Delta and Omicron in 2021 which appear to be very different, both genetically and phenotypically, from the three original VOCs [7, 8]. The VOCs are characterised by a large number of mutations relative to the genetic background from which they first emerged, and exhibit altered phenotypes resulting in varying combinations of increased transmissibility, virulence, and immune evasion [6, 9-11].

Phylogenetic analyses show that a large number of mutations, mostly located in the spike protein, have independently evolved in multiple lineages of SARS-CoV-2 including the Alpha, Beta and Gamma variants and are likely playing a key role in the adaptive evolution of the SARS-CoV-2 [7, 12]. Experimental measurements and molecular dynamics simulations also show that some of these mutations have synergistic interactions for important functional traits [13, 14], indicating that they may have greater combined fitness benefit to the virus. Some of the distinctive mutations in the VOCs, including the E484K and N501Y mutations found in the first three VOCs, have also been observed in chronic infections such as those in certain immunocompromised individuals [15-17], suggesting that the VOCs may have arisen from such infections. Some of the other possible explanations for the emergence of VOCs include prolonged circulation of the virus in areas of the world with poor genomic surveillance or reverse-zoonosis from other animals such as rodents followed by sustained transmission and adaptive evolution within the animal population and a spill over back to the humans (see [18] for a recent review on the possible origins of variants of SARS-CoV-2).

While finding the evolutionary process(es) that may have led to the emergence of VOCs has profound consequences for understanding the fate of the SARS-CoV-

2 pandemic, there have currently been no systematic investigations to assess the likelihood of any particular evolutionary pathway that would lead to the emergence of VOCs. In this work we investigate whether the emergence of VOCs was the result of evolution via sustained transmission chains between acutely infected individuals or prolonged infections, and evaluate plausible fitness landscapes. We also discuss the potential implications of our results for the future of the pandemic and potential measures that might lower the rate at which new VOCs emerge.

5.3 Results

5.3.1 Emergence of VOCs: an evolutionary dilemma

The Alpha, Beta, and Gamma VOCs arose independently and in quick succession, with several shared mutations, in three different countries and began to spread globally (Figure 5.1). This long waiting time followed by clustered emergence of a handful of lineages was not predicted by any simple evolutionary theories. Typically, one would assume that either the beneficial mutation supply is small, in which case one expects a long waiting time for the first VOC but also long gaps before subsequent VOCs, or the mutation supply is large, in which case one expects many VOCs with only a short waiting time [19]. Moreover, each VOC had >6-10 mutations distinguishing it from then-dominant genotypes, which was also unexpected. One of the key evolutionary questions is whether VOCs evolved over the course of many acute infections or within single chronic infected hosts. Both possibilities have serious issues. The many-acute-infections hypothesis needs to explain how the virus acquired so many changes, as the mutant lineages would have had to remain at frequencies below the detection threshold in different countries for several months. The chronic-infection hypothesis needs to explain both why adaptation to the within-host environment led to a transmission advantage between hosts, and

why there was no 'leakage' of some intermediate mutations at the between-host level before the emergence of the VOCs, i.e., why genotypes with some of the VOC mutations did not escape from the chronically infected patients earlier.

Table 5.1: **Model parameters.**

Symbol	Description	Value (range)	Source
t	Time in units of generations	assuming 5.2 days per generation	[52]
IFR	Global median infection fatality rate of COVID-19	0.5% (0.2% - 1.5%)	[47]
N	Number of daily infectious individuals worldwide	daily confirmed deaths / median global IFR	--
μ	Mutation rate per nucleotide per generation	1.0 (0.87 - 2.0) x 10^{-5}	[53]
s	Selective advantage of the VOCs	(0.3 - 1.1)	--
k	Dispersion in distribution of number of secondary infections	0.1 (0.05 - 0.2)	[25]
T_{obs}	Time to the emergence of the first VOC (number of days since 2020-01-03)	(180 - 317) days	[4-6, 54]
ΔT_{obs}	Time between the emergence of the first and second VOC	(0 - 137) days	[4-6, 54]
P_f	Probability of a chronic SARS-CoV-2 infection in an ICI producing a VOC	--	--
μ_C	Within-host fixation rate of VOC mutations per generation	--	--

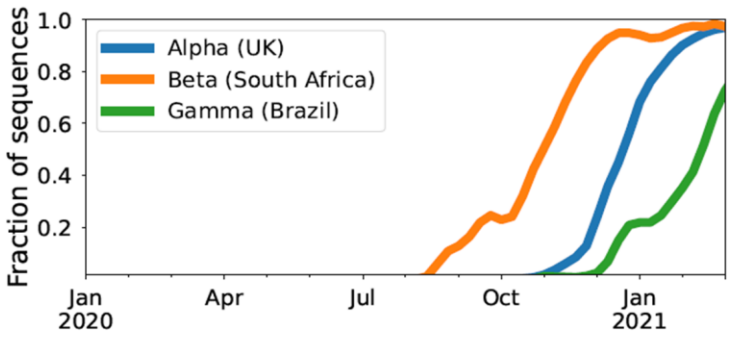
5.3.2 Between-host model of VOC emergence

We assume the effective virus population size is $N_e=N/\sigma^2$ where N is the number of infectious individuals worldwide and σ^2 is the variance in offspring number (secondary cases). We treat each acute infection as one generation, assuming a tight transmission bottleneck of a single virion [20-22]. Viruses mutate at rate μ per base per generation (see Methods section). For a mutant virus population with selective advantage s relative to the background, the average number of secondary cases increases by a factor $1+s$. We also assume that the number of secondary cases approximately follows a negative binomial distribution with mean R_t and dispersion parameter k, so that $\sigma^2 \approx R_t(1+R_t/k)$. There is substantial uncertainty in the amount of overdispersion in the pandemic, and consequently similar uncertainty in the effective population size. Therefore, we consider a range of values for k to see if any would be consistent with the observed dynamics of the VOC emergence. We

also note that while the importance of spatial structure is clearly visible in the spatially restricted initial spread of the VOCs from real-world data, we expect that we can neglect it when analysing their *emergence*. This is because spatial structure should not have a large impact on viral dynamics until a lineage becomes locally common, and the specific mutations differentiating the VOCs were all locally rare prior to their emergence.

5.3.3 Within-host model of VOC emergence

Unlike tracking the between-host evolution of SARS-CoV-2 where there has been an unprecedented effort has led to huge numbers of consensus genome sequences [23], our current knowledge of the within-host evolutionary dynamics of SARS-CoV-2 is still very limited, particularly in those with chronic infections. Because there is very limited data with which to constrain the within-host evolutionary dynamics of chronic infections with SARS-CoV-2, we simply treat it as a 'black box' and assume with some probability, P_f, that a new infection is chronic and may lead to the production of a VOC (Table 5.1; Methods section). We also assume that within-host substitutions required for the production of the VOC occur at a constant rate μ_C per generation (see Table 5.1). (Here a generation is still defined as the typical length of an *acute* infection.) Given that we know only three VOC lineages emerged by late 2020, we expect $T_{obs}N\,P_f \sim 3$ where T_{obs}~180-317 days is the expected time to the emergence of the first VOC since the beginning of the pandemic based on phylogenetic estimates (see Table 5.1). Therefore, given the typical variation in the population size throughout the pandemic for biologically relevant parameter combinations $N \sim 1\text{x}10^6\text{-}1\text{x}10^7$, we expect that values of $P_f \sim 5\text{x}10^{-9}\text{-}1\text{x}10^{-7}$ will maximize the likelihood of the within-host model and focus on these.



Variant\Site	Spike: 18	Spike: 417	Spike: 484	Spike: 501	Spike: 614	NSP-6: 106	N: 203-204
Alpha			E*	Y	G	deletion	K-R
Beta	F	N	K	Y	G	deletion	
Gamma	F	T	K	Y	G	deletion	K-R

*E484K has been reported in some of the Alpha variant genomes sampled within the UK and elsewhere.

Figure 5.1: **The three initial Variants of Concern arose in quick succession after a long period of limited adaptation.** For each VOC, the curve shows its frequency among the SARS-CoV-2 sequences collected each week from its country of origin. The table shows the amino acid changes across the SARS-CoV-2 genome that are shared between at least two of the three VOCs [7].

5.3.4 Fitness landscapes

One possible explanation for the temporal clustering of VOCs with large numbers of mutations is that the underlying fitness landscape may have some structure that causes the dynamics to deviate from our usual expectations. Unfortunately, the full space of possible fitness landscapes is enormous and impossible to explore exhaustively. To investigate the possible effects of the landscape on the dynamics, we therefore focus on three limiting local fitness landscapes that span a range of biologically plausible scenarios (Figure 5.2A). Importantly, these landscapes describe only between-host fitness, which could be very different from within-host fitness. As mentioned above, we treat within-host dynamics implicitly using an effective substitution rate and so do not need an explicit fitness landscape for it. In all three landscapes, the peak is a VOC phenotype with fitness advantage s over the ancestor. We assume that Alpha, Beta, and Gamma are similar enough that they can be approximately described by the same landscape and the same value of s, which we infer from the early rate of increase of the VOCs (see Methods). Landscape 1 is the simplest possibility: a single mutation on the ancestral background is sufficient to confer the full advantage. In Landscape 2, we test whether simply increasing the number of mutations involved can explain the temporal clustering. In this landscape, the VOC phenotype is produced by a combination of $K > 1$ mutations, each providing an independent fitness benefit s/K. In Landscape 3, we test whether epistasis may have an effect: the VOC phenotype again requires K mutations, but we now assume that they provide no fitness benefit until the full combination is acquired, i.e., the population must cross a fitness plateau. As mentioned above, there is experimental evidence for this form of epistasis among the VOC mutations [13, 14]. We expect that shallow fitness valleys

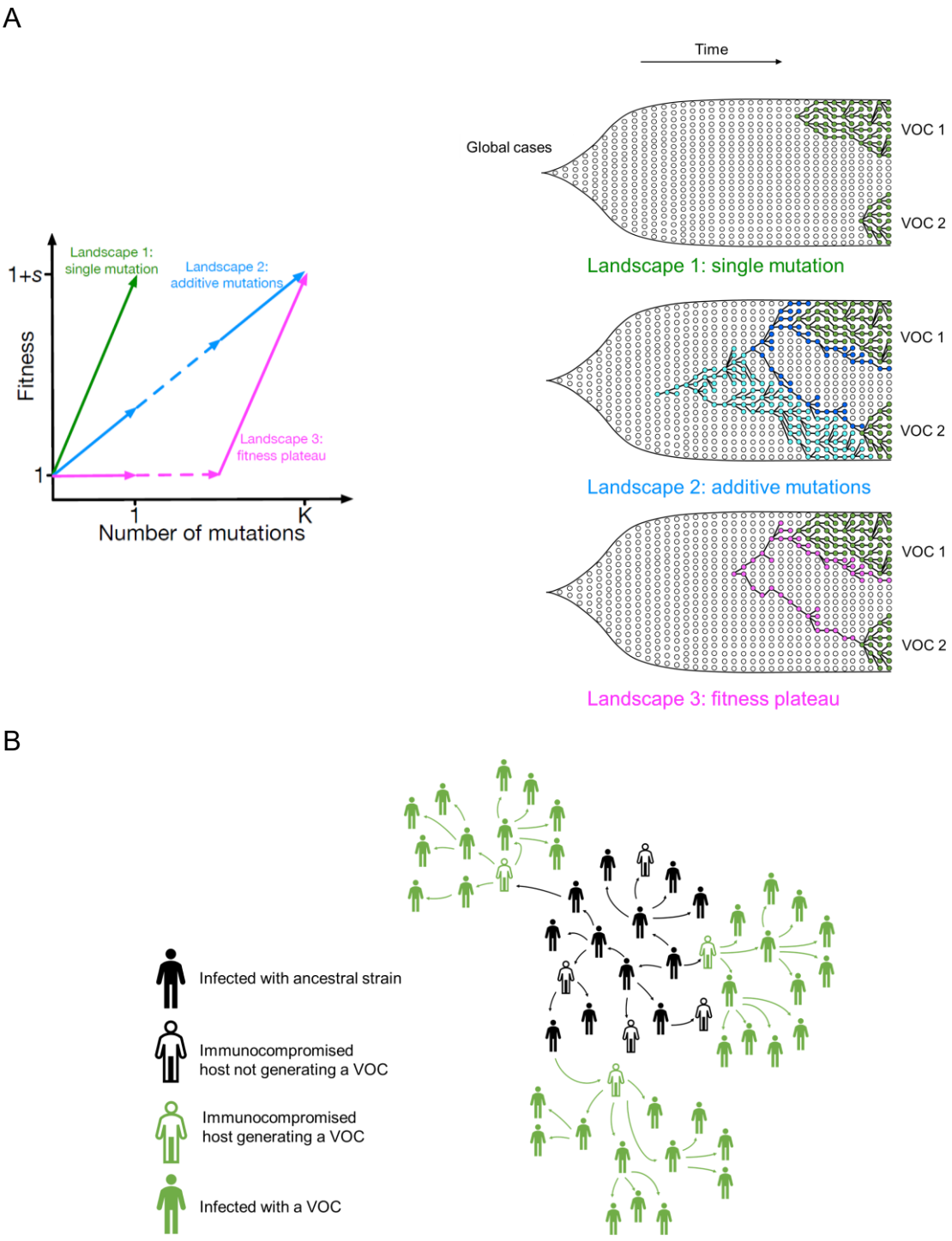


Figure 5.2: **Possible evolutionary pathways to the emergence of SARS-CoV-2 VOCs. (A**, left**)** The three limiting fitness landscapes for the emergence of VOCs as a function of the relevant number of mutations required, *K*. **(A**, right**)** VOCs can emerge from either a single advantageous mutation (green) or multiple mutations that each contribute independently to increasing fitness (blue) or only in combination (magenta). **(B)** Emergence of VOCs via the within-host evolutionary path such that an infectious individual passes on a wild-type variant of the virus to an immunocompromised individual where the virus may acquire the relevant mutations during the chronic phase of the infection and later be passed on to the rest of the population.

will produce similar dynamics to Landscape 3, as will shallow upward slopes with a large jump in fitness at the end [24]. Note that mutations in all the three landscapes can be acquired via the between- or within-host evolutionary pathways (Figure 5.2B). For each evolutionary scenario, we find a range of parameter values consistent with the data on the timing of the emergence of Alpha, Beta, and Gamma variants of SARS-CoV-2 (see Methods; Table 5.1). For these parameter values, we further investigate whether they correspond to biologically reasonable scenarios in terms of the frequencies of the intermediate mutations prior to the emergence of VOCs, total number of mutations required to produce VOCs, total number of successful VOC lineages produced over time, and the timing between the emergence of different VOC lineages.

5.3.5 Landscape 1: single mutations

We start with the simplest possible fitness landscape, in which a single mutation conferring a fitness advantage s relative to the genetic background of circulating lineages is required for the emergence of VOCs. We first consider the between-host evolutionary pathway. As long the effective population size of the pandemic was not much smaller than the census size (i.e., overdispersion was not too large), the mutation supply $N_e \mu$ became large early in 2020. At this point, numerous lineages would have emerged over a short period of time (Figure S1), inconsistent with the observed dynamics.

If overdispersion were very large, it could have kept $N_e \mu$ low through the establishment of the VOCs. Figure 5.3A,B shows that under extremely high levels of overdispersion (k=0.005) this model can match the long waiting time for the emergence of the first VOC. Such high levels of overdispersion are not supported

by any existing epidemiological studies on SARS-CoV-2 transmission [25]. Moreover, Figure 5.3 shows that this model cannot fit the joint waiting time distribution for all three VOCs. Under these mutation-limited conditions, there is an approximately exponential waiting time for the arrival of each VOC lineage (once the growth in the total pandemic slows). Thus, it predicts similarly long waiting times for the emergence of Alpha, Beta, and Gamma, inconsistent with the observed clustering. Therefore, there is no biologically reasonable combination of parameters that result in the clustered emergence of VOCs in late 2020 via Landscape 1 between-host evolutionary pathway.

On the other hand, if VOCs arose from chronic infections, then their emergence was a two-step process: first, chronic infections had to occur, and then the VOC mutation had to arise in them. The waiting time for the first step is determined by NP_f; note that the number of chronic infections depends on the census size N rather than N_e, i.e., it is insensitive to the amount of overdispersion. The second step follows an exponential distribution within each chronic host, with rate μ_C. Once the VOC mutation had arisen, it would establish in the population with probability s. Figure 5.3C-F shows that as long the level of overdispersion is very high where there is very little chance of between-host evolutionary pathway actively contributing towards the production of VOCs (i.e., very low mutation supply, $N_e\mu$) such that chronic infections are frequently produced in the population (i.e., $NP_f \sim 1$) and there is a relatively long waiting time before the production of each VOC mutation ($\mu_C \sim 0.01$), the within-host pathway can create evolutionary dynamics that resemble the clustered emergence of 2-6 VOC lineages by late 2020. However, like the between-host pathway, this scenario requires very high levels of overdispersion which makes Landscape 1 within-host evolutionary pathway also an unlikely

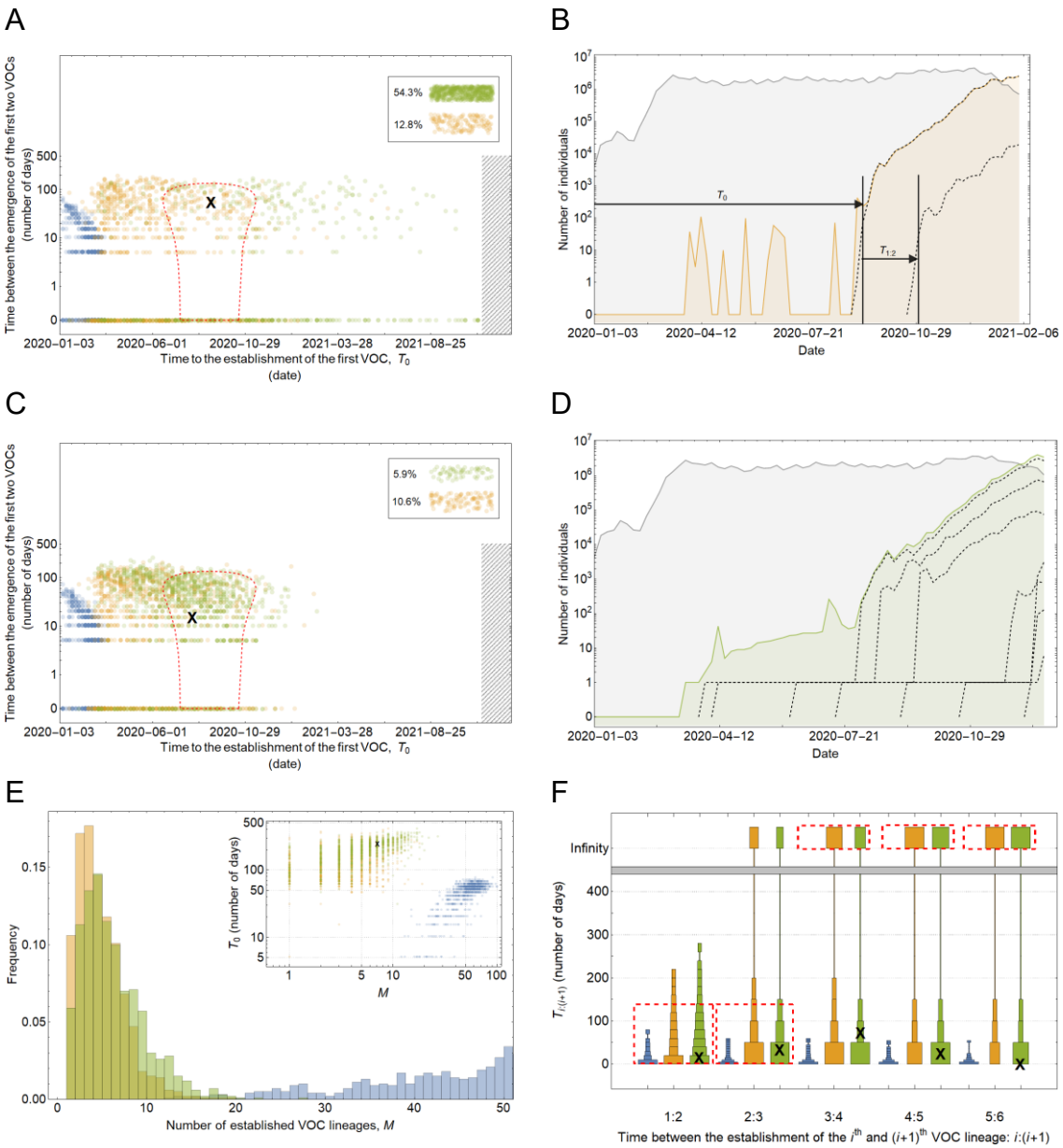


Figure 5.3: **Evolutionary dynamics of VOC emergence assuming a fitness landscape with a single adaptive mutation. (A)** The establishment time of the first two VOCs that emerged via the between-host pathway is measured under varied levels of overdispersion, k, such that IFR=1.5%, μ=0.87x10^{-5}, K=1, and s=0.4. For each simulation run, represented by a point on the graph, we measure the time that it takes for a single adaptive mutation to establish in the population and the time difference between the establishment of the first and second successful VOC lineage assuming k=0.2 (blue) k=0.005 (orange), and k=0.001 (green). For each of the three scenarios, there are 1,000 simulation runs. The inset shows that 12.8% and 54.3% of the runs for the k=0.005 and k=0.001 scenarios only have either zero or one established VOC lineage by the end of the simulation period. Each run stops once the frequency of the VOC population reaches 75%. The hatched area shows the region that goes beyond the last possible time point in any simulation run, 695 days. **(B)** A typical evolutionary trajectory of the k=0.005 scenario highlighted with a cross sign ("**X**") in (A). The graph shows the VOC population (orange) emerging from the background population (gray). The time that it takes for the establishment of the first VOC, T_0, and the time between the establishment of the first and second VOC, $T_{1:2}$, are highlighted. The black dashed lines show the trajectories of all established VOC lineages over time. **(C)** The establishment time of the first two VOCs

that emerged via the within-host pathway is measured assuming k=0.2, P_f=5x10^{-10} and μ_C=0.001 (blue), k=0.005, P_f=6x10^{-8} and μ_C=0.1 (orange), and k=0.001, P_f=6x10^{-6} and μ_C=0.01 (green), such that μ=0.87x10^{-5}, IFR=1.5%, K=1, and s=0.4. **(D)** A typical evolutionary trajectory of the within-host model of VOC emergence with P_f=6x10^{-6} and μ_C=0.01 (green) highlighted with a cross sign ("**X**") in (C). The graph shows the VOC population (green) along with the individual VOC lineages (dashed lines) emerging from the background population (gray). **(E)** The total number of VOC lineages established (M) for each of the three scenarios in (C). The inset shows a graph of M with respect to T_0 for each simulation point. **(F)** The distribution of time that it takes between the establishment of the ith and $(i+1)^{th}$ lineage, $T_{i:(i+1)}$, for the first 6 established VOC lineages in each of the three scenarios in (C). For each of the simulation runs where the $(i+1)^{th}$ lineage does not emerge after the establishment of the ith lineage, we assume $T_{i:(i+1)}$ is infinity. In graphs (A), (C), and (F), the enclosed red dashed lines show the region of the parameter space corresponding to the emergence the first three SARS-CoV-2 VOCs. In graphs (E) and (F), the cross signs ("**X**") highlight the values corresponding to the simulation run shown in (D).

explanation for the emergence of VOCs.

5.3.6 Landscape 2: additive mutations

Landscape 2 corresponds to an evolutionary pathway in which there were $K>1$ major mutations involved in the emergence of VOCs, each making an additive contribution of $\approx s/K$ to fitness. If evolution occurred at the whole-population level, Figure 5.4A,B and Figure S2 show that, for a range of parameter combinations, the additive fitness landscape requiring up to four mutations can create evolutionary dynamics with long waiting times before the arrival of the first successful VOC lineage—for combinations of more than 4 mutations, VOC lineages do not emerge by late 2020 under any biologically reasonable parameter combinations for effective population size, mutation rate, and selective coefficient. However, this long waiting time is accompanied by the establishment of more than a few dozen VOC lineages that emerge in quick succession most cases (see Figure S2A,B). The sheer number of VOC lineages established under these scenarios is not supported by the observation of only 3 VOC lineages emerging in late 2020. A closer examination of a typical $K=3$ evolutionary trajectory that matches the long waiting time before the establishment of the first VOC reveals that the intermediate single- and double-mutants reach high frequencies before the emergence of the first successful (triple-mutant) VOC lineage (Figure 5.4B) -- we observe similar dynamics for $K=2$ and 4 scenarios (Figure S2D,F). The sequential fixation of adaptive mutations at the population-level implies that the intermediate mutations were detectable many months prior to the emergence of VOCs which is also not supported by the genomic surveillance data from around the world. The sequential fixation dynamics creates a ladder-like phylogenetic relationship between the background and mutant populations whereby every new VOC mutation becomes dominant in the population

before giving rise to lineages with additional mutations. Even though such phylogenetic relationships may emerge in SARS-CoV-2 over longer evolutionary timescales (as have been observed in human coronaviruses [26]), they do not resemble the underlying phylogenetic topology of the VOC clades of SARS-CoV-2, which tend to be more star-like.

For the chronic-infection pathway, on the other hand, the intermediate mutants could have fixed within the host while remaining at undetectable frequencies at the between-host level until the production of the VOCs. Figure 5.4C-F and Figure S3 show that for a combination of parameters requiring K=3 and K=6 mutations where the mutation supply is low and the strength of selection is relatively weak such that the intermediate mutants cannot reach fixation before the emergence of the VOC population, Landscape 2 within-host pathway can lead to the clustered emergence of a few VOC lineages by late 2020. However, like we saw with the Landscape 2 between-host pathway, if the selective coefficient s is too high, the intermediate mutants can reach fixation from very early on in the pandemic. This can also result in new chronic infections being seeded by those intermediate mutants which have already acquired some of the VOC mutations, further expediting the emergence of VOC lineages over a short time span. One way to prevent any intermediate mutants from reaching fixation is by increasing the number of VOC mutations, K. This has a net effect of reducing the per capita strength of selection on each mutant genotype, thereby reducing the rate at which they grow in size over time (see Figure 5.4D and Figure S3).

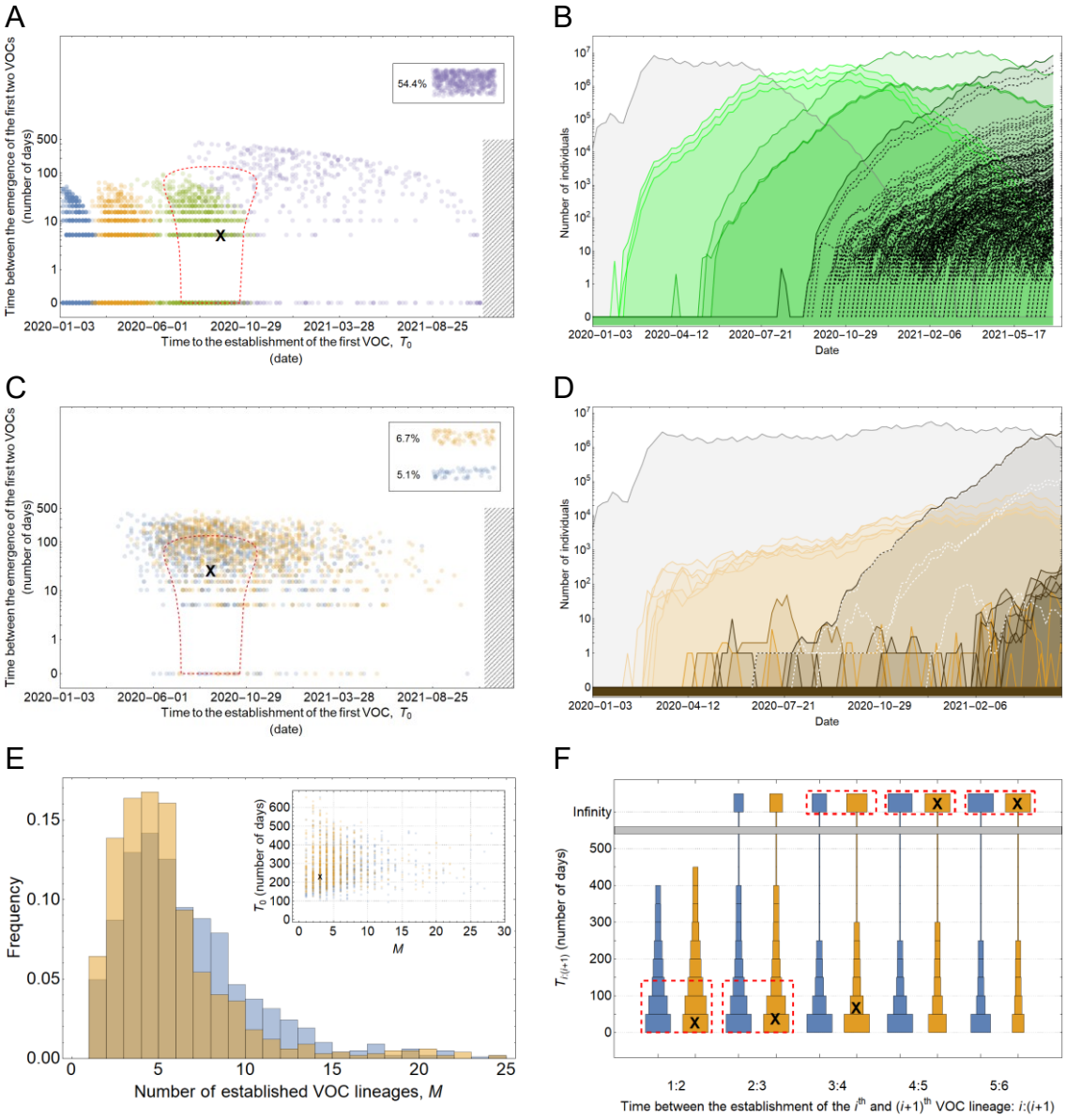


Figure 5.4: **Evolutionary dynamics of VOC emergence assuming a fitness landscape with additive mutations. (A)** The establishment time of the first two VOCs that emerged via the between-host pathway is measured for K=1 (blue), K=2 (orange), K=3 (green), and K=4 (purple) such that IFR=0.5%, μ=1x10^{-5}, k=0.2, and s=1.0. For each simulation run, represented by a point on the graph, we measure the time that it takes for the first VOC lineage to establish in the population and the time difference between the establishment of the first and second successful VOC lineages. For each scenario, there are 1,000 simulation runs. The inset shows the percentage of simulation runs that either produced no or 1 VOC lineage by the end of each simulation period. Each run stops once the frequency of the VOC population reaches 75%. The hatched area shows the region that goes beyond the last possible time point in any simulation, 695 days. **(B)** A typical evolutionary trajectory of the K=3 between-host model of VOC emergence (green) highlighted with a cross sign ("**X**") in (A). The graph shows the background population in gray and the single-, double-, and triple-mutant (VOC) populations in different shades of green from light to dark, respectively. Note that for the K=3 scenario, there are three single-mutant, three double-mutant, and one triple-mutant genotypes. The dashed lines show the dynamics of all the established VOC lineages over time. **(C)** The establishment time of the first two VOCs that emerged via the within-host pathway is measured for K=3,

P_f=3.5x10^{-8}, and μ_C=0.15 (blue) and K=6, P_f=3x10^{-8}, and μ_C=0.3 (orange), such that μ=0.87x10^{-5}, IFR=1.5%, k=0.05, and s=0.3. **(D)** A typical evolutionary trajectory of the K=6 within-host model of VOC emergence (orange) highlighted with a cross sign ("**X**") in (C). The graph shows the background population in gray and the i-mutant populations ($1<i\leq K$) in different shades of orange from light (fewer mutations) to dark (more mutations). **(E)** The total number of VOC lineages established (M) for each of the two scenarios in (C). The inset shows a graph of M with respect to T_0 for each simulation point. **(F)** The distribution of time that it takes between the production of the ith and $(i+1)^{th}$ lineage, $T_{i:(i+1)}$, for the first 6 established VOC lineages in each of the two scenarios in (C). For each simulation run where the $(i+1)^{th}$ lineage does not emerge after the establishment of the ith lineage, we assume $T_{i:(i+1)}$ is infinity. In graphs (A), (C), and (F), the enclosed red dashed lines show the region of the parameter space corresponding to the emergence of the first three SARS-CoV-2 VOCs. In graphs (E) and (F), the cross signs ("**X**") highlight the values corresponding to the simulation run shown in (D).

5.3.7 Landscape 3: fitness plateau crossing

As in Landscape 2, Landscape 3 describes an evolutionary pathway where there are $K>1$ major mutations involved in the generation of VOCs, but, in this case, only the full K-mutant VOC genotype has a selective advantage s relative to the background population. This does not necessarily imply that the selective coefficients of the intermediate genotypes are small in the standard weak-selection sense (small relative to $1/N_e$), but only that they are too small to substantially affect the dynamics of the production of the first successful K-mutant VOC lineage [24].

For the between-host model of VOC emergence, our analysis suggests that only a plateau-crossing of size K=2 may be consistent with the timing of the emergence of SARS-CoV-2 VOCs (Figure 5.5A). Extended plateaus requiring K>2 mutations take much longer to cross and for most parameter combinations either zero or 1 VOC lineage is produced before the end of 2020. For a typical K=2 plateau-crossing trajectory, single-mutant genotypes grow linearly over time and reach a frequency of $\lessapprox$0.1% before producing ~1-5 successful VOC lineages that emerge in quick succession (see Figure 5.5B; Figure S3A,B). Therefore, unlike the between-host evolutionary pathway in Landscape 2, a fitness plateau could have led to the clustered emergence of several VOCs after a long waiting time during which none of the intermediate mutations reached high frequency. However, the fact that for biologically plausible parameter values only a narrow plateau of K=2 mutations can be crossed seems inconsistent with the high number of mutations found in the VOCs and particularly with the high number of similar mutations shared across unrelated VOC lineages. This inconsistency may be partly reconciled with the possibility of compounded evolutionary effects following the plateau-crossing event such as the

emergence of hyper-mutability traits across certain sites or strong within-host selection following the acquisition of the *K* mutations.

On the other hand, if the VOCs arose from chronic infections, the intermediate (neutral) VOC mutations can fix within a host much more quickly, allowing extended plateaus to be crossed compared to the between-host evolutionary pathway. Also, unlike Landscape 2 within-host pathway, the early leakage of intermediate mutations to the population is much less likely as they have no strong selective advantage over the background population, nor will there be long exponential delays between the emergence of each VOC lineage like Landscape 1. The latter is because unlike the within-host pathway requiring a single adaptive mutation at the population level (i.e., K=1) with rate parameter μ_C, for any evolutionary scenario requiring K>1, the waiting time for the fixation of K mutations within a host during a chronic infection is an Erlang-distributed number mean K/μ_C and variance K/μ_C^2. Therefore, for a fixed μ_C, the variance of the Erlang distribution shrinks inversely proportional to K. This implies that the more extended the fitness-plateau is, the less variable the emergence time of the full K-mutant becomes. Figure 5.5C-F reveals that the within-host evolutionary pathway of Landscape 3 creates evolutionary trajectories that are consistent with the clustered emergence of ~3 VOCs in late 2020. There is also less seeding of new chronic infections with intermediate mutations, leading to fewer VOC lineages compared to Landscape 2 (also see Figure S4C).

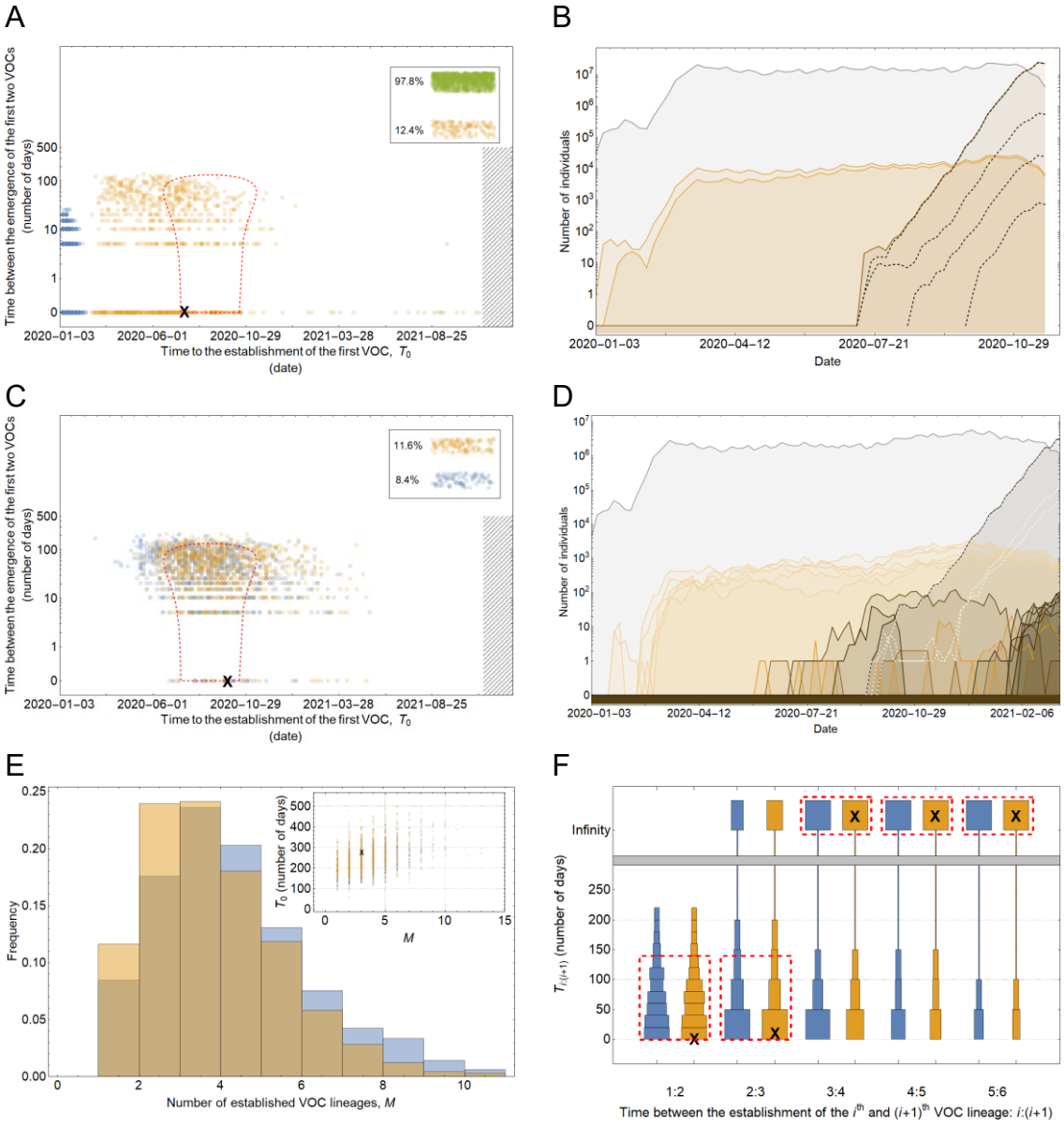


Figure 5.5: **Evolutionary dynamics of VOC emergence assuming a fitness plateau crossing. (A)** The establishment time of the first two VOCs that emerged via the between-host pathway is measured for K=1 (blue), K=2 (orange), and K=3 (green) such that IFR=0.2%, μ=2x10^{-5}, k=0.2, and s=1.0. For each simulation run, represented by a point on the graph, we measure the time that it takes for the first VOC lineage to establish in the population and the time difference between the establishment of the first and second successful VOC lineages. For each scenario, there are 1,000 simulation runs. The inset shows the percentage of simulation runs that either produced no or 1 VOC lineage by the end of each simulation run. Each run stops once the frequency of the VOC population reaches 75%. The hatched area shows the region that goes beyond the last possible time point of any simulation, 695 days. **(B)** A typical evolutionary trajectory of a K=2 plateau crossing scenario (orange) of the between-host model of VOC emergence highlighted with a cross sign ("**X**") in (A). The graph shows the background population in gray, single-mutants in light orange, and double-mutants in dark orange. Note that for the K=2 scenario, there are two single-mutant and one double-mutant genotypes. **(C)** The establishment time of the first two VOCs that emerged via the within-host pathway is measured for K=3, P_f=2x10^{-8}, and μ_C=0.1 (blue) and K=6, P_f=4.5x10^{-8}, and μ_C=0.25 (orange), such that μ=1x10^{-5}, IFR=0.5%, k=0.1, and s=0.7. **(D)** A typical evolutionary

trajectory of the within-host model of VOC emergence requiring a K=3 fitness plateau crossing (orange) and highlighted with a cross sign ("**X**") in (C). The graph shows the background population in gray and the i-mutant populations ($1<i\leq K$) in different shades of orange from light (fewer mutations) to dark (more mutations). **(E)** The total number of VOC lineages established (M) for each of the two scenarios in (C). The inset shows a graph of M with respect to T_0 for each simulation point. **(F)** The distribution of time that it takes between the production of the ith and $(i+1)^{th}$ lineage, $T_{i:(i+1)}$, for the first 6 established VOC lineages in each of the two scenarios in (C). For each simulation run where the $(i+1)^{th}$ lineage does not emerge after the establishment of the ith lineage, we assume $T_{i:(i+1)}$ is infinity. In graphs (A), (C), and (F), the enclosed red dashed lines show the region of the parameter space corresponding to the emergence of the first three SARS-CoV-2 VOCs. In graphs (E) and (F), the cross signs ("**X**") highlight the values corresponding to the simulation run shown in (D).

Finally, it is important to note that in all of the within-host evolutionary pathways (i.e., Landscapes 1, 2, and 3), we found parameter combinations that can re-create the clustered emergence of the first three SARS-CoV-2 VOCs. In particular, we showed that either because of having very few mutations that are selectively beneficial at the population level (i.e., Landscape 1) or the low prevalence of intermediate mutations before the emergence of the VOCs (i.e., Landscapes 2 and 3), we would expect the phylogenetic relationship between the VOC lineages and background populations to manifest itself with a long evolutionary distance branch connected to deeper internal nodes of the tree with each VOC clade independently emerging from a unique genetic background and be subsequently replaced by another VOC clade from an entirely different background (Figure S4D). This creates a phylogenetic relationship between VOC clades that is similar to what we observe for Alpha, Beta, and Gamma variants [4-6].

5.4 Discussion

The global spread of the Omicron variant of SARS-CoV-2 has given a renewed attention to the underlying evolutionary mechanisms that lead to the emergence of VOCs. Practically, we would like to know whether to expect future VOCs to arise, and if so when and whether there will be early warning signs. Answering this question is not only important for understanding the fate of the pandemic but also may have major public health implications for how to best develop strategies for controlling the spread of the disease. In this study, we provided a quantitative framework for investigating the likelihood of different evolutionary pathways that can give rise to VOCs of SARS-CoV-2. We found that VOCs are unlikely to be driven by a single adaptive change at the population level as this would require significantly high levels of overdispersion which is not supported by any existing epidemiological

study on SARS-CoV-2 tranmission [25]. We also showed that if multiple VOC mutations combine additively for advantage, they can only emerge on the background of a chronic infection, otherwise individual VOC mutations would reach high frequencies from the early stages of the pandemic and, therefore, would have been picked up from genomic surveillance data. If individual VOC mutations were acquired during chronic infections and had a strong advantage relative to the then-dominant genotypes, they may have still been leaked to the population at large before the emergence of VOCs. Therefore, we showed that only additive mutations with relatively small fractional contribution to VOC fitness may yield evolutionary dynamics that resembles the clustered emergence of SARS-CoV-2 VOCs in late 2020. On the other hand, we showed that cryptic circulation of a mutant lineage for sustained periods of time before producing VOCs is possible via a fitness plateau-crossing landscape. While at the between-host level such a landscape may not yield more than 2 mutations in excess of the background population over a period of 7-12 months under biologically relevant parameter combinations, many more mutations can be accumulated during a chronic infection, for example such as those found in certain immunocompromised individuals, without ever being leaked to the rest of the population. We found that the pattern of the timing of VOC emergence via the fitness plateau-crossing landscape under both the within- and between-host pathways are aligned with the timing of the clustered emergence of Alpha, Beta, and Gamma variants in late 2020.

5.4.1 Cryptic transmission of VOCs in humans

Another possibility for why the VOCs were not detected until mid to late 2020 is that they may have been circulating cryptically in areas of the world with poor genomic surveillance before becoming globally dominant. While variants of SARS-CoV-2

with multiple spike mutations have been detected through travel surveillance from passengers travelling from areas with little to no genomic surveillance [27, 28], if they were highly transmissible variants and had a potential to become a VOC, given the interconnectivity of the human interactions, it should not take very long before they become globally dominant. Therefore, as we showed in our analysis of Landscape 1, this scenario of VOC emergence seems to be only possible under significant levels of overdispersion such that it prohibits the selectively beneficial mutations from immediately taking off globally relative to other variants.

5.4.2 Possibility of reverse zoonosis

A somewhat similar idea to the cryptic transmission of variants in human populations is the possibility of a lineage (or multiple lineages) of SARS-CoV-2 jumping from humans to other mammals such as white-tailed deer, mink, hamster, and mouse where they circulate and evolve without being detected for a relatively long period before they jump back to the human population [29-32]. In particular, some recent studies have reported the detection of multiple spillovers of SARS-CoV-2 from humans and onward transmission in deer population with highly divergent genomes being detected in deer population with potential deer-to-human transmission [33, 34]. However, the genomic composition of these divergent genomes in deer population are different from the VOCs with a much lower ratio of non-synonymous to synonymous changes which suggests they may be following a completely different evolutionary path. Nevertheless, these studies indicate that it is possible for a highly divergent set of genomes to evolve in another species with a potential for deer-to-human transmission without ever being detected. Mink and hamster sequences offer some of the more compelling examples of transmission from humans to a non-human species and back, supported by phylogenetic evidence

[30, 31]. None of the currently identified sequences from animals appear as sister taxa to any of the circulating VOCs. While we cannot rule out evolution in an animal reservoir, one might expect the contribution of animals to human transmission chains to be dwarfed by the amount of human-to-human transmission currently happening.

5.4.3 Role of recombination

Recombination can bring together mutations from different backgrounds, potentially expediting the rate of adaptation by creating viable and more pathogenic hybrid new variants of a pathogen. Coronaviruses are also known to recombine with one another during mixed infections [35, 36]. While during the early stages of the pandemic SARS-CoV-2 sequences typically differed by only a handful of mutations from each other thereby making the effects of recombination indistinguishable from those of recurrent mutation [37], as more viral genetic diversity built-up in the population, the generation and transmission of interlineage recombinants of SARS-CoV-2 in humans were reported in multiple studies [38, 39]. Even though there is currently no definitive evidence for recombination being involved in the emergence of VOCs, including the Omicron variant of SARS-CoV-2 [40], we would expect the role of recombination to be more pronounced as the virus continues accumulate more genetic diversity by sustained circulation around the world.

5.4.4 Shifting landscape

We have assumed a static fitness landscape prior to the emergence of the first three VOCs; here we consider the plausibility of that assumption. During the first year of the pandemic, a novel virus was spreading in an immunologically naïve population [12]. As more individuals became infected and developed natural immunity, it is

possible that the fitness landscape for the virus shifted as selection for immune escape increased [41]. However, by the time the first three VOCs emerged in late 2020, the majority of the world's population were still susceptible to the disease and may not have even been exposed to it. Therefore, it is unlikely that the build-up of natural immunity alone was the reason behind their increased selective advantage. In contrast, the global dominance of Omicron in late 2021 was largely due to its immune escape properties relative to previous variants of SARS-CoV-2 [42] and, therefore, its emergence was likely the result of a changing viral fitness landscape.

5.4.5 Future VOCs

Emergence of new VOCs with increased transmissibility, immune evasion properties, and virulence poses a great challenge to managing SARS-CoV-2. Based on our findings, one of the major implications of chronic infections being the main source of generating VOCs is that the higher the global prevalence of the disease, the more likely it is for some individuals to develop chronic infections and possibly produce new VOCs. One of the main challenges with assessing the likelihood of VOC emergence during chronic infections would be to quantify the prevalence of immunosuppressed individuals within a population and determine which forms of immunosuppression are associated with chronic infections.

Several studies have now shown evidence of recurrent SARS-CoV-2 mutations in immunocompromised patients [15-17], with some suggesting the detection of a variant-like lineage which arose from a chronic infection that spilled over into a local population [43]. Another major implication of our work is that we can now quantitatively explain the possibility of such events and find the expected time that it takes for a possible new VOC to emerge from a within-host evolutionary pathway

that involves any number of mutations. We showed that a typical within-host plateau-crossing or additive mutation pathway involving 3-6 mutations requires a within-host fixation rate of μ_C~0.1-0.3 per generation which corresponds to a period of 50-300 days since the start of a chronic infection. The timing of such events aligns with the time window over which some of the major mutations involved in VOCs have been observed in patients with chronic infections [15-17]. This implies that if a VOC emerges from the within-host evolutionary pathway, it is more likely to reflect the genetic diversity of the virus population from several months ago. This can explain why the Omicron variant was not a descendent of Delta which was the most prevalent variant at the time of emergence of Omicron. We also note that, just like we saw three VOCs with shared mutations in late 2020, it is possible to see a clustered emergence of multiple new VOCs around the same time in the future. However, if it takes longer than a few months before a new VOC emerges, it is less likely to belong to the same fitness landscape. Some of the key remaining questions involve how much more of the fitness landscape will the virus be able to explore as more chronic cases accumulate and existing chronic cases last longer? For instance, if they crossed a 4-mutant fitness-plateau so far, how long until they will have explored possible 10-mutant combinations?

5.5 Methods

5.5.1 Between-host model of VOC emergence

Effective population size

We approximate the between-host evolution of SARS-CoV-2 as a haploid population of size $N(t)$ which is equal to the number of daily infectious individuals with SARS-CoV-2 worldwide. Since the number of confirmed cases is often a significant underestimation of the true number of infections (e.g., see: [44, 45]), we use the number of daily confirmed deaths [46] to back-calculate the number of infectious individuals, $N(t)$, from the global median infection fatality rate (IFR) of COVID-19 [47]. We note this approach is still subject to several potential sources of bias including variation in IFR over time (e.g. due to various pharmaceutical interventions) and across different demographics [48]. Using confirmed COVID-19-related deaths may still underestimate the true number of deaths associated with the disease due to under-reporting of deaths particularly in areas of the world where there is limited testing from suspect cases [49]. Nevertheless, by allowing for a wide variation in global IFR (from 0.2% to 1.5%), we can capture most of the uncertainty in the number of infectious individuals worldwide. We also note that for the timespan of interest in our work (i.e., start of the pandemic until the emergence of the first three VOCs), the impact of pharmaceutical interventions such as vaccination on lowering the global IFR is likely to have been negligible given that vaccination campaigns mostly started in 2021. The confirmed global deaths started being reported from 2020-01-23. Assuming a 20-day delay from the onset of symptoms to death [50], we set 2020-01-03 as the first timepoint in the simulation.

Epidemic spread

Due to a high degree of individual-level variation in the transmission of SARS-CoV-2 (i.e., overdispersion) [25, 51], we use a Dirichlet-multinomial (instead of a multinomial) distribution to assign offspring in generation t+1 to parents from generation t. The variance in offspring number is therefore $\sigma^2=R_t(1+R_t/k)$ where $R_t=N(t+1)/N(t)$ and k is the dispersion parameter of the negative binomial that approximates an individual's offspring distribution. To adjust the variance of the Dirichlet-multinomial distribution to match with the dispersion in the number of secondary infections, we set it to be equal to the variance of the negative binomial, σ^2, for every individual in the population. Therefore, we have $\sigma^2=\frac{N(t+1)}{N(t)}\left(1-\frac{1}{N(t)}\right)\frac{N(t+1)+A}{1+A}$ and $A\approx kN(t+1)$ where $\vec{\alpha}=A\vec{p}$, $\vec{\alpha}$ is the parameter vector of a Dirichlet-multinomial distribution and $\vec{p}$ is the probability vector of a multinomial distribution with $N(t+1)$ number of trials.

Mutation rate

Assuming a constant generation time 5.2 days for all variants of SARS-CoV-2 over time [52], we use the phylogenetically estimated substitution rate per site per year [53] to calculate the mutation rate per site per generation time, μ. We also note that generation time may vary over time depending on the behavioural changes in the population and emergence of variants, that is why we allow for some variation in the mutation rate parameter in our model (0.87 - 2.0) x 10^{-5} based on phylogenetic estimates. We assume each site has two states: wild-type and mutant. Therefore, for a group of K sites, there are 2^K genotypes.

Advantage of mutants

The selective coefficient of a mutant individual depends on its number of mutations and the fitness landscape (see Figure 5.2A). For instance, in the case of an additive fitness landscape of size K=3, the fitness advantage of the single-, double-, and triple-mutants are $s/3$, $2s/3$, and s relative to the wild-type population, respectively. During one generation, the frequency f_i of individuals with genotype i and selective advantage s_i relative to the wild-type increases by a factor $(1 + s_i)$, along with further adjustments to its frequency due to mutations from/to other genotypes. Upon normalisation ($\sum_i f_i = 1$), these frequencies are used for the Dirichlet-multinomial sampling step such that the parameter vector only depends on the population size as $\vec{\alpha} = A\vec{p}$. After the sampling step, the numbers of cases are converted to frequencies for sampling in the next generation.

Inferring the selective advantage of VOCs

Finally, the selective advantage s of the VOCs is determined by fitting an exponential function, $f(t)$, of the form, $f(t)=ae^{st}$, to the proportion of Alpha, Beta, and Gamma variants sampled in the country where they were first detected (i.e., UK, South Africa, and Brazil) using the NonlinearModelFit function in Mathematica 11.0. We find that the selective advantage s for Alpha, Beta, and Gamma are 0.37 (95% confidence interval: 0.33 - 0.41), 0.74 (95% confidence interval: 0.65 - 0.83), and 0.84 (95% confidence interval: 0.58 - 1.08), respectively (see Figure S5). The confidence interval is obtained by multiplying the standard error by the value of Student's t for the given confidence level and degrees of freedom. Given the uncertainty in our estimates due to noise in the observations, potential sampling bias, and spatio-temporal heterogeneities, we make the assumption that the value of s is roughly the same for the different VOCs and use the same estimate for all three trajectories (Table 6.1).

5.5.2 Within-host model of VOC emergence

Each VOC mutation is fixed within the host at rate μ_C such that the fixation time is an exponentially distributed number with mean $1/\mu_C$. Each mutation may then spread to the rest of the population with a probability that is proportional to its fitness as determined by the Dirichlet-multinomial sampling. At any time during the pandemic, a chronic infection can be seeded by other infectious individuals within the population, $N(t)$, with a probability P_f. Therefore, at every generation, the number of chronic infections is given by a binomial distribution with success probability P_f. Once a chronic infection emerges, it remains in the population for the remainder of the simulation period.

5.5.3 Simulation setup

For both within-host and between-host models of VOC emergence, we run each evolutionary scenario for a given combination of model parameters 1,000 times. We then measure total number of established VOC lineages, M, the time that it takes for the establishment of the first VOC, T_0, and the time between the establishment of the ith and $(i+1)^{th}$ VOC, $T_{i:(i+1)}$, for the first 6 established VOC lineages in each scenario. An *established* VOC lineage is defined as a lucky lineage with selective advantage s that survives drift upon reaching a size $1/s$. Similarly, the establishment time of a VOC lineage is defined as the time that it takes for that lineage to reach size $1/s$. Each run stops once the frequency of the VOC population reaches 75%.

We assume the population follows a generalised Wright-Fisher model with discrete, nonoverlapping generation and variable size (see section 5.5.1 for more information about the sampling scheme). Each generation is defined as the duration of an acute infection, 5.2 days (see Table 5.1). The population size, $N(t)$, is defined

as the number of daily infectious individuals worldwide and is determined by dividing the total number of daily confirmed deaths worldwide by the median global infection IFR (see section 5.5.1 for more information of the effective population size). We assume all infected individuals are immune to reinfection over the course of the simulation and that there is no population structure (i.e., panmictic population) and sampling noise (i.e., perfect detection of the state of the system). This is because at the early stages of the pandemic until the end of 2020, the virus was still spreading in a largely immunologically naïve population. Furthermore, while spatial structure is clearly visible in the spatially restricted nature of initial spread of the VOCs, it should not have a large impact on viral dynamics until that lineage becomes locally common (N.B., the specific mutations differentiating the VOCs were all locally rare prior to their emergence). We assume perfect detection of variant frequencies because we would like to fully reconstruct all the possible evolutionary trajectories that can lead to the emergence of the first three VOCs.

Software implementation

All software code and analysis scripts to reproduce the figures are available at https://github.com/weissmanlab/Valley_Crossing.

5.6 References

1. L. Z. Zhang, et al., *SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity.* Nature Communications, 2020. **11**: p. 6013.
2. Hodcroft, E.B., *CoVariants: SARS-CoV-2 Mutations and Variants of Interest.* 2021, URL: https://covariants.org/.
3. T. P. Peacock, et al., *SARS-CoV-2 one year on: evidence for ongoing viral adaptation.* Journal of General Virology, 2021. **102**(4): p. 001584.
4. A. Rambaut, et al., *Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations.* Virological, 2020. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.
5. H. Tegally, et al., *Detection of a SARS-CoV-2 variant of concern in South Africa.* Nature, 2021. **592**(7854): p. 438-443.
6. N. R. Faria, et al., *Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil.* Science, 2021. **372**(6544): p. 815-821.
7. K. M. Tao, et al., *The biological and clinical significance of emerging SARS-CoV-2 variants.* Nature Reviews Genetics, 2021. **22**(12): p. 757-773.
8. D. Planas, et al., *Considerable escape of SARS-CoV-2 Omicron to antibody neutralization.* Nature, 2022. **602**: p. 671-675
9. T. N. Starr, et al., *Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding.* Cell, 2020. **182**(5): p. 1295-1310.
10. N. G. Davies, et al., *Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England.* Science, 2021. **372**: p. 149-154.
11. S. Cele, et al., *Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma.* Nature, 2021. **593**: p. 142-146.
12. D. P. Martin, et al., *The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages.* Cell, 2021. **184**(20): p. 5189-5200.e7.
13. J. Zahradnik, et al., *SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution.* Nature Microbiology, 2021. **6**(9): p. 1188-1198.
14. G. Nelson, et al., *Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant.* bioRxiv, 2021. https://doi.org/10.1101/2021.01.13.426558.
15. B. Choi, M. Cernadas, and J.Z. Li, *Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host.* New England Journal of Medicine, 2020. **383**(23): p. 2291-2293.
16. S. A. Kemp, et al., *SARS-CoV-2 evolution during treatment of chronic infection.* Nature, 2021. **592**: p. 277-282.
17. F. Karim, et al., *Persistent SARS-CoV-2 infection and intra-host evolution in association with advanced HIV infection.* medRxiv, 2021. https://doi.org/10.1101/2021.06.03.21258228.
18. S. P. Otto, et al., *The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic.* Current Biology, 2021. **31**(14): p. R918-R929.

19. T. Karasov, P.W. Messer, and D.A. Petrov, *Evidence that Adaptation in Drosophila Is Not Limited by Mutation at Single Sites.* Plos Genetics, 2010. **6**(6): p. e1000924.
20. M. A. Martin and K. Koelle, *Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2".* Science Translational Medicine, 2021. **13**(617): p. eabe2555.
21. K. A. Lythgoe, et al., *SARS-CoV-2 within-host diversity and transmission.* Science, 2021. **372**(6539): p. eabg0821.
22. K. M. Braun, et al., *Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks.* Plos Pathogens, 2021. **17**(8): p. e1009849.
23. J. Hadfield, et al., *Nextstrain: real-time tracking of pathogen evolution.* Bioinformatics, 2018. **34**(23): p. 4121-4123.
24. D. B. Weissman, et al., *The rate at which asexual populations cross fitness valleys.* Theoretical Population Biology, 2009. **75**(4): p. 286-300.
25. A. Endo, et al., *Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China.* Wellcome Open Research, 2020. **5**: p. 67.
26. R. T. Eguia, et al., *A human coronavirus evolves antigenically to escape antibody immunity.* Plos Pathogens, 2021. **17**(4): p. e1009453.
27. T. de Oliveira, et al., *A novel variant of interest of SARS-CoV-2 with multiple spike mutations detected through travel surveillance in Africa.* medRxiv, 2021. https://doi.org/10.1101/2021.03.30.21254323.
28. P. Colson, et al., *Emergence in Southern France of a new SARS-CoV-2 variant of probably Cameroonian origin harbouring both substitutions N501Y and E484K in the spike protein.* medRxiv, 2021. https://doi.org/10.1101/2021.12.24.21268174.
29. J. C. Chandler, et al., *SARS-CoV-2 exposure in wild white-tailed deer (Odocoileus virginianus).* Proceedings of the National Academy of Sciences of the United States of America, 2021. **118**(47): p. e2114828118.
30. B. B. O. Munnink, et al., *Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans.* Science, 2021. **371**(6525): p. 172-177.
31. H. L. Yen, et al., *Transmission of SARS-CoV-2 delta variant (AY.127) from pet hamsters to humans, leading to onward human-to-human transmission: a case study.* Lancet, 2022. **399**(10329): p. 1070-1078.
32. C. S. Wei, et al., *Evidence for a mouse origin of the SARS-CoV-2 Omicron variant.* Journal of Genetics and Genomics, 2021. **48**(12): p. 1111-1121.
33. S. V. Kuchipudi, et al., *Multiple spillovers from humans and onward transmission of SARS-CoV-2 in white-tailed deer.* Proceedings of the National Academy of Sciences of the United States of America, 2022. **119**(6): p. e2121644119.
34. B. Pickering, et al., *Highly divergent white-tailed deer SARS-CoV-2 with potential deer-to-human transmission.* bioRxiv, 2022. https://doi.org/10.1101/2022.02.22.481551.
35. R. L. Graham and R.S. Baric, *Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission.* Journal of Virology, 2010. **84**(7): p. 3134-3146.

36. S. Lytras, et al., *Exploring the Natural Origins of SARS-CoV-2 in the Light of Recombination.* Genome Biology and Evolution, 2022. **14**(2): p. evac018.
37. G. McVean, P. Awadalla, and P. Fearnhead, *A coalescent-based method for detecting and estimating recombination from gene sequences.* Genetics, 2002. **160**(3): p. 1231-1241.
38. B. Jackson, et al., *Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic.* Cell, 2021. **184**(20): p. 5179-5188.
39. B. Gutierrez, et al., *Emergence and widespread circulation of a recombinant SARS-CoV-2 lineage in North America.* medRxiv, 2021. https://doi.org/10.1101/2021.11.19.21266601.
40. R. Viana, et al., *Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa.* Nature, 2022. **603**(7902): p. 679-682.
41. R. D. H. Barrett and D. Schluter, *Adaptation from standing genetic variation.* Trends in Ecology & Evolution, 2008. **23**(1): p. 38-44.
42. D. Planas, et al., *Considerable escape of SARS-CoV-2 Omicron to antibody neutralization.* Nature, 2022. **602**(7898): p. 671-675.
43. S.A. Wilkinson, et al., *Recurrent SARS-CoV-2 Mutations in Immunodeficient Patients.* medRxiv, 2021. https://doi.org/10.1101/2022.03.02.22271697.
44. A. M. Caliendo, et al., *Better Tests, Better Care: Improved Diagnostics for Infectious Diseases.* Clinical Infectious Diseases, 2013. **57**: p. S139-S170.
45. M. Ghafari, et al., *Lessons for preparedness and reasons for concern from the early COVID-19 epidemic in Iran.* Epidemics, 2021. **36**: p. 100472.
46. *World Health Organization COVID-19 Dashboard. .* 2021, URL: https://covid19.who.int.
47. A.T. Levin, et al., *Assessing the Burden of COVID-19 in Developing Countries: Systematic Review, Meta-Analysis, and Public Policy Implications.* medRxiv, 2021. https://doi.org/10.1101/2021.09.29.21264325.
48. M. O'Driscoll, et al., *Age-specific mortality and immunity patterns of SARS-CoV-2.* Nature, 2020. **590**: p. 140–145.
49. A. Karlinsky and D. Kobak, *Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset.* Elife, 2021. **10**: p. e69336.
50. J. T. Wu, et al., *Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China.* Nature Medicine, 2020. **26**(4): p. 506-510
51. J. O. Lloyd-Smith, et al., *Superspreading and the effect of individual variation on disease emergence.* Nature, 2005. **438**(7066): p. 355-359.
52. L. Ferretti, et al., *Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing.* Science, 2020. **368**(6491): p. 619-625.
53. M. Ghafari, et al., *Purifying Selection Determines the Short-Term Time Dependency of Evolutionary Rates in SARS-CoV-2 and pH1N1 Influenza.* Molecular Biology and Evolution, 2022. **39**(2): p. msac009.
54. K.A. Lythgoe, et al., *Lineage replacement and evolution captured by the United Kingdom Covid Infection Survey.* medRxiv, 2022. https://doi.org/10.1101/2022.01.05.21268323.

Statement of Authorship for joint/multi-authored papers for PGR thesis.

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	Investigating the evolutionary origins of the first three SARS-CoV-2 variants of concern
Publication Status	Unpublished and unsubmitted work written in a manuscript style
Publication Details	Ghafari M, Liu Q, Dhillon A, Katzourakis A, and Daniel B Weissman. Investigating the evolutionary origins of the first three SARS-CoV-2 variants of concern. (Unpublished)

Student confirmation

Student Name	Mahan Ghafari
Contribution to the paper	MG wrote the chapter and conducted all the analysis under the supervision from AK and DBW. All authors provided editorial assistance.
Signature 	Date 11/05/2022

Supervisor confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title	Professors Aris Katzourakis and Oliver G Pybus	
Supervisor comments		
Signature 		Date 11/05/2022

Chapter 6: Lessons for preparedness and reasons for concern from the early COVID-19 epidemic in Iran

6.1 Abstract

Many countries with an early outbreak of SARS-CoV-2 struggled to gauge the size and start date of the epidemic mainly due to limited testing capacities and a large proportion of undetected asymptomatic and mild infections. Iran was among the first countries with a major outbreak outside China. We constructed a globally representative sample of 802 genomes, including 46 samples from patients inside or with a travel history to Iran. We then performed a phylogenetic analysis to identify clades related to samples from Iran and estimated the start of the epidemic and early doubling times in cases. We leveraged air travel data from 36 exported cases of COVID-19 to estimate the point-prevalence and the basic reproductive number across the country. We also analysed the province-level all-cause mortality data during winter and spring 2020 to estimate under-reporting of COVID-19-related deaths. Finally, we use this information in an SEIR model to reconstruct the early outbreak dynamics and assess the effectiveness of intervention measures in Iran. By identifying the most basal clade that contained genomes from Iran, our phylogenetic analysis showed that the age of the root is placed on 2019-12-21 (95% HPD: 2019-09-07 – 2020-02-14). This date coincides with our estimated epidemic start date on 2019-12-25 (95%CI: 2019-12-11 – 2020-02-24) based air travel data from exported cases with an early doubling time of 4.0 (95%CI: 1.4 – 6.7) days in cases. Our analysis of all-cause mortality showed 21.9 (95% CI: 16.7 – 27.2) thousand excess deaths by the end of summer. Our model forecasted the second epidemic peak and suggested that by 2020-08-31 a total of 15.0 (95%CI: 4.9 – 25.0)

million individuals recovered from the disease across the country. These findings have profound implications for assessing the stage of the epidemic in Iran despite significant levels of under-reporting. Moreover, the results shed light on the dynamics of SARS-CoV-2 transmissions in Iran and central Asia. They also suggest that in the absence of border screening, there is a high risk of introduction from travellers from areas with active outbreaks. Finally, they show both that well-informed epidemic models are able to forecast episodes of resurgence following a relaxation of interventions, and that NPIs are key to controlling ongoing epidemics.

6.2 Introduction

With significant levels of mortality and morbidity, the ongoing pandemic of Coronavirus Disease 2019 (COVID-19) continues to have a major impact on many regions around the world [1]. With many nations struggling to contain the spread of its causative pathogen SARS-CoV-2, it is paramount to re-evaluate our knowledge of the epidemic in search for useful lessons. Different countries represent different realisations of SARS-CoV-2 epidemics in different contexts; therefore, accurate country-specific analyses can provide key insights into the dynamics of the epidemic and the effects of interventions.

The pandemic was first detected in the city of Wuhan, in the Chinese province of Hubei in December 2019, with isolated cases recorded in European and East Asian countries until the second half of February, when large outbreaks were detected at the same time in Lombardy, Italy, and in Qom, Iran. While the early phases of the pandemic in China and Italy have been widely studied, the early Iranian epidemic has not attracted the same attention, possibly due to the lower amount of data available. As a consequence, very little is known about the origins

of the outbreak and pattern of spread across the country. Some early reports suggested that the first two confirmed cases in Qom, one of the thirty-one provinces of Iran, could have been infected by a merchant who had reportedly travelled from China [2]. However, with COVID-19, identifying the so-called ‘patient zero’ is problematic due to high rates of asymptomatic and pauci-symptomatic infections . As an example, the COG-UK consortium has shown that there have likely been more than one thousand unique introductions of the virus into the UK, possibly many of which are from individuals who are asymptomatic or show mild symptoms [4]. Therefore, the most likely route of virus spread from China to Iran, and indeed many other countries, was via asymptomatic, presymptomatic, or mildly symptomatic travellers. Infected individuals typically show no symptoms for about 5 days [5-7] or sometimes no symptoms at all [8-10], while silently spreading the virus [11]. Using aggregate data from the influenza surveillance networks has been proposed as one of the ways to detect early arrival of SARS-CoV-2 into a country [12]. However, a major flu outbreak in Iran which reportedly stretched from November 2019 to early January 2020 may have conflicted with the early diagnosis of COVID-19 patients across the country, thereby hindering control and diagnosis efforts [13, 14]. By the time the first cases were reported, the outbreak was large enough to overrun the testing capacity of the Iranian health system, similar to a pattern seen in other countries that were caught off-guard by the rapid spread of the virus outside China.

The Iranian COVID-19 epidemic is an interesting case study because Iran has not only been one of the first countries to face an outbreak, but also acts as a major source of the spread in Central Asia and the Middle East during the early stages of the pandemic. While struggling with under-reporting issues, the Iranian health system has been able to collect informative clinical and epidemiological data, which

can be combined with external sources to improve our understanding of the likely epidemic dynamics in other Central Asian countries. It has also been the first country to relax public health measures by reopening businesses and public facilities. This untimely relaxation of NPIs led to a second COVID-19 incidence, showing, for the first time, that the so-called “second wave” was not just a theoretical prediction.

In this study, we provide a detailed analysis of the COVID-19 outbreak in Iran. By gathering genetic and air travel data from passengers with a link to Iran, we estimate the start date of the epidemic and its prevalence across the country. Coupling this with the estimated number of excess deaths and other key clinical and epidemiological information, such as the age-stratified infection fatality ratios and delays from onset of infection to death, enables us to reconstruct the full transmission dynamics of COVID-19 despite significant levels of under-reporting in prevalence and deaths.

First, we use epidemiological and genetic data from air travellers with a travel history to Iran along with the first whole genomes from inside the country to determine the start date and early growth rate of the epidemic. Next, we provide a province-level analysis on the pattern of spread during winter and spring, highlighting the degree of under-reporting in both point prevalence and deaths. We also assess the risk of importation of infected individuals from China into the country. Finally, we evaluate the impact of non-pharmaceutical interventions (NPIs) on the growth rate of the epidemic and use an SEIR model to reconstruct the early transmission dynamics in the country.

6.3 Results

6.3.1 The situation in Iran

First reported cases of COVID-19 in Iran were of two infected individuals in the province of Qom on 19 February. In only a few days, reported cases grew rapidly in almost a dozen provinces (Figure 6.1A) which raised concerns that the virus had already been widespread by the time the first two cases were detected in Qom. By 28 February, WHO reported nearly 100 confirmed exported cases of COVID-19 from Iran to several countries [38] while, at the time, the MoHME reported a total of 388 confirmed cases across the country [39]. Here, we present a brief summary of the key epidemiological data that was collected during the early phase of the epidemic in Iran.

During the early days of the epidemic, Pasteur Institute of Iran (PI) was the primary source of testing suspect COVID-19 cases and had enough resources to perform only a few hundred RT-PCR tests per day. By mid-March, WHO and the Chinese Government delivered several shipments of emergency medical supplies along with more than 200,000 additional test kits which helped with the diagnosis of more hospitalised patients [40]. Nevertheless, some of the preliminary analyses showed that Iran was only reporting less than 10% of its symptomatic cases during its first peak in late March [41]. The extent of underreporting of COVID-19 cases and deaths also led to speculations on the possibility of data manipulations, which are however not substantiated by an analysis based on Benford's law (see supplementary materials). From early April, as the country ramped up its testing capacity, MoHME started to report outpatients and carried out limited levels of contact tracing in several provinces (see Figure 6.1B). However, despite the growing testing capacity, since the government announced plans for re-opening

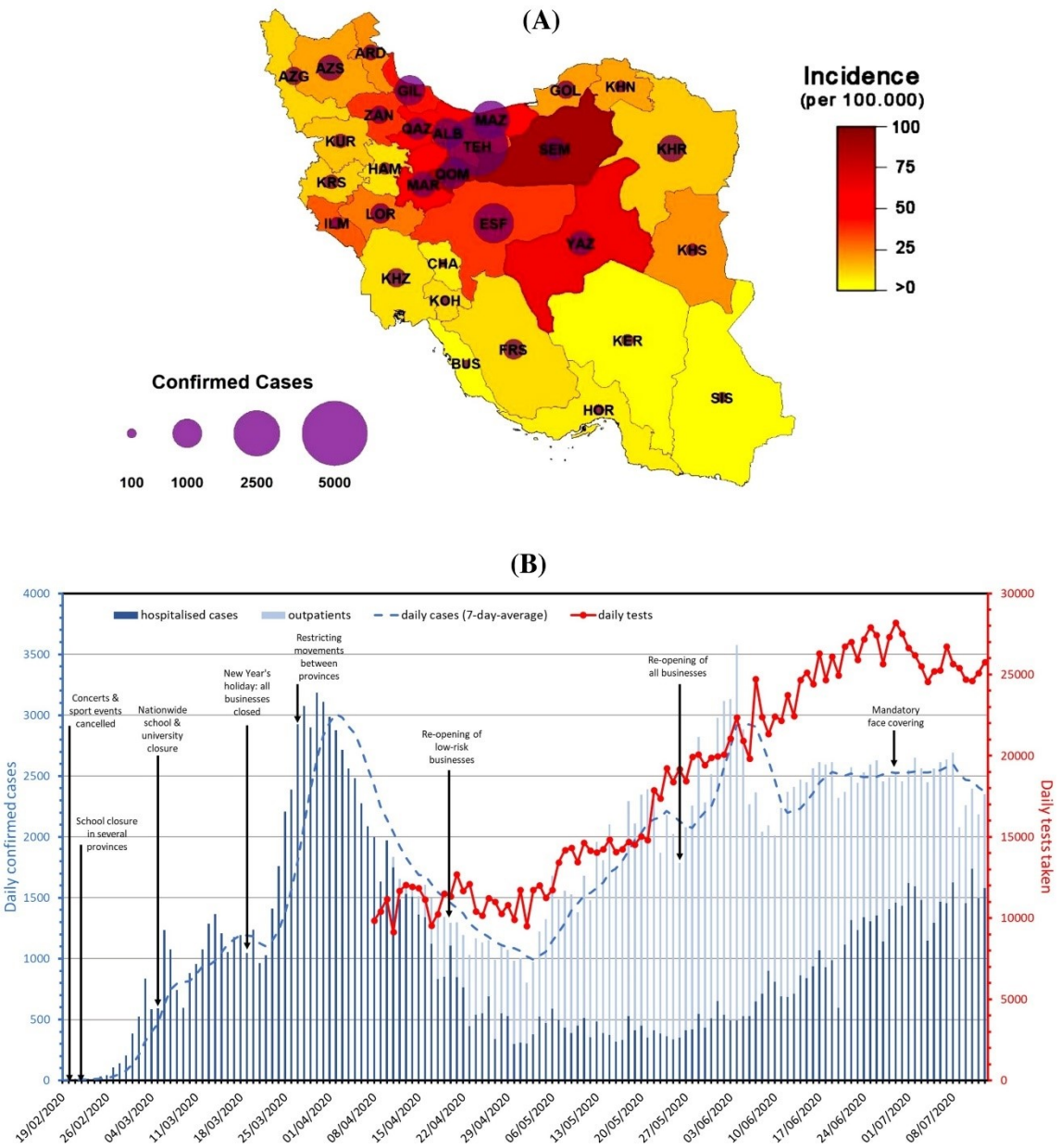


Figure 6.1: **Epidemiological and demographic characteristics of confirmed COVID-19 cases in Iran through time.** (**A**) Total incidence of COVID-19 by province on 22 Mar 2020 (last day that MoHME released province-level daily cases). (**B**) Number of confirmed hospitalised patients (dark blue) and outpatients (light blue). Red lines show the number of daily tests taken over time. Vertical arrows indicate some of the major interventions or changes of policies.

high-risk activities in late May, the second peak in new cases emerged and the number of outpatients started to shrink over time. This raised concerns that, like the first peak when hospitals in many provinces were at near maximum capacity, the degree of under-reporting in both prevalence and deaths have risen substantially higher again. Another point of concern is that as the delay from the onset of symptoms to testing increases, the probability that an infected individual tests positive likely decrease. This is because by the time tests are taken from hospitalised cases of COVID-19, there could be typically 3 to 7 days past since the onset of their symptoms, and many could be clearing the infection by then. As a result, a significant portion of their tests (up to 50%) may come out as negative [5, 42].

6.3.2 Origins of the epidemic in Iran: a phylogenetic analysis

Since the virus spread undetected during the initial phase of the Iranian epidemic, no epidemiological information is available during this period. Molecular epidemiology provides a powerful tool to reconstruct the early epidemic trajectory a posteriori, especially when direct epidemiological data are absent, incomplete or biased. The report on 2020-03-16 by NextStrain and findings from returned travellers to Australia and New Zealand showed that sequences from cases with a travel history to Iran correspond to a distinct clade of SARS-CoV-2 [43, 44]. We first construct a maximum likelihood phylogenetic tree of all sequences that were linked to Iran (see Figure 6.2A) to identify the most basal clade (i.e., earliest introduction) related to genomes from Iran. After identifying the basal clade, we use these genomic samples to determine the TMRCA of the clade and characterise the initial epidemic growth rate in Iran. Consistent with other studies [20], our inferred substitution rate is 1.06×10^{-3} (95% Highest Posterior Density (HPD):

$9.9\times10^{-5} - 2.14\times10^{-3}$) per site per year and the exponential growth rate of 41 in units of years, corresponding to a doubling time of around 6.14 (95% HPD: 3.34 - 38.33) days. The age of the root is placed on 2019-12-21 (95% HPD: 2019-09-07 – 2020-02-14), nearly two months before when the first cases in Qom were reported.

6.3.3 Estimating the under-reporting of incidence using air-travel data

High levels of under-reporting have been noticed in several countries during the early phase of the COVID-19 pandemic. Such under-reporting creates biases that hinders direct estimates of the actual incidence of the disease and are likely to be present in MoHME's report as well. A more reliable estimate of the size of the Iranian epidemic can be obtained from the number of exported cases detected abroad [45]. In late February, a total of 8 internationally exported cases of COVID-19 with direct flights from Iran were detected in Lebanon, UAE, Oman, and Kuwait. Similarly, in early March, an additional 28 cases were identified in China (see Table 6.1). By finding the approximate number of passengers per week travelling from Iran to each of these countries, we use a binomial likelihood estimation (see Methods section) to find the nationwide incidence on the dates when the exported case was reported (see Figure 6.2B). We then use an exponential regression fit to the estimated nationwide incidence and find that the outbreak size by 2020-02-25 and 2020-03-06 to be 45,981 (95%CI: 1 – 96,840) and 258,661 (95%CI: 171,443 – 345,879), respectively, which is aligned with estimated outbreak size in late February from other studies [46, 47] and is more than 100 times higher than confirmed cases that MoHME reported at the time – see the sensitivity analysis on the parameters of our model in Table S1. These estimates also suggest a doubling time of 4.0 (95%CI: 1.4 – 6.7) days, which is close to the growth rate in the initial phase of European COVID-19 epidemics [48]. We then use this range of doubling times to extrapolate

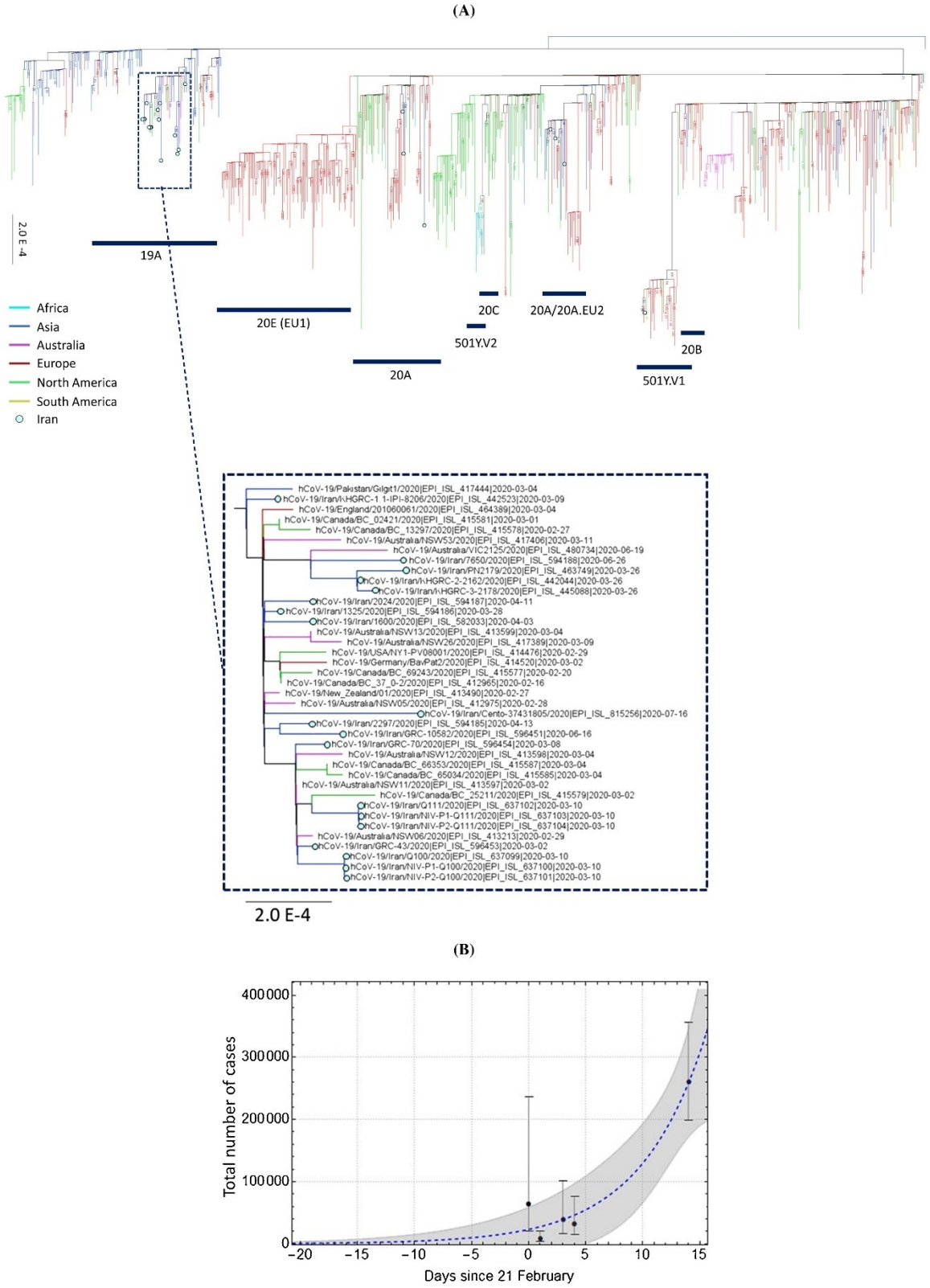


Figure 6.2: **Phylogenetic and epidemiological analysis on cases with a travel link to Iran.** (**A**) Maximum likelihood phylogenetic tree of a globally representative sample of 802 genomes, including 26 genomes from Iran (highlighted in light blue) - the tree is rooted with respect to earliest sample from Wuhan, Wuhan/IPBCAMS-WH-01/2019. The shown samples from Iran belong to clades 19A, 20A, and 501Y.V1. The inset shows the clade of returned travellers which includes 20 genomes from passengers with a travel link to Iran in addition to 19 samples from inside the country (also see Figure S1). The scale bar

shows the mean number of substitutions per site. (**B**) Estimating the start date of the epidemic and its initial growth trajectory using an exponential regression fit based on the likelihood analysis on exported cases to UAE, Lebanon, Kuwait, Oman, and China (see Table 6.1). The blue dashed line and gray area show the line of best fit and its corresponding mean prediction bands, respectively.

the number of cases back in time and find 2019-12-25 (95%CI: 2019-12-11 – 2020-02-24) to be the approximate start date of the epidemic, which is also well-aligned with the estimated root age of our phylogenetic analysis.

Table 6.1: **Estimated number of infectious individuals based on air travel data.** This includes a list of all countries with confirmed cases of COVID-19 travelling from Iran via direct flights (see Data S3 for more information).

Date	Country	Passengers/week	Reported cases	Estimated active cases (95% CI)
2020-02-21	Lebanon[1]	800	1	64,000 (19,700 – 236,900)
2020-02-22	UAE[2]	13,430	2	7,700 (3,200 – 19,900)
2020-02-24	Oman[3]	2,660	2	39,000 (16,100 – 101,000)
2020-02-25	Kuwait[4]	4,025	3	32,200 (14,400 – 76,000)
2020-03-06	China[5]	6,700	28	259,500 (199,200 – 356,200)

*[1]Source, [2]Source, [3]Source, [4]Source,[5]Source

6.3.4 Estimating the under-reporting of cumulative deaths using excess mortality data

Given the high levels of under-reporting, excess mortality data represent a more reliable source to estimate for COVID-19-related deaths. By analysing the seasonal reports from NOCR on province-level all-cause deaths, we find that five provinces in winter and twenty-eight provinces in spring 2020 had significantly higher recorded deaths compared to previous years (see Figure 6.3A,B). During winter, the five provinces in central and northern Iran (Qom, Gilan, Golestan, Qazvin, and Mazandaran) showed a 24% rise in mortality with 3,558 (95% CI: 3,171 – 3,949) deaths in excess of previous years. In spring, our estimates show a 22% increase with a total of 18,359 (95% CI: 13,506 – 23,212) excess deaths in twenty-eight provinces compared to previous years. Qom and Gilan, two of the hardest-hit provinces in winter, show a lower percentage of excess deaths in spring possibly

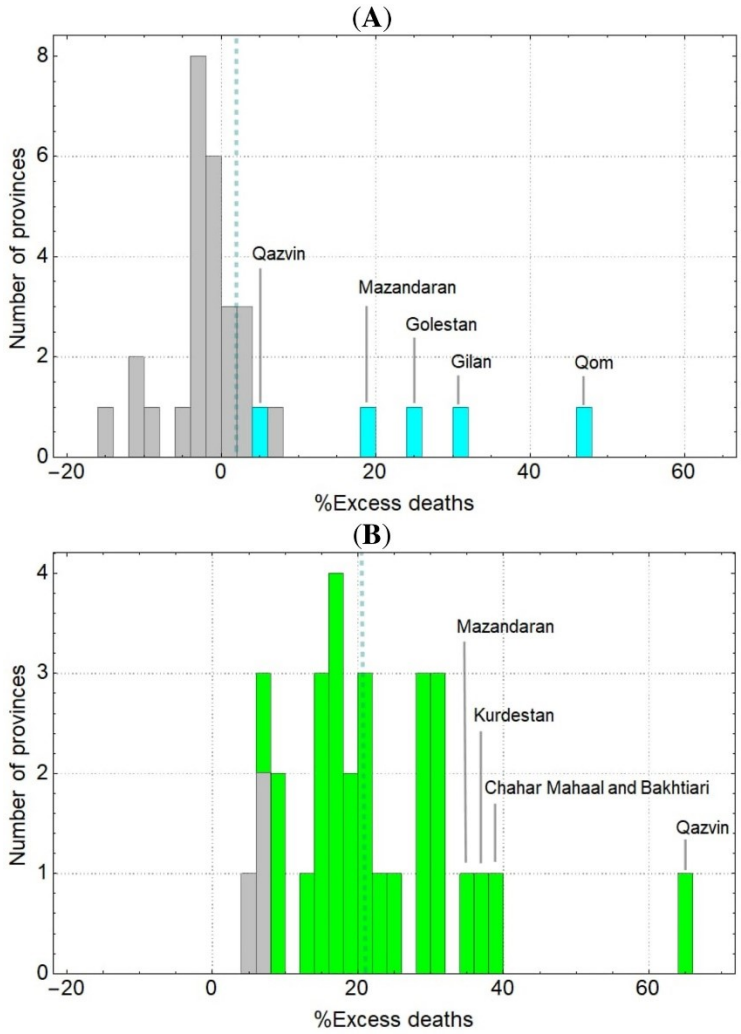


Figure 6.3: **Percentage of excess deaths (with respect to the five-year average) during winter and spring 2020 in 31 provinces of Iran.** (**A**) and (**B**) show the percentage of excess deaths during winter (from 2019-12-22 to2020-03-19) and spring (from 2020-03-20to2020-06-20), respectively. Four provinces (highlighted in cyan) during winter and twenty-six provinces (highlighted in green) during spring show significant levels of excess mortality. Gray bars represent provinces with no significant deviations from their five-year average (based on 95% confidence interval). The vertical dashed lines show the mean percentage of excess mortality in each season. During winter, the mean excess mortality is only 2% while in spring this grows to 21% across all provinces.

indicating that the largest part of their outbreak occurred back in winter while in many other provinces such as Mazandaran, Khuzestan, and Qazvin the excess deaths continued to increase during spring. This is also corroborated by numerous reports from various provinces across the country and is aligned with the fact that the first nationwide epidemic peak occurred in late-March.

6.3.5 Early-stage importation risk from China

Given the debate on travel restrictions, it is worth assessing what was the actual early-stage risk of COVID-19 importation to Iran. Our analysis shows that the expected number of imported cases into the country from 2020-01-22 to 2020-02-14 (the upper bound of our TMRCA estimate for the start date of the epidemic in the country) was 0.17 (95%CI: 0.04 - 0.28) which suggests that the Iranian epidemic was most likely seeded by one or a few imported cases flying from China. The fact that the epidemic in Iran started despite the low number of importations suggests a strong overdispersion in transmission and that the infected individuals likely belonged to a high-risk subgroup of the population, such as businesspeople, politicians, or other individuals who frequently travel to the country and tend to meet a large number of contacts [49].

6.3.6 Evaluating effectiveness of intervention measures and the dynamics of the outbreak

We first evaluate the effectiveness of key NPIs (see Methods section) implemented on set dates in Iran and their impact on reducing the effective reproduction number, Rt (Table 6.2 and Figure 6.4A). Our analyses indicate that while intervention measures, particularly the nationwide lockdown during the New Year's holiday month, were effective in curbing the spread (Rt < 1), the reopening of high-risk

businesses in May led to the emergence of a new peak of infection across the country (Rt > 1). They also suggest that mandatory face covering in public spaces coupled with social distancing measures and other health and safety protocols can be effective enough to stop the exponential growth pattern (Rt <1) and reduce the transmission rates by late-July/early-August conditioned on having no other major change in NPIs or social behaviour (e.g., travelling for holidays or mass gatherings) during this time period.

Table 6.2: **Timeline of intervention policy announcements and their impact on effective reproduction number.**

Intervention measures	Description	Date effective†	Relative %reduction in Rt (range)	Constrain on relative reduction in Rt [Ref.]
School & university closure with ban on public events	Implemented in multiple provinces: Tehran, Mazandaran, Qom, Qazvin, Golestan, Gilan, and several others.	2020-02-22	35% (25-40)	(0% - 40%) [34]
Nationwide lockdown*	Closure of all schools, universities, sport centres, & holy sites.	2020-03-05	80% (70-85)	(70% - 90%) [34]
Re-opening of low-risk businesses	Partial re-opening of certain businesses such as banks and governmental agencies with limited staff.	2020-04-18	65% (60-70)	(40% - 70%) [35, 36]
Re-opening of all businesses	Re-opening of all businesses including shops and restaurants while observing social distancing rules	2020-05-26	55% (50-65)	(40% - 70%) [35, 36]
Mandatory face-covering	Mandatory in all public spaces including public transports, restaurants, & shops	2020-07-05	70% (60-85)‡	(60% - 85%)‡ [50]

*This intervention is followed by the closure of all business across the country due to the Iranian New Year's holiday period which started in the week leading to 2020-03-19 and has the highest impact on reducing Rt.
†Older interventions end as soon as a new one takes into effect.
‡We assume that the relative reduction in Rt because of this NPI is aligned with the estimated odds ratio of 0.6 (95%CI: 0.5 - 0.8) in reducing risk of transmission for mask wearers compared to a control group from other studies [50].

We show that if this intervention is 70% (95%CI: 60 - 85) effective, i.e. at least 60% effective, it can reduce Rt from 1.4 (1.2 - 1.5) to 0.9 (0.7 - 1.0) which is enough to

curb the spread. This corresponds to a relative reduction of 0.64 (95%CI: 0.58 - 0.80) in Rt which is in agreement with other studies reporting the reduced risk of transmission for mask wearers compared to a control group [50].

Our model successfully recreates the nationwide trend of the outbreak until mid-July and forecasted the insurgence of the second epidemic peak (see Figure 6.4C,D). We predict that a total of 15.0 (95%CI: 4.9 – 25.0) million people have recovered and 51.1 (95%CI: 17.0 – 88.1) thousand died as of 2020-08-31 (end of the simulation period). Both the first peak in March and the second peak in July had very likely overwhelmed the healthcare system with the number of severely ill patients reaching 8.76 (95%CI: 2.97 – 22.03) thousand individuals on 2020-03-18 and 16.68 (95%CI: 3.76 – 17.16) thousand individuals on 2020-07-11 (see Data S5).

6.4 Discussion

Countries in different regions of the world with a lack of adequate epidemic surveillance systems can suffer greatly from under-reporting of prevalence and deaths associated with an infectious disease epidemic and, as a result, may not have a timely public health response to contain the spread or closely monitor local outbreaks. In this work, we provided a number of approaches that can help to understand the pattern of spread and to gauge the level of under-reporting in a region, in this case Iran, where detailed epidemiological data from local surveillance is insufficient.

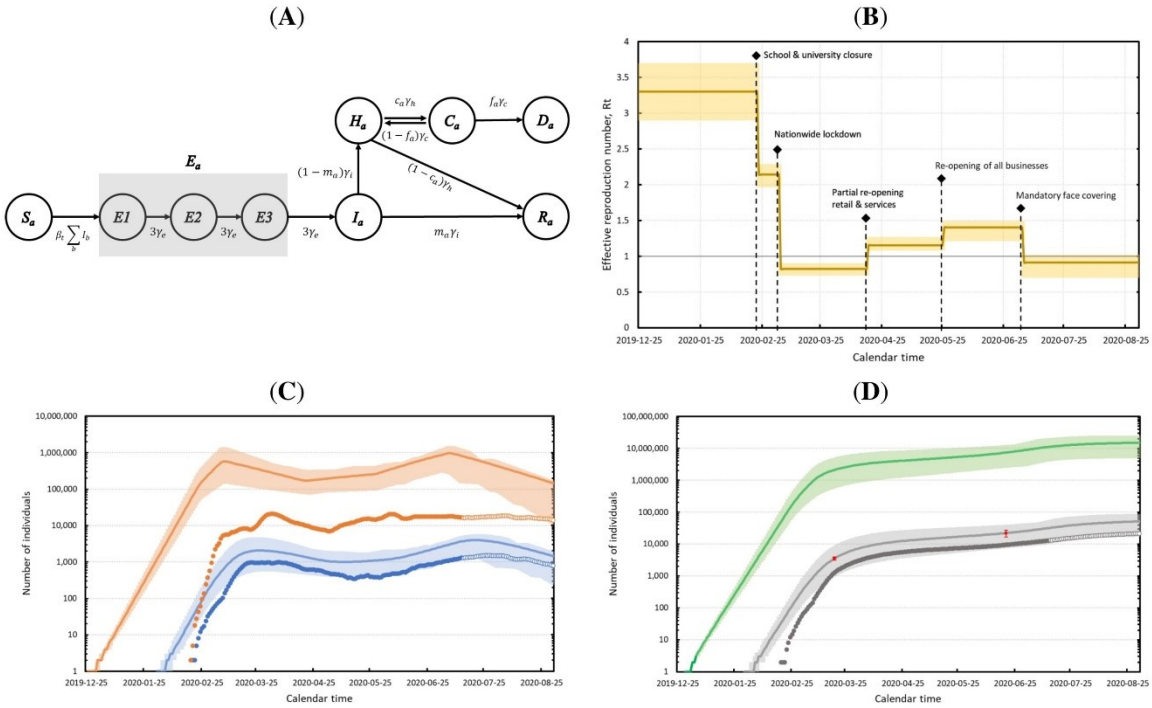


Figure 6.4: **Reconstructing the epidemic in Iran using an SEIR model.** Country-level estimation of infections, deaths, and proportion recovered based on the effect of NPIs on effective reproduction number. (**A**) Diagram of the SEIR model (this figure is adapted from Figure 6.1 of Noll et al. [51]. For every age-group, *a*, we have susceptible (S), exposed (E), infected (I), hospitalised (H), critical (C), dead (D), and recovered (R) individuals -- we assume no separate compartment for the ICU overflow. β_t is the time-dependent infection rate which is determined by the basic reproduction number R0 and the time period of patient infectivity γ_i^{-1}. The rate of transition out of the exposed, infectious, hospitalised, and critical compartments are given by $\gamma_e, \gamma_i, \gamma_h,$ and γ_c, respectively. The age-specific fractions $m_a, c_a,$ and f_a correspond to mild, critical, and fatal infections, respectively. (**B**) Time-varying reproduction number, Rt. Vertical dashed lines correspond to the dates when major intervention measures were implemented (see Table 6.2 and Table S4). (**C**) Shows the estimated number of weekly infections (orange line), deaths (blue line), confirmed cases (orange dots), and confirmed deaths (blue dots). (**D**) Shows the estimated cumulative deaths (gray line), total number of individuals recovered (green line), cumulative reported deaths (gray dots), estimated excess mortality during winter and spring (red dots). Lines represent model predictions and shaded areas are their 95% confidence intervals. Open circles show the data points that are not used for fitting to the model.

We combined air travel and genomic data from confirmed cases with a history of travel to the country to estimate the true prevalence and growth rate of the epidemic. Our analysis shows that it is possible to reliably infer the epidemic size and its early growth dynamics by combining indirect sources of information. Phylogenetic analysis (clades with genomes related to Iran), coupled with epidemiological (cases with a recent travel history to Iran in late February and early March) and clinical data (date of symptom onset), further helped with identifying the earliest introduction of SARS-CoV-2 into Iran and estimating the start date and early growth rate of the epidemic in the country. The overall low importation rate of new cases from China suggests that importation to Iran likely occurred via a high-risk individual with frequent travels to the country. It also suggests that risk of importation from travellers is higher than expected, most likely because of their mobility patterns and high number of social contacts.

By further combining our epidemiological and phylogenetic analyses, we were able to fully reconstruct the outbreak of SARS-CoV-2 across the country. The first version of this manuscript made a clear prediction on 2020-04-23 about the risk of a second peak due to relaxed NPI measures, based on well-accepted models for nowcasting and forecasting of COVID-19 epidemics [51]. Iran was in fact the first country to experience a second epidemic peak from the beginning of May (Figure 6.1), confirming both the validity of our epidemic modelling and the relevance of NPIs to control the spread. Our SEIR modelling analysis shows that by 2020-06-10 7.7% (95%CI: 2.18 – 13.9) of the population have recovered from COVID-19 which is in general agreement with seroprevalence measurements of SARS-CoV-2 in the general population of 18 cities across the country by early June [52]. This also implies that a large fraction of the population is still vulnerable to contracting COVID-

19which has significant implications both for the possibility of the virus becoming endemic to the country and its likely return during winter this year which, if coupled with seasonal flu, can significantly overwhelm the hospitals. Furthermore, the continuation of under-reporting in prevalence due to limited testing of suspect cases and tracing their contacts will likely lead to several undetected superspreading events that can spark new outbreaks in different parts of the country making nowcasting and forecasting of the COVID-19 epidemic in Iran extremely challenging. Furthermore, it highlights the major risk imposed by exported cases from any region of the world with an uncontrolled outbreak of COVID-19 in seeding new epidemics in different countries. Thus, it is prudent for countries that have successfully controlled the epidemic to implement more careful screening of travellers from such regions of the world to avoid new bursts of epidemic in their countries.

The combination of air travel and genomic data with the all-cause seasonal mortality data enabled us to assess the level of under-reporting in cases and deaths across the country by comparing them to surveillance data. We assessed the internal consistency of reported numbers by MoHME (both in daily reported cases and deaths due to COVID-19) using Benford's law. However, this test alone cannot be used to rule out some systematic or random patterns of absence of data on case or death counts, e.g. from certain hospitals. The number of confirmed COVID-19 cases in each country may vary depending on the transparency to report correct statistics and the capacity of the healthcare system to detect new cases. The latter also depends on the accuracy of laboratory test kits and accessibility of diagnostic and screening tests. Indeed, many countries with limited testing capacities may have to prioritise testing that informs policy decisions, e.g. they may not test

suspected cases with mild symptoms or those who are asymptomatic and are a close contact of a confirmed case. As a result, the true number of cases is always significantly higher than those reported by the health agencies. In addition, the burden of the epidemic can significantly impact the performance of the healthcare system to properly allocate cause of deaths in individuals with underlying health conditions such as diabetes or heart disease. Therefore, tracking all-cause registered deaths and estimating excess mortality during the outbreak provides a more sensitive measure of COVID-19–associated deaths than would be recorded by counting confirmed deaths. We note that all-cause deaths may also include factors that are not causally associated with SARS-CoV-2 that might affect death rates such as the circulation of the 2019–20 seasonal influenza. According to the United Nations Statistical Division the estimated coverage of registered deaths in Iran is 92% which could potentially be suffering even more from under-counting during the peak of the outbreak when there are likely going to be further delays in death registrations. Also, excess seasonal mortality does not take into account the catalysing role of COVID-19 in deaths among individuals with underlying comorbidities who would have likely died during a particular season even without contracting COVID-19 (as part of the ‘background’ deaths).

Our data have many limitations. We investigated air travel data only to countries with direct flights to Iran and discarded information from detected cases in countries such as Qatar and Canada since we were not able to independently verify the fraction of passengers on board the planes that travelled from Iran to those countries. Also, given the lack of mobility data from Iran, we were unable to investigate possible international exportation of cases to Afghanistan, Iraq, Syria, Azerbaijan, Turkey, and other countries with significant flow of ground transportation

(i.e. trains, buses, and cars) from Iran. We did not have access to the province-level number of confirmed COVID-19 deaths which is a significant source of information to assess excess deaths in the winter and spring 2020. Also, there is likely extreme heterogeneity in the geographical spread of COVID-19 across the country due to various factors such as demographic structure of the provinces' populations, the pattern of social contacts between age groups, and the quality of healthcare and effectiveness of NPIs in different districts. Despite these drawbacks, our approach successfully integrated multiple sources to obtain a more reliable picture of the COVID-19 epidemic. This kind of study should be extended to countries with unreliable local epidemic surveillance systems in order to evaluate the local stage of the epidemic, inform local policies and highlight countries in need of international support to control the epidemic.

6.5 Methods

6.5.1 Data on aggregate number of cases and deaths

We use the time series data for the number of confirmed cases and deaths (see Data S2) from the Johns Hopkins University Centre for Systems Science and Engineering COVID-19 GitHub repository (accessed on 2020-06-04). We also obtain time series data on confirmed cases in all the 31 provinces of Iran from the Ministry of Health website (behdasht.gov.ir). We note that the Ministry stopped releasing province data from 2020-03-23 onward. They also did not release province data on 2020-03-02 and 2020-03-03. We obtain seasonal mortality statistics from the National Organization for Civil Registration (NOCR) of Iran (sabteahval.ir).

6.5.2 Phylogenetic analysis

From a globally representative sample of 802 genomes, including 20 from passengers with a travel link to Iran and additional three genomes that we sequenced from inside Iran, we construct a maximum-likelihood tree using RAxML [15] to examine the number of independent introductions into the country from the start of the pandemic until January 2021. The samples are randomly selected from 11 major clades as identified by NextStrain in 2020 (i.e., 19A, 19B, 20A, 20B, 20C, 20D, 20E, 20F, 20G, 501Y.V1, and 501Y.V2). We enrich the sampled set with 143 samples from clade 19A and 157 samples from 20A to ensure that there is enough resolution in identifying the lineages of interest related to samples from Iran. We bootstrap the maximum-likelihood tree 100 times and use an HKY+G substitution model [16] to construct the tree. All the genomes were downloaded on 2021-01-21 from GISAID's EpiCoV^{TM} Database (gisaid.org), a public database for rapid data sharing hosted by the global initiative on sharing all Influenza data – a complete metadata table with detailed information about all the samples and acknowledgements of all of the authors and institutions involved in sequencing is available in Data S1. We use MUSCLE v3.8.425 [17] for sequence alignment and mask sites in the first and last 130 bp of the alignments. We only include complete sequences (>29 Kbp) with high coverage as determined by GISAID's default search options. To jointly infer the Time to the Most Recent Common Ancestor (TMRCA) of the earliest genomes linked to Iran and the early doubling times in cases from a monophyletic tree using BEAST [18], we first remove genomes that are epidemiologically linked. We then use the known sampling times with a Continuous-Time Markov Chain reference prior on the substitution rate, an exponential population coalescent model with a log-normal prior on size (mean=1, SD=2),

Laplace prior on growth rate (scale=100), and an HKY+G substitution mode [19, 20] and allow the Markov chain Monte Carlo to run for a hundred million steps and discard the first 10% steps as burn-in to find the parameters of interest. Effective sample sizes of all parameters are >1,000 ensuring that they are well-sampled [21]. We use FigTree for tree visualisations [22].

6.5.3 Estimating point-prevalence using air travel data

To find the probability, $p(t)$, that a traveller on board the plane is infected on a given day, t, we assume only asymptomatic or pre-symptomatic cases are on board the plane. This probability is given by

$$p(t) = \alpha \sum_{k=0}^{+\infty} (1 - S(k)) i(t+k) / M$$

where α is the inverse of the proportion of ascertained infections (e.g. $\alpha = 1$ implies complete ascertainment), $S(k)$ is the CDF of the time from infection to symptoms (incubation period) or to detection (incubation period plus delay from onset of symptoms to having a positive RT-PCR test in an infected individual), $i(t)$ is the incidence on day t, and M is the catchment population size. The probability that the infected individual is detected upon travelling is given by

$$\pi = \sum_{t'} n(t') / (\alpha \lambda(t)) = DT / M$$

where $n(t)$ is the number of infected air-travellers that are detected on day t, $\lambda(t) = \sum_{t'} i(t')$ is the total number of cases on day t, $D = n(t) / p(t)$ is the daily number of outbound international travellers, and T is the mean time from exposure to detection. Given that the incidence over time in Iran is unknown, we made the

approximation that $p(t)$ only depends on the incidence on day t such that $p(t) \approx \alpha Ti(t)/M$. Finally, we estimate the total number of cases, λ_i, by finding the maximum likelihood of a binomial function with the number of successful outcomes, N_i, and success probability, π_i, for country *i*, given by

$$L = \arg\max_{\lambda_i} \binom{\lambda_i}{N_i} \pi_i^{N_i} (1-\pi_i)^{\lambda_i - N_i}$$

where $N_i = (1+P)\sum_{t'} n_i(t')$ is the total number of imported cases in country i and P is the proportion of undetected asymptomatic and mildly symptomatic air travellers. By taking the point-prevalence estimates for each country, we can find a line of best fit for the prevalence over time, assuming an exponential regression model, to estimate the start date of the epidemic and early doubling time in cases.

We consider air travel data from passengers in four of the busiest airports in Iran (Tehran, Mashhad, Isfahan, and Shiraz) who flew to Oman, Lebanon, UAE, Kuwait, and China (see Data S3). We allow for 50% (20% - 100%) of the 55 million residents in provinces near the airports to represent their catchment population size (see Data S4). There is typically 5 (4-6) days of incubation period [5-7, 23, 24] and an additional 5 (3-7) days of delay from symptom onset to detection [24, 25]. We further assume that an additional 45% (30%-55%) of asymptomatic and 15% (10%-25%) mildly symptomatic cases were among the undetected exported carriers of COVID-19 [26, 27].

6.5.4 Estimating seasonal excess mortality

We use the publicly available data from the NOCR which records the all-cause registered deaths in Iran per province per season according to Solar Hijri (SH)

calendar [28]. We assemble this data for the last 5 years to compare the excess mortality during winter and spring 2020 to previous years. Excess deaths refer to the number of deaths above expected seasonal baseline levels, regardless of the reported cause of death, and can be used as a nonspecific measure of the severity of the epidemic and provide a more accurate measure of its burden on the healthcare system [29]. We analyse the NOCR data from 1394 SH to 1399 SH (from 2015-12-22 to 2020-06-20 in the Gregorian calendar). For each province, we apply a least-square regression model on the seasonal deaths from previous years to find the 'expected' seasonal death in 2020 (see Figures S2 and S3) and then calculate the difference with respect to the observed death toll for each province to calculate excess mortality (see Tables S2 and S3). We attribute excess deaths in those provinces with significant deviations (i.e., two standard deviation units) with respect to their expected seasonal value to COVID-19-related deaths.

6.5.5 Evaluating the importation risk from China in January

We define the importation risk as the mean number of infectious individuals travelling from China to Iran over the span of approximately one month, from when China started reporting cases on 2020-01-22 until when the epidemic in Iran started. We assume importations to Iran only occur via air travel from infected areas in China. The main routes of travel to Iran are from airports in Beijing, Shanghai, and Guangdong. We allow a fraction, f, of passengers to come from the Hubei province. To calculate f we take the ratio of total passenger traffic from airports in Hubei to 100 busiest airports across the country according to the 2018 report from the Civil Aviation Administration of China [30].

To calculate the probability, $p(t)$, that an individual on board the plane from a given province is infected, we correct for ascertainment bias in reported cases in China during the early stages of the pandemic, $\alpha^{-1} = 0.23\left(95\%\text{CI}: 0.14 - 0.42\right)$ [31]. We further assume that the catchment population size is equal to the population of the province, the incubation period is log-normally distributed (mean=4.8 days and s.d.=1.9 days) [24], and the delay from the onset of symptoms to having a positive RT-PCR test in an infected individual is Gamma distributed (shape parameter=2.12 and scale parameter=0.39) [24]. Finally, to calculate the expected number of imported cases, N_i, to Iran from 2020-01-22, t=0, to a later date, t*, we have

$$N(t) = \sum_{i=\{\text{Beijin,Shanghai,Guangdong}\}} \sum_{t=0}^{t^*} (1-f) D_i p_i(t) + f D_{\text{Hubei}} p_{\text{Hubei}}(t)$$

where D_i is the average number of passengers flying from province i per day.

6.5.6 SEIR model

We use a generalised SEIR model with age-stratified compartments to reconstruct the early transmission dynamics of the epidemic in Iran (covid19-scenarios.org) [31]. In this model, susceptible individuals can be exposed to the virus through contact with an infected individual. They then progress towards the infectious stage where they can either recover without hospitalisation or progress towards severe illness that requires hospitalisation. For the latter group, individuals either recover or transition to the ICU stage at which point they either die or return back to the hospitalisation state (see the schematic diagram of the model in Figure 6.4A). We use the line of best fit for prevalence over time based on our air travel analysis on exported cases to parameterise the basic reproduction number, R_0, for the SEIR model. To do so, we fix the start date of the epidemic according to the line of best

fit and re-fit the exponential regression model to the point-prevalence estimates to capture the variation in growth rate and estimate R_0 using the following equation:

$$R_0 = \frac{r(T_G - T_E)}{\sinh r(T_G - T_E)} E^{rT_G}$$

where r is the doubling time, T_E is the latency period, $T_G = T_E + 1/2T_I$ is the generation time, and T_I is the infectious period (see Tables 6.2 and 6.3) [32, 33].

Table 6.3: **SEIR model parameters.** We use this information, along with the effectiveness of intervention measures in reducing Rt (Table 6.1), to run the simulations on *covid19-scenarios.org* dashboard [51].

Age group	Age distribution*	% of all infections that are fatal†	Model parameters	Expected value	Reference
0-9	14,434,000	0.00075	Basic reproduction number	3.3 (2.9-3.6)	[34-36]
10-19	11,812,000	0.0045	Latency period‡	2.5 days	[53, 54]
20-29	12,254,000	0.009	Infectious period‡	5 days	[53, 54]
30-39	16,786,000	0.02	Hospital stay	4.5 days	Fig. S3
40-49	11,926,000	0.072	ICU stay	6 days	Fig. S3
50-59	8,281,000	0.25	Imports per day	0	
60-69	5,264,000	1.1	Initial number of cases	1	
70-79	2,264,000	3.1	Simulation start date	2019-12-25	
+80	1,015,000	6.9	Simulation end date	2020-08-31	

*based on annual reports on vital statistics from the Statistical Center of Iran (*amar.org.ir*).
†We use the statistics from the Chinese CDC [55] for age-stratified IFR which are also broadly compatible with estimates from other studies [56]
‡Latency period is the delay from infection to onset of infectiousness. Infectiousness typically starts from 2.5 days and peaks at 1 day before the onset of symptoms. Together with infectious period, latency period sets the serial interval [11].

The effective reproduction number, Rt, is modelled as a piecewise constant function that changes immediately upon a new NPI. The expected reduction in Rt upon the implementation of a new NPI is bound to be within the same range as estimates (of the same NPI) from previous studies [34-36] and the exact value of the mean Rt is

determined by finding the line of best fit that matches the mean cumulative deaths with our estimates for COVID-19-related deaths based on excess seasonal mortality in winter and spring. The lower bound in the effectiveness of each NPI is set such that the modelled cumulative deaths do not drop below the confirmed COVID-19-related deaths as reported by the Ministry of Health and Medical Education (MoHME) from the first day of case reporting, 2020-02-19, until the day that the last NPI, mandatory face-covering, went into effect on 2020-07-05. The SEIR package, covid19-scenarios.org, allows the user to set the range of transmission reduction during each intervention. Therefore, by fixing the mean and lower bound of Rt, the upper bound can also be determined. From the time of the announcement of the last NPI (i.e., mandatory face covering) until the last day of the simulation run, the reconstructed transmission dynamics are solely based on an assumed effectiveness of mandatory face covering which is derived from previous studies (see Table 6.2). Thus, any inference beyond 2020-07-05 should be treated as part of the model's projection of the actual dynamics. The duration of hospital/ICU stay is taken from the 2020-03-14 report by MoHME (Figure S3) [37]. We also assume all infected individuals are immune to reinfection over the course of the simulation.

6.5.7 Resource availability

All data and codes used for the analysis are available online on our GitHub repository (github.com/mg878/Iran_study).

6.6 References

1. *Coronavirus disease 2019 (COVID-19): situation report, 176.* 2020, World Health Organization.
2. Wright, R., *How Iran became a new epicenter of the coronavirus outbreak*, in *The newyorker*.
3. Li, R.Y., et al., *Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2).* Science, 2020. **368**(6490): p. 489-494.
4. Pybus, O.G., Rambaut, A., and COG-UK-Consortium, *Preliminary analysis of SARS-CoV-2 Importation & Establishment of UK Transmission Lineages.* Virological.org, 2020.
5. Li, Q., et al., *Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia.* New England Journal of Medicine, 2020. **382**(13): p. 1199-1207.
6. Lauer, S.A., et al., *The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application.* Annals of Internal Medicine, 2020. **172**(9): p. 577-581.
7. Ganyani, T., et al., *Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020.* Eurosurveillance, 2020. **25**(17): p. 12-19.
8. Oran, D.P. and E.J. Topol, *Prevalence of Asymptomatic SARS-CoV-2 Infection A Narrative Review.* Annals of Internal Medicine, 2020. **173**(5): p. 362-368.
9. Buitrago-Garcia, D.C., et al., *The Role of Asymptomatic SARS-CoV-2 Infections: Rapid Living Systematic Review and Meta-Analysis.* medRxiv, 2020.
10. Bai, Y., et al., *Presumed Asymptomatic Carrier Transmission of COVID-19.* Jama-Journal of the American Medical Association, 2020. **323**(14): p. 1406-1407.
11. Ferretti, L., et al., *Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing.* Science, 2020. **368**(6491): p. 619-622.
12. Silverman, J.D., N. Hupert, and A.D. Washburne, *Using influenza surveillance networks to estimate state-specific prevalence of SARS-CoV-2 in the United States.* Science Translational Medicine, 2020. **12**(554).
13. *World Health Organization, Influenza situation in the Eastern Mediterranean Region, W11/2020.*; Available from: https://applications.emro.who.int/docs/EMCSR272E.pdf?ua=1.
14. Ghafari, M. and K. Madani, *In search of the murder: Making sense of Iran's reported deaths*, in *Medium*. 2020.
15. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-1313.

16. Hasegawa, M., H. Kishino, and T.A. Yano, *Dating of the Human Ape Splitting by a Molecular Clock of Mitochondrial-DNA.* Journal of Molecular Evolution, 1985. **22**(2): p. 160-174.

17. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Research, 2004. **32**(5): p. 1792-1797.

18. Suchard, M.A., et al., *Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10.* Virus Evolution, 2018. **4**(1).

19. Rambaut, A., *Phylogenetic analysis of nCoV-2019 genomes.* Virological.org, 2020.

20. Boni, M.F., et al., *Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic.* Nature Microbiology, 2020. **5**(11): p. 1408-1411.

21. Rambaut, A., et al., *Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7.* Systematic Biology, 2018. **67**(5): p. 901-904.

22. Rambaut, A. *Figtree v1.4.4*. Available from: http://tree.bio.ed.ac.uk/software/figtree/.

23. He, X., et al., *Temporal dynamics in viral shedding and transmissibility of COVID-19.* Nature Medicine, 2020. **26**(5): p. 672-678.

24. Bi, Q.F., et al., *Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study.* Lancet Infectious Diseases, 2020. **20**(8): p. 911-919.

25. *Ministry of Health, Labour and Welfare Japan, 新型コロナウイルスに関連した肺炎の患者の発生について（1例目).* 2020; Available from: https://www.mhlw.go.jp/stf/newpage_08906.html.

26. Lavezzo, E., et al., *Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'.* Nature, 2020. **584**(7821): p. 425-428.

27. Casey, M., et al., *Estimating pre-symptomatic transmission of COVID-19: a secondary analysis using published data.* medRxiv, 2020.

28. *National Organization for Civil Registration, Registered Deaths by Province.* 1 June 2020]; Available from: https://www.sabteahval.ir/.

29. Olson, D.R., et al., *Preliminary Estimate of Excess Mortality During the COVID-19 Outbreak - New York City, March 11-May 2, 2020.* Mmwr-Morbidity and Mortality Weekly Report, 2020. **69**(19): p. 603-605.

30. *Civil Aviation Administration of China, 年民航机场生产统计公报.* . 5 March 2020]; Available from: http://www.caac.gov.cn/XXGK/XXGK/TJSJ/201903/P020190305338562571372.xls?COLLCC=3782305670&.

31. Hao, X.J., et al., *Reconstruction of the full transmission dynamics of COVID-19 in Wuhan.* Nature, 2020. **584**(7821): p. 420-423.

32. Roberts, M.G. and J.A.P. Heesterbeek, *Model-consistent estimation of the basic reproduction number from the incidence of an emerging infection.* Journal of Mathematical Biology, 2007. **55**(5-6): p. 803-816.

33. Park, S.W., et al., *A practical generation-interval-based approach to inferring the strength of epidemics from their speed.* Epidemics, 2019. **27**: p. 12-18.

34. Flaxman, S., et al., *Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe.* Nature, 2020. **584**(7820): p. 257-260.

35. Mellan, T.A., et al., *Estimating COVID-19 cases and reproduction number in Brazil.* medRxiv, 2020.

36. Unwin, H.J.T., et al., *State-level tracking of COVID-19 in the United States.* Nature Communications, 2020. **11**(1): p. 6189.

37. *National Committee on COVID-19 Epidemiology, Daily Situation Report on Coronavirus disease (COVID-19) in Iran.* 14 March 2020]; Available from: Redirecting you to http://corona.behdasht.gov.ir/files/site1/files/Factsheet-12.24-En1.%20pdf.

38. *WHO Director-General's opening remarks at the media briefing on COVID-19.* 28 February 2020; Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---28-february-2020.

39. *behdasht.gov*. 28 February 2020; Available from: https://behdasht.gov.ir/%d9%be%d9%88%d8%b3%d8%aa%d9%87-%d8%b1%db%8c%d8%b3%d9%be%d8%a7%d9%86%d8%b3%db%8c%d9%88-%d8%af%d8%a7%d8%ae%d9%84%db%8c-94/%d8%b4%d9%86%d8%a7%d8%b3%d8%a7%db%8c%db%8c-143-%d9%85%d9%88%d8%b1%d8%af-%d8%ac%d8%af%db%8c%d8%af-%d8%a7%d8%a8%d8%aa%d9%84%d8%a7%db%8c-%d9%82%d8%b7%d8%b9%db%8c-%d8%a8%d9%87-%da%a9%d9%88%d9%88%db%8c%d8%af19-%d8%af%d8%b1-%da%a9%d8%b4%d9%88%d8%b1.

40. *WHO ships emergency medical supplies to the Islamic Republic of Iran.* 12 March 2020]; Available from: https://www.who.int/news-room/feature-stories/detail/who-ships-emergency-medical-supplies-to-the-islamic-republic-of-iran.

41. Russell, T.W., et al., *Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections.* Bmc Medicine, 2020. **18**(1).

42. Wikramaratna, P.S., et al., *Estimating the false-negative test probability of SARS-CoV-2 by RT-PCR.* Eurosurveillance, 2020. **25**(50): p. 2000568.

43. Hodcroft, E., Müller, N., Wagner, C., Ilcisin, M., Hadfield, J., Bell, S. M., Neher, R., Bedford, T. *Genomic analysis of COVID-19 spread (Nextstrain Situation Report).* 2020-03-13; Available from: https://nextstrain.org/narratives/ncov/sit-rep/2020-03-13.

44. Eden, J.S., et al., *An emergent clade of SARS-CoV-2 linked to returned travellers from Iran.* Virus Evol, 2020. **6**(1): p. veaa027.

45. Imai, N., Dorigatti, I., Cori, A., Donnelly, C., Riley, S., and Ferguson, N., *Report 2: Estimating the potential total number of novel Coronavirus cases in Wuhan City, China.* Imperial College London COVID-19 Response Team, 2020.

46. Tuite, A.R., et al., *Estimation of Coronavirus Disease 2019 (COVID-19) Burden and Potential for International Dissemination of Infection From Iran.* Annals of Internal Medicine, 2020. **172**(10): p. 699-704.

47. Zhuang, Z., et al., *Preliminary estimation of the novel coronavirus disease (COVID-19) cases in Iran: A modelling analysis based on overseas cases and air travel data.* International Journal of Infectious Diseases, 2020. **94**: p. 29-31.

48. Pellis, L., Scarabel, F., Stage, H.B., Overton, C.E., Chappell, L.H., Lythgoe, K.A., Fearon, E., Bennett, E., Curran-Sebastian, J., Das, R. and Fyles, M., *Challenges in control of Covid-19: short doubling time and long delay to effect of interventions.* arXiv:2004.00117, 2020.

49. Sun, K., et al., *Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2.* Science, 2021. **371**(6526).

50. Reiner, R.C., et al., *Modeling COVID-19 scenarios for the United States.* Nature Medicine, 2021. **27**(1): p. 94-99.

51. Noll, N.B., Aksamentov, I., Druelle, V., Badenhorst, A., Ronzani, B., Jefferies, G., Albert, J., Neher, R., *COVID-19 Scenarios: an interactive tool to explore the spread and associated morbidity and mortality of SARS-CoV-2.* medRxiv: 2020.05.05.20091363, 2020.

52. Poustchi, H., et al., *SARS-CoV-2 antibody seroprevalence in the general population and high-risk occupational groups across 18 cities in Iran: a population-based cross-sectional study.* Lancet Infect Dis, 2020.

53. He, X., et al., *Temporal dynamics in viral shedding and transmissibility of COVID-19.* Nature Medicine, 2020. **26**(5).

54. Lavezzo, E., et al., *Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo' (vol 584, pg 425, 2020).* Nature, 2021.

55. Zijian Feng, et al., *The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China.* Chinese Center for Disease Control and Prevention, 2020. **2** (8): p. 113-122.

56. Verity, R., et al., *Estimates of the severity of coronavirus disease 2019: a model-based analysis.* Lancet Infect Dis, 2020. **20**(6): p. 669-677.

57. Rambaut, A. and A. Drummond. *TreeAnnotator v2.6.0*. 2018; Available from: http://beast.bio.ed.ac.uk/.

58. Benford, F., *The law of anomalous numbers.* Proceedings of the American philosophical society., 1938. **31**: p. 551-572.

59. Cerioli, A., et al., *Newcomb-Benford law and the detection of frauds in international trade.* Proceedings of the National Academy of Sciences of the United States of America, 2019. **116**(1): p. 106-115.

60. Pericchi, L. and D. Torres, *Quick Anomaly Detection by the Newcomb-Benford Law, with Applications to Electoral Processes Data from the USA, Puerto Rico and Venezuela.* Statistical Science, 2011. **26**(4): p. 502-516.

61. Durtschi, C., W. Hillison, and C. Pacini, *The effective use of Benford's law to assist in detecting fraud in accounting data.* Journal of forensic accounting, 2004. **5**: p. 17-34.

62. Idrovo, A.J. and E.F. Manrique-Hernandez, *Data Quality of Chinese Surveillance of COVID-19: Objective Analysis Based on WHO's Situation Reports.* Asia-Pacific Journal of Public Health, 2020. **32**(4): p. 165-167.

63. Crocetti, E. and G. Randi, *Using the Benford's Law as a First Step to Assess the Quality of the Cancer Registry Data.* Frontiers in Public Health, 2016. **4**: p. 225.

Statement of Authorship for joint/multi-authored papers for PGR thesis.

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

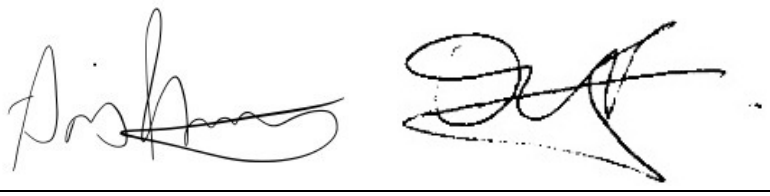
Title of Paper	Lessons for preparedness and reasons for concern from the early COVID-19 epidemic in Iran
Publication Status	Published
Publication Details	Ghafari M, Hejazi B, Karshenas A, Dascalu S, Kadvidar A, Khosravi MA, Abbasalipour M, Heydari M, Zeinali S, Ferretti L, Ledda A, and Katzourakis A. Lessons for preparedness and reasons for concern from the early COVID-19 epidemic in Iran. Epidemics 36: 100472, 2021.

Student confirmation

Student Name	Mahan Ghafari
Contribution to the paper	MG wrote the chapter and conducted all the analysis under the supervision from AK. MAK, MA, and SZ prepared the genomic samples from Iran. All authors provided editorial assistance.
Signature 	Date 11/05/2022

Supervisor confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title	Professors Aris Katzourakis and Oliver G Pybus	
Supervisor comments		
Signature 		Date 11/05/2022

Chapter 7: Inferring SARS-CoV-2 transmission dynamics using excess mortality data

Part a: Excess deaths associated with the Iranian COVID-19 epidemic: A province-level analysis

7.1 Abstract

There has been no province-level data on the number of COVID-19-related deaths in Iran since the start of the pandemic. We estimate the number of COVID-19 deaths and population-level exposure per province using seasonal all-cause mortality data. We gather time series data on the seasonal all-cause mortality from spring 2015 to summer 2020 (from 21 March 2015 to 21 September 2020), in accordance with the Solar Hijri (SH) calendar, from the National Organization for Civil Registration (NOCR) to estimate the expected number of seasonal deaths for each province using a piecewise linear regression model. We then apply a population-weighted infection fatality ratio to estimate the level of exposure per province during this period. We find that from the start of winter to the end of summer (from 22 December 2019 to 21 September 2020), there were a total of 58.9K (95%CI: 46.9K-69.5K) excess deaths across all the 31 provinces with 27% (20%-34%) estimated nationwide exposure to SARS-CoV-2. In particular, Qom and Golestan, had among the hardest-hit provinces with nearly 57% exposure while another 27 provinces showed significant levels of excess mortality in at least one season with >20% population-level exposure to the virus. We also detected unexpectedly high levels of excess mortality during fall 2019 (from 23 September to 21 December 2019) across 18 provinces, unrelated and prior to the start of the COVID-19 pandemic. Our study quantifies the pattern of spread of the COVID-19 across the country for

the first time and identifies areas with the largest epidemic growth requiring the most immediate interventions.

7.2 Introduction

7.2.1 Historical trends of all-cause mortality in Iran

Information on all-cause mortality in Iran and its changing pattern over time provides a major source for evaluating the burden of the COVID-19 epidemic across the country. In order to estimate the number of deaths in excess of previous years, i.e. the mortality attributable to a public health crisis, we can compare the total number of registered deaths during the time of the crisis and compare them to some 'background' level which is measured during periods with no major crisis [1]. Like many other developing countries, Iran has incomplete records of registered deaths. A recent report in 2015 (United Nations Statistical Division, 2015) shows that the estimated coverage of registered births and deaths in Iran is 97% and 92%, respectively (coverage information was obtained from country representative(s) either from the National Statistical Office or the Civil Registration Authority attending one of the Nations Statistics Division workshops) [2]. The National Organization for Civil Registration (NOCR) is responsible for the registration of births, marriages, divorces, and deaths across the country. It initially updates the number of registered (crude) deaths at the end of every season for all the 31 provinces in Iran (in accordance with the Solar Hijri calendar date) [2]. More detailed data on the number of registered deaths per month will then follow at the end of each calendar year (on the March equinox) along with an annual report from the Ministry of Health and Medical Education (MoHME) which includes an aggregate record of causes of deaths (based on the international statistical classification of diseases and related

health problems, 10th revision) according to individuals age, sex, and place of residence [4].

Upon examining the historical trends of all-cause mortality in Iran from the 1960s until present, we find significant levels of variations over time (Figure 7.1). In particular, there is a spike in 1995 with more than 2.7 million registered deaths in just one year (a record-high with almost 15 times higher death toll than previous years). This was part of an ‘emergency operation’ to record significant levels of under-counting from previous years, such as the casualties of the Iran-Iraq war in the 1980s, and a redesign in NOCR’s data collection strategy [4]. Between 1966 and 1995, the mortality data based on cemetery records were collected only in a sample of 24 cities. In 1995, the system was redesigned to cover the entire country. Since then, the average number of registered deaths has roughly doubled with approximately 389 thousand deaths compared to 175 thousand from previous years. However, even after the completion of the emergency operation (from 1994 to 1999), the annual numbers continue to fluctuate significantly over time with amplitude variations as large as 100 thousand deaths between consecutive years. Such variations are so high that even some of the major natural catastrophes in Iran such as the Manjil-Rudbar earthquake with more than 40 thousand reported deaths in 1990 or the Bam earthquake with more than 50 thousand deaths in 2003 would not have a detectable footprint in the annual all-cause mortality data. Previous works have shown despite the progress since 1995 in reporting the true number of registered deaths, such variations may persist due to substantial delays in registration and recording of deaths [5]. Examining the underlying cause(s) of these large variations is beyond the scope of this study, but we note that this is an area that requires further in-depth investigation.

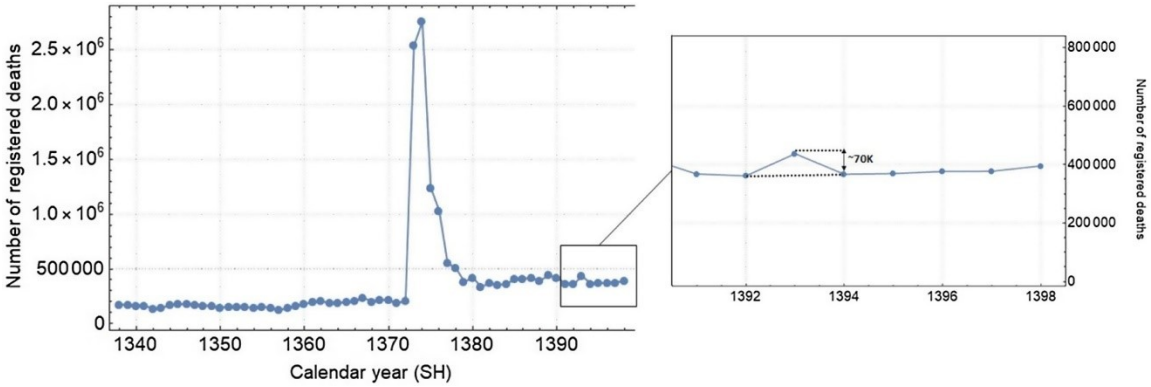


Figure 7.1: **Annual number of registered deaths in Iran over the last 70 years (data from NOCR).** The inset shows the variations in registered deaths during the more recent years which seems to have stabilised with the exception of year 1393 SH with approximately 70 thousand deaths above baseline.

MoHME never released provincial data on the number of confirmed COVID-19 deaths and only announced the number of cases during the first 33 days since the start of the epidemic (from 19 February to 22 March 2020, with the exception of 2 and 3 March). Thus, seasonal excess mortality is the only proxy to assess the geographic spread of the epidemic across the country. Similar works have also been carried out to monitor excess mortality in Africa during a measles outbreak [6]. Despite significant fluctuations in NOCR's data over the years, they seem to have remained stable from 1394 Solar Hijri (SH) onward (Figure 7.1). Therefore, we only incorporate the data from the beginning of spring 1394 SH (21 March 2015) onward to estimate excess mortality rates (Methods section).

7.3 Results

7.3.1 Did the COVID-19 epidemic in Iran start during fall 2019?

Upon examining the seasonal trends in mortality during fall in our earlier reports [7,8], we noticed, for the first time, that there was a spike across the country in 1398 SH (from 23 September to 21 December 2019) compared to previous years. This was particularly unexpected as it occurred before the start of the COVID-19 pandemic in December [9,10] and its emergence in Iran during January [11] which led to some speculations about the cryptic transmission of the virus across the country in fall 2019.

Our analysis shows an 8% rise in nationwide mortality rate during fall with a total of 6.04K (95%CI: 3.48K - 8.60K) extra recorded deaths in 18 provinces -- This is also aligned with estimates from other studies [12]. For these extra deaths in fall to be associated with COVID-19, we would expect, due to the absence of any non-pharmaceutical interventions at the time, that similar (or higher) levels of excess

mortality to be detected in the same provinces during winter. However, we find only 5 provinces with significantly higher levels of mortality in winter 1398 SH (22 December 2019 to 19 March 2020) (Table 7.1). These provinces include Qom, Gilan (Guilan), Golestan, Mazandaran, and Qazvin which were reportedly hit the hardest by an early wave of the COVID-19 epidemic with a total of 3.48K (95%CI: 3.17K - 3.78K) combined extra deaths [13]. This observation rejects any possible association between the excess mortality in fall and significant levels of undetected COVID-19-related deaths across the country.

In search of possible causes of the excess deaths during fall, we investigated the WHO annual influenza situation reports for the Eastern Mediterranean Region in the past two years and found a 20% increase in the number of positive flu cases during fall 1398 SH (with the co-circulation of both H1N1 and flu type B in Iran) [7]. However, in the absence of surveillance data for influenza-like illnesses or severe acute respiratory infections in Iran it is not possible to assess the burden of the flu epidemic across the country.

7.3.2 The geographical spread of COVID-19 in Iran

While the mean nationwide excess mortality was only at approximately 2% during winter, it rose to 21% during spring 1399 SH (from 20 March to 20 June 2020) with a total of 18.36K (95%CI: 13.51K - 23.21K) excess deaths in 28 provinces and 39% during summer 1399 SH (from 21 June to 21 September 2020) with a total of 36.28K (95%CI: 30.26K - 42.30K) excess deaths in 30 provinces (Figure 7.2).

We also find that the 5 provinces in central north part of Iran with an early outbreak of COVID-19 in winter (i.e., Qom, Gilan, Golestan, Mazandaran, and Qazvin) continued to have significantly high levels of mortality during spring and

Table 7.1: **Seasonal trend of provinces with (+ sign) or without (- sign) significant levels of excess mortality from fall 1398 SH to summer 1399 SH.** (F=fall, W=winter, Sp=spring, Su=summer).

Provinces	F 98	W 98	Sp 99	Su 99
East Azerbaijan	+	−	+	+
West Azerbaijan	+	−	+	+
Ardabil	−	−	+	+
Isfahan	+	−	+	+
Alborz	−	−	+	+
Ilam	+	−	+	+
Bushehr	−	−	−	+
Tehran	+	−	+	+
South Khorasan	−	−	+	−
Razavi Khorasan	−	−	+	+
North Khorasan	+	−	+	+
Khuzestan	−	−	+	+
Zanjan	+	−	+	+
Semnan	+	−	+	+
Sistan and Baluchistan	−	−	−	+
Fars	+	−	−	+
Qazvin	+	+	+	+
Qom	+	+	+	+
Kurdistan	+	−	+	+
Kerman	−	−	+	+
Kermanshah	−	−	+	+
Kohgiluyeh and Boyer-Ahmad	−	−	+	+
Lorestan	+	−	+	+
Mazandaran	+	+	+	+
Markazi	+	−	+	+
Hormozgan	−	−	+	+
Hamedan	−	−	+	+

Yazd	+	−	+	+
Chaharmahal and Bakhtiari*	−	−	+	−
Golestan	+	+	+	+
Gilan (Guilan)	+	+	+	+
Nationwide	+	−	+	+

*We omitted the number of registered deaths for summer 1394 SH out of the analysis for this province as it showed a record-high number with nearly three times more deaths compared to any of the following years.

summer while 3 provinces in the south (Fars, Bushehr, and Sistan-Baluchistan) only showed a rise in mortality during summer (Figure 7.3). The remaining provinces had significant levels of excess mortality in spring followed by a higher increase during summer (with the exception of South Khorasan which only showed significant excess deaths in spring) which indicates that a sustained circulation of the virus had been going on for at least six months in most provinces (Figure 7.3). We also note that the level of excess mortality in Gilan continued to decrease after its initial peak during winter while Qom seemed to have increased again in summer after a temporary drop during spring.

7.3.3 Estimating the level of exposure across the population

Assuming that all the excess deaths since winter 1398 SH onward are directly associated with the COVID-19 epidemic, we applied a population-weighted Infection Fatality Ratio (IFR) [14] to estimate the level of exposure in each province (Methods section) by the end of summer 1399 SH (Table 7.2). We find that 13 (out of 31) provinces had nearly 30% exposure or above with Qom and Golestan having the highest percentage of individuals exposed at 57% (95%CI: 44% - 69%) and 56% (95%CI: 44% - 69%), respectively (Figure 7.4).

By combining the estimated number of deaths associated with COVID-19 based on excess mortality and the number of individuals exposed to SARS-CoV-2 in each province, we can find the nationwide number of COVID-19-associated infections and deaths (Figure 7.5). We find that the corrected nationwide COVID-19-related deaths during each season is approximately 2.5 times higher than the number of confirmed deaths from MoHME and that both estimates follow a similar trajectory over time with particularly elevated levels of excess mortality during winter (i.e.

approximately 2.8 times higher than the confirmed deaths at the time) which may indicate that a larger portion of deaths during the early stages of the epidemic were under-counted as also noted in our previous analysis [13].

Table 7.2: **Excess mortality associated with COVID-19 and estimated level of exposure in each province.** We use the age-structure of each province from the Statistical Center of Iran [29] to calculate their corresponding population-weighted IFR. Numbers in (.) show the 95% confidence intervals.

Provinces	Excess deaths	%IFR	Number of individuals exposed	%Exposed
East Azerbaijan	3.0K (2.2K - 3.8K)	0.30 (0.28 - 0.32)	1.01M (0.72M - 1.29M)	25 (18 - 32)
West Azerbaijan	2.0K (1.7K - 2.3K)	0.24 (0.22 - 0.25)	0.83M (0.64M - 1.03M)	24 (18 - 30)
Ardabil	1.3K (1.0K - 1.6K)	0.26 (0.25 - 0.28)	0.49M (0.36M - 0.61M)	37 (28 - 47)
Isfahan	3.3K (2.6K - 4.0K)	0.29 (0.27 - 0.31)	1.13M (0.81M - 1.45M)	21 (15 - 27)
Alborz	2.0K (1.7K - 2.3K)	0.24 (0.22 - 0.25)	0.84M (0.64M - 1.05M)	29 (22 - 36)
Ilam	0.4K (0.3K - 0.5K)	0.23 (0.22 - 0.25)	0.17M (0.13M - 0.21M)	28 (21 - 34)
Bushehr	0.7K (0.6K - 0.7K)	0.18 (0.17 - 0.19)	0.38M (0.30M - 0.45M)	30 (24 - 36)
Tehran	8.8K (7.8K - 9.7K)	0.28 (0.26 - 0.30)	3.09M (2.24M - 3.94M)	22 (16 - 28)
South Khorasan	0.1K (0.0K - 0.1K)	0.26 (0.24 - 0.29)	0.03M (0.02M - 0.04M)	4 (3 - 5)
Razavi Khorasan	4.6K (3.9K - 5.4K)	0.23 (0.22 - 0.25)	2.00M (1.53M - 2.47M)	29 (22 - 36)
North Khorasan	0.4K (0.2K - 0.6K)	0.22 (0.21 - 0.24)	0.19M (0.15M - 0.24M)	22 (17 - 27)
Khuzestan	4.5K (3.9K - 5.1K)	0.19 (0.28 - 0.20)	2.33M (1.85M - 2.81M)	47 (37 - 57)
Zanjan	1.0K (0.8K - 1.1K)	0.28 (0.26 - 0.30)	0.35M (0.26M - 0.44M)	32 (23 - 40)
Semnan	0.3K (0.2K - 0.5K)	0.27 (0.25 - 0.29)	0.12M (0.09M - 0.16M)	16 (12 - 21)
Sistan and Baluchestan	0.9K (0.6K - 1.2K)	0.13 (0.12 - 0.13)	0.73M (0.62M - 0.84M)	24 (20 - 27)
Fars	1.8K (1.5K - 2.2K)	0.26 (0.24 - 0.28)	0.70M (0.52M - 0.88M)	14 (10 - 17)
Qazvin	1.6K (1.1K - 2.1K)	0.25 (0.23 - 0.27)	0.64M (0.48M - 0.80M)	48 (36 - 60)
Qom	1.6K (1.4K - 1.8K)	0.21 (0.19 - 0.22)	0.79M (0.62M - 0.96M)	57 (44 - 69)
Kurdistan	1.7K (1.1K - 2.3K)	0.25 (0.23 - 0.26)	0.68M (0.51M - 0.85M)	41 (31 - 51)
Kerman	1.5K (1.0K - 2.0K)	0.22 (0.20 - 0.23)	0.72M (0.56M - 0.87M)	21 (13 - 26)

Kermanshah	1.2K (0.9K - 1.5K)	0.27 (0.25 - 0.29)	0.45M (0.33M - 0.56M)	22 (16 - 28)
Kohgiluyeh and Boyer-Ahmad	0.2K (0.1K - 0.4K)	0.20 (0.19 - 0.22)	0.12M (0.09M - 0.15M)	16 (12 - 19)
Lorestan	1.2K (0.9K - 1.4K)	0.25 (0.23 - 0.27)	0.47M (0.35M - 0.58M)	26 (20 - 32)
Mazandaran	3.7K (3.3K - 4.1K)	0.31 (0.29 - 0.34)	1.18M (0.83M - 1.53M)	35 (25 - 45)
Markazi	1.0K (0.8K - 1.3K)	0.32 (0.29 - 0.34)	0.33M (0.23M - 0.42M)	22 (16 - 29)
Hormozgan	1.4K (1.1K - 1.6K)	0.17 (0.16 - 0.18)	0.81M (0.66M - 0.96M)	42 (34 - 49)
Hamedan	1.3K (0.9K - 1.7K)	0.30 (0.28 - 0.32)	0.43M (0.31M - 0.55M)	24 (17 - 31)
Yazd	0.6K (0.5K - 0.8K)	0.24 (0.22 - 0.26)	0.27M (0.20M - 0.33M)	21 (16 - 27)
Chaharmahal and Bakhtiari	0.6K (0.3K - 0.9K)	0.24 (0.22 - 0.26)	0.25M (0.19M - 0.31M)	26 (19 - 32)
Golestan	2.3K (1.9K - 2.6K)	0.20 (0.19 - 0.22)	1.12M (0.87M - 1.36M)	56 (44 - 69)
Gilan (Guilan)	3.1K (2.5K - 3.8K)	0.36 (0.34 - 0.39)	0.86M (0.57M - 1.14M)	33 (22 - 44)
Nationwide	58.9K (46.9K - 69.5K)	0.25 (0.24 - 0.27)	23.0M (17.2M - 28.7M)	27 (20 - 34)

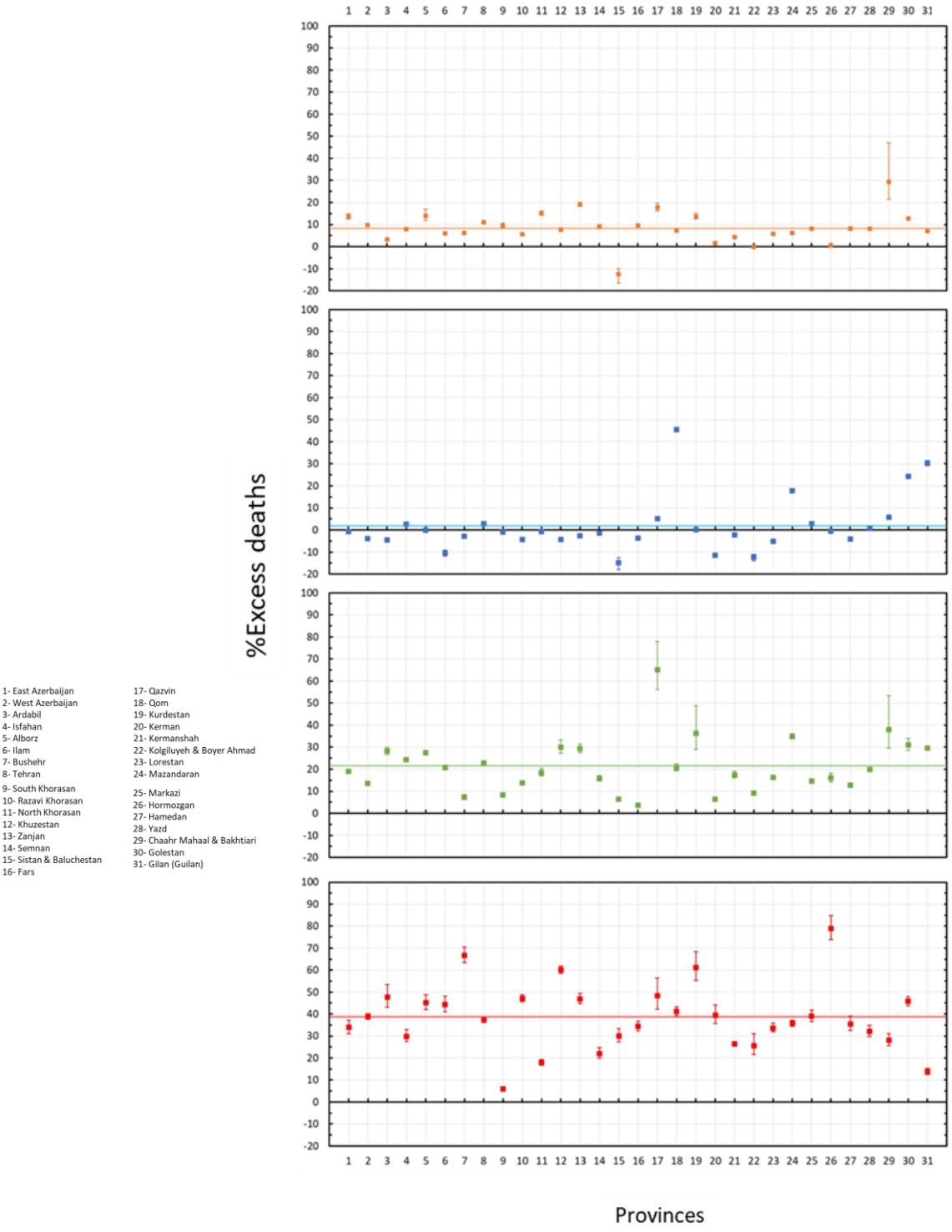


Figure 7.2: **Percentage of excess deaths (compared to the background level)**. Estimated percentage of excess mortality in fall 1398 SH (from 23 September to 21 December 2019), winter 1398 SH (from 22 December to 19 March 2020), spring 1399 SH (from 20 March to 20 June 2020), and summer 1399 SH (from 21 June to 21 September 2020). The horizontal solid lines show the mean percentage of excess mortality in fall (orange), winter (blue), spring (green), and summer (red).

7.4 Discussion

In this work, we compared mortality rates in 31 provinces of Iran from winter 1398 SH to summer 1399 SH with the aim to detect and measure excess deaths related to the COVID-19 epidemic. Our findings suggest a total of 58.9 (95%CI: 46.9 - 69.5) thousand COVID-19-related deaths and up to 23.0 (95%CI: 17.2 - 28.7) million individuals exposed to SARS-CoV-2 by 21 September 2020 with some of the hardest-hit provinces such as Qom and Golestan reaching up to approximately 57% (95%CI: 44% - 69%) exposure levels.

Given that the epidemic was never fully brought under control with three major peaks in late-March, mid-July, and late-November that overwhelmed hospitals (as of 30 November 2020), some provinces are experiencing among the highest known population-level prevalence in the World, comparable to Manaus, Brazil [15,16]. Recent reports from the Iranian health officials indicate that up to half of suspect hospitalised cases may not be tested due to shortage of test kits but are being treated as COVID-19 patients based on their clinical symptoms. This highlights the large degree of under-reporting in the actual number of COVID-19 cases and deaths which also seems to be aligned with our findings that the actual number of COVID-19-related deaths could be up to around 2.5 times higher than the number of confirmed deaths.

The earliest nationwide serology report [20] estimated that around 10-15 million Iranians (11.9%-17.9% of the population) were infected by late spring. Later in mid-summer, a statement from the then deputy minister of research and technology of MoHME was released by the press [21] that approximately 25 million (29.8% of the population) have so far been exposed to the virus based on recent

seroepidemiological analyses. In addition, a detailed antibody seroprevalence report from 18 cities conducted between 17 April to 2 June 2020 showed greatly varied levels of exposure in the general population during this period. This report also showed that 17.1% (95%CI: 14.6 - 19.5) of the population in these cities were infected by the end of April 2020. The only systematic province-level was conducted in Gilan during April 2020 showed that the adjusted prevalence of SARS-CoV-2 seropositivity was 22.2% (95%CI: 16.4 - 28.5) at the time. This is aligned with our estimated exposure levels for Gilan based on the seasonal excess mortality trends in spring, i.e., 28% (95%CI: 19 - 36). Also, our provincial IFR estimates ranging from lowest, 0.13%, in Sistan and Baluchistan to highest, 0.36%, in Gilan are well-aligned with those of many other low-income countries [28].

In our earlier analysis [13], we reconstructed the early transmission dynamics of the Iranian epidemic and projected that as the second peak emerges in summer there will be a total of 9 million individuals (10.7% of the population) recovered by 14 July 2020 which is very close to our current estimate, 10.3%, for the nationwide exposure levels based on excess mortality data by 20 June 2020. Although we cannot independently verify the validity of any of the nationwide seroprevalence studies released to the press, they seem to be in general agreement with our estimates based on seasonal excess mortality data during winter and spring. Assuming that the ratio between excess mortality and confirmed COVID-19-related deaths remains the same over time (i.e. approximately 2.5), by 25 October 2020 there would have been a total of approximately 84 thousand COVID-19-related deaths (based on the 32,616 confirmed deaths at the time) and 33 million Iranians exposed to SARS-CoV-2 which is very close to the 35 million estimate based on the nationwide seroprevalence studies at the time [22].

We also demonstrated that the mean excess mortality rate across the country has increased by more than tenfold since the start of the epidemic and that all provinces had high mortality rates for at least one season as a consequence of the COVID-19 epidemic. In particular, most of the provinces that had a significant number of excess deaths in more than one season, experienced a surge in the death toll following the initial surge back in winter and/or spring which indicates that the epidemic was never fully brought under control. Given the record-high number of confirmed COVID-19-related deaths during fall 2020, we predict that the excess mortality counts will continue to increase across the country once NOCR updates the data for fall 1399 SH.

We previously showed, through an epidemiological and phylogenetic analysis on individuals with a travel link to Iran, that there were significant levels of under-reporting of cases and deaths, particularly at the early stages of the epidemic which started back in mid- to late-January [11]. By comparing the mortality trends in provinces with significant number of recorded deaths both in fall (13 out of 31 provinces) and winter (5 out of 31 provinces) 2019-20, we determined that the unusually high mortality rates from 23 September to 21 December 2019 were not linked to any potential cryptic transmission of the virus across the country. More detailed investigation into the causes of deaths and time-resolved monitoring for acute respiratory infections are required to determine the key source(s) of the spike during this period.

7.4.1 Limitations

While COVID-19 is directly responsible for the majority of excess deaths during the epidemic, it is possible that not all excess deaths are directly attributable to infection from the disease [17, 30]. We also note that NOCR's registration system might be strained when facing a potentially massive backlog of uncounted death certificates and prone to temporary under-counting. Our earlier analysis shows that deaths due to other causes such as traffic injuries have only dropped by a few hundred compared to previous years (data can be obtained from the Iranian Legal Medicine Organization [18]) and is not a major contributor to the changing trend in excess mortality during the epidemic [19].

Given the limited number of data points per year, i.e. one data point for every season, we avoid using more complex models such as the Generalised Linear Models (GLMs) which may demand more parameters to fit the data. In particular, the issue with using GLMs for our analysis is that such models make no attempt to remove redundant parameters from the nonlinear terms. As a result, fitting such models to only a few time points risks overparametrisation. However, GLMs are desirable for datasets with high temporal resolution (e.g., when excess mortality is recorded on a weekly or biweekly basis). Nevertheless, using a standard 5-year average to calculate the expected number of deaths may also bias our estimates as it does not correctly account for significant changes in the excess mortality trend over time due to, for instance, a changing population size or other secular trends that are present in NOCR's data. We also note that although some provinces show a seasonal trend in mortality with higher death rates during fall and winter compared to spring and summer, such patterns are not readily detectable for every province (i.e. variations during each season are high). Therefore, we also avoid using a

periodic regression model as it may bias our estimates for expected seasonal deaths [12]. Such wide variations during each season also makes it hard to find a baseline for the expected seasonal mortality in each province. We note that to estimate the mean IFR for each province, we re-weighted the IFR of each age group by its respective share of the total population in each province. This procedure relies on the assumption that contagion was homogeneous between age groups. We made this assumption because neither the all-cause mortality data from NOCR nor the reported daily cases and deaths from MoHME provide any information about the age distribution of COVID-19 infections. We also do not have any information about the social contact patterns across the country. Thus, this assumption may be partly violated due to reasons such as the heterogeneity in the burden of the epidemic in nursing homes where a subgroup of the population are interacting more frequently among themselves. However, as outlined in the methods section, we can correct for such heterogeneities by adjusting the age-specific IFR estimates [14]. In addition, a recent analysis shows that the increase in mortality by age from COVID-19 strongly resembles the age pattern of all-cause mortality [27]. On the other hand, potential age dependency of susceptibility to infection by SARS-CoV-2, may result in certain age-groups becoming more prone to getting infected [26].

7.5 Methods

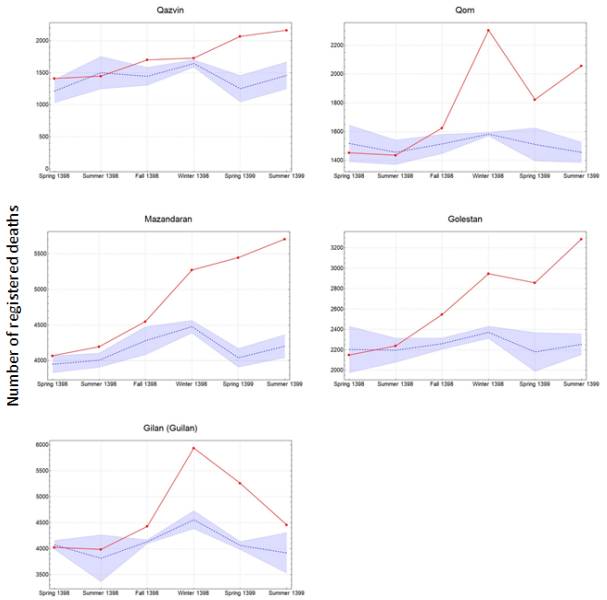
To calculate the expected number of deaths in one season of a particular year for one province we apply a linear regression model using the reported deaths from that particular season starting from the year 1394 SH until one year prior to the year of interest and consider anything above 2 standard deviations units from the mean (95% confidence interval) to be a significant level of excess mortality. We use the age demographics of each province (Statistical Center of Iran [29]) to calculate their

population-weighted IFR [14]. For a province with a total population size, P, this is given by $\mathrm{IFR} = \sum_i^M w_i \, \mathrm{IFR}_i$; where *M* is the total number of age-groups, IFR_i is the infection fatality rate of age-group *i* with population size p_i, and $w_i = p_i/P$. Thus, assuming homogenous social contact patterns across the population, all the deaths attributable to COVID-19 are weighted proportionally among different age-groups according to their respective population size. To find the number of individuals exposed to the virus in each province, *Z*, we took our estimated number of COVID-19-associated deaths based on excess mortality, *X*, and divide that by the IFR, *Y*, such that the corresponding variance (Var) in *Z* is given by

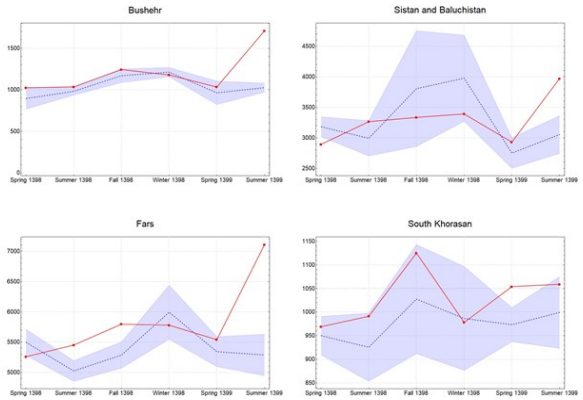
$$Var(Z) = Var(X/Y) = E(X^2)E(1/Y^2) - E^2(X)E^2(1/Y)$$

where *X* is a normally distributed number, and the standard deviation of *Y* is given by $\sigma_Y = \sqrt{\sum_i^M w_i^2 \sigma_i^2}$. The standard deviation of each age-group, *i*, is computed using the Delta method and is given by $\sigma_i = (95\%\ upper\ limit - 95\%\ lower\ limit) \,/\, z_{1-\alpha/2}$ where $z_{1-\alpha/2}$is the $(1-\alpha/2)$ quantile of the standard normal distribution. Thus, the inverse of *Y* is a random number drawn from the reciprocal normal distribution.

A


Qazvin
Qom
Mazandaran
Golestan
Gilan (Guilan)
Number of registered deaths

B


Bushehr
Sistan and Baluchistan
Fars
South Khorasan

C

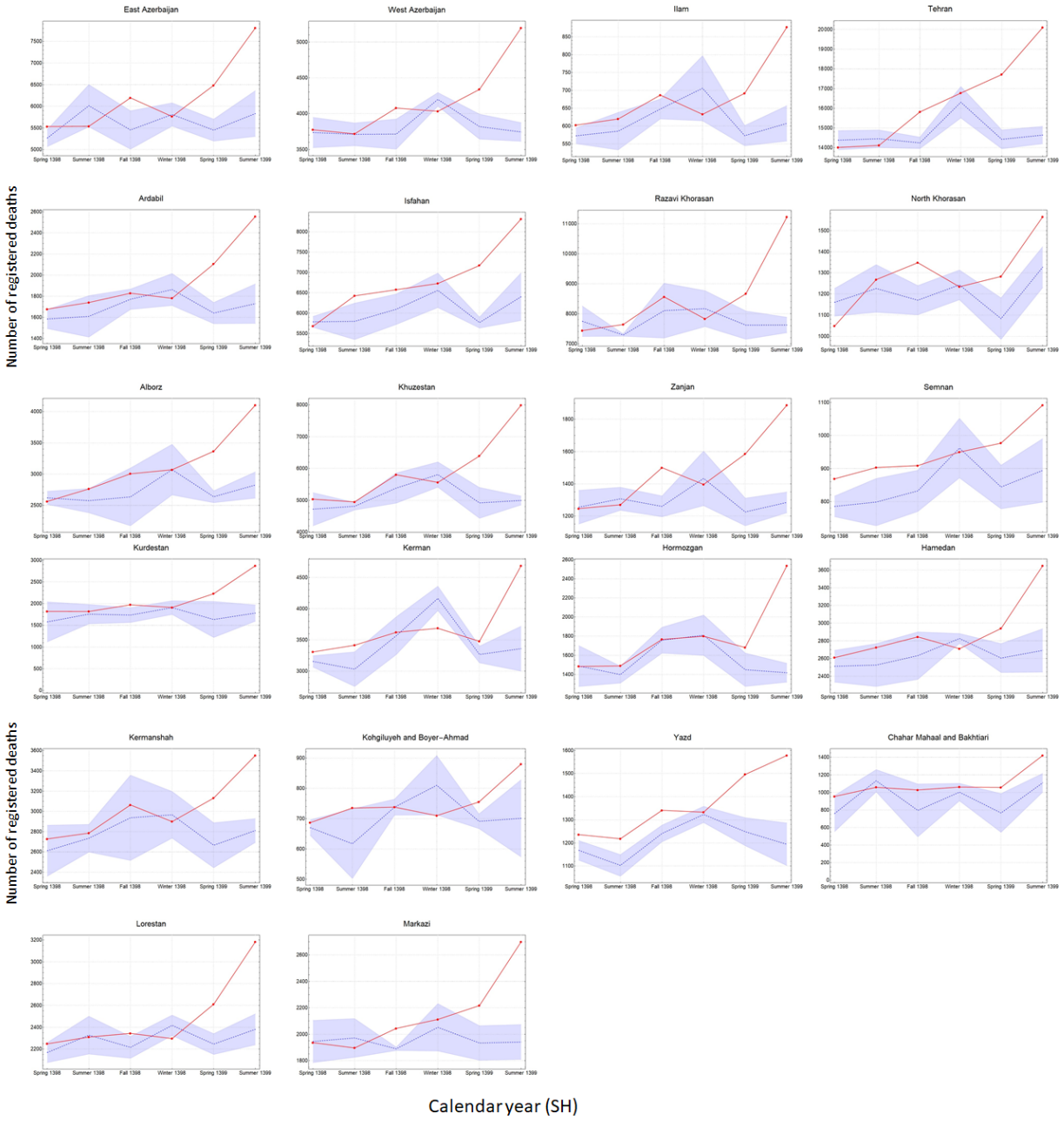
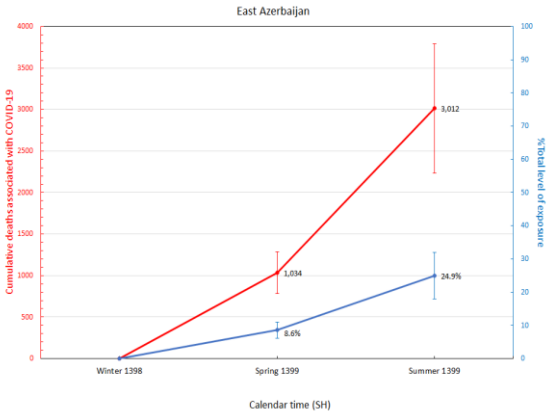
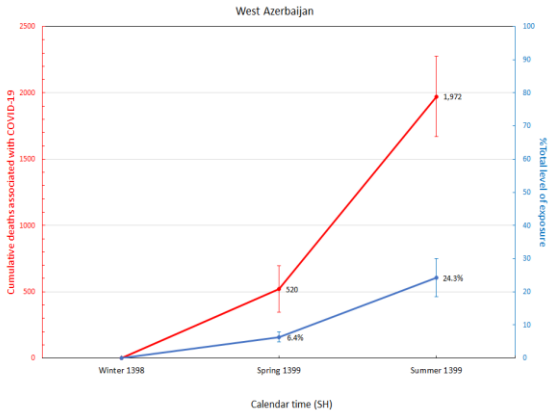
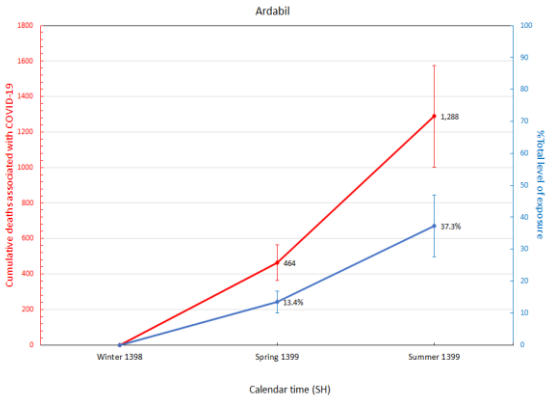
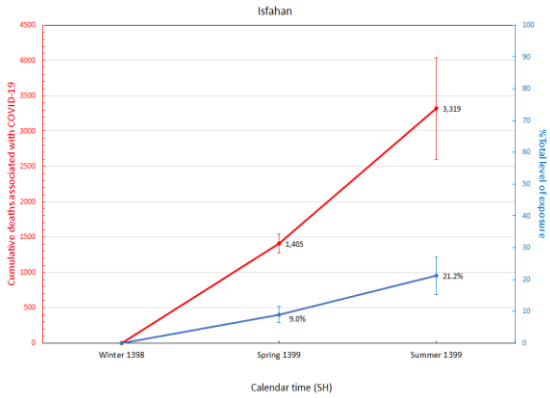


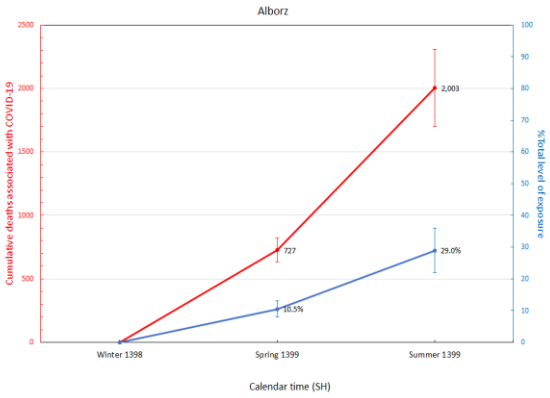
Figure 7.3**: The pattern of excess mortality from spring 1398 SH (21 March 2019) to summer 1399 SH (21 September 2020)**. **(A)** 5 provinces that experienced significant levels of excess mortality from winter 1398 SH to summer 1399 SH, **(B)** 4 provinces that experienced high mortality rates in either spring or summer 1399 SH, and **(C)** the 22 remaining provinces that experienced a growing excess mortality trend from spring 1399 SH to summer 1399 SH.

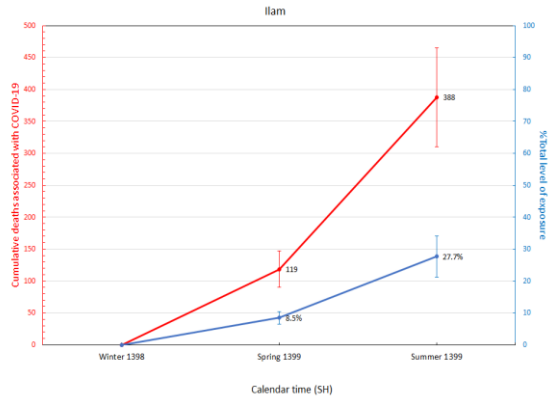

East Azerbaijan
Cumulative deaths associated with COVID-19
%Total level of exposure
3,012
1,034
24.9%
8.6%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

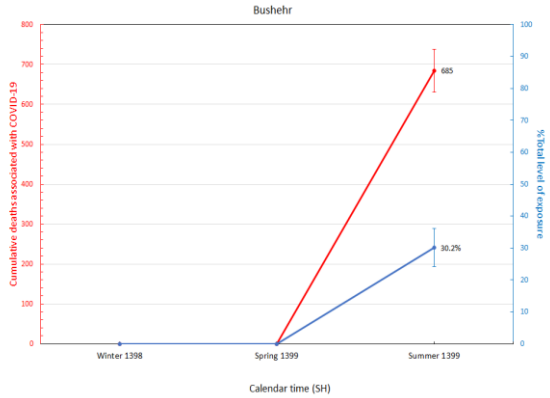

West Azerbaijan
Cumulative deaths associated with COVID-19
%Total level of exposure
1,972
520
24.3%
6.4%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

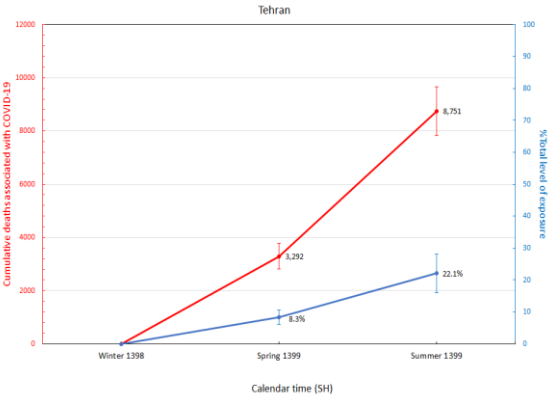

Ardabil
Cumulative deaths associated with COVID-19
%Total level of exposure
1,288
464
37.3%
13.4%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

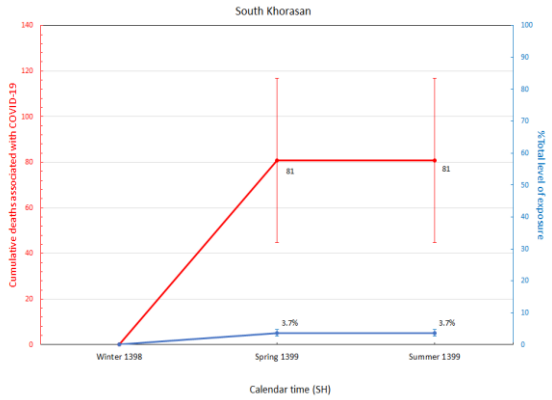

Isfahan
Cumulative deaths associated with COVID-19
%Total level of exposure
3,319
1,405
21.2%
9.0%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

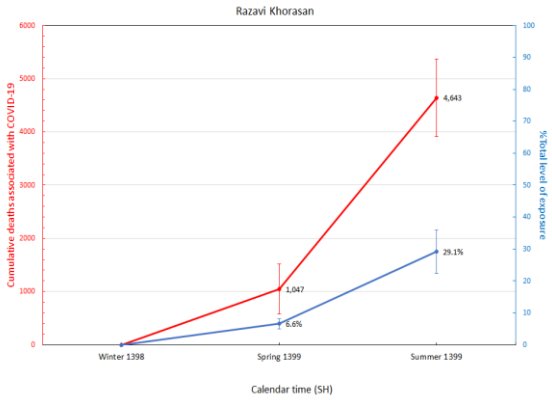

Alborz
Cumulative deaths associated with COVID-19
%Total level of exposure
2,003
727
29.0%
10.5%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

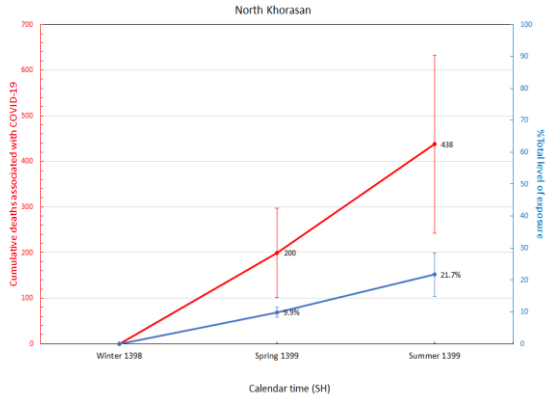

Ilam
Cumulative deaths associated with COVID-19
%Total level of exposure
388
119
27.7%
8.5%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

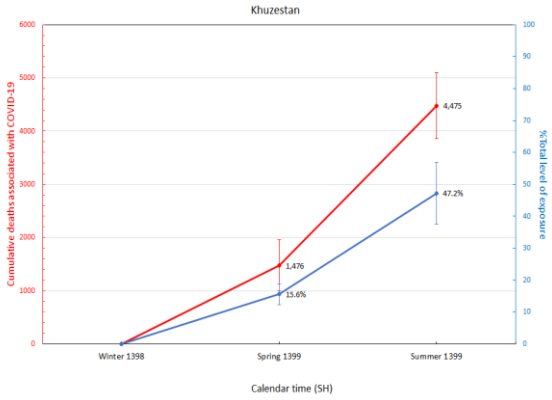

Bushehr
Cumulative deaths associated with COVID-19
%Total level of exposure
685
30.2%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

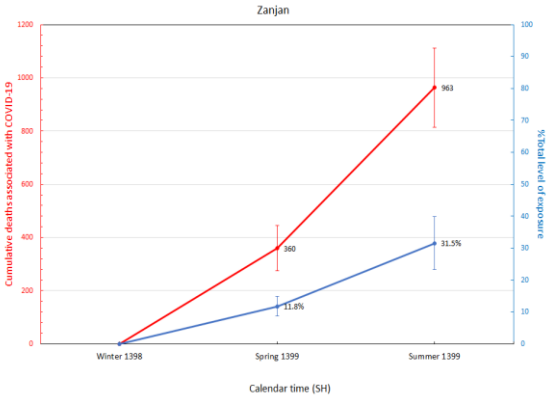

Tehran
Cumulative deaths associated with COVID-19
%Total level of exposure
8,751
3,292
22.1%
8.3%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

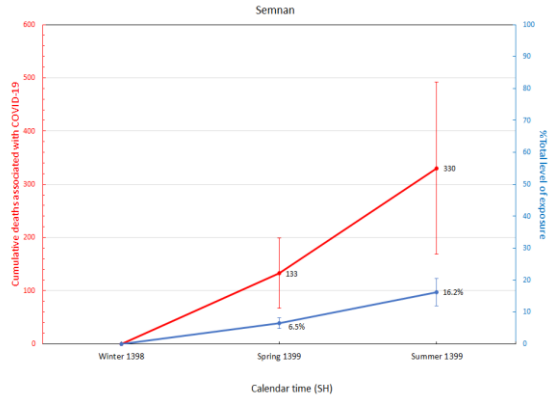

South Khorasan
Cumulative deaths associated with COVID-19
%Total level of exposure
81
81
3.7%
3.7%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

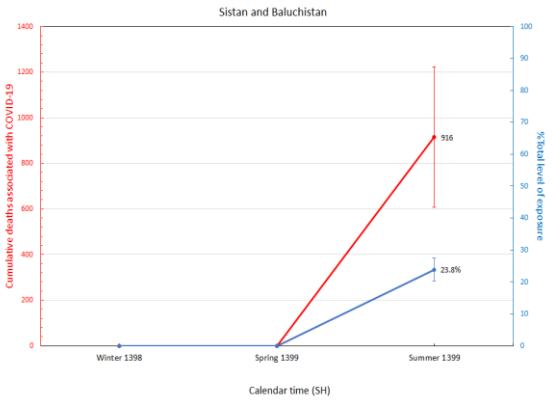

Razavi Khorasan
Cumulative deaths associated with COVID-19
%Total level of exposure
4,643
1,047
29.1%
5.6%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

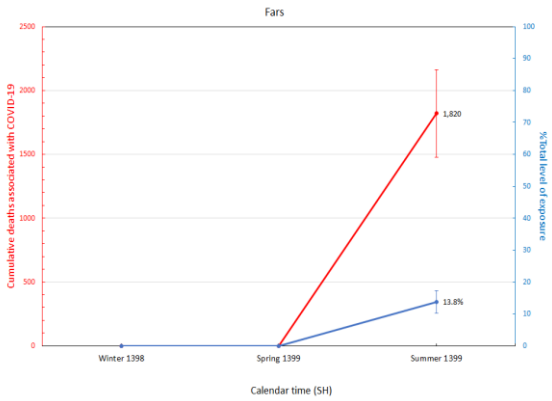

North Khorasan
Cumulative deaths associated with COVID-19
%Total level of exposure
438
200
21.7%
9.9%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

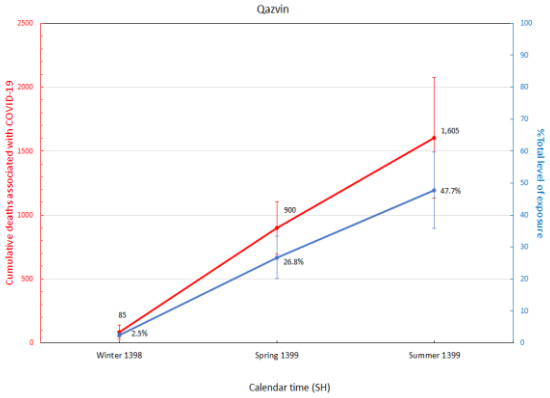

Khuzestan
Cumulative deaths associated with COVID-19
%Total level of exposure
4,475
1,476
47.2%
15.6%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

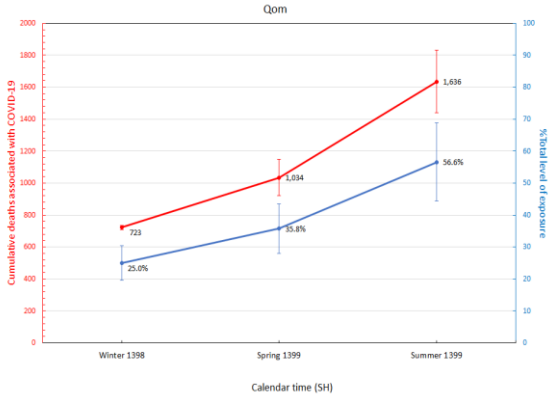

Zanjan
Cumulative deaths associated with COVID-19
%Total level of exposure
963
360
31.5%
11.8%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

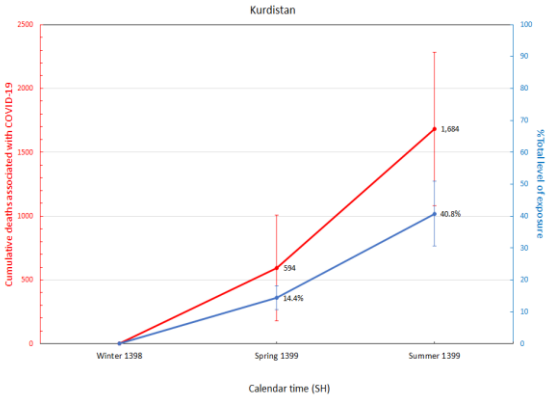

Semnan
Cumulative deaths associated with COVID-19
%Total level of exposure
330
133
16.2%
6.5%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

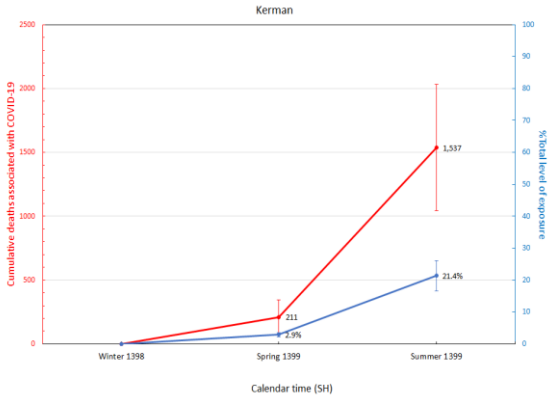

Sistan and Baluchistan
Cumulative deaths associated with COVID-19
%Total level of exposure
916
23.8%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

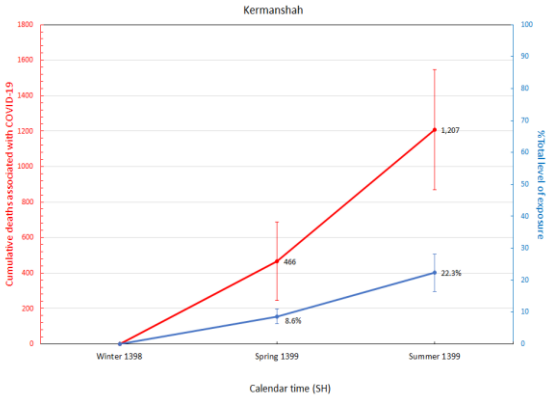

Fars
Cumulative deaths associated with COVID-19
%Total level of exposure
1,820
13.8%
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)

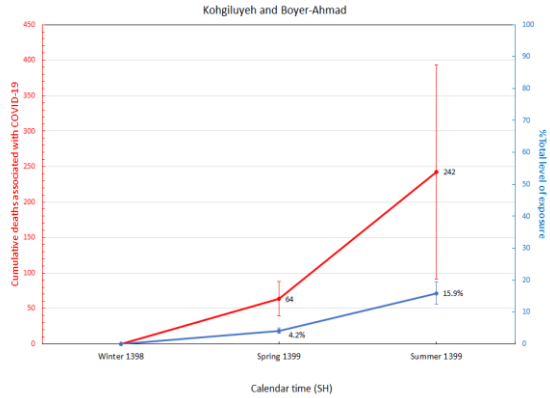

Qazvin
Cumulative deaths associated with COVID-19
% Total level of exposure
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)
85
900
1,605
2.5%
26.8%
47.7%

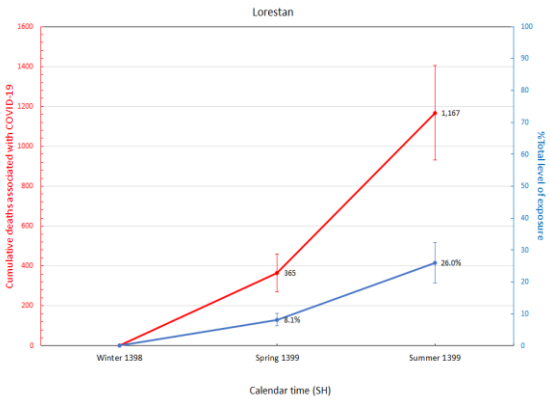

Qom
Cumulative deaths associated with COVID-19
% Total level of exposure
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)
723
1,034
1,636
25.0%
35.8%
56.6%

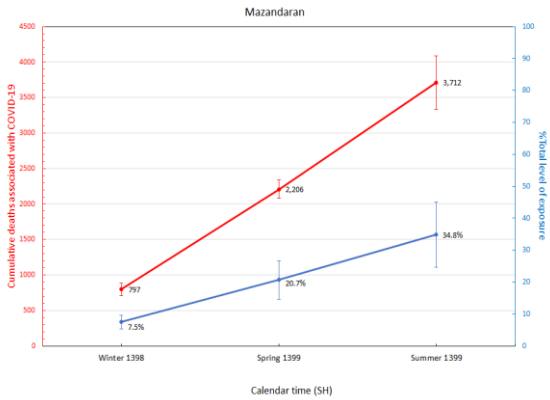

Kurdistan
Cumulative deaths associated with COVID-19
% Total level of exposure
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)
594
1,684
14.4%
40.8%


Kerman
Cumulative deaths associated with COVID-19
% Total level of exposure
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)
211
1,537
2.9%
21.4%


Kermanshah
Cumulative deaths associated with COVID-19
% Total level of exposure
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)
466
1,207
8.6%
22.3%


Kohgiluyeh and Boyer-Ahmad
Cumulative deaths associated with COVID-19
% Total level of exposure
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)
64
242
4.2%
15.9%


Lorestan
Cumulative deaths associated with COVID-19
% Total level of exposure
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)
365
1,167
8.1%
26.0%


Mazandaran
Cumulative deaths associated with COVID-19
% Total level of exposure
Winter 1398
Spring 1399
Summer 1399
Calendar time (SH)
797
2,206
3,712
7.5%
20.7%
34.8%

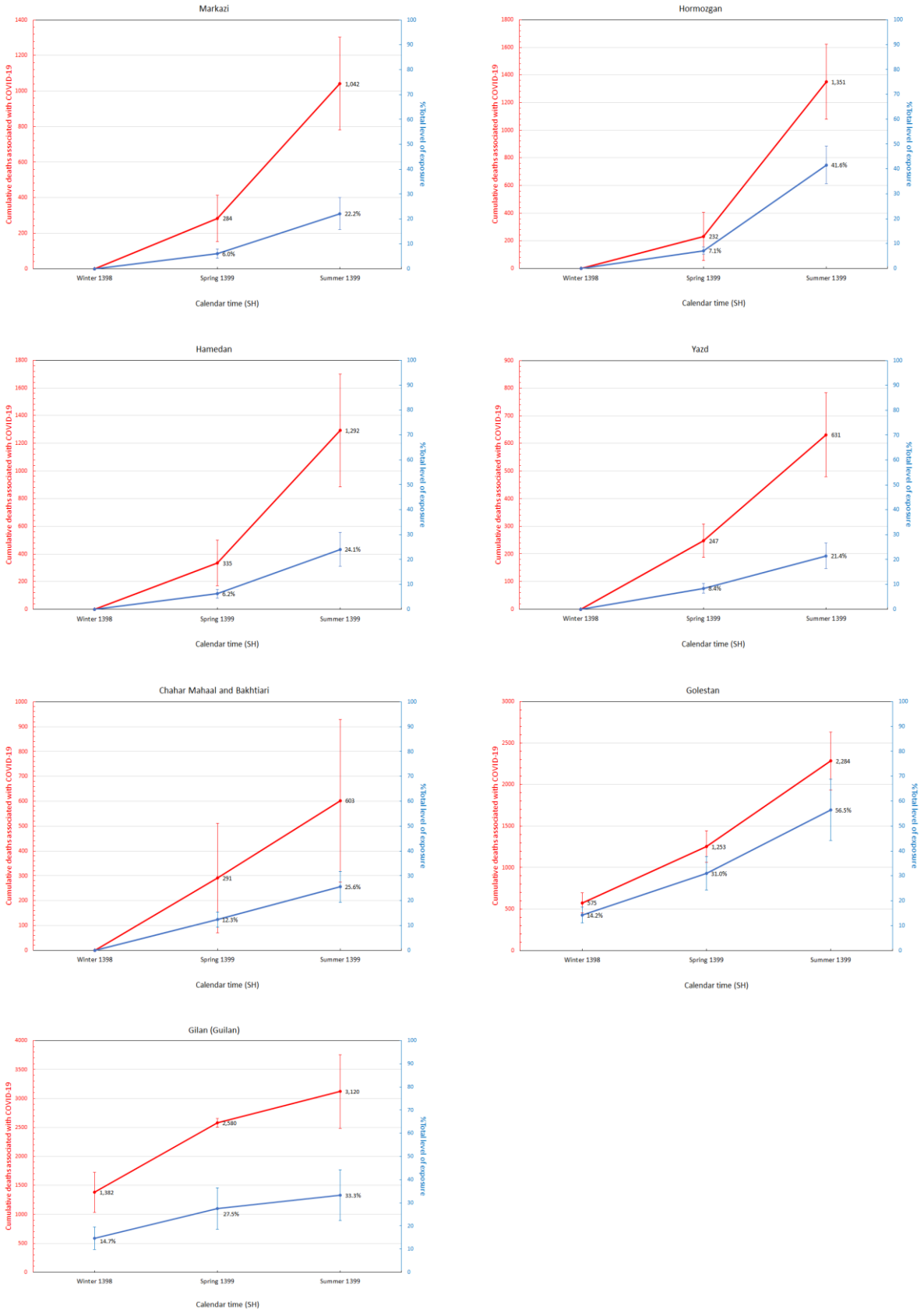


Figure 7.4**: Cumulative number of COVID-19-related deaths and percentage of people exposed to the virus across all provinces.** Cumulative deaths (red) and cumulative fraction of individuals exposed to SARS-CoV-2 (blue) in each province from winter 1398 SH to summer 1399 SH.

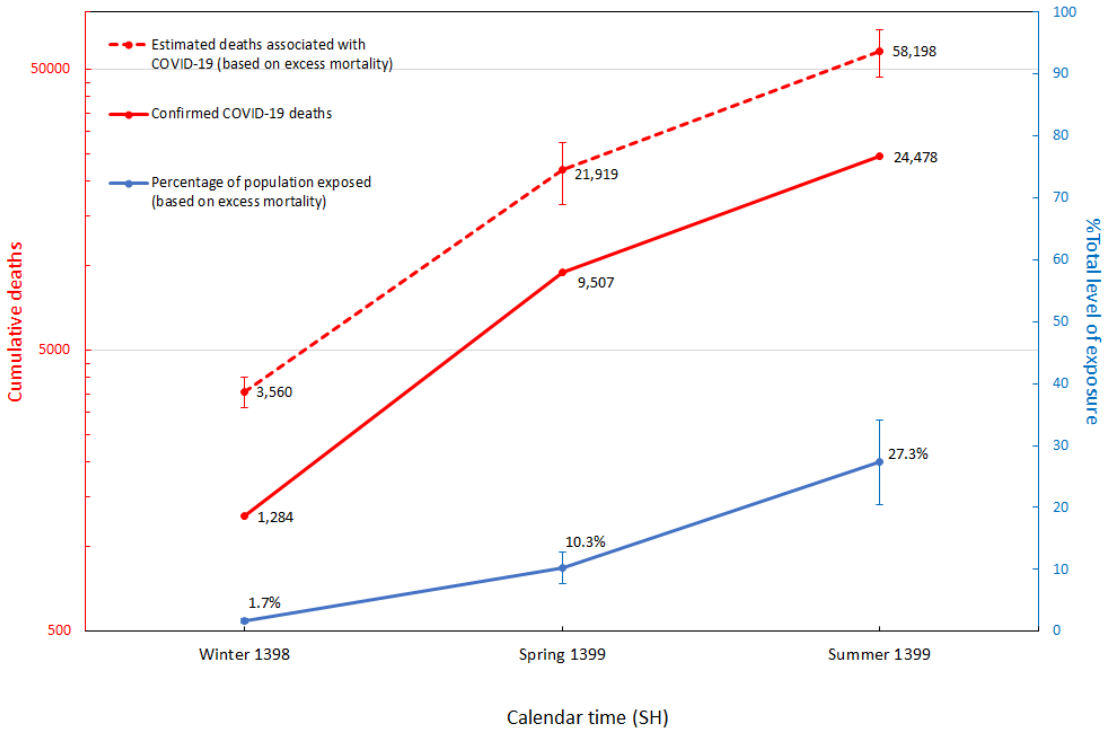


Figure 7.5**: Estimated number of deaths and exposures to COVID-19 based on excess mortality data from winter 1398 SH to summer 1399 SH.** Confirmed COVID-19 deaths are collected from the daily reports by MoHME.

7.6 References

1. Checchi F, Roberts L. Interpreting and using mortality data in humanitarian emergencies. Humanitarian Practice Network. 2005;52. Available: https://www.odi.org/sites/odi.org.uk/files/odi-assets/events-presentations/776.pdf

2. United Nations Statistics Division. Demographic and Social Statistics. [cited 26 Nov 2020]. Available: https://unstats.un.org/unsd/demographic-social/crvs/

3. National Organization for Civil Registration of the Islamic Republic of Iran. Available: https://www.amar.org.ir

4. Khosravi A, Taylor R, Naghavi M, Lopez AD. Mortality in the Islamic Republic of Iran, 1964-2004. Bull World Health Organ. 2007;85: 607–614.

5. Khosravi A, Taylor R, Naghavi M, Lopez AD. Differential mortality in Iran. Popul Health Metr. 2007;5: 7.

6. Rasambainarivo F, Rasoanomenjanahary A, Rabarison JH, Ramiadantsoa T, Ratovoson R, Randremanana R, et al. Monitoring for outbreak associated excess mortality in an African city: Detection limits in Antananarivo, Madagascar. Int J Infect Dis. 2020. doi:10.1016/j.ijid.2020.11.182

7. Ghafari M, Madani K. In search of the murder: Making sense of Iran's reported deaths. In: Medium [Internet]. 25 May 2020 [cited 27 Nov 2020]. Available: https://medium.com/@kavehmadani/in-search-of-the-murder-making-sense-of-irans-reported-deaths-4279d2b03175

8. Kadivar A. Corona's footsteps in civil registration statistics [Persian]. 18 May 2020 [cited 27 Nov 2020]. Available: https://blog.statsminute.ir/%d8%b1%d8%af-%d9%be%d8%a7%db%8c-%da%a9%d8%b1%d9%88%d9%86%d8%a7-%d8%af%d8%b1-%d8%a2%d9%85%d8%a7%d8%b1-%d8%ab%d8%a8%d8%aa-%d8%a7%d8%ad%d9%88%d8%a7%d9%84/

9. Lu J, du Plessis L, Liu Z, Hill V, Kang M, Lin H, et al. Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. Cell. 2020;181: 997–1003.e9.

10. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020;26: 450–452.

11. Ghafari M, Hejazi B, Karshenas A, Dascalu S, Ferretti L, Ledda A, et al. Ongoing outbreak of COVID-19 in Iran: challenges and signs of concern. medRxiv. 2020. Available: https://www.medrxiv.org/content/10.1101/2020.04.18.20070904v1.

12. Tadbiri H, Moradi-Lakeh M, Naghavi M. All-cause excess mortality and COVID-19-related deaths in Iran. Med J Islam Repub Iran. 2020. Available: http://mjiri.iums.ac.ir/browse.php?a_code=A-10-6056-1&slc_lang=en&sid=1

13. Ghafari M, Hejazi B, Karshenas A, Dascalu S, Ferretti L, Ledda A, et al.

Ongoing outbreak of COVID-19 in Iran: challenges and signs of concern with under-reporting of prevalence and deaths. Research Square. 2020. Available: https://www.researchsquare.com/article/rs-32500/latest.pdf

14. O'Driscoll M, Dos Santos GR, Wang L, Cummings DAT, Azman AS, Paireau J, et al. Age-specific mortality and immunity patterns of SARS-CoV-2. Nature. 2020. doi:10.1038/s41586-020-2918-0

15. Buss LF, Prete CA Jr, Abrahim CMM, Mendrone A Jr, Salomon T, de Almeida-Neto C, et al. COVID-19 herd immunity in the Brazilian Amazon. medRxiv. 2020. doi:10.1101/2020.09.16.20194787

16. Hallal PC, Hartwig FP, Horta BL, Silveira MF, Struchiner CJ, Vidaletti LP, et al. SARS-CoV-2 antibody prevalence in Brazil: results from two successive nationwide serological household surveys. The Lancet Global Health. 2020;8: e1390–e1398.

17. Viglione G. THE TRUE TOLL OF THE PANDEMIC. Nature. 2020; 22–24.

18. Iranian Legal Medicine Organization. Motor Vehicle Crash Statistics. Tehran, Iran. In: lmo.ir [Internet]. 2020 [cited 1 Dec 2020]. Available: https://www.lmo.ir/web_directory/53999-%D8%AA%D8%B5%D8%A7%D8%AF%D9%81%D8%A7%D8%AA.html

19. Kadivar A. Increase in the number of registered deaths during spring [Persian]. In: statsminute.ir [Internet]. 7 Jul 2020 [cited 29 Nov 2020]. Available: https://blog.statsminute.ir/%D8%A7%D9%81%D8%B2%D8%A7%DB%8C%D8%B4-%D8%A7%D8%B3%D9%86%D8%A7%D8%AF-%D9%81%D9%88%D8%AA-%D8%AB%D8%A8%D8%AA-%D8%A7%D8%AD%D9%88%D8%A7%D9%84-%D8%AF%D8%B1-%D8%A8%D9%87%D8%A7%D8%B1-99/

20. National Committee on COVID-19 Epidemiology. Summary of the scattered seroepidemiological reports from across the country [Persian]. Ministry of Health and Medical Education. 15 May 2020. Available: behdasht.gov.ir

21. Zeinalipour M. Serology tests; the basis for the report that 25 million Iranians have been infected to Corona [Persian]. In: ISNA News Agency [Internet]. 19 Jul 2020 [cited 29 Nov 2020]. Available: https://www.isna.ir/news/99042921012/%D8%AA%D8%B3%D8%AA%D9%87%D8%A7%DB%8C-%D8%B3%D8%B1%D9%88%D9%84%D9%88%DA%98%DB%8C-%D9%85%D8%A8%D9%86%D8%A7%DB%8C-%D8%A2%D9%85%D8%A7%D8%B1-%DB%B2%DB%B5-%D9%85%DB%8C%D9%84%DB%8C%D9%88%D9%86%DB%8C-%D8%A7%D8%A8%D8%AA%D9%84%D8%A7%DB%8C-%D8%A7%DB%8C%D8%B1%D8%A7%D9%86%DB%8C%D8%A7%D9%86-%D8%A8%D9%87-%DA%A9%D8%B1%D9%88%D9%86%D8%A7

22. New statistics on Coronavirus infections in Iran 35 million people! [Persian]. In: Hamshahri Online [Internet]. 24 Oct 2020 [cited 29 Nov 2020]. Available: https://www.hamshahrionline.ir/news/559881/%D8%A2%D9%85%D8%A7%D

8%B1-%D8%AC%D8%AF%DB%8C%D8%AF-
%D9%85%D8%A8%D8%AA%D9%84%D8%A7%DB%8C%D8%A7%D9%86-
%DA%A9%D8%B1%D9%88%D9%86%D8%A7-%D8%AF%D8%B1-
%D8%A7%DB%8C%D8%B1%D8%A7%D9%86-%DB%B3%DB%B5-
%D9%85%DB%8C%D9%84%DB%8C%D9%88%D9%86-
%D9%86%D9%81%D8%B1-%DB%B5%DB%B0-
%D9%85%DB%8C%D9%84%DB%8C%D9%88%D9%86

23. Shakiba M, Nazari SSH, Mehrabian F, Rezvani SM, Ghasempour Z, Heidarzadeh A. Seroprevalence of COVID-19 virus infection in Guilan province, Iran. medRxiv. 2020. Available: https://www.medrxiv.org/content/10.1101/2020.04.26.20079244v1.abstract

24. Bobrovitz N, Arora RK, Cao C, Boucher E, Liu M, Rahim H, et al. Global seroprevalence of SARS-CoV-2 antibodies: a systematic review and meta-analysis. medRxiv. 2020. Available: https://www.medrxiv.org/content/10.1101/2020.11.17.20233460v2.full-text

25. Meyerowitz-Katz G, Merone L. A systematic review and meta-analysis of published research data on COVID-19 infection fatality rates. International Journal of Infectious Diseases. 2020;101: 138–148.

26. Davies NG, Klepac P, Liu Y, Prem K, Jit M, Eggo RM, et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. Nat Med. 2020;26: 1205–1211.

27. Goldstein JR, Lee RD. Demographic perspectives on the mortality of COVID-19 and other epidemics. Proc Natl Acad Sci U S A. 2020;117: 22035–22041.

28. Brazeau N, Verity R, Jenks S, Fu H, Whittaker C, Winskill P, et al. Report 34: COVID-19 infection fatality ratio: estimates from seroprevalence. 2020. Available: https://dsprdpub.cc/handle/10044/1/83545

29. Statistical Center of Iran. In: amar.org [Internet]. [cited 29 Nov 2020]. Available: https://www.amar.org.ir/english

30. Beaney T, Clarke JM, Jain V, Golestaneh AK, Lyons G, Salman D, Majeed A. Excess mortality: the gold standard in measuring the impact of COVID-19 worldwide?. Roy Soc Med. 2020; 113(9):329-34.

Part b: A framework for reconstructing SARS-CoV-2 transmission dynamics using excess mortality data

7b.1 Abstract

The transmission dynamics and burden of SARS-CoV-2 in many regions of the world is still largely unknown due to the scarcity of epidemiological analyses and lack of testing to assess the prevalence of disease. In this work, we develop a quantitative framework based on excess mortality data to reconstruct SARS-CoV-2 transmission dynamics and assess the level of underreporting in infections and deaths. Using weekly all-cause mortality data from Iran, we are able to show a strong agreement between our attack rate estimates and seroprevalence measurements in each province and find significant heterogeneity in the level of exposure across the country with 11 provinces reaching near 100% attack rates. Despite having a young population, our analysis reveals that incorporating limited access to medical services in our model, coupled with undercounting of COVID-19-related deaths, leads to estimates of infection fatality rate in most provinces of Iran that are comparable to high-income countries.

7b.2 Introduction

Assessing the true burden of COVID-19 requires a comprehensive surveillance system to monitor outbreaks and a large capacity for diagnostic testing. However, this is particularly challenging for many low- and middle-income countries due to limited diagnostic capacities and access to healthcare [1-6]. An alternative approach to measuring the burden of COVID-19 is to monitor changes in all-cause mortality trends with respect to baseline levels based on historic trends (i.e., measuring excess mortality). While excess mortality during the current pandemic may not

exactly match the true number of COVID-19-related deaths due to various reasons such as disruptions in treatment for other fatal diseases and socio-economic disparities in health care in different regions of a country, several works have shown that a significant portion of excess deaths in many countries are directly attributable to COVID-19 [1, 7-11]. This enables us to use excess mortality as a proxy for estimating the number of infections and fatalities from COVID-19 and correct for under-reporting of cases and deaths that may arise due to limited testing from suspected cases and uncertainty in the number of fatalities attributable to COVID-19 [12, 13].

Iran was among the first countries outside mainland China to report a large outbreak of SARS-CoV-2 in February and March 2020 and also acted as a major source for its spread in several countries in the Middle East and elsewhere [11, 14-18]. It was also one of the first countries to experience a second wave of infection after the relaxation of non-pharmaceutical interventions (NPIs) in July 2020 [14, 19]. Since then, the country experienced three additional waves, two of which were driven by the Alpha and Delta variants of SARS-CoV-2, in May and September 2021. Despite a recent increase in the number of daily administered vaccine doses, Iran had a very slow start on its immunisation programme with only ~3% of the population being fully vaccinated by the end of July 2021, during the surge in cases with the Delta variant [20]. Because the Ministry of Health and Medical Education (MoHME) of Iran stopped releasing province-level data on the number of confirmed COVID-19 cases and deaths from 2020-03-22, the transmission dynamics across the country have remained largely unknown. One way to estimate the number of infections and account for under-reporting of cases is by converting deaths attributable to COVID-19 to infections using a population-weighted infection fatality

ratio (IFR) estimate (i.e., infections = deaths / IFR), taking into account the demographics of the population of interest [21]. However, such estimation of attack rates can be inaccurate or misleading if it does not take into account the impact of limited access to health care in municipalities with low socioeconomic status, uneven adoption of pharmaceutical and non-pharmaceutical intervention in different areas, and the increased chance of re-infection over time. In particular, it is clear from the SARS-CoV-2 epidemics arounds the world during 2020 and 2021 that the demand for intensive care beds and mechanical ventilators can place a severe strain on health systems. This has likely more profound consequences for low-income settings where the availability and quality of healthcare and related resources is typically limited [22]. Neglecting to account for these factors can result in an overestimate of SARS-CoV-2 attack rates and underestimation its potential for continued transmission, with immediate implications for assessments of the potential burden of SARS-CoV-2 variants of concern such as Omicron in a population.

In this work, we use the newly updated province-level age-stratified weekly all-cause mortality data from the National Organization for Civil Registration (NOCR) of Iran and develop a mathematical framework to fully reconstruct the Iranian epidemic with the aim to estimate the attack rate, number of daily hospital admissions, deaths, and re-infection rates of SARS-CoV-2 across the country.

7b.3 Results

We obtain age-stratified weekly all-cause mortality data from NOCR and apply a statistical model to calculate the weekly excess mortality for all 31 provinces and across 17 age-groups during the Iranian epidemic (see Methods section). Figure

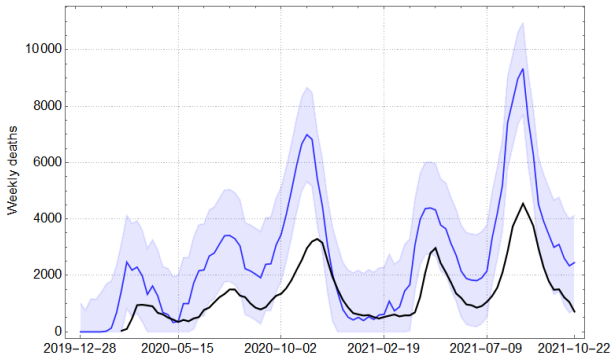
7.6A shows 5 distinct peaks in excess mortality with a temporal trend that is strongly associated with the nationwide reported COVID-19 deaths over time. Several provinces show significant levels of excess mortality from the first week of February 2020 suggesting that the Iranian epidemic likely started at least a month before that (Figure 7.6B). This is in agreement with previous phylogenetic and epidemiological studies suggesting the epidemic started in late December 2019 to early January 2020 [14, 23]. Furthermore, during the first two waves of the Iranian epidemic, the excess mortality was roughly 2.5 times higher than reported COVID-19 fatalities. However, this dropped to 2.1 at the later stages suggesting that the testing capacity of the country to record infections and deaths gradually improved over time, but not to the extent to fully capture the majority of COVID-19-related deaths. This is also in agreement with statements from the Iranian health officials suggesting that Iran's COVID-19-related deaths could be twice the official numbers and also previous studies showing elevated levels of under-reporting of COVID-19 fatalities in Iran, particularly in the early stages of the pandemic [14, 15]. We also find that the total nationwide excess deaths by 2021-10-22 is 259,641 (95% CI: 230,493 - 288,790) which is nearly twice the 125,000 reported COVID-19 deaths at the time.

Figure 7.6B shows the intensity of excess mortality as quantified by the number of excess deaths per 100,000 persons in each province over time. During the first wave in spring 2020, only a few provinces showed significant levels of excess mortality with Gilan, Qom, Mazandaran, and Golestan among the hardest-hit provinces. This is likely because these provinces usually attract a large number of tourists and pilgrims around the New Year's holiday in February/March which led to large superspreading events and uncontrolled outbreaks due to limited NPIs before the national lockdown on 2020-03-05 [14]. As the epidemic progressed, the timing

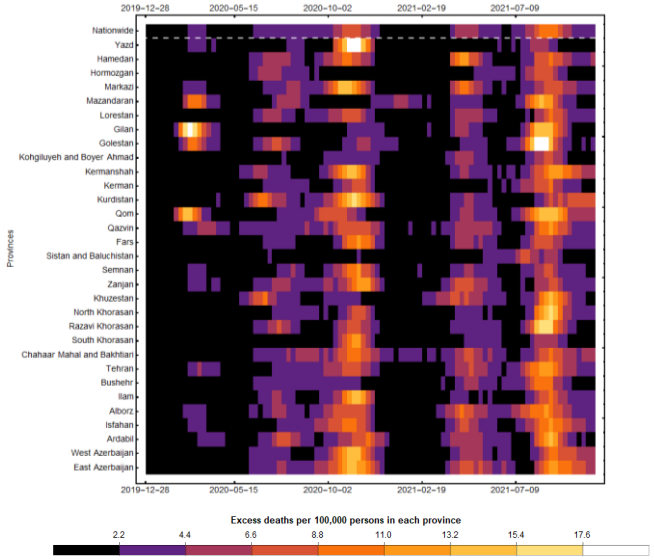
of the peak excess mortality between different provinces were more simultaneous (Figure 7.6B; Figure S1). Most notably, during the third wave in October/November 2020 and the fifth wave in August/September 2021 with the Delta variant of SARS-CoV-2, almost all provinces experienced significant number of fatalities around the same time.

We also carry out a similar analysis to estimate the temporal pattern of COVID-19 fatalities for different age-groups (Figure 7.6C). We find that the majority of excess deaths are concentrated in the older age categories (>55) and that the third and fifth epidemic waves had the largest impact on younger age groups. Figure 7.6D shows the per capita ratio of excess mortality per age group with respect to the 40–49 age group. The reason for choosing the 40-49 age group as a baseline is because, unlike younger age groups, excess mortality in this age group is subject to relatively low fluctuations over time due to experiencing more deaths. It also serves as a baseline for measuring the impact of vaccination on lowering excess deaths in older age groups who have received vaccination at an earlier date (see Figure S2). We see that while there is a gradual increase in the ratio of deaths in the older age-groups during the first three waves, there is a two-fold drop in the 60-74 and 75+ age-groups before the start of the fifth wave. This is likely due to a higher vaccination coverage in these age-groups. While the exact number of vaccinated individuals per age-group in Iran is not known, given that the immunisation programme for individuals aged >75 started in May 2021 (first phase priority groups for vaccination in Iran are frontline healthcare workers as well as those aged above 80; see [20] for more information), and that Iran fully vaccinated roughly 5 times the population size of individuals aged 75 and above by 2021-09-07, it is likely that the

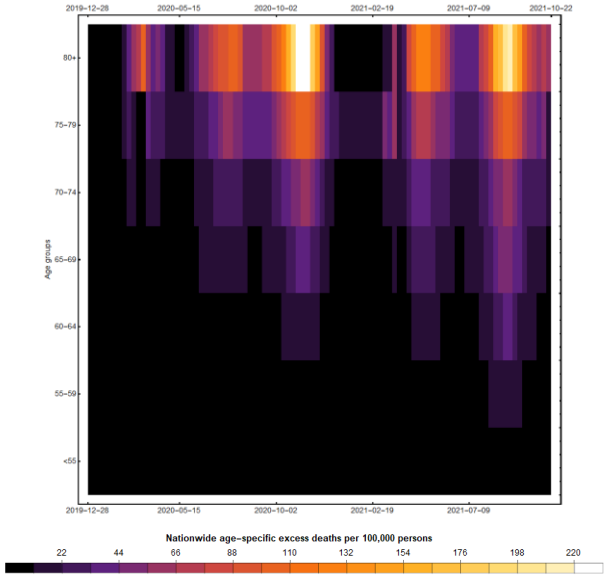
A


Weekly deaths
10000
8000
6000
4000
2000
0
2019–12–28
2020–05–15
2020–10–02
2021–02–19
2021–07–09
2021–10–22

B


Provinces
Nationwide
Yazd
Hamedan
Hormozgan
Markazi
Mazandaran
Lorestan
Gilan
Golestan
Kohgiluyeh and Boyer Ahmad
Kermanshah
Kerman
Kurdistan
Qom
Qazvin
Fars
Sistan and Baluchistan
Semnan
Zanjan
Khuzestan
North Khorasan
Razavi Khorasan
South Khorasan
Chahaar Mahal and Bakhtiari
Tehran
Bushehr
Ilam
Alborz
Isfahan
Ardabil
West Azerbaijan
East Azerbaijan
2019–12–28
2020–05–15
2020–10–02
2021–02–19
2021–07–09
Excess deaths per 100,000 persons in each province
2.2
4.4
6.6
8.8
11.0
13.2
15.4
17.6

C


Age groups
80+
75–79
70–74
65–69
60–64
55–59
<55
2019–12–28
2020–05–15
2020–10–02
2021–02–19
2021–07–09
2021–10–22
Nationwide age–specific excess deaths per 100,000 persons
22
44
66
88
110
132
154
176
198
220

D

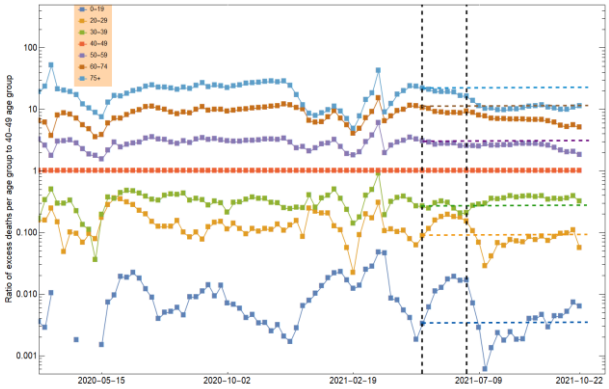


Figure 7.6: **Excess deaths per province and age group in Iran. (A)** One-month average of weekly excess mortality (blue) and reported COVID-19 deaths (black) in Iran from 2019-12-28 to 2021-10-22. Shaded area shows the 95% forecast interval in weekly excess mortality. **(B)** One-month average of weekly excess deaths per 100,000 persons per province over time. **(C)** Nationwide excess deaths per 100,000 persons per age group over time. **(D)** Ratio of weekly excess per age-group to the 40-49 age-group. Vertical dashed lines show the start of the immunisation programme for individuals aged above 75 on 2021-05-18 and above 60 on 2021-07-14. The horizontal dashed lines show the ratio of excess deaths at the start of the immunisation programme for individuals aged above 75.

drop in excess mortality in the older age-groups before and during the fifth wave is due to their high vaccination coverage.

Assuming that lack of testing is the primary reason behind the discrepancy between excess mortality and official COVID-19 deaths [1, 24, 25], we attribute the mean excess mortality during the pandemic period to COVID-19. This enables us to reconstruct the SARS-CoV-2 transmission dynamics across the country and infer the population fatality rate (PFR), IFR, and attack rate per province (see Methods section). We first estimate the national and provincial attack rates of SARS-CoV-2 under two different sets of assumptions of the IFR by age relationship and select the better-fitting model by comparing the modelled seroprevalence estimates to data (Figure S3; Data S1-S6). By tracking the number of people who require a general hospital or ICU bed over time, we were able to account for the impact of bed shortages on increasing IFR for COVID-19 patients who do not receive a bed in each province (Methods section). We find that the model with a higher IFR per age-group (and an overall lower estimated attack rate) that explicitly accounts for an increase in IFR when hospital capacity is reached provides a better fit to seroprevalence data. With a median value of approximately 0.50% (0.32-0.64%), our inferred province-level IFR is generally higher than a population-weighted IFR which only takes into account the demographics of a given province (Table 7.3; Methods section) and is roughly twice as high as the IFR estimates derived from serological studies in predominantly high-income countries [21]. This discrepancy is expected because disparities in healthcare access can be offset by the younger population age structure of many developing countries resulting in population IFRs that are similar to high-income countries [7, 26, 27]. Figure 7.7A and Figure 7.7B show a nearly 75% attack rate of SARS-CoV-2 across the country with nearly 15%

Table 7.3: **Evaluation of the burden of COVID-19 in each province of Iran as of 2020-10-22**. The overall excess mortality, PFR, estimated IFR (estimated from the transmission model), population-weighted IFR (based on age distribution of each province), and percentage of the population ever exposed to SARS-CoV-2 in each province.

Province	Excess deaths	% PFR	% Estimated IFR	% Population-weighted IFR	% Exposed
East Azerbaijan	17,035 (8,887 - 28,237)	0.438 (0.228 - 0.726)	0.50 (0.41 - 0.62)	0.41 (0.29 - 0.53)	84.92 (69.06 - 103.05)
West Azerbaijan	11,609 (6,289 - 19,204)	0.346 (0.187 - 0.573)	0.47 (0.36 - 0.57)	0.33 (0.23 - 0.43)	75.66 (60.25 - 98.53)
Ardabil	4,538 (1,974 - 8,053)	0.361 (0.157 - 0.641)	0.38 (0.31 - 0.46)	0.37 (0.26 - 0.47)	92.55 (76.19 - 112.14)
Isfahan	18,281 (9,612 - 30,412)	0.355 (0.187 - 0.591)	0.52 (0.41 - 0.64)	0.41 (0.29 - 0.53)	65.90 (51.80 - 81.46)
Alborz	11,068 (5,477 - 18,219)	0.388 (0.192 - 0.639)	0.49 (0.38 - 0.61)	0.35 (0.24 - 0.46)	64.09 (51.07 - 81.97)
Ilam	1,719 (611 - 3,743)	0.294 (0.104 - 0.640)	0.47 (0.38 - 0.56)	0.32 (0.22 - 0.42)	61.35 (51.01 - 73.74)
Bushehr	3,023 (1,241 - 5,649)	0.244 (0.100 - 0.45)	0.43 (0.31 - 0.55)	0.26 (0.17 - 0.35)	61.29 (44.63 - 82.10)
Tehran	49,320 (27,155 - 78,900)	0.365 (0.201 - 0.585)	0.47 (0.39 - 0.58)	0.40 (0.28 - 0.52)	77.06 (65.08 - 93.99)
Chahaar Mahal & Bakhtiari*	3,624 (1,089 - 7,452)	0.376 (0.113 - 0.774)	0.48 (0.37 - 0.59)	0.33 (0.23 - 0.43)	76.28 (63.01 - 101.11)
South Khorasan	1,736 (473 - 4,466)	0.210 (0.059 - 0.558)	0.50 (0.43 - 0.59)	0.36 (0.26 - 0.46)	41.27 (34.65 - 48.50)
Razavi Khorasan	20,847 (9,945 - 36,946)	0.300 (0.147 - 0.549)	0.36 (0.29 - 0.43)	0.33 (0.23 - 0.42)	90.19 (71.76 - 113.32)
North Khorasan	2,478 (985 - 5,250)	0.283 (0.112 - 0.600)	0.36 (0.27 - 0.44)	0.32 (0.22 - 0.42)	77.73 (61.94 - 100.78)
Khuzestan	15,366 (8,235 - 25,189)	0.317 (0.170 - 0.520)	0.32 (0.26 - 0.40)	0.28 (0.19 - 0.37)	98.48 (79.82 - 120.59)
Zanjan	3,938 (1,442 - 7,797)	0.367 (0.134 - 0.720)	0.50 (0.42 - 0.61)	0.38 (0.27 - 0.49)	73.11 (60.85 - 87.84)
Semnan	2,057 (742 - 4,226)	0.275 (0.099 - 0.566)	0.49 (0.43 - 0.58)	0.37 (0.26 - 0.48)	57.26 (49.85 - 65.92)
Sistan & Baluchistan	4,401 (1,104 - 12,829)	0.140 (0.036 - 0.419)	0.37 (0.25 - 0.49)	0.18 (0.12 - 0.24)	46.60 (35.38 - 67.60)
Fars	13,825 (6,361 - 25,441)	0.282 (0.130 - 0.520)	0.50 (0.40 - 0.62)	0.36 (0.25 - 0.47)	56.63 (42.64 - 71.82)
Qazvin	5,255 (1,793 - 9,699)	0.404 (0.138 - 0.747)	0.47 (0.38 - 0.59)	0.35 (0.24 - 0.45)	85.97 (69.52 - 106.44)
Qom	5,519 (3,193 - 8,925)	0.399 (0.231 - 0.646)	0.49 (0.33 - 0.62)	0.30 (0.20 - 0.39)	84.74 (66.12 - 116.36)
Kurdistan	6,421 (3,015 - 11,243)	0.394 (0.185 - 0.69)	0.47 (0.37 - 0.58)	0.34 (0.24 - 0.45)	80.05 (66.82 - 103.86)
Kerman	7,340 (2,761 - 14,621)	0.224 (0.084 - 0.446)	0.46 (0.36 - 0.57)	0.30 (0.21 - 0.40)	50.58 (40.64 - 63.22)
Kermanshah	6,844 (3,662 - 11,734)	0.356 (0.190 - 0.610)	0.51 (0.39 - 0.63)	0.38 (0.27 - 0.50)	66.67 (54.07 - 86.87)
Kohgiluyeh & Boyer Ahmad	1,417 (301 - 3,569)	0.191 (0.040 - 0.481)	0.32 (0.26 - 0.39)	0.28 (0.19 - 0.37)	59.41 (48.25 - 74.93)
Golestan	6,920 (3,138 - 12,352)	0.357 (0.162 - 0.630)	0.35 (0.27 - 0.43)	0.30 (0.20 - 0.39)	105.43 (87.31 - 125.82)
Gilan	8,788 (3,888 - 17,078)	0.362 (0.160 - 0.704)	0.64 (0.51 - 0.75)	0.50 (0.36 - 0.65)	52.62 (45.21 - 67.60)
Lorestan	4,942 (1,895 - 9,850)	0.280 (0.100 - 0.566)	0.48 (0.40 - 0.59)	0.35 (0.24 - 0.45)	56.51 (47.10 - 68.36)
Mazandaran	10,500 (4,372 - 19,563)	0.324 (0.130 - 0.603)	0.55 (0.46 - 0.64)	0.44 (0.31 - 0.57)	58.36 (48.01 - 69.78)
Markazi	5,200 (2,273 - 9,601)	0.367 (0.160 - 0.678)	0.55 (0.46 - 0.67)	0.43 (0.31 - 0.55)	64.71 (53.01 - 75.59)
Hormozgan	3,753 (1,369 - 7,864)	0.194 (0.070 - 0.407)	0.42 (0.33 - 0.54)	0.24 (0.16 - 0.32)	48.52 (38.76 - 64.10)
Hamedan	5,952 (2,868 - 10,475)	0.300 (0.168 - 0.615)	0.53 (0.44 - 0.65)	0.41 (0.29 - 0.53)	63.97 (53.45 - 76.72)
Yazd	3,287 (1,572 - 6,475)	0.270 (0.129 - 0.533)	0.35 (0.30 - 0.40)	0.33 (0.23 - 0.43)	80.50 (71.43 - 93.68)

*Year 1394 Solar Hijri is excluded from the calculation of background mortality for this province due to a record-high excess mortality in that year [15].

of all infections due to reinfections by 2021-10-22. We also find that the age-distribution of the modelled COVID-19 deaths are well-aligned with the pattern of excess mortality per age-group and both follow a clear log-linear relationship with respect to age, a signature characteristic of COVID-19 fatalities [28], suggesting that the model successfully reconstructs age-dependent contact patterns for disease spread in Iran (Figure 7.7C). There is also a strong association between the modelled and observed daily hospital admissions over time and across provinces, further indicating that the majority of excess deaths during the pandemic period are directly attributable to COVID-19 (Figure S4). There is, however, a notably lower number of modelled daily hospital admissions compared to observed hospitalisations in some provinces (e.g., South Khorasan, Lorestan, and Yazd) during the last wave in comparison to earlier waves. To explore these patterns, we use a mixed-effects model to examine the relationship between the number of hospitalisations inferred from our model fits and the observed confirmed COVID-19 hospital admissions over time and across epidemic waves while controlling for province-specific variation (see Methods section). Using this framework, we show that there was no significant decrease in the ratio of model inferred hospitalisations to observed hospitalisations during the last wave compared to previous waves, and that the majority of variation in hospitalisation patterns observed is explained by province-level random effects, most likely due to specific changes in treatment seeking or the age demographics being most affected in later waves (Figure S5). However, a significant decrease in the ratio of model inferred hospitalisations to observed hospitalisations was observed since the start of the pandemic, which suggests that case fatality rates from hospitalised infections have decreased during

this period, likely due to a combination of improved case management, increased number of infections resulting from reinfections, and higher vaccination coverage.

Using our transmission model, we also calculate the cumulative attack rate by 2021-10-22 for each province (Figure 7.7D). Our results show a high degree of heterogeneity in the estimated attack rates across the country with 11 provinces reaching close to and higher than 100% cumulative attack rates after the fifth wave (Table S1). While in most regions, the majority of infections took place during the third and fifth waves, the Delta variant of SARS-CoV-2 was responsible for the highest re-infection rates during the fifth wave (Figure S6). Qazvin, Qom, and East Azerbaijan had the highest per capita mortality rate with cumulative PFRs of nearly 0.4%, while Golestan, Khuzestan, and Qom had the highest re-infection rate with nearly 20% of infections being due to re-infection (Table 7.3; Figure S6). Figure S1 and Figure S4 also show that while some provinces such as Ardabil, Alborz, Semnan, Golestan, Hamedan, and Mazandaran experienced five distinct epidemic peaks, others like Bushehr, Qazvin, Gilan, Kohgiluyeh and Boyer Ahmad, and Hormozgan experienced sustained transmission of SARS-CoV-2 for an extended period of time throughout the pandemic. In late September 2021, official reports from MoHME indicate that COVID-19 deaths in Sistan and Baluchistan, Ardabil, Ilam, Bushehr, Chahaar Mahal and Bakhtiari, South Khorasan, North Khorasan, Zanjan, Kohgiluye and Boyer Ahmad, Lorestan, Hormozgan, and Hamedan have all dropped to below 5 individuals per day in recent weeks, an almost record-low since the first wave in March 2020 [29-31]. This is congruent with our estimates of high attack rates in these provinces which may have resulted in a (temporary) herd-immunity to be reached after the fifth wave.

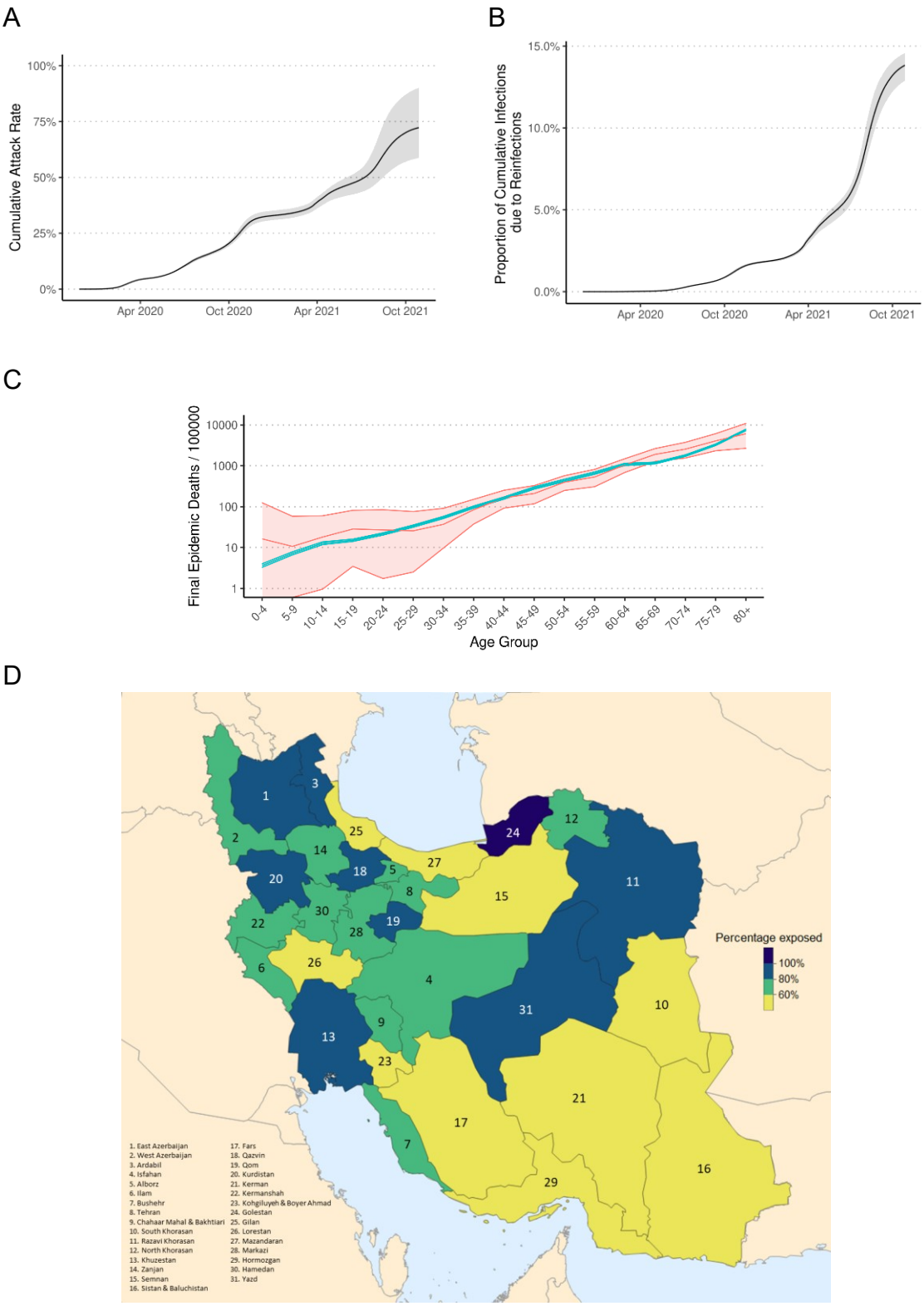


Figure 7.7: **Cumulative attack rate, reinfection rate, and deaths due to COVID-19 in Iran. (A)** Modelled cumulative nationwide attack rate of SARS-CoV-2 and **(B)** percentage of individuals reinfected over time. Shaded area in black shows the estimated quantities from the three sets of model assumptions used for estimating attack rates. **(C)** Per-capita excess mortality per age-group according to the transmission model (green) and observed data (red). Shaded areas in green and red show the 95% confidence interval in modelled and observed excess mortality, respectively. **(D)** Cumulative attack rate of SARS-CoV-2 per province as of 2021-10-22.

7b.4 Discussion

Since the start of the pandemic, the global heterogeneity in COVID-19 reporting systems globally has resulted in substantial debate about the true extent to which COVID-19 has spread in each country [4, 32, 33]. International comparisons of COVID-19 death tolls have subsequently relied on excess mortality due to under-reporting of COVID-19 deaths in many parts of the world [2]. This study is the first to formally explore the suitability of excess mortality as a proxy for estimating under-reporting of COVID-19 infections and deaths, confirming the information contained within reliable excess mortality statistics through the broad agreement with representative seroprevalence estimates. Additionally, this analysis lends further support to evidence that IFR patterns estimated in high income countries may underestimate the IFR in settings with stretched healthcare capacity [27], as experienced in many provinces in Iran at periods during the pandemic [34, 35]. Even though in some provinces we find that the transmission model with a lower estimated IFR provides a better fit to seroprevalence estimates, such a model offers a poor fit across most provinces and often results in epidemiologically unrealistic outcomes such that the modelled transmission dynamics outruns the pool of susceptible individuals. We also note that while some factors such as misidentification of baseline all-cause mortality and deaths attributable to COVID-19, particularly in the younger age-groups with fewer COVID-19-related deaths, can contribute to varied attack rates in some provinces, other factors such as nonrepresentative serosurveys, varied time to seroreversion using different assays, and uncertainty in inferring the true contact matrix between different age-groups within the population can impact the inferred attack rates [36, 37]. Absence of accurate vaccination data could be another potential source of bias that may

influence our modelled attack rates in some provinces, particularly in provinces that experienced high transmission levels due to Delta during the last wave. For instance, if we overestimate the level of protection from vaccines, more individuals need to be infected in a given wave in order to produce an epidemic that correctly matches the observed mortality. Similarly, if more vaccines are distributed to younger individuals than we currently assume in the model, it could lead to an artificial increase in the IFR with those highest at risk not being protected.

Prior to the emergence of the variants of concern, the basic reproduction number of the SARS-CoV-2 virus was estimated to be around 2.8 which corresponds to a herd immunity threshold of about 64% [38]. This threshold does not represent the point at which the outbreak immediately stops, but rather the point where, on average, fewer secondary infections are produced. During unmitigated outbreaks with such reproduction numbers, the attack rate can reach as high as 90% [39, 40]. Similarly, during unmitigated outbreaks with variants of concerns such as the Delta variant with increased reproduction numbers compared to the ancestral SARS-CoV-2 virus, the attack rate can reach even higher levels [41]. While the duration of antibody responses after coronavirus infection over longer timescales remains largely unknown, recent surveillance data and comparative studies have shown that re-infection with SARS-CoV-2 is possible after 3-6 months, and that older age-groups have a much lower protection against reinfection [42, 43]. Previous studies on human coronaviruses and SARS-CoV have also shown that protective immunity typically starts declining after 5-8 months post infection [44-46]. Our findings of high attack rates in several provinces show that herd immunity through natural infection has not been achieved in the population even after nearly 20 months since the start of the Iranian epidemic. This is likely due to substantial reduction in protection

against repeat infection over time either due to waning immunity, increased chance of re-infection with variants of concern, or a combination of both. We note that while systematic biases such as identifying the baseline level of all-cause mortality and overestimation of deaths attributable to COVID-19, particularly in the younger age-groups, can contribute to increased attack rates in some provinces, other factors such as heterogeneities in immune protection, limited access to health care in municipalities with low socioeconomic status, and uneven adoption of public health intervention in different regions can also contribute to varied attack rates across the country.

7b.5 Methods

7b.5.1 Excess mortality

We collect the weekly time-series data on all-cause mortality per province per age group from NOCR. We use data from the beginning of year 1394 to the end of summer 1398 in Solar Hijri calendar (SH) (from 2015-03-15 to 2019-09-22 in Common Era) to calculate background mortality. We find significant levels of excess mortality during several weeks (mostly concentrated in weeks 35 to 45) of autumn 2019 prior to the start of the COVID-19 pandemic (**Figure S7**). This has also been reported in a previous study which investigated some of the potential underlying causes of this significant excess mortality [15]. Therefore, in order to avoid biasing our expected mortality estimates from years prior to the pandemic, we exclude mortality data during autumn 2019 (from 2019-09-23 to 2019-12-21) from the calculation of background mortality as it may shift the baseline mortality upwards and, hence, result in the underestimation of excess mortality during the pandemic years in autumn 2020 and 2021. We calculate the expected mortality using a linear regression model that accounts for seasonality and trend, previously developed to

track excess mortality across more than 100 countries and territories around the world [1]. As a robustness test, we also calculate expected mortality with the same model under the assumption of an over-dispersed Poisson to see if different methods of calculating excess mortality yields significantly different results. We find that the estimated excess mortality using both methods are virtually identical over time across all provinces (**Figure S8**). For example, we find that by 2021-10-22, the estimated total excess mortality at the national level based on the linear regression is 259,641 (95% CI: 230,493 - 288,790) persons while, based on the quasi-Poisson model, it is 258,472 (95% CI: 230,369 - 286,574) persons which shows that the two estimates are very similar to each other.

7b.5.2 Transmission model

We use a previously published age-structured COVID-19 transmission model [47] and fitting framework [25] to fit the weekly estimated excess deaths in each province in Iran. In overview, the model is a population-based age-structured Susceptible-Exposed-Infected-Recovered model, which explicitly represents disease severity, passage through different healthcare levels and the roll out of vaccination. Model fitting was carried out within a Bayesian framework using a Metropolis-Hastings Markov Chain Monte Carlo (MCMC) based sampling scheme, which estimates the epidemic start date, R_0 and the time varying reproduction number, R_t, using a series of pseudo-random walk parameters, which alter transmission every 2-weeks, given by:

$$R_t = R_0 . f(- \rho_1 - \rho_2 \dots - \rho_n)$$

where $f(x) = 2.exp(x)/(1+exp(x))$, i.e., twice the inverse logit function. Each parameter is introduced two weeks after the previous parameter, serving to capture changes in transmission every two weeks. The last change in transmission, ρ_n, is maintained for the last 4 weeks prior to the current day to reflect our inability to estimate the effect size of this parameter due to the approximate 21-day delay between infection and death [48]. The MCMC sampling scheme includes adaptive tuning of the proposal implemented during sampling using the Johnstone-Chang optimisation algorithm [49]. All parameter inference results are based on 100,000 iterations, 10,000 of which are discarded as burn-in. MCMC chains are checked for convergence using standard Gelman-Rubin diagnostics [50].

Each model fit used province-specific demography, with the population size in 5-year age bands sourced from the Statistical Center of Iran [51]. The effective number of general hospital beds and intensive care beds were sourced from the Iran statistical yearbook published in 1396 in Solar Hijri calendar [52]. At the time of writing, MoHME has not published subnational daily vaccination data prior to 2021-10-26, nor the proportion of which vaccines are administered. As a result, we used multiple data sources to approximate the roll of vaccines subnationally. From 2021-10-26, MoHME started to release the total number of first and second vaccine doses subnationally, which we use in combination with national vaccination data [53] to infer the daily vaccination rate per province. A mix of vaccine types have been administered in Iran, with 80% of the population vaccinated with Sinopharm, 10% with AZ and the remaining 10% either Sputkink-V (1.5%) and CoViran Barekat (8.5%). Due to the mix of vaccine types with uncertain efficacy estimates, particularly related to Delta, we use vaccine efficacy estimates for first and second dose efficacy sourced from estimates for the Sinopharm vaccine and other

inactivated whole virus vaccine platforms (Table 7.4). Efficacy estimates are weighted to produce an overall population efficacy by the proportion of vaccinated individuals who have received either a first dose or second dose, and by the assumed proportion of the Delta variant in the population, sourced from the sampling dates of Delta sequences in Iran from GISAID (*gisaid.org*), lagged by 14 days to reflect the delay from administering vaccine to protection being conferred.

Imposing a single IFR across the whole country may lead to biased attack rate estimates as it does not account for various factors such healthcare capacity or demographics of each province. To avoid this, we assume that the IFR in each province is dependent on the demographics of the province, such that provinces with older populations have, on average, higher IFRs. The model also tracks the number of people who require a general hospital bed or an ICU bed so that if there is a bed shortage during an epidemic peak in a given province then the mortality outcomes for individuals who do not receive a bed will increase. We explore three sets of model assumptions, which scan across the likely range of parameter estimates for key model parameters that impact the inferred attack rate. These parameters relate to the assumed number of general and ICU hospital beds available, the change in hospitalisation related to the Delta variant and the level of immune escape conferred by the Delta variant relative to previous infection by the Alpha variant and the original wild type strains. Here, the level immune escape only refers to immunity from previous infection, which acts by making a proportion of people who would have been protected against reinfection by the same variant susceptible to infection with Delta. With regards to immunity from vaccination, we assume that the impact of Delta is to reduce the level of protection conferred by the vaccines. These ranges are described in Table 7.4. Given uncertainty in the IFR

and its impact of the inferred attack rate, we explore two alternative estimates of IFR by age [21, 54] and conduct model comparison between the two references through comparison against province level estimated seroprevalence data [16], with seroprevalence assumed to be described by a Binomial distribution. Lastly, to explore if IFR is changing over time, we explore the ratio of model inferred hospitalisations and observed daily hospitalisations against time using a mixed-effects model, controlling for variation between provinces using a random intercept and slope with respect to time.

To assess the impact of healthcare capacity constraints on increasing the mortality from COVID-19, we also calculate the mean population-weighted IFR [15], $\alpha = \sum_{i=1}^{L} \alpha_i w_i$, and its corresponding standard error, $\sigma_\alpha = \sqrt{\sum_{i=1}^{L} w_i^2 \sigma_i^2}$, for each province using the estimates of IFR by age from Brazeau et al [54], where *L* is the total number of age-groups, α_i is the age-stratified IFR for age-group *i*, σ_i is its standard error, and w_i is the proportion of individuals within age-group *i* in the population. Unlike IFR estimates based on our transmission model which takes into account the impact of various factors such as province-level demographics, impact of healthcare capacity constraints on increased mortality rates, and increased chance of hospitalisation with Delta the population-weighted IFR only accounts for the province-level demographics which enables us to assess the influence of other factors on increasing IFR.

7b.5.3 Data availability

All the data used for the analysis are available online on our GitHub repository (https://github.com/mg878/Iran_WeeklyMortality), with all model fitting code and

analysis available at (https://github.com/OJWatson/iran-ascertainment). For regular updates on excess mortality in Iran, you can visit: https://github.com/akarlinsky/world_mortality.

Table 7.4: **Vaccine efficacy estimates, and parameter uncertainty related to hospitalisation and the Delta variant.**

Description\Scenarios	Optimistic	Central	Worst	Notes & references
Vaccine efficacy against infection	Alpha: 50%; 67% Delta: 10%; 60%	Alpha: 50%; 67% Delta: 10%; 60%	Alpha: 50%; 67% Delta: 10%; 60%	50%; 67% | Dose 1; Dose 2 [55-57]
Vaccine efficacy severe disease	Alpha: 50%; 80% Delta: 20%; 70%	Alpha: 50%; 80% Delta: 20%; 70%	Alpha: 50%; 80% Delta: 20%; 70%	50%; 80% | Dose 1; Dose 2 [55-57]
Immune escape by the Delta variant	10%	27%	50%	Median immune escape estimated at 27% with 50% Bayesian Credible Interval spanning 10% - 50% [58].
Increased risk of hospitalization from infections with the Delta variant	1.08	1.45	1.95	Increased probability of hospital - adjusted HR 1.45 (95% 1.08, 1.95) [59] - or emergency care attendance due to Delta variant within 14 days informed by UK Public Health England data.
Hospital Capacity	No shortage of beds at any point.	Bed numbers available equal to estimated bed capacity.	50% of estimated total beds available due to occupancy with non-COVID-19 patients.	Total bed numbers sourced from Iran statistical yearbook published in 1396 in Solar Hijri calendar [52]. Baseline occupancy previously estimated at ~70% [60], which during COVID-19 we assume will be lower to reflect pandemic prioritisation as observed in other countries [61].

7b.6 References

1. Karlinsky, A. and D. Kobak, *Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset.* Elife, 2021. **10**.
2. Whittaker, C., et al., *Under-reporting of deaths limits our understanding of true burden of covid-19.* Bmj-British Medical Journal, 2021. **375**.
3. Lau, H., et al., *Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters.* Pulmonology, 2021. **27**(2): p. 110-115.
4. Meyerowitz-Katz, G. and L. Merone, *A systematic review and meta-analysis of published research data on COVID-19 infection fatality rates.* International Journal of Infectious Diseases, 2020. **101**: p. 138-148.
5. Caliendo, A.M., et al., *Better Tests, Better Care: Improved Diagnostics for Infectious Diseases.* Clinical Infectious Diseases, 2013. **57**: p. S139-S170.
6. Giri, A.K. and D.R. Rana, *Charting the challenges behind the testing of COVID-19 in developing countries: Nepal as a case study.* Biosafety and Health, 2020. **2**(2): p. 53-56.
7. Mena, G., et al., *Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile.* Science, 2021. **372**(6545): p. 934-937.
8. Bradshaw, D., et al., *Tracking mortality in near to real time provides essential information about the impact of the COVID-19 pandemic in South Africa in 2020.* Samj South African Medical Journal, 2021. **111**(8): p. 732-740.
9. Woolf, S.H., et al., *Excess Deaths From COVID-19 and Other Causes, March-April 2020.* Jama-Journal of the American Medical Association, 2020. **324**(5): p. 510-513.
10. Karlinsky, A., *National Excess Mortality from Sub-National data: Method and Application for Argentina.* medRxiv, 2021. https://doi.org/10.1101/2021.08.30.21262814.
11. Ghafari, M., A. Kadivar, and A. Katzourakis, *Estimates of anti-SARS-CoV-2 antibody seroprevalence in Iran.* Lancet Infectious Diseases, 2021. **21**(5): p. 602-603.
12. Modi, C., et al., *Estimating COVID-19 mortality in Italy early in the COVID-19 pandemic.* Nat Commun, 2021. **12**(1): p. 2729.
13. Chitwood, M.H., et al., *Reconstructing the course of the COVID-19 epidemic over 2020 for US states and counties: results of a Bayesian evidence synthesis model.* medRxiv, 2021.
14. Ghafari, M., et al., *Lessons for preparedness and reasons for concern from the early COVID-19 epidemic in Iran.* Epidemics, 2021. **36**: p. 100472
15. Ghafari, M., A. Kadivar, and A. Katzourakis, *Excess deaths associated with the Iranian COVID-19 epidemic: A province-level analysis.* International Journal of Infectious Diseases, 2021. **107**: p. 101-115.
16. Khalagi, K., et al., *Prevalence of COVID-19 in Iran: results of the first survey of the Iranian COVID-19 Serological Surveillance programme.* Clinical Microbiology and Infection, 2021. **27**(11): p. 1666-1671.
17. Poustchi, H., et al., *SARS-CoV-2 antibody seroprevalence in the general population and high-risk occupational groups across 18 cities in Iran: a*

population-based cross-sectional study. Lancet Infect Dis, 2021. **21**(4): p. 473-481.
18. *WHO Director-General's Opening Remarks at the Media Briefing on COVID-19*. 20 February 2020; Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---28-february-2020.
19. Han, E., et al., *Lessons learnt from easing COVID-19 restrictions: an analysis of countries and regions in Asia Pacific and Europe.* Lancet, 2020. **396**(10261): p. 1525-1534.
20. Ghafari, M., et al., *Iran's covid-19 vaccination programme: using transparency to build public trust in immunisation* BMJ Opinion, 2021. https://blogs.bmj.com/bmj/2021/08/03/irans-covid-19-vaccination-programme-using-transparency-to-build-public-trust-in-immunisation/.
21. O'Driscoll, M., et al., *Age-specific mortality and immunity patterns of SARS-CoV-2.* Nature, 2020.
22. Walker, P.G.T., et al., *The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries.* Science, 2020. **369**(6502): p. 413-422.
23. Fattahi, Z., et al., *SARS-CoV-2 outbreak in Iran: The dynamics of the epidemic and evidence on two independent introductions.* Transboundary and Emerging Diseases, 2021.
24. Buchmann, B., et al., *Estimating COVID-19 mortality in Italy early in the COVID-19 pandemic.* Nature Communications, 2021. **12**(1).
25. Watson, O.J., et al., *Leveraging community mortality indicators to infer COVID-19 mortality and transmission dynamics in Damascus, Syria.* Nature Communications, 2021. **12**(1).
26. Ghisolfi, S., et al., *Predicted COVID-19 fatality rates based on age, sex, comorbidities and health system capacity.* Bmj Global Health, 2020. **5**(9).
27. Levin, A.T., et al., *Assessing the Burden of COVID-19 in Developing Countries: Systematic Review, Meta-Analysis, and Public Policy Implications.* medRxiv, 2021. https://doi.org/10.1101/2021.09.29.21264325.
28. Levin, A.T., et al., *Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications.* European Journal of Epidemiology, 2020. **35**(12): p. 1123-1138.
29. *BBC Persian. Coronavirus in Sistan and Baluchistan; attorny general of the province: mortuary fridges are filled with dead bodies (Farsi)*. 11 July 2021 (Access date: 22 September 2021)]; Available from: https://www.bbc.com/persian/iran-57793914.
30. *IRNA. No new coronavirus deaths have been recorded in Sistan Baluchistan over the last 24 hours.* 29 August 2021 [Access date: 22 September 2021]; Available from: https://www.irna.ir/news/84452428.
31. *IRNA. Daily coronavirus deaths were below 5 in 16 provinces*. 23 September 2021 [Access date: 22 September 2021]; Available from: https://www.irna.ir/news/84485785.
32. Sood, N., et al., *Seroprevalence of SARS-CoV-2-Specific Antibodies Among Adults in Los Angeles County, California, on April 10-11, 2020.* Jama-Journal of the American Medical Association, 2020. **323**(23): p. 2425-2427.
33. Ioannidis, J.P.A., *Infection fatality rate of COVID-19 inferred from seroprevalence data.* Bulletin of the World Health Organization, 2021. **99**(1): p. 19-22.

34. *IRNA. Hospitals reaching full capacity and highest daily coronavirus deaths*. 14 November 2020 [Access date: 22 September 2021]; Available from: https://www.irna.ir/news/84109005.
35. *IRNA. Hospitals at full capacity and measures of healthcare providers in South Khorasan*. 7 August 2021 [Access date: 22 September 2021].
36. Peluso, M.J., et al., *SARS-CoV-2 antibody magnitude and detectability are driven by disease severity, timing, and assay.* Science Advances, 2021. **7**(31).
37. Mistry, D., et al., *Inferring high-resolution human mixing patterns for disease modeling.* Nature Communications, 2021. **12**(1).
38. Billah, M.A., M.M. Miah, and M.N. Khan, *Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence.* Plos One, 2020. **15**(11).
39. Handel, A., I.M. Longini, and R. Antia, *What is the best control strategy for multiple infectious disease outbreaks?* Proceedings of the Royal Society B, 2007. **274**(1611).
40. de Souza, W.M., et al., *Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil.* Nat Hum Behav, 2020. **4**(8): p. 856-865.
41. Liu, Y. and J. Rocklöv, *The reproductive number of the Delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus* Journal of Travel Medicine. https://doi.org/10.1093/jtm/taab124, 2021.
42. Buss, L.F., et al., *Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic.* Science, 2021. **371**(6526): p. 288-291.
43. Hansen, C.H., et al., *Assessment of protection against reinfection with SARS-CoV-2 among 4 million PCR-tested individuals in Denmark in 2020: a population-level observational study.* Lancet, 2021. **397**(10280): p. 1204-1212.
44. Huang, A.T., et al., *A systematic review of antibody mediated immunity to coronaviruses: kinetics, correlates of protection, and association with severity.* Nat Commun, 2020. **11**(1): p. 4704.
45. Cao, W.C., et al., *Disappearance of antibodies to SARS-associated coronavirus after recovery.* New England Journal of Medicine, 2007. **357**(11): p. 1162-1163.
46. Townsend, J.P., et al., *The durability of immunity against reinfection by SARS-CoV-2: a comparative evolutionary study.* Lancet Microbe, 2021. https://doi.org/10.1016/S2666-5247(21)00219-6.
47. Hogan, A.B., et al., *Within-country age-based prioritisation, global allocation, and public health impact of a vaccine against SARS-CoV-2: A mathematical modelling analysis.* Vaccine, 2021. **39**(22): p. 2995-3006.
48. Walker, P.G.T., et al., *The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries.* Science, 2020. **369**(6502): p. 413-419.
49. Johnstone, R.H., et al., *Uncertainty and variability in models of the cardiac action potential: Can we build trustworthy models?* Journal of Molecular and Cellular Cardiology, 2016. **96**: p. 49-62.
50. Gelman, A. and D. Rubin, *Inference from Iterative Simulation Using Multiple Sequences.* Statistical Science, 1992. **7**: p. 457-511
51. *Statistical Center of Iran*. 2020 [Access date: 29 November 2020]; Available from: https://www.amar.org.ir/english.

52. *Iran Open Data: Number of active beds in hospital wards by province in year 1396*. 2018 [Access date: 29 November 2020]; Available from: https://iranopendata.org/en/dataset/number-of-active-beds-in-hospital-wards-by-province-in-year-1396/resource/73e29d0b-7cc9-47c0-8b4e-b3e61cc0b800.
53. Mathieu, E., et al., *A global database of COVID-19 vaccinations.* Nature Human Behaviour, 2021. **5**(7): p. 947-953.
54. Brazeau, N., et al., *Report 34: COVID-19 infection fatality ratio: estimates from seroprevalence.* Imperial College London COVID-19 Response Team, 2020.
55. Imai, N., et al., *Interpreting estimates of coronavirus disease 2019 (COVID-19) vaccine efficacy and effectiveness to inform simulation studies of vaccine impact: a systematic review.* Wellcome Open Research, 2021. **6**: p. 185.
56. Kang, M., et al., *Effectiveness of inactivated COVID-19 vaccines against COVID-19 pneumonia and severe illness caused by the B. 1.617. 2 (Delta) variant: evidence from an outbreak in Guangdong, China.* SSRN, 2021.
57. Li, X.N., et al., *Effectiveness of inactivated SARS-CoV-2 vaccines against the Delta variant infection in Guangzhou: a test-negative case-control real-world study.* Emerging Microbes & Infections, 2021. **10**(1): p. 1751-1759.
58. Dhar, M.S., et al., *Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India.* Science, 2021. **374**(6570): p. 995-999.
59. Twohig, K.A., et al., *Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: a cohort study.* Lancet Infect Dis, 2021.
60. Moradi, G., et al., *Assessment of the Efficiency of Hospitals Before and After the Implementation of Health Sector Evolution Plan in Iran Based on Pabon Lasso Model.* Iranian Journal of Public Health, 2017. **46**(3): p. 389-395.
61. McCabe, R., et al., *Modelling intensive care unit capacity under different epidemiological scenarios of the COVID-19 pandemic in three Western European countries.* International Journal of Epidemiology, 2021. **50**(3): p. 753-767.

Statement of Authorship for joint/multi-authored papers for PGR thesis.

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	Part a: Excess deaths associated with the Iranian COVID-19 epidemic: A province-level analysis Part b: A framework for reconstructing SARS-CoV-2 transmission dynamics using excess mortality data
Publication Status	Part a: Published Part b: Accepted for Publication
Publication Details	Part a: Ghafari M, Kadivar A, and Katzourakis A. Excess deaths associated with the Iranian COVID-19 epidemic: A province-level analysis. International Journal of Infectious Diseases 107: 101-115, 2021. Part b: Ghafari M, Watson OJ, Karlinsky A, and Katzourakis A. Excess deaths associated with the Iranian COVID-19 epidemic: A framework for reconstructing SARS-CoV-2 transmission dynamics using excess mortality data. (Accepted in Nature Communications)

Student confirmation

Student Name	Mahan Ghafari
Contribution to the paper	Part a: MG wrote the paper and conducted all the analysis under the supervision from AK. All authors provided editorial assistance. Part b: MG wrote the paper under the supervision from AK. MG and OJW contributed equally to the analysis. AKar and MG performed the calculations for weekly excess mortality. OJW and AKar provided assistance for generating the figures.
Signature 	Date 11/05/2022

Supervisor confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title	Professors Aris Katzourakis and Oliver G Pybus	
Supervisor comments		
Signature 		Date 11/05/2022

Chapter 8: Discussion

8.1 Conclusions and final thoughts

Viruses evolve incredibly rapidly, and this feature has enabled detailed reconstructions of the emergence and molecular epidemiology of several major pathogens including HIV-1, Hepatitis C (HCV), pandemic H1N1 influenza, and SARS-CoV-2 [1-4]. However, these rapid evolutionary rates mean that longer term reconstructions of their evolutionary history are problematic, as accumulated sequence changes would create completely unrecognisable viral genome sequences if maintained over thousands of years. This simple observation stands in stark contrast to apparent conservatism in virus evolution, where ancient sequences from the endogenous virus 'genomic fossil' record, and from archaeological samples are often substantially similar to those of their modern-day relatives which would imply that viruses are evolving at an apparently much slower rates over longer timescales [5-8]. Furthermore, the rising popularity of time-scaled phylogenies and availability of viral genetic data, means that, more than ever before, researchers are estimating rates of molecular evolution of viruses, particularly over shorter timescales (e.g., during the course of an epidemic), and reporting rates estimates that seems to be decreasing with the timespan of their measurements which would again suggest that the apparent rate of virus evolution is continuously decreasing over time.

The aim of this thesis was to better understand such conundrums relating to evolutionary rate relativity where the estimated rates of molecular evolution are dependent on the timescale of measurement, and to quantitatively understand the biological mechanisms influencing rate estimates. This problem, also known as

time-dependent rate phenomenon (TDRP), has been largely investigated empirically in viruses by Aiewsakun and Katzourakis who showed that TDRP over long timescales is best explained by a power-law regression fit [9]. While such a top-down explanation of TDRP provides an empirical tool to better estimate dates of origin of viruses when extrapolating substitution rates further back in time, without an underlying biological or evolutionary basis, its predictive value is only as good as the data available on known dates of virus origins to calibrate it [10]. It is also not clear whether a single power-law fit should provide the best fit across every evolutionary timescale because different biological processes may influence the rate of molecular evolution differently, leading to divergent patterns of TDRP.

The significance of having a mechanistic evolutionary model that explains TDRP in viruses which I described in chapter 2 is that it offers a bottom-up explanation to what Aiewsakun and Katzourakis found empirically. The mathematical formulation of this model was based on the idea that the host imposes absolute or epistatic constraints on the evolution of some sites on virus genome while other sites may evolve more rapidly. I showed that this model can readily explain TDRP over long timescales (any timescale greater than 100-200 years before present). One of the major implications of this is that it will not only allow for the estimation of timescales and rates of evolution with ever greater precision, but also provides a tool to uncover key aspects of virus evolution such as the proportion of freely evolving, epistatic or constrained sites, and the rates at which they evolve.

In chapters 2 and 3, I applied this model to re-date the diversification of HCV virus to nearly 400,000 years before present, a time preceding the dispersal of humans out of Africa, and provided a radical revision of our understanding of HCV origins and an explanation of their current endemic genotype distributions. I also

showed that the PoW model can recover known divergence times in Foamy viruses (based on co-speciation evidence) and SIVs (based on biogeographic calibrations and observation of endogenous lentiviruses in primates) which date back to millions of years before present without requiring any external calibration information. I also deployed the PoW model to sarbecoviruses, showing that they date back to approximately 23,500 years ago, indicating that humanity may have been exposed to these viruses since the Paleolithic. Understanding the evolutionary history of sarbecoviruses is particularly important as they have crossed into humans twice in the last two decade, leading to the deadly SARS1 outbreak and the COVID-19 pandemic. With this new mechanistic evolutionary model, we can look much more widely at other viruses; re-evaluate the timescales of their deeper evolution and gain insights into host relationships that are key to understanding their ability to cause disease.

In chapters 4 and 5, I showed that TDRP is not a phenomenon that is unique to longer evolutionary timescales but can also be observed over a short evolutionary time span of a few months in SARS-CoV-2 and pH1N1 influenza. In chapter 4, I found that unlike longer evolutionary timescales where site saturation is the dominating factor in creating TDRP, over a short time frame of 3-6 months, the overabundance of segregating deleterious mutations that are not yet purged from the population in both SARS-CoV-2 and pH1N1 influenza sequences are responsible for giving rise to TDRP. I found evidence in support of purifying selection based on the strong correlation between the number of segregating low-frequency deleterious mutations and the rate estimates at the terminal branches of both SARS-CoV-2 and pH1N1 trees.

It is also important to emphasise the TDRP explained in chapters 2 and 4 stem from completely different underlying biological mechanisms. The former is caused by saturation of sites (i.e., multiple mutations at the same site) which creates a power-law rate decay over time, while the latter is explained by the exponential waiting time (inversely proportional to the strength of purifying selection) before deleterious mutations are purged from the population which creates a pattern of exponential decay in rate estimates over time (see section 1.4 in chapter 1 for more detail on the quantitative explanation). Another difference between the two TDRP mechanisms is that while we observe a rate decay in both cases, other inferred evolutionary parameters such as the time to the most recent common ancestor does not seem to be influenced significantly by purifying selection over short timescales conditioned on picking the right growth coalescent model. This is because as long as we have a large enough time-stamped data set, we can reliably infer evolutionary parameters such as effective population size, growth rate and time to the most recent common ancestor of the sequences, whereas, over longer timescales, such information will be completely lost due to site saturation in any phylogenetic analysis unless there is external knowledge about how these quantities change with respect to the timeframe of measurement.

Similarly, the mechanism behind the TDRP I investigated in chapter 5 is quite different from either site saturation or purifying selection. As I explained in chapter 5, the sheer number of mutations observed in the first three variants of SARS-CoV-2, Alpha, Beta, and Gamma, that emerged in late 2020 is much higher than what we would expect under any standard phylogenetic estimate of the nucleotide evolutionary rate of SARS-CoV-2. The nucleotide substitution rate of nearly 0.7-1.1x10^{-3} per site per year during the first year of evolution of SARS-CoV-2 which I

reported in chapter 4 implies that the virus can acquire, on average, 2-3 new mutation along its genome for every additional month that it evolves in human hosts. Therefore, the observation of, for instance, 14 extra mutations in Alpha would require the molecular clock rate to be roughly twice as fast for Alpha relative to the background lineages and be sustained for a period of at least 6 months. This would require most of the variant-defining mutations to remain undetected for several months. And the way the virus is likely doing this is by acquiring most, if not all, of those extra mutations during chronic infections where it can persist inside a host while evolving more rapidly for a relatively long period without ever being detected at the population level.

In chapters 6 and 7, I demonstrated how to use excess mortality and mobility data to reconstruct the early transmission dynamics of SARS-CoV-2 in Iran and estimate the start of the epidemic as well as its growth rate in the absence of high-quality genome sequence data from the virus and province-level epidemiological data on the number of cases and deaths. In chapter 6, I found a broad agreement between the estimated starting date and growth rate of the Iranian epidemic using genome sequence data and air travel data from exported cases of COVID-19 to other countries with a travel link to Iran. In chapter 7, I found that these estimates are also in agreement with the significant rise of excess mortality in a handful of provinces in northern and central Iran at the beginning of the pandemic in early January 2020 which, after correcting for the delay between the onset of symptoms to deaths, would point towards approximately the same dates for when the Iranian epidemic started and how fast it grew over time based on the phylogenetic estimates carried out in chapter 6.

Moreover, after the publication of my findings on seasonal excess mortality (chapter 7 part a), I discovered that there is a strong correlation between the level of excess mortality in each province and the degree of population-level exposure to the virus based on seroprevalence measurements taken from 17 provinces across the country (see appendix 5). This led to the development of the systematic framework for reconstructing SARS-CoV-2 transmission dynamics using excess mortality data (chapter 7 part b) which is particularly insightful for understanding the burden of disease in regions with significant underreporting of COVID-19-related deaths. There was an excellent agreement between the model estimates and representative, province-level seroprevalence measurements across the country. There was also great agreement between the model estimates and pattern of weekly hospital admissions in all of the 31 provinces of Iran throughout the pandemic.

One of the other key lessons from the analysis of the Iranian epidemic, which is also broadly applicable to almost all other countries that experienced a major COVID-19 epidemic, was that the number of confirmed cases often significantly underestimates the true number of infections. This is less pronounced when it comes to estimating the number of true COVID-19-related hospital admissions and deaths. That is one of the reasons why in chapter 6 when I estimated the virus population size worldwide, I used the number of daily confirmed deaths to back-calculate the number of infectious individuals from the global median infection fatality rate of COVID-19 instead of directly using the number of confirmed cases to estimate this quantity.

8.2 Future directions

Perhaps one of the most immediate methodological advances that can be made on the basis of the work that has been developed in this thesis is to improve the ad-hoc implementation of the mechanistic evolutionary model in chapter 2 for estimating divergence times by fully integrating it into a Bayesian phylogenetic framework such as BEAST [11]. This would allow for the joint inference of short-term substitution rate, $< \mu >$, and fastest-evolving rate group, $\mu_{\max}$, two key parameters of the PoW model, for any given data set. One of the advantages of doing this would be that we can relax the assumption about using a strict molecular clock when estimating node heights in units of substitutions as a first step and potentially incorporate various sources of uncertainty in the estimated divergence time for each node due to rate variation across different lineages or having samples from vastly different timescales (e.g., endogenous viral elements or ancient DNA).

While the empirical model of TDRP based on the power-law relation between rate estimates and timescale of rate measurements has already been incorporated in BEAST [10], without deep calibration points, this approach is unable to reconstruct the evolutionary history of any given data set for which there is no deep evolutionary calibration points. Therefore, one of the main advantages of incorporating the PoW model into a software package like BEAST would be to shed light on the evolutionary history of viruses or other organisms for which we do not have a deep calibration point to correct for the time-dependency of evolutionary rate estimates. This can also help resolve the great discrepancy between viral evolutionary rates inferred from paleovirological evidence, such as endogenous viral elements, and modern-day viral sequences which, as we explained in chapters 2 and 3, are vastly different from each other.

From a mathematical point of view, PoW model is simply a transformation function that takes any estimate of evolutionary distance derived from standard substitution models and converts that into time. However, the PoW model is constructed in such a way that it can also be turned into a substitution model itself as long as we define a distance metric for it [12-14]. This is because the transformation provides a mechanistic framework for how a given site evolves over time. The challenge would be to turn it into a computationally tractable form for measuring distance. Developing such a substitution model will provide a more internally consistent way of estimating evolutionary timescales using the PoW model and will allow for the model parameters, $< \mu >$ and $\mu_{\max}$, to be directly inferred from data.

Another area of possible methodological advance would be to develop quantitative methods to identify the nucleotide bases or loci that have experienced substitution saturation. One of the commonly used approaches for this purpose is to compare the sites in a sequence alignment with those that have fully saturated according to the entropy principle [15, 16]. However, it is not clear how such saturation tests would perform under the PoW model where each site, depending on whether it belongs to a rapidly- or slowly-evolving rate group, may reach the saturation point at vastly different times. Identifying saturation of sites evolving at different rates would also help us directly estimate the numerical value of the fastest-evolving rate group, $\mu_{\max}$, and number of rate groups, *M*, in the PoW model.

The PoW model also shines a light on the importance of understanding the biological factors influencing site saturation in viruses [17-19]. A recent work on HIV-1 pandemic strain has shown that rapid within-host reversion mutations are under positive selection and can push the viral sequences towards a preferred

evolutionary state that is closer to consensus sequence at the between-host level which results in saturation effects influencing substitution rate estimates over a time span of a few decades [20]. Therefore, it is important to investigate such mechanisms in other viruses and assess the degree to which they can inform our parameter estimation for the fastest-evolving rate group, $\mu_{\max}$.

Our analysis of the strength of purifying selection in SARS-CoV-2 only spanned the first years of the evolution of the virus, just before when the variants of concern (VOCs) became globally dominant. Therefore, we were unable to investigate whether purifying selection continued to influence clock rate estimates after the emergence of VOCs or not. Recent studies have shown that the emergence of VOCs only results in the elevation of clock rates at the stem branch leading to a VOC lineage while the actual VOC lineages have a similar clock rate as the ancestral strain [21, 22]. This suggests that we would likely see a similar trend of deleterious mutations getting purged from the population due to purifying selection even after the emergence of VOCs. However, the exact strength of purifying selection and how it might vary over time is yet to be determined.

Based on the predictions from nearly neutral theory, we would expect to see stronger purifying selection in larger population sizes with faster rates of evolution [23]. It would be interesting to investigate this prediction in other viruses or even other organisms to see the degree to which the strength of purifying selection determines TDRP depending on the effective population size, mutation rate, and other evolutionary characteristics of a particular organism [24-27].

One of the main take-aways from the analysis of the Iranian epidemic was that reported cases and deaths only represent a partial count of the true numbers. As a

result, detailed understanding of the SARS-CoV-2 transmission dynamics and assessment of its burden is particularly challenging for many low- and middle-income countries where there is limited diagnostic capacities and access to healthcare [28, 29]. Consequently, the number of studies quantifying the dynamics and impacts of the pandemic in lower income settings have been significantly lower than for high income settings and our understanding of the true mortality caused by the pandemic in non-high-income countries is limited [30-33].

By providing the first quantitative framework for investigating the suitability of excess mortality as a proxy for understanding the spread of COVID-19 in lower income settings through confirmatory comparisons against seroprevalence measurements, we opened the possibility for investigating the spread of the virus in many more regions of the world where there is limited epidemiological and clinical data, but all-cause mortality data is available. These findings have far-reaching implications for understanding the true burden of the pandemic and provide further evidence for the investment in scaling vital registration infrastructure for pandemic monitoring and preparedness [28, 34-36].

Another benefit of using excess mortality to reconstruct transmission dynamics is to estimate the number of infectious individuals which, as we showed in chapter 5, can be used to infer the effective virus population size over time. Inferring population dynamics of a virus is also possible by applying phylogenetic models based on the coalescent theory using viral sequence data [37, 38]. It would be interesting to see the degree to which these two estimates of effective population size are similar to each other and under what demographic conditions (e.g., increasing, constant, or decreasing populations). This may also enable us to use

estimates of effective population size based on excess mortality when there is limited genetic data available to use phylogenetic methods.

8.3 Concluding remarks

The main goal of this thesis was to put forward quantitative and mechanistic frameworks for understanding the empirical observation of time-dependent rate estimates in viruses across timescales. This included understanding various biological processes that are involved in creating TDRP over different timescales. We showed that TDRP is very complex and is governed by several biological processes, many of which are yet to be determined. Some of the underlying causes of the TDRP highlighted in this thesis included site saturation, purifying selection, and rapid within-host evolution during chronic infections. We demonstrated that a mechanistic evolutionary model which we developed in chapter 2 provides a simple and sufficient explanation for TDRP over evolutionary timescales of around hundreds of years before present and older. We also found purifying selection to be the dominating factor influencing TDRP over shorter timescales. However, the latter does not seem to affect the estimates for the time to the most recent common ancestor of viral sequences as long as sufficient temporal signal is present in the data. Further work is required to understand how these two dominating factors, namely purifying selection and site saturation, come together to create a continuous TDRP effect over time. Other possible biological factors influencing TDRP may include the bottlenecking nature of viral transmission, variation in population size over time, and natural selection pressures. It is unclear how these biological processes interact to give rise to TDRP, or indeed whether they can give rise to the TDRP at all, and if so, under which conditions.

8.4 References

1. P. V. Markov, et al., *Phylogeography and molecular epidemiology of hepatitis C virus genotype 2 in Africa.* Journal of General Virology, 2009. **90**: p. 2086-2096.
2. A. G. Meyer, et al., *Time dependence of evolutionary metrics during the 2009 pandemic influenza virus outbreak.* Virus Evolution, 2015. **1**(1): p. vev006.
3. S. Duchene, et al., *Temporal signal and the phylodynamic threshold of SARS-CoV-2.* Virus Evolution, 2020. **6**(2): p. veaa061.
4. P. Lemey, A. Rambaut, and O.G. Pybus, *HIV evolutionary dynamics within and among hosts.* Aids Reviews, 2006. **8**(3): p. 125-140.
5. A. Katzourakis, et al., *Discovery and analysis of the first endogenous lentivirus.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(15): p. 6261-6265.
6. A. Katzourakis and R.J. Gifford, *Endogenous viral elements in animal genomes.* PLoS Genet, 2010. **6**(11): p. e1001191.
7. R. J. Gifford, et al., *A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution.* Proc Natl Acad Sci U S A, 2008. **105**(51): p. 20362-7.
8. P. Simmonds, P. Aiewsakun, and A. Katzourakis, *Prisoners of war - host adaptation and its constraints on virus evolution.* Nat Rev Microbiol, 2019. **17**(5): p. 321-328.
9. P. Aiewsakun and A. Katzourakis, *Time-Dependent Rate Phenomenon in Viruses.* Journal of Virology, 2016. **90**(16): p. 7184-7195.
10. J. V. Membrebe, et al., *Bayesian Inference of Evolutionary Histories under Time-Dependent Substitution Rates.* Molecular Biology and Evolution, 2019. **36**(8): p. 1793-1803.
11. M. A. Suchard, et al., *Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10.* Virus Evolution, 2018. **4**(1): p. vey016.
12. J. Felsenstein, *Distance Methods for Inferring Phylogenies: A Justification.* Evolution, 1984. **38**(1): p. 16-24.
13. J. Felsenstein, *Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods.* Methods Enzymol, 1996. **266**: p. 418-27.
14. J. Felsenstein, *Taking variation of evolutionary rates between sites into account in inferring phylogenies.* J Mol Evol, 2001. **53**(4-5): p. 447-55.
15. X. H. Xia, et al., *An index of substitution saturation and its application.* Molecular Phylogenetics and Evolution, 2003. **26**(1): p. 1-7.
16. D. A. Duchene, et al., *Excluding Loci With Substitution Saturation Improves Inferences From Phylogenomic Data.* Systematic Biology, 2021: p. syab075.
17. Hilton, S.K. and J.D. Bloom, *Modeling site-specific amino-acid preferences deepens phylogenetic estimates of viral sequence divergence.* Virus Evolution, 2018. **4**(2): p. vey033.
18. J. D. Bloom, *An Experimentally Determined Evolutionary Model Dramatically Improves Phylogenetic Fit.* Molecular Biology and Evolution, 2014. **31**(8): p. 1956-1978.
19. P. C. Jain and R. Varadarajan, *A rapid, efficient, and economical inverse polymerase chain reaction-based method for generating a site saturation mutant library.* Analytical Biochemistry, 2014. **449**: p. 90-98.

20. V. Druelle and R.A. Neher, *Reversions to consensus are positively selected in HIV-1 and bias substitution rate estimates.* 2022, bioRxiv. https://doi.org/10.1101/2022.02.13.480259.
21. J. H. Tay, et al., *The Emergence of SARS-CoV-2 Variants of Concern Is Driven by Acceleration of the Substitution Rate.* Molecular Biology and Evolution, 2022. **39**(2): p. msac013.
22. K.A. Lythgoe, et al., *Lineage replacement and evolution captured by the United Kingdom Covid Infection Survey.* medRxiv, 2022. https://doi.org/10.1101/2022.01.05.21268323.
23. T. Ohta, *The Nearly Neutral Theory of Molecular Evolution.* Annual Review of Ecology and Systematics, 1992. **23**: p. 263-286.
24. E. C. Holmes, *Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus.* Journal of Virology, 2003. **77**(20): p. 11296-11298.
25. S. Lequime, et al., *Genetic Drift, Purifying Selection and Vector Genotype Shape Dengue Virus Intra-host Genetic Diversity in Mosquitoes.* Plos Genetics, 2016. **12**(6): p. e1006111.
26. G. V. S. Jerzak, et al., *Genetic diversity and purifying selection in West Nile virus populations are maintained during host switching.* Virology, 2008. **374**(2): p. 256-260.
27. C. T. T. Edwards, et al., *Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection.* Genetics, 2006. **174**(3): p. 1441-1453.
28. P. G. T. Walker, et al., *The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries.* Science, 2020. **369**(6502): p. 413-422
29. D. E. McMahon, et al., *Global resource shortages during COVID-19: Bad news for low-income countries.* Plos Neglected Tropical Diseases, 2020. **14**(7): p. e0008412.
30. I. Chitungo, et al., *COVID-19: Unpacking the low number of cases in Africa.* Public Health Pract (Oxf), 2020. **1**: p. 100038.
31. J. Hopman, B. Allegranzi, and S. Mehtar, *Managing COVID-19 in Low- and Middle-Income Countries.* JAMA, 2020. **323**(16): p. 1549-1550.
32. L. L. Liang, et al., *Covid-19 mortality is negatively associated with test number and government effectiveness.* Scientific Reports, 2020. **10**(1): p. 1-7.
33. M. Dahab, et al., *COVID-19 control in low-income settings and displaced populations: what can realistically be done?* Conflict and Health, 2020. **14**(1): p. 1-6.
34. A. B. Suthar, et al., *Evaluation of approaches to strengthen civil registration and vital statistics systems: A systematic review and synthesis of policies in 25 countries.* Plos Medicine, 2019. **16**(9): p. e1002929.
35. C. AbouZahr, et al., *The COVID-19 Pandemic: Effects on Civil Registration of Births and Deaths and on Availability and Utility of Vital Events Data.* American Journal of Public Health, 2021. **111**(6): p. 1123-1131.
36. A. A. Gebru, et al., *Global burden of COVID-19: Situational analyis and review.* Hum Antibodies, 2021. **29**(2): p. 139-148.
37. B. Rothinger and P. Tautu, *On the Genealogy of Large Cell-Populations.* Workshop on Stochastic Modelling in Biology, 1990: p. 166-235.

38. R. C. Griffiths and S. Tavare, *Sampling Theory for Neutral Alleles in a Varying Environment.* Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 1994. **344**(1310): p. 403-410.

Appendices

Appendix 1: Supplementary material for chapter 2

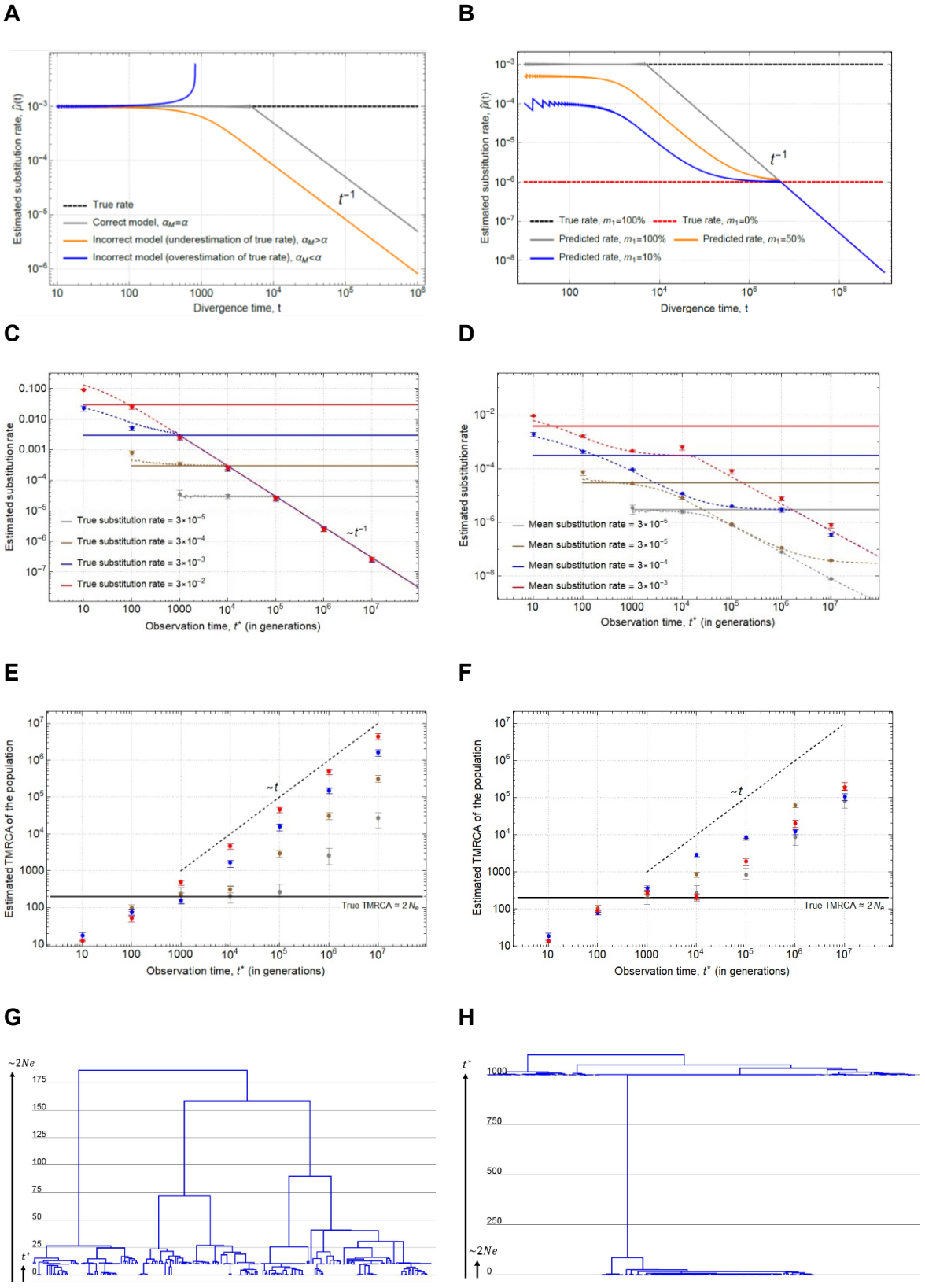


Figure S1: **Estimated substitution rates, time to the most recent common ancestors, and phylogenetic trees of simulated datasets under different substitution rates. Related to Figure 2.1. (A)** Estimating the substitution rate $\mu = 10^{-3}$ (black dashed line) using Equation S2 for a pair of sequences that have diverged from each other t

generations ago under the correct substitution model (gray), a model that over-estimates the true saturation frequency (orange), α, and a model that under-estimates the true saturation frequency (blue). **(B)** Estimating the expected substitution rate, $< \mu >$, when a fraction m_1 of sites evolve at rate $\mu_1 = 10^{-3}$ (black dashed line) and the remaining fraction, $1 - m_1$, at rate $\mu_2 = 10^{-6}$ (red dashed line). The expression t^{-1} in both plots shows the dominating term in rate decay with respect to divergence time, t, corresponding to slope -1 on the graphs. (**C**) A population of size N_e evolving neutrally with $L_n = 100$ sites evolving at the same rate, μ, and (**D**) a model with rate heterogeneity across sites such that $L_1 = 100$ sites evolve at rate μ and the remaining $L_2 = 900$ sites at rate μ^2. The rate is estimated as a function of observation gap, t^*, between when the first and second sampling point from the population. Dashed lines show the theoretical prediction according to Equation S6 and solid lines show the mean (expected) rates used for the simulations. (**E**) and (**F**) show the estimated TMRCA for the first group of sampled sequences, i.e. at $t^* = 0$, and solid lines show the mean TMRCA according to neutral theory. The rates and TMRCAs are estimated using BEAST under a strict clock assumption. Dots represent the median values taken from 100 independent runs and error bars show the interquartile region. (**G**) and (**H**) show the maximum clade credibility trees for one simulation run corresponding to $\mu = 3 \times 10^{-5}$ when the observation gap is $t^* = 10,\ t^* \ll 2N_e$ and $t^* = 1000,\ t^* \gg 2N_e$, respectively.

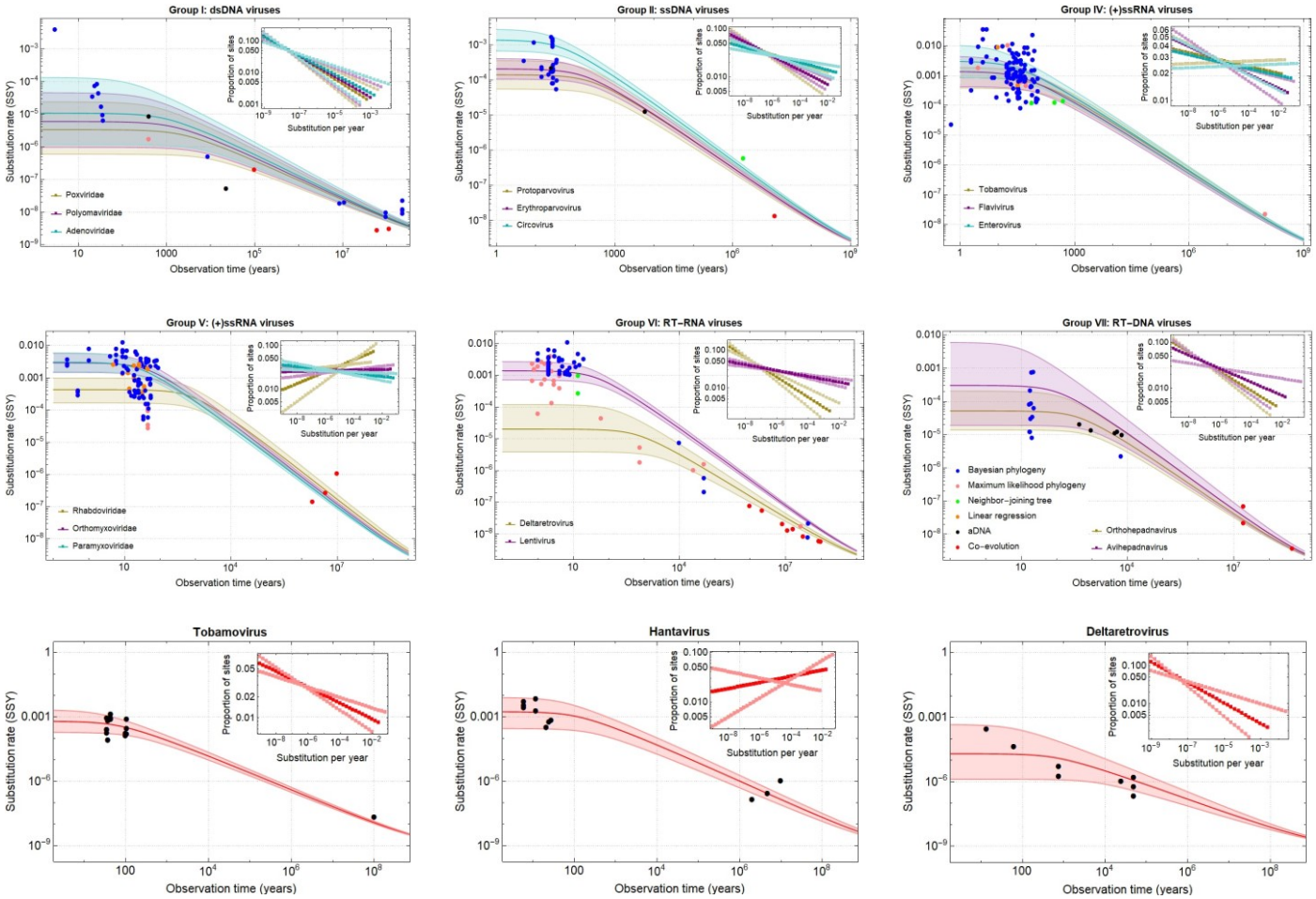
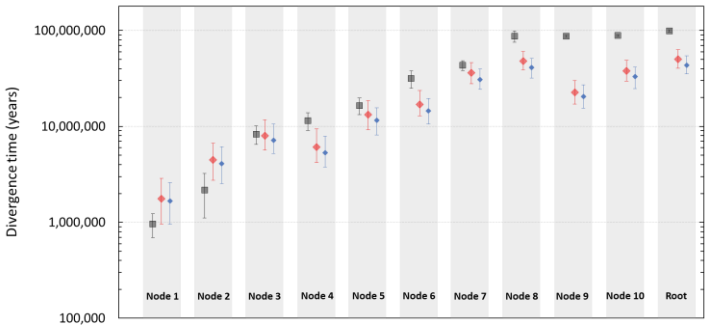
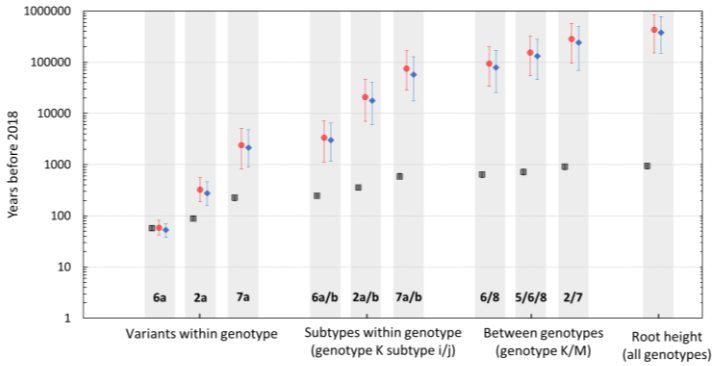


Figure S2: **Estimated time-dependent rate curves for each viral group and three selected genera according to the PoW model and their corresponding distribution of rate groups (inset). Related to Figure 2.2.** The first two rows show two or three distinct mean substitution rates (coloured in gold, purple, and blue) that are selected for each virus family to estimate rate curves according to the PoW model. The bottom row shows the same rate curve estimations for Tobamoviruses with $\langle\mu\rangle = 0.6(0.2-2)\times10^{-3}$ SSY and $\mu_{max} = 3(1-6)\times10^{-2}$ SSY, Hantaviruses with $\langle\mu\rangle = 2(0.3-8)\times10^{-3}$ SSY and $\mu_{max} = 2(0.6-4)\times10^{-2}$ SSY, and Deltaretroviruses with $\langle\mu\rangle = 2(0.1-50)\times10^{-5}$ SSY and $\mu_{max} = 3(2-30)\times10^{-3}$ SSY. These genera are selected as they have the widest timespan of rate measurement. The solid lines show the best fit and shaded areas the 95% confidence interval ($\Delta M = 1.58$ and $\alpha_M = \alpha = 3/4$).

A



B



C

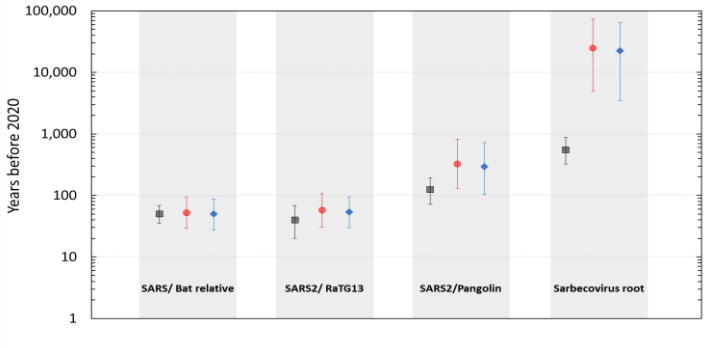


Figure S4: **Estimated divergence times for FV, HCV, and sarbecovirus datasets using standard substitution models and PoW transformation with uncertainty in the underlying clock model. Related to Figure 2.3.** Same legend as Figure 2.3. (**A**) Estimated divergence times between labelled internal nodes of FV (see Figure 2.3). For the PoW transformation using the HKY85 and JC69 substitution models, the short-term substitution rate, $\langle \mu \rangle$, was randomly sampled from the Normal distribution, $N(\text{mean} = 6.1\times10^{-6}, \text{stdev} = 2.0\times10^{-6})$. (**B**) For HCV, we compare the results between a standard HKY+G model and two POW-transformed estimates using the HKY85 and JC69 substitution models. The short-term substitution rate, $\langle \mu \rangle$, was randomly sampled from the posterior rate distribution of the standard HKY+G model. (**C**) Similarly, for Sarbecovirus, we compare the results between a standard HKY+G model

and two POW-transformed estimates using the HKY85 and JC69 substitution models. The short-term substitution rate, $\langle \mu \rangle$, was randomly sampled from the posterior rate distribution of the standard HKY+G model. In all the three datasets, the fastest-evolving rate group, $\mu_{\max}$, was sampled from the Normal distribution, $N(\text{mean} = 3.65 \times 10^{-2}, \text{stdev} = 5.0 \times 10^{-3})$.

Table S1: **Estimated short-term substitution rate and fastest-evolving rate group across 6 viral groups. Related to Table 2.1.** Only short-term rate estimates (measured over time scales of <100 years) – along with the long-term rate estimates (>100 years) from the entire data set – from a particular virus family or genus (rather than the entire viral group) is used for rate calibration in the PoW model. Parentheses correspond to 95% confidence intervals.

Viral group	Virus family/genus used for calibration	Short-term substitution rate, $\langle \mu \rangle$	Fastest rate group, $\mu_{\max}$
Group I	Poxviridae	$0.3(0.06-2)\times10^{-5}$	$0.6(0.2-3)\times10^{-3}$
	Polyomaviridae	$0.6(0.1-5)\times10^{-5}$	$1(0.3-4)\times10^{-3}$
	Adenoviridae	$1(0.1-10)\times10^{-5}$	$2(0.3-10)\times10^{-3}$
Group II	Protoparvovirus	$1(0.6-4)\times10^{-4}$	$0.6(0.4-1)\times10^{-2}$
	Erythroparvovirus	$2(1-4)\times10^{-4}$	$1(0.6-2)\times10^{-2}$
	Circovirus	$10(7-300)\times10^{-4}$	$4(3-6)\times10^{-2}$
Group IV	Tobamovirus	$0.4(0.9-2)\times10^{-3}$	$2(1-3)\times10^{-2}$
	Flavivirus	$1(0.4-5)\times10^{-3}$	$4(2-10)\times10^{-2}$
	Enterovirus	$3(0.9-10)\times10^{-3}$	$6(3-10)\times10^{-2}$
Group V	Rhabdoviridae	$0.4(0.2-1)\times10^{-3}$	$0.3(0.2-0.4)\times10^{-2}$
	Orthomyxoviridae	$3(1-6)\times10^{-3}$	$4(3-6)\times10^{-2}$
	Paramyxoviridae	$3(2-6)\times10^{-3}$	$6(4-10)\times10^{-2}$
Group VI	Deltaretrovirus	$0.02(0.004-0.1)\times10^{-3}$	$0.3(0.06-1)\times10^{-2}$
	Lentivirus	$1(0.7-3)\times10^{-3}$	$4(3-6)\times10^{-2}$
Group VII	Orthohepadnavirus	$5(1-20)\times10^{-5}$	$4(2-10)\times10^{-3}$
	Avihepadnavirus	$30(2-600)\times10^{-5}$	$20(2-100)\times10^{-3}$

Appendix 2: Supplementary material for chapter 4

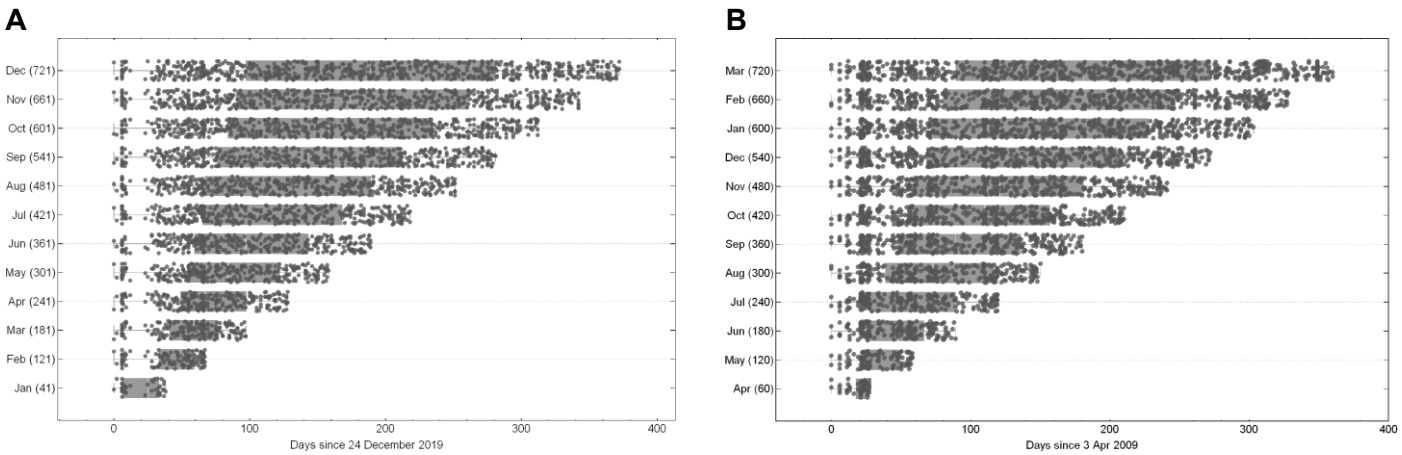


Figure S1: **Sampling intervals and numbers of sequences in the different datasets used for SARS-CoV-2 and pH1N1 analyses.** The sampling interval is measured from the day when the first genome was collected.

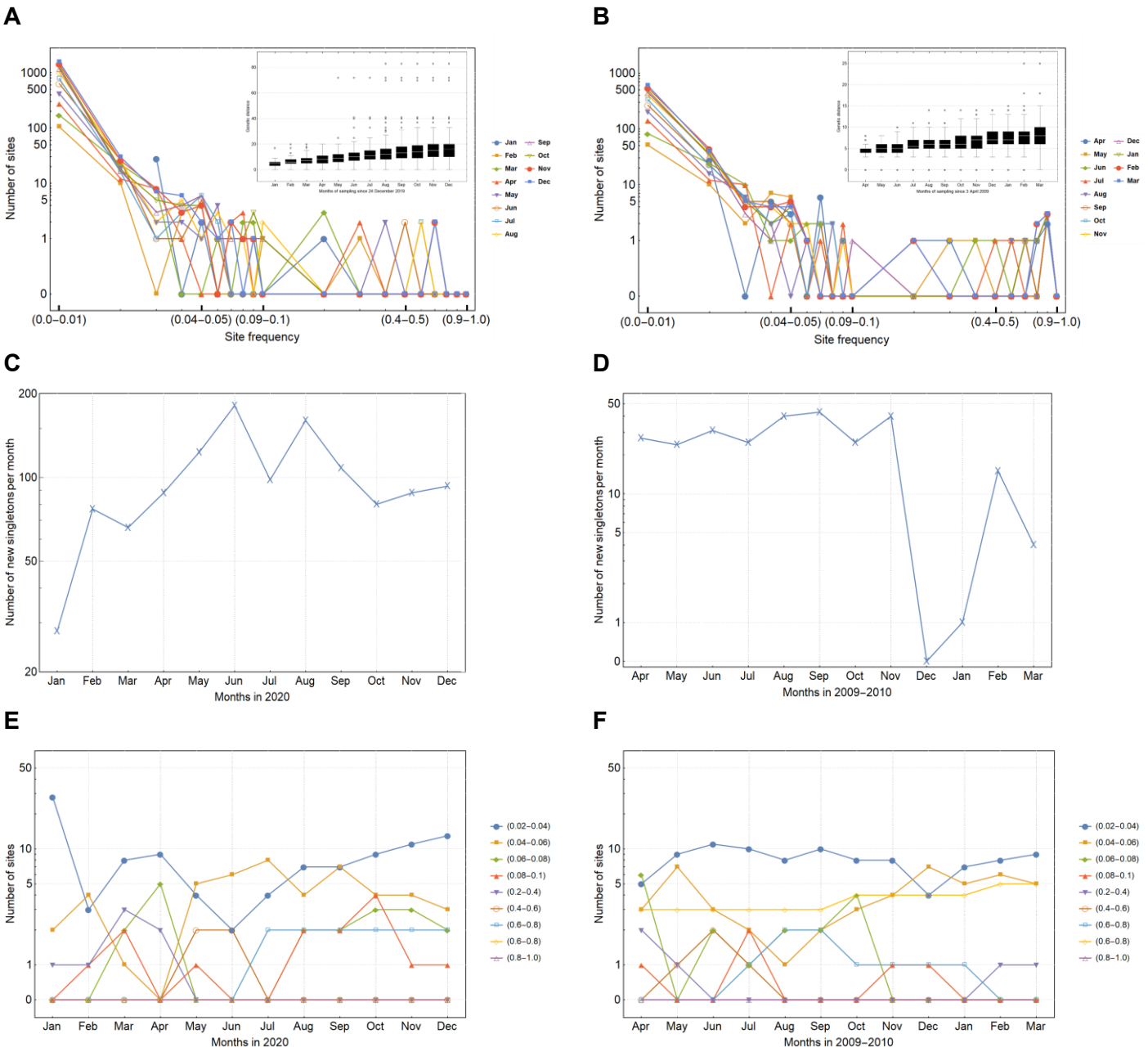


Figure S2: **(A, B)** Site frequency spectrum and genetic distance (in the inset) over time for SARS-CoV-2 (left column) and pH1N1 (right column). By comparing the main and ancestral alignments, each site in the main alignment is defined as invariant if it is identical to the ancestral nucleotide, or variant if it is not. Number of sites are counted based on the number of differences between the ancestral and main nucleotide positions. **(C, D)** Number of new single-nucleotide differences (singletons) added to the dataset per month. **(E, F)** Number of polymorphic sites in each frequency range over time.

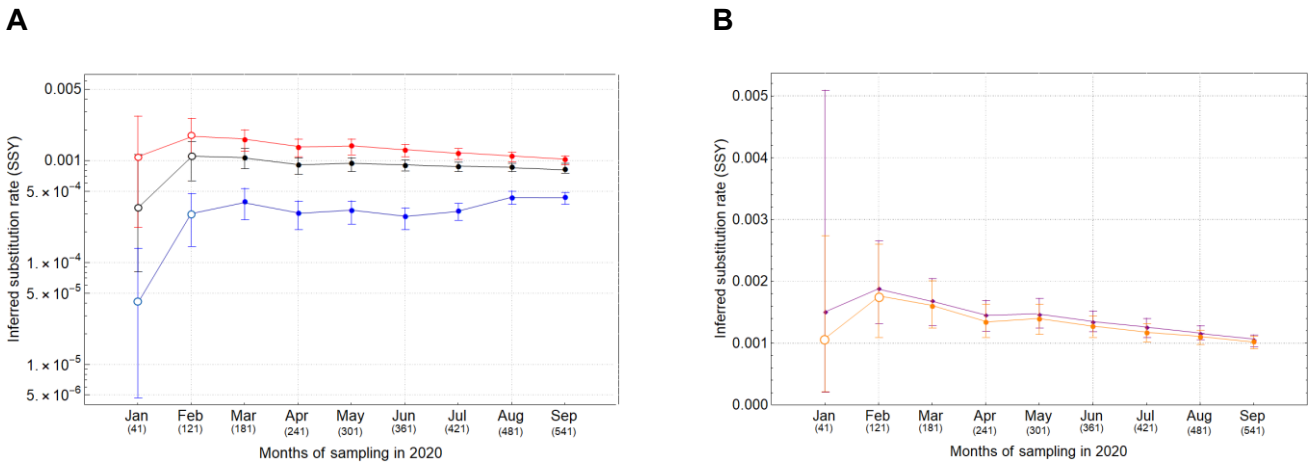


Figure S3: **(A)** Mean clock rate (black), the rates at the terminal (red) and internal (blue) branches of SARS-CoV-2 alignments with masked sites. **(B)** Inferred substitution rate at terminal branches of SARS-CoV-2 alignments without (purple) and with (orange) masked sites. Open circles represent nonconvergence for at least one parameter in the Bayesian analysis.

Appendix 3: Supplementary material for chapter 5

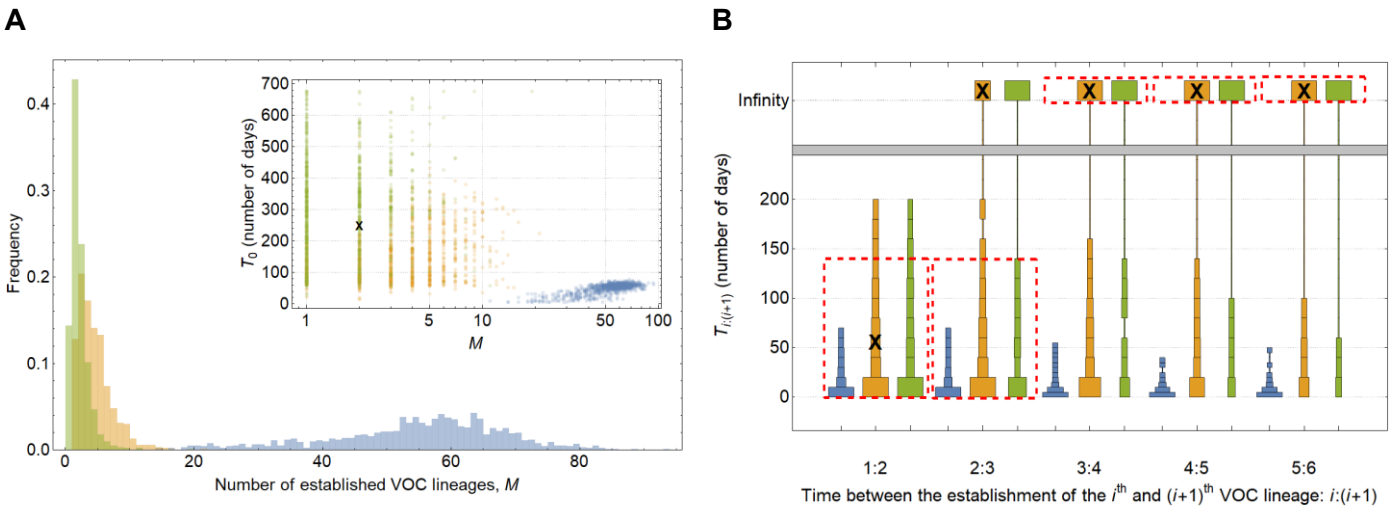


Figure S1: **Evolutionary dynamics of VOC emergence via the between-host model of evolution with a single adaptive mutation (Landscape 1). (A)** The total number of VOC lineages established (*M*) and **(B)** the distribution of time it takes between the production of the ith and $(i+1)^{th}$ lineage, $T_{i:(i+1)}$, under the between-host evolutionary pathway corresponding to Landscape 1 with levels of overdispersion, *k*, where *k*=0.2 (blue) *k*=0.005 (orange), and *k*=0.001 (green). In all three scenarios, IFR=1.5%, μ=0.87x10^{-5}, *K*=1, and *s*=0.4. These three scenarios correspond to the same set of simulations shown in Figure 3A. The inset in (A) shows a graph of *M* with respect to T_0 for each simulation point. The cross signs ("**X**") highlight the values corresponding to the simulation run shown in Figure 3D.

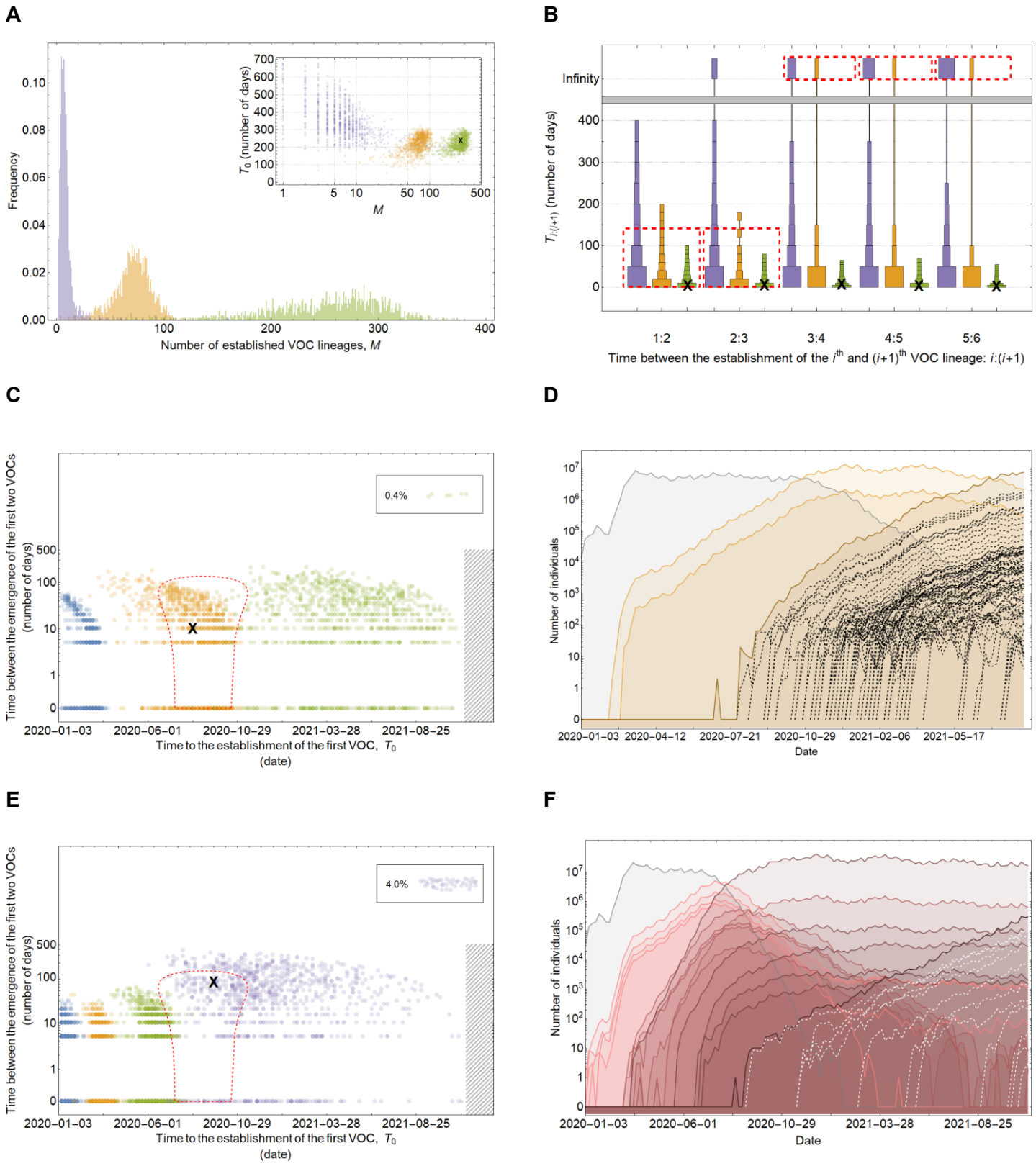


Figure S2: **Evolutionary dynamics of VOC emergence via the between-host model of evolution with additive mutations (Landscape 2). (A, B)** The total number of VOC lineages established (M) and the distribution of time it takes between the production of the ith and $(i+1)^{th}$ lineage, $T_{i:(i+1)}$, via the between-host pathway requiring K=4 (purple), K=3 (green), and K=2 (orange) mutations. For the K=4 scenario, IFR=0.2%, μ=2x10^{-5}, k=0.2, and s=1.0. For the K=3 scenario, IFR=0.5%, μ=1x10^{-5}, k=0.2, and s=1.0. For the K=3 scenario, IFR=0.5%, μ=0.87x10^{-5}, k=0.05, and s=0.4. These three scenarios correspond to the same set of simulations

shown in Figure 4A. The cross signs ("**X**") highlight the values corresponding to the simulation run shown in Figure 4B. **(C, D)** The establishment time of the first two VOC lineages and a typical evolutionary trajectory of the between-host pathway, highlighted with cross signs ("**X**"), corresponding to a K=2 (orange) scenario with IFR=0.2%, μ=2x10^{-5}, k=0.2, and s=1.0. **(E, F)** The establishment time of the first two VOC lineages and a typical evolutionary trajectory of a between-host pathway, highlighted with cross signs ("**X**"), corresponding to a K=4 (purple) scenario such that IFR=0.2%, μ=2x10^{-5}, k=0.2, and s=1.0.

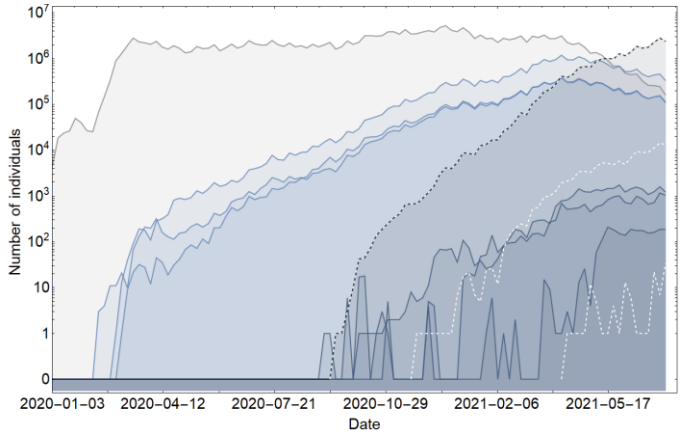


Figure S3: **Evolutionary dynamics of VOC emergence via the within-host model of evolution with additive mutations (Landscape 2).** A typical evolutionary trajectory of the within-host pathway corresponding to a K=3, P_f=3.5x10^{-8}, and μ_C=0.15 (blue) scenario with μ=0.87x10^{-5}, IFR=1.5%, k=0.05, and s=0.3 shown in Figure 4C. The graph shows the background population in gray and the single-, double-, and triple-mutant (VOC) populations in different shades of blue from light to dark, respectively.

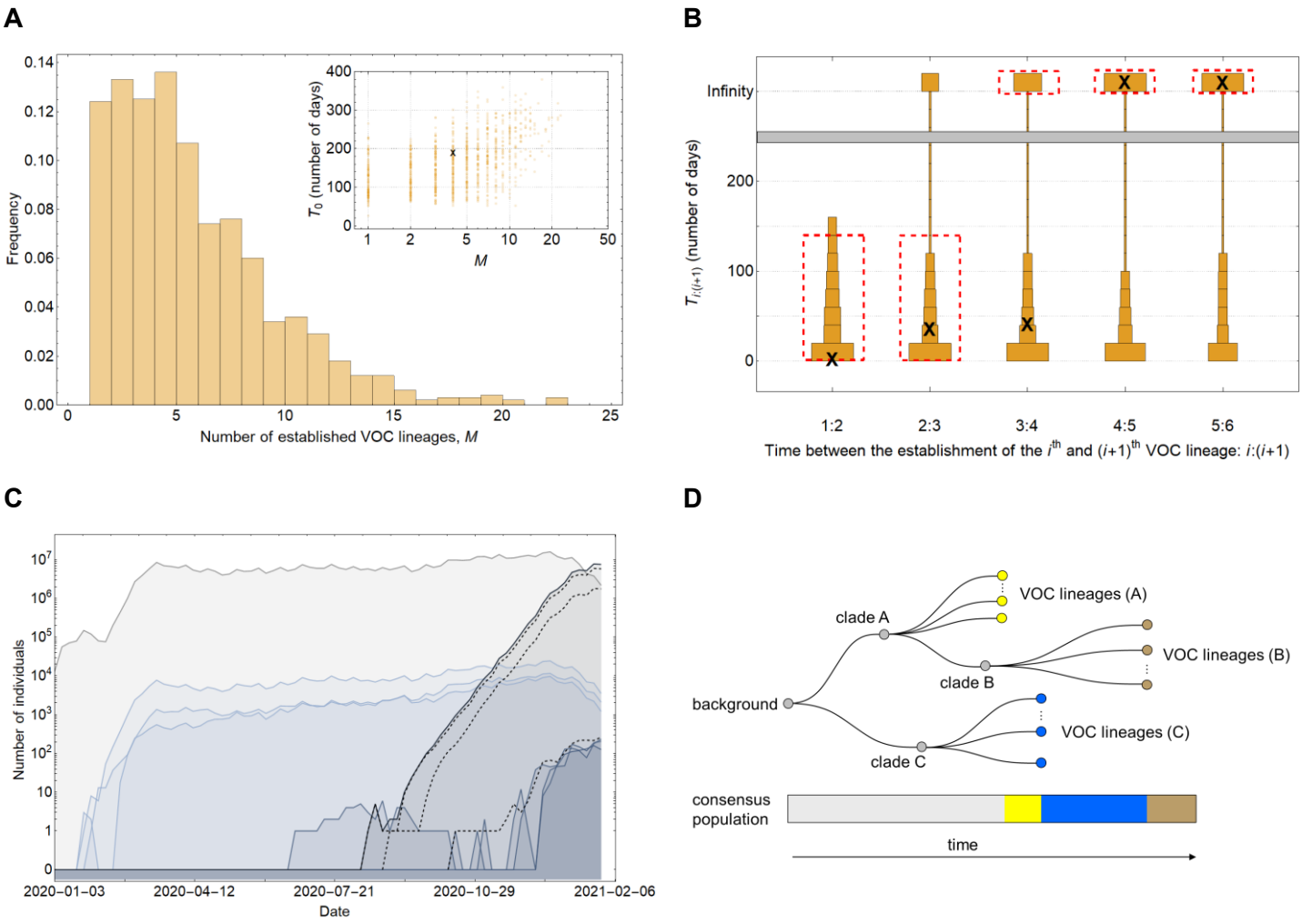


Figure S4: **Evolutionary dynamics of VOC emergence via the within- and between-host models of evolution with fitness plateau crossing (Landscape 3). (A, B)** The total number of VOC lineages established (*M*) and the distribution of time it takes between the production of the ith and $(i+1)^{th}$ lineage, $T_{i:(i+1)}$, via the between-host pathway corresponding to the *K*=2 scenario shown in Figure 5A. **(C)** A typical evolutionary trajectory of the within-host model of VOC emergence corresponding to the *K*=3 scenario shown in Figure 5C where P_f=2x10^{-8}, and μ_C=0.1, μ=1x10^{-5}, IFR=0.5%, *k*=0.1, and *s*=0.7. The graph shows the background population in gray and the single-, double-, and triple-mutant (VOC) populations in different shades of blue from light to dark, respectively. **(D)** Phylogenetic relation between different VOC lineages. For a wide range of parameter combinations in Landscape 1, 2, and 3, we showed that the within-host evolutionary dynamics can become virtually uncoupled from the the evolution at the between-host level enabling each VOC lineage to arise independently on the background of a different clade which leads to a star-like tree topology.

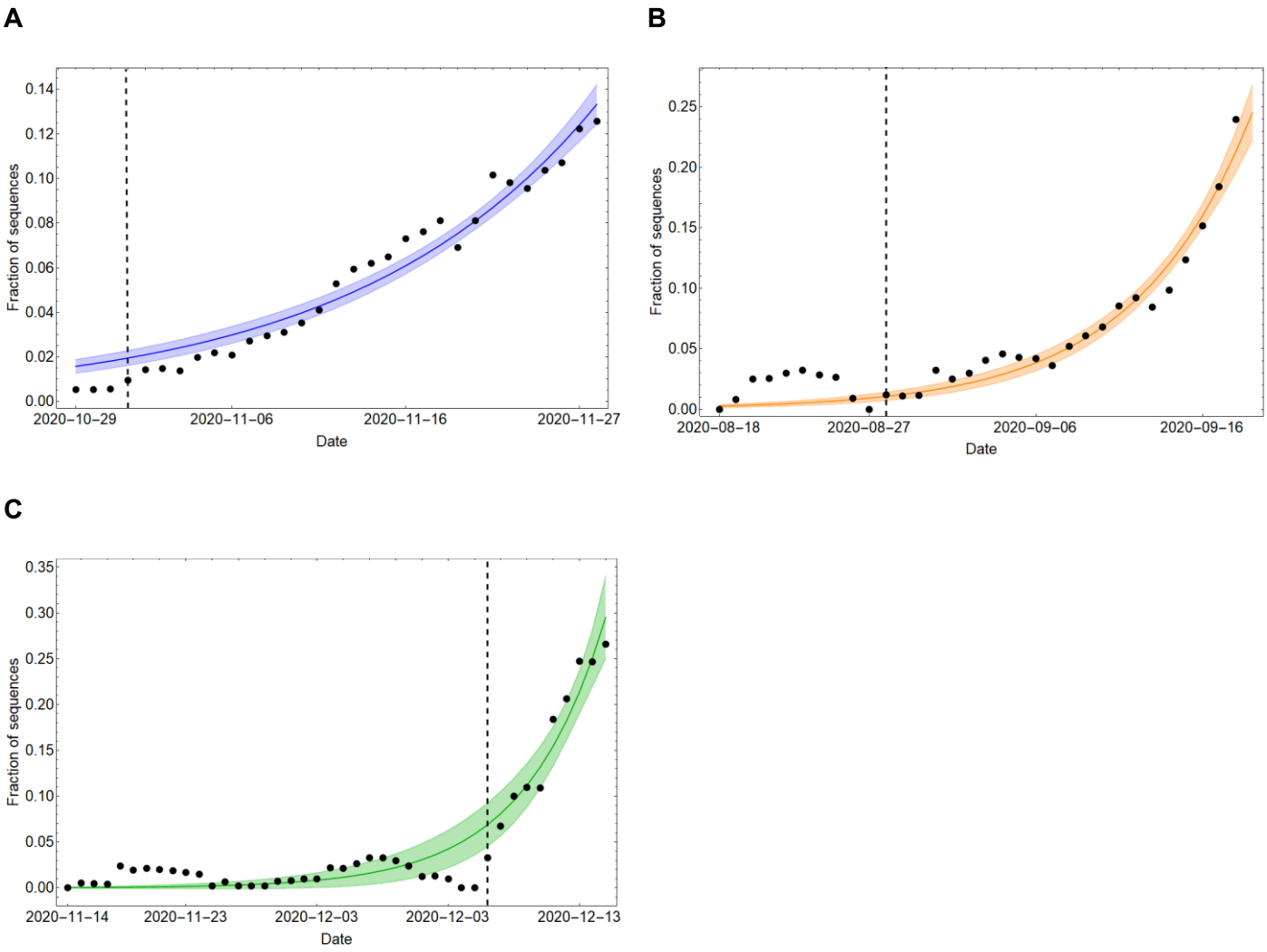


Figure S5: **Exponential model fits to the frequency of individual SARS-CoV-2 VOC sequences sampled in its country of origin. (A-C)** Fitting an exponential function of the form, $f(t)=ae^{bt}$, to the frequency of Alpha, Beta, and Gamma sequences sampled in the UK, South Africa, and Brazil, respectively. Vertical dashed line shows the starting timepoint used for the fitting. The shaded area shows the mean prediction bands.

Appendix 4: Supplementary material for chapter 6

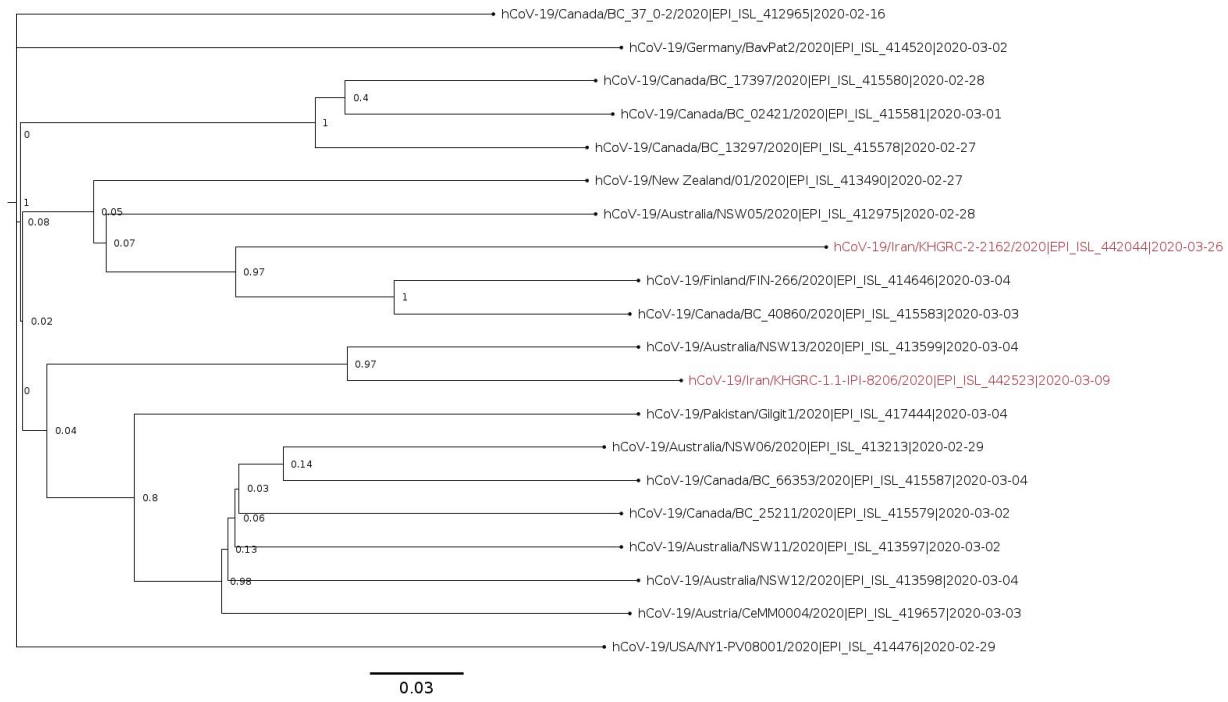


Figure S1: **Maximum a posteriori tree of SARS-CoV-2 genomes linked to returned travellers from Iran.** Following our analysis on BEAST to estimate the TMRCA and early doubling times for this clade, we identify the MCMC step with the highest posterior probability and use TreeAnnotator v2.6.0 to summarise the information from a sample of trees onto the maximum a posteriori tree (i.e., the 'target' tree). Numbers on each node represent their posterior support and the scale bar shows the unit of time in years. Genomes from Iran are highlighted in red.

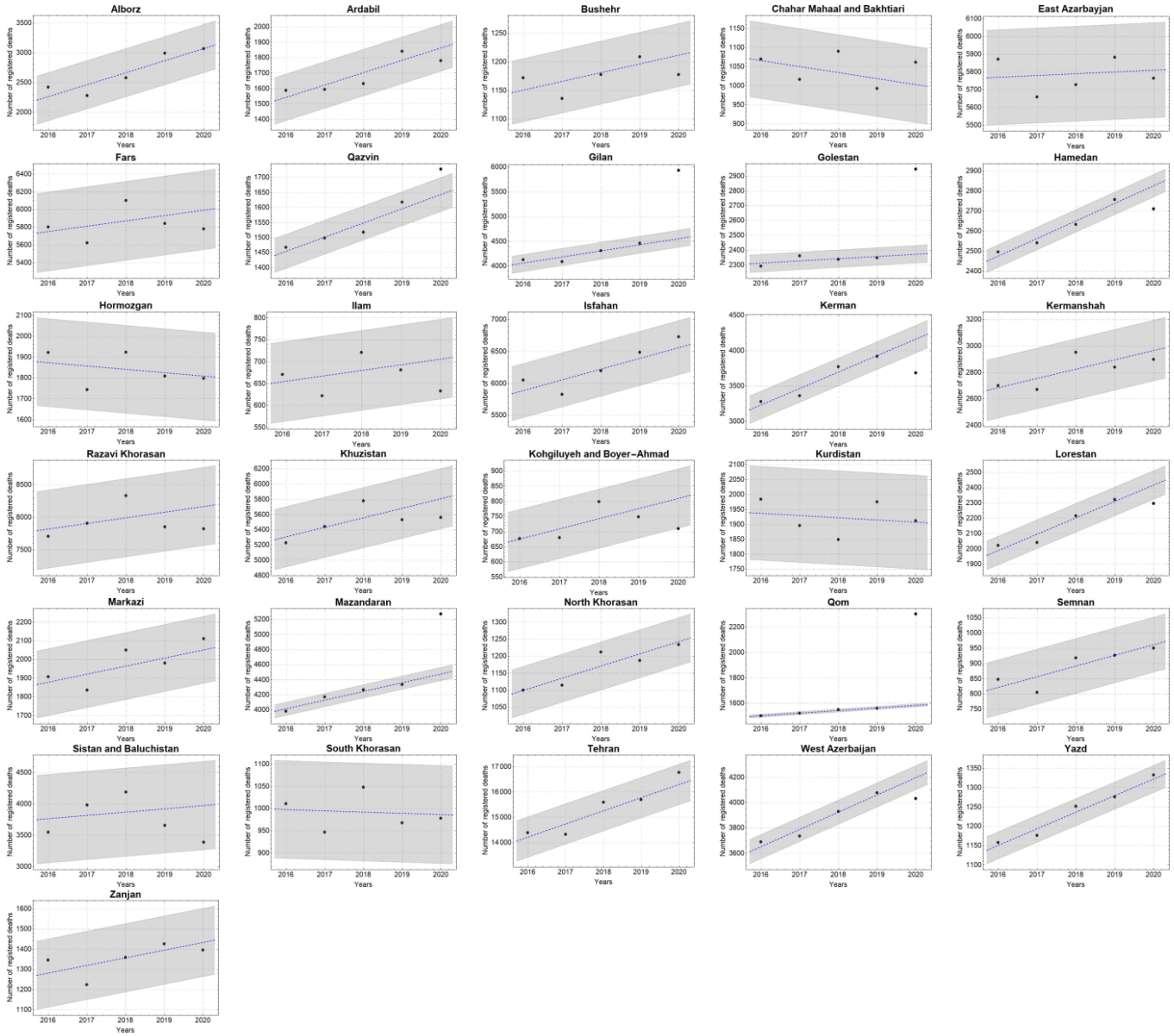


Figure S2: **Province-level pattern of excess mortality during winter** This record covers every registered death over the last five years including last winter (from 2019-12-22 to2020-03-19). Gray areas show the 95% confidence interval.

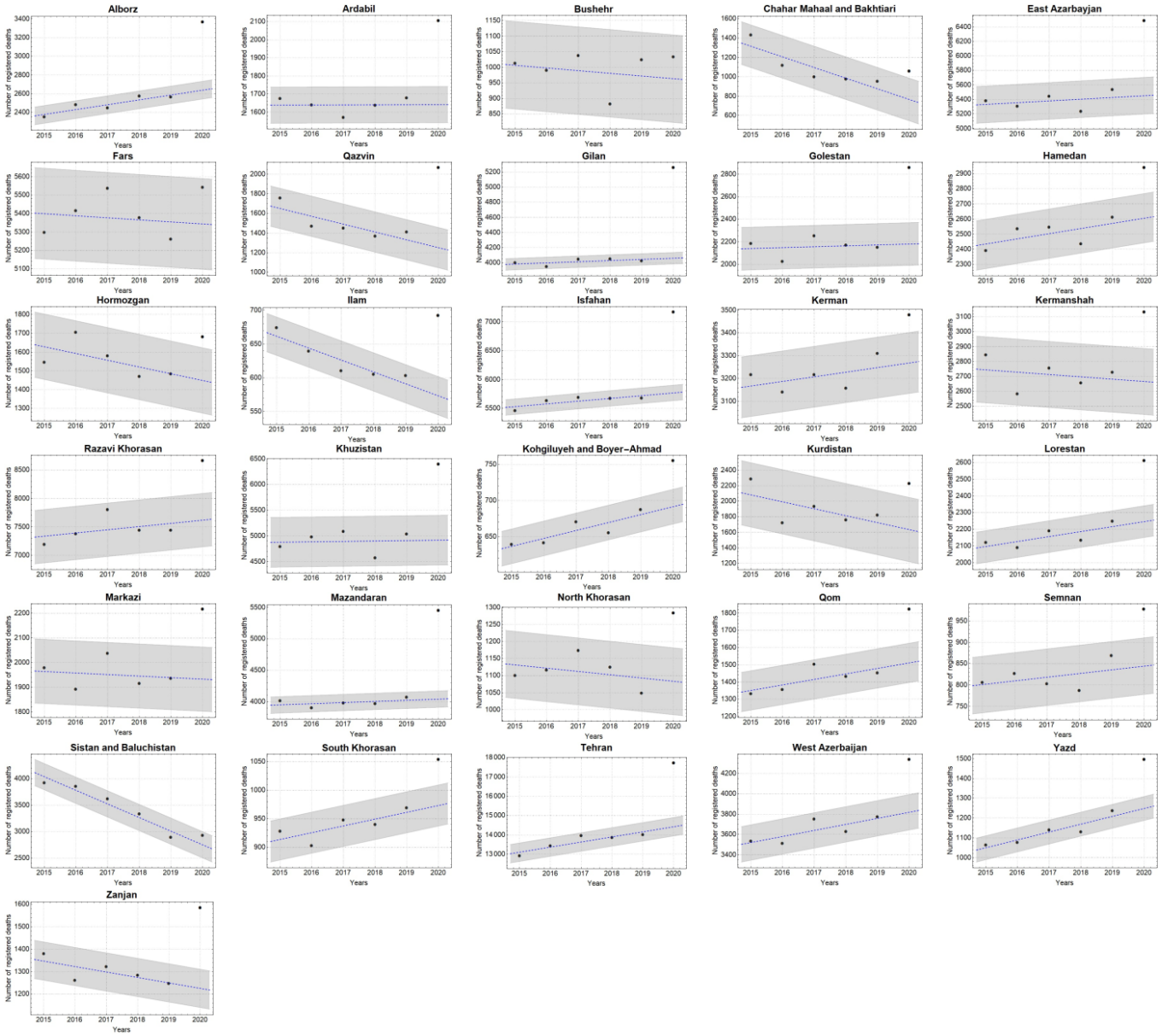


Figure S3: **Province-level pattern of excess mortality during spring.** This record covers every registered death over the last five years including last spring (from 2020-03-20 to2020-06-20). Gray areas show the 95% confidence interval.

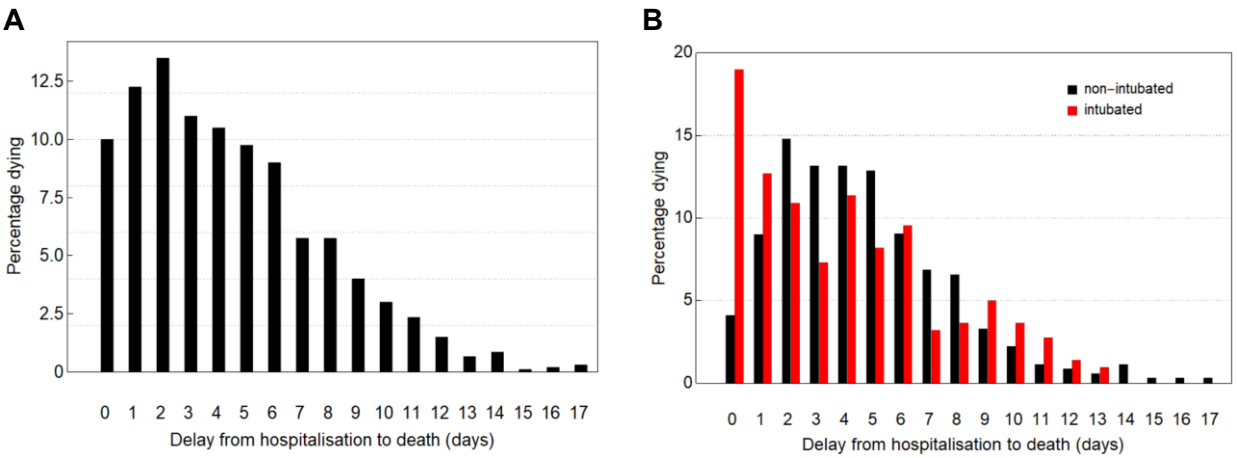


Figure S4: **Delay from hospitalization to death in Iran.** (**A**) Delay from hospital admission to hospital death [50]. (**B**) Comparison of delay from hospitalization to hospital death for patients with (red) and without tracheal intubation (black).

Table S1: **Sensitivity analysis for the number of active cases based on confirmed exported cases.**

Lebanon* (21 Feb)				Catchment population (M= million individuals)			
				min = 11.1M		max = 55.6M	
				Airplane capacity			
				min = 50%	max = 90%	min = 50%	max = 90%
Percentage of undetected pre- & asymptomatic cases	min =40%	Exposure time (days)	min = 7	54,700 (16,700 – 202,900)	30,000 (8,900 – 112,300)	277,100 (87,600 – 1,018,400)	153,500 (48,200 – 565,400)
			max = 13	29,000 (8,600 – 108,800)	15,700 (4,300 – 60,000)	148,800 (46,700 – 547,900)	82,300 (25,500 – 304,000)
	max = 80%		min = 7	54,700 (16,700 – 202,900)	30,000 (8,900 – 112,300)	277,100 (87,600 – 1,018,400)	153,500 (48,200 – 565,400)
			max = 13	29,000 (8,600 – 108,800)	15,700 (4,300 – 60,000)	148,800 (46,700 – 547,900)	82,300 (25,500 – 304,000)

UAE (22 Feb)				Catchment population (M= million individuals)			
				min = 11.1M		max = 55.6M	
				Airplane capacity			
				min = 50%	max = 90%	min = 50%	max = 90%
Percentage of undetected pre- & asymptomatic cases	min =40%	Exposure time (days)	min = 7	5,000 (1,800 – 14,600)	2,800 (1,000 – 8,000)	24,800 (9,500 – 73,200)	13,800 (5,200 – 40,600)
			max = 13	2,700 (900 – 7,800)	1,500 (500 – 4,300)	13,400 (5,100 – 39,400)	7,400 (2,800 – 21,800)
	max = 80%		min = 7	6,600 (2,700 – 17,000)	3,700 (1,500 – 9,400)	33,100 (14,000 – 85,500)	18,400 (7,700 – 47,400)
			max = 13	3,600 (1,400 – 9,100)	2,000 (1,500 – 5,000)	17,800 (7,500 – 46,000)	9,900 (4,100 – 25,500)

Oman (24 Feb)				Catchment population (M= million individuals)			
				min = 11.1M		max = 55.6M	
				Airplane capacity			
				min = 50%	max = 90%	min = 50%	max = 90%
Percentage of undetected pre- & asymptomatic cases	min =40%	Exposure time (days)	min = 7	25,100 (9,200 – 74,200)	13,900 (5,100 – 41,200)	125,400 (46,600 – 371,400)	69,700 (25,800 – 206,300)
			max = 13	13,500 (4,900 – 39,900)	7,500 (2,700 – 22,100)	67500 (25000 - 199900)	37,500 (13,900 – 111,000)
	max = 80%		min = 7	33,400 (13,800 – 86,600)	18,600 (7,600 – 48,100)	167200 (69200 - 433400)	92,900 (38,400 – 240,700)
			max = 13	18,000 (7,400 – 46,600)	10,000 (4,000 – 25,800)	90,000 (37,200 – 233,300)	50,000 (20,600 – 129,600)

Kuwait (25 Feb)				Catchment population (M= million individuals)			
				min = 11.1M		max = 55.6M	
				Airplane capacity			
				min = 50%	max = 90%	min = 50%	max = 90%
Percentage of undetected pre- & asymptomatic cases	min =40%	Exposure time (days)	min = 7	5,000 (1,800 – 14,600)	2,800 (1,000 – 8,000)	24,800 (9,500 – 73,200)	13,800 (5,200 – 40,600)
			max = 13	2,700 (900 – 7,800)	1,500 (500 – 4,300)	13,400 (5,100 – 39,400)	7,400 (2,800 – 21,800)
	max = 80%		min = 7	6,600 (2,700 – 17,000)	3,700 (1,500 – 9,400)	33,100 (14,000 – 85,500)	18,400 (7,700 – 47,400)
			max = 13	3,600 (1,400 – 9,100)	2,000 (1,500 – 5,000)	17,800 (7,500 – 46,000)	9,900 (4,100 – 25,500)

China (6 Mar)				Catchment population (M= million individuals)			
				min = 11.1M		max = 55.6M	
				Airplane capacity			
				min = 50%	max = 90%	min = 50%	max = 90%
Percentage of undetected pre- & asymptomatic cases	min =40%	Exposure time (days)	min = 7	197,700 (149,600 – 277,000)	109,800 (83,100 – 153,800)	988,400 (748,400 – 1,385,300)	549,100 (415,700 – 769,600)
			max = 13	106,400 (80,500 – 149,100)	59,100 (44,700 – 82,800)	532,200 (402,900 – 745,900)	295,700 (223,800 – 414,300)
	max = 80%		min = 7	252,100 (196,400 – 339,100)	140,000 (109,100 – 188,400)	1,260,300 (982,300 – 1,696,100)	700,100 (545,700 – 942,200)
			max = 13	135,700 (105,700 – 182,600)	75,400 (58,700 – 101,400)	678,600 (528,900 – 91,3200)	377,000 (293,800 – 507,300)

Table S2: **Excess mortality in 31 provinces during Winter 2020.** This record covers every registered death from 22 December 2019 to 19 March 2020. Provinces with significantly higher excess deaths compared to previous years are highlighted in cyan.

Province	Expected deaths (regression)	Registered deaths	Excess deaths	%Excess deaths	SD Excess deaths	Excess deaths lower bound (CI 95%)	Excess deaths upper bound (CI 95%)	Significant excess death
Qom	1,583		723	46%	6	710	735	+
Gilan	4,556		1,382	30%	86	1,209	1,554	+
Golestan	2,373		575	24%	30	515	634	+
Mazandaran	4,477		797	18%	44	709	885	+
Charmahal o Baxtiyari	1,003		58	6%	50	-41	157	-
Qazvin	1,644		85	5%	28	28	141	+
Markazi	2,052	2,112	60	3%	89	-119	239	-
Tehran	16,306	16,773	467	3%	399	-331	1,265	-
Isfahan	6,561	6,729	169	3%	212	-255	592	-
Yazd	1,323	1,333	10	1%	18	-25	45	-
Kordestan	1,908	1,912	5	0%	78	-152	161	-
Alborz	3,072	3,070	-2	0%	202	-406	403	-
Hormozgan	1,810	1,799	-11	-1%	105	-221	200	-
North Khorasan	1,243	1,234	-9	-1%	35	-79	61	-
East Azarbayjan	5,811	5,765	-46	-1%	134	-314	222	-
South Khorasan	987	978	-9	-1%	55	-118	101	-
Semnan	962	950	-12	-1%	45	-102	78	-
Kermanshah	2,967	2,899	-68	-2%	115	-297	162	-
Zanjan	1,434	1,396	-38	-3%	84	-206	131	-
Bushehr	1,212	1,178	-34	-3%	28	-89	21	-
Fars	5,994	5,780	-214	-4%	222	-658	231	-
West Azarbayjan	4,198	4,032	-166	-4%	48	-261	-71	-
Hamedan	2,827	2,711	-116	-4%	28	-172	-59	-
Khorasan Razavi	8,168	7,825	-343	-4%	299	-941	256	-
Khuzestan	5,809	5,562	-247	-4%	198	-643	149	-
Ardabil	1,864	1,781	-83	-4%	76	-234	68	-
Lorestan	2,418	2,296	-122	-5%	47	-215	-28	-
Ilam	706	633	-73	-10%	45	-164	18	-
Kerman	4,166	3,687	-479	-11%	97	-673	-284	-
Kohgilueh o Boyerahmad	810	710	-100	-12%	49	-198	-2	-
Sistan o Baluchestan	3,976	3,390	-586	-15%	352	-1290	118	-

Province	Expected deaths (regression)	Registered deaths	Excess deaths	%Excess deaths	SD Excess deaths	Excess deaths lower bound (CI 95%)	Excess deaths upper bound (CI 95%)
5 Provinces	14,633	18,191	3,558	24%	196	3,171	3,949

Table S3: **Excess mortality in 31 provinces during 2020.** This record covers every registered death from 20 March to 20 June 2020. Provinces with significantly higher excess deaths compared to previous years are highlighted in green.

Province	Expected deaths (regression)	Registered deaths	Excess deaths	%Excess deaths	SD Excess deaths	Excess deaths lower bound (CI 95%)	Excess deaths upper bound (CI 95%)	Significant excess death
Qazvin	1,251	2,067	816	65%	103	610	1,021	+
Charmahal o Baxtiyari	766	1,057	291	38%	110	71	511	+
Kordestan	1,635	2,229	594	36%	207	180	1,008	+
Mazandaran	4,040	5,449	1,409	35%	65	1,279	1,539	+
Golestan	2,180	2,858	678	31%	95	489	868	+
Khuzestan	4,919	6,395	1,476	30%	242	993	1,959	+
Gilan	4,059	5,258	1,199	30%	38	1,123	1,275	+
Zanjan	1,225	1,585	360	29%	43	274	445	+
Ardabil	1,641	2,105	464	28%	50	364	564	+
Alborz	2,638	3,365	727	28%	47	632	822	+
Isfahan	5,767	7,172	1,405	24%	67	1,271	1,540	+
Tehran	14,421	17,713	3,292	23%	239	2,814	3,769	+
Ilam	573	692	119	21%	14	90	147	+
Qom	1,511	1,823	312	21%	57	198	426	+
Yazd	1,249	1,496	247	20%	30	187	308	+
East Azarbayjan	5,451	6,485	1,034	19%	126	781	1,287	+
North Khorasan	1,083	1,283	200	18%	49	102	298	+
Kermanshah	2,666	3,132	466	17%	111	244	688	+
Lorestan	2,246	2,611	365	16%	47	270	460	+
Hormozgan	1,449	1,681	232	16%	87	58	406	+
Semnan	844	977	133	16%	33	67	199	+
Markazi	1,933	2,217	284	15%	65	153	414	+
Khorasan Razavi	7,618	8,665	1,047	14%	235	577	1,517	+
West Azarbayjan	3,818	4,338	520	14%	87	345	694	+
Hamedan	2,605	2,940	335	13%	82	172	498	+
Kohgilueh o Boyerahmad	691	755	64	9%	12	40	88	+
South Khorasan	973	1,054	81	8%	18	45	117	+
Bushehr	963	1,033	70	7%	70	-71	210	-
Sistan o Baluchestan	2,750	2,929	179	7%	124	-69	427	-
Kerman	3,268	3,479	211	6%	67	77	344	+
Fars	5,343	5,541	198	4%	123	-49	444	-

Province	Expected deaths (regression)	Registered deaths	Excess deaths	%Excess deaths	SD Excess deaths	Excess deaths lower bound (CI 95%)	Excess deaths upper bound (CI 95%)
28Provinces	82,522	100,881	18,359	22%	2427	13,506	23,212

Table S4: **Estimated reproductive number, R_t.** Upper and lower bound correspond to 95% confidence intervals.

Time interval	Rt, mean	Rt, lower bound	Rt, upper bound
2019-12-25, 2020-02-21	3.3	2.9	3.7
2020-02-22, 2020-03-04	2.14	1.97	2.29
2020-03-05, 2020-04-17	0.82	0.73	0.9
2020-05-18, 2020-05-25	1.15	1.08	1.27
2020-05-26, 2020-07-04	1.4	1.21	1.49
2020-07-05, 2020-08-31	0.91	0.7	1.02

Supplementary methods: Investigating data manipulations using Benford's law

Benford's law (BL) defines a probability distribution for the leading digits in a determined set of numbers. The probability that a digit, $d = 1, 2, ..., 9$, is a leading number is given by $P(d) = \log_{10}(1 + (1/d))$ where numbers with leading digit 1 have the highest probability of appearance and this probability steadily decreases as the starting digit becomes larger. BL is used in a variety of fields such as accounting, trade, and election results to study possible fraud and irregularities with data [59-61] and is also frequently used to assess the quality of epidemiological and clinical data.

In the early phase of an outbreak when the number of reported cases and deaths grow exponentially a function of the form $2^{t/T}$ exactly obeys the BL distribution where t is the unit of time (measured in days) and T is the doubling time in new cases/deaths. If we combine two data sets with different growth rates into one larger data set, the numbers in the larger data set still obey the BL distribution. In the context of the COVID-19 outbreak, we can examine possible manipulation of data by comparing the probability distribution of the leading digits in the reported number of cases and deaths from different countries to BL distribution. The distribution of the leading digits of a manipulated data can take various shapes depending on how it is manipulated. A common way to check deviations from the BL is to calculate the Pearson correlation between the data and the BL - stronger correlations result in the correlation coefficient approaching closer to 1. We note that while this method can be used to test if data manipulation has occurred, it does not provide any information about deliberate absence of data by, for instance, not reporting deaths from certain hospitals.

By comparing the reported prevalence and deaths in several countries with sizable outbreaks from mid-February to early-April, we find that many countries, including Iran, were initially on an exponential growth trajectory (see Figures S5A and S5B). Depending on the testing strategy of each country, i.e. whether they only test hospitalised patients or also allow testing for outpatients, we would expect the reported death toll to follow the same trend as confirmed cases with a few days of delay. However, since 22 March, after passing a total of 1,000 confirmed deaths, Iran's reported death toll changed course to a linear growth despite the fact that the number of cases were still growing exponentially up to 2 April. This appeared in stark contrast with many other countries with large outbreaks and raised concerns over the credibility of the reports from Iran and possible manipulation of data by MoHME. We investigate the latter using Benford's Law. We compare the distribution of leading digits in reported cases and deaths during the exponential phase of the outbreak in Iran, USA, and UK to the Benford distribution (Figure S5C). The result shows that the distribution of leading digits in all the three countries are similar to Benford's distribution with the correlation coefficient >0.95 and that there is no evidence to suggest a manipulation of data has occurred in any of the examined data sets. It is likely that the apparent discrepancy in MoHME's reported numbers is due to delayed turnaround times at the peak of the outbreak when hospitals in several provinces were at near maximum capacity. Prioritising testing for active cases over post-mortem testing of suspect cases is also likely to have contributed to the observed difference in trends.

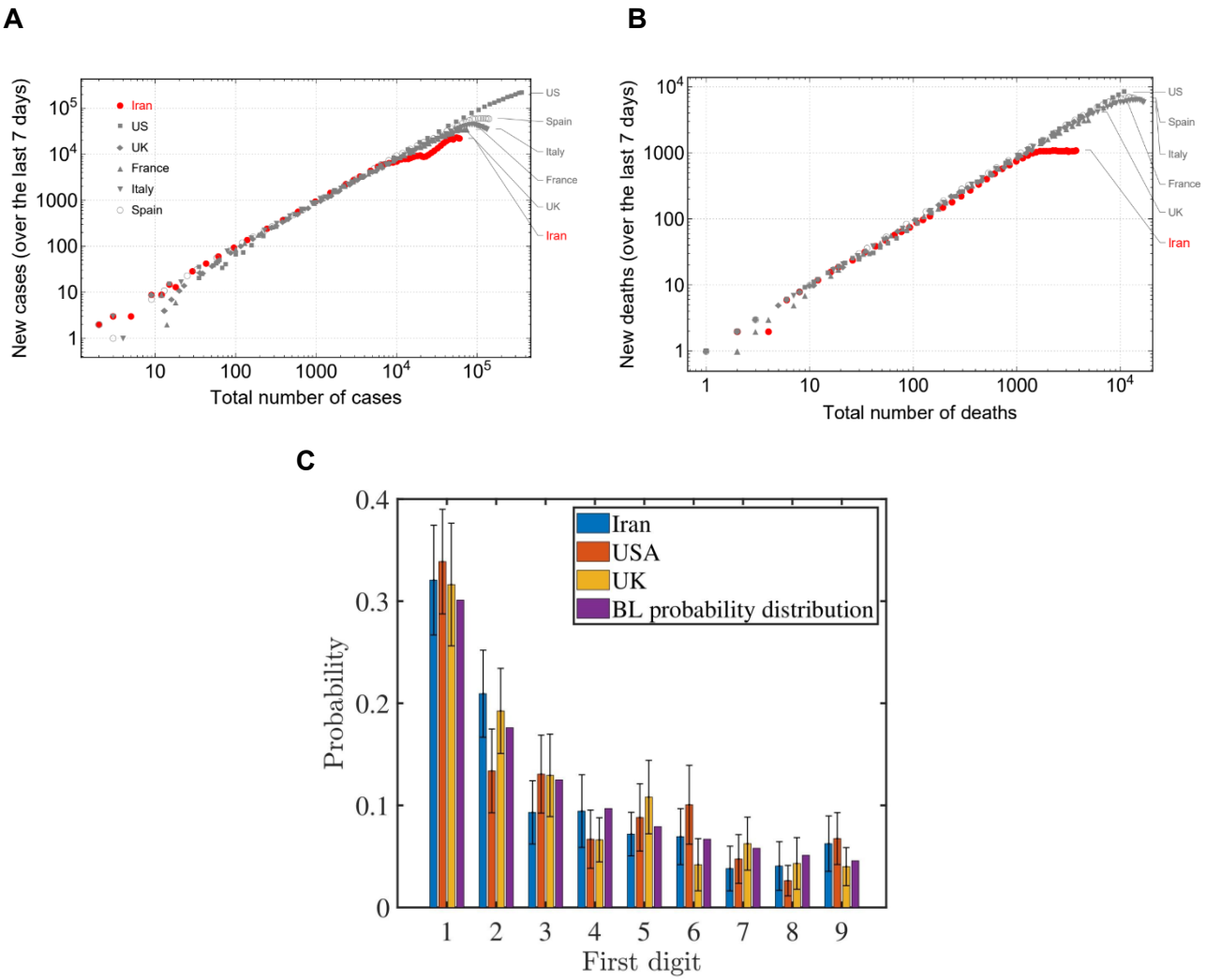


Figure S5: **Investigating data manipulation on reported number of cases and deaths.** (**A**) and (**B**) show the number of new cases and deaths over the last seven days with respect to the total number of cases and deaths, respectively. (**C**) Shows the probability distribution of leading digits in the total number of confirmed cases and deaths from the start of the epidemic in each country during the exponentially growing phase and compare this to the distribution given by BL. We sample 40 random numbers from this data set for each country and count the number of occurrences for the leading digits. To ensure convergence, we repeat this process 50 times. Error bars represent one standard deviation unit away from the mean.

Appendix 5: Estimates of anti-SARS-CoV-2 antibody seroprevalence in Iran

Iran was among the first countries outside China to report a large outbreak of COVID-19, but the transmission dynamics across the country have largely remained unknown due to the scarcity of serological, epidemiological, and genomic data. One of the main barriers has been the fact that Iran's Ministry of Health and Medical Education (MoHME) stopped releasing province-level data on the number of confirmed COVID-19 cases from March 22, 2020, onward. Furthermore, provincial data on the number of confirmed COVID-19-related deaths were never released. Instead, MoHME reports the percentage change in the number of cases with respect to previous days as an indicator of the state of the epidemic in each province and colour-codes them from blue (low incidence) to yellow (medium incidence), orange (high incidence), and red (very high incidence).

Despite the significant implications of understanding the Iranian epidemic for the country and the Eastern Mediterranean region as a whole, research investigations have largely been hindered due to the lack of epidemiological data on the number of cases and deaths, age-stratified and sex-stratified data, both at the national and province level, and seroepidemiological analysis [1].

The study by Hossein Poustchi and colleagues [2], sponsored by MoHME and carried out by the then Deputy Minister of Research and Technology of the Ministry of Health Reza Malekzadeh and his team, to measure SARS-CoV-2 antibody seroprevalence in the general population across 18 cities of Iran was the first systematic investigation into the geographical spread of COVID-19 across the country nearly a year after the first two cases were reported in Qom on Feb 19,

2020. Their analysis showed greatly varied levels of exposure in different cities, with some reaching very high levels (>50% in Qom and Rasht) by late April to early June.

Before the study by Poustchi and colleagues, we did a similar province-level analysis using seasonal all-cause mortality data to estimate the excess mortality in all 31 provinces of Iran from winter to summer, 2020 . Our findings corroborate the results by Poustchi and colleagues (Table A5.1), with an overall significant correlation (R^2=0·67 and p<0·001; Figure A5.1). Our results further suggest that most provinces would continue to have a two to four times increase in exposure until the end of summer (Sept 21, 2020), with Qom and Golestan reaching approximately 57% (95% CI 44–69) population-level exposure .

In the absence of more recent serology or province-level data, our estimates provide the most recent indicator of prevalence. This comparison is of immediate epidemiological importance as it highlights areas with the largest epidemic growth, which require the most immediate interventions. The continued availability of province-level data would be of paramount public health importance in a country that is facing such a heavy toll from COVID-19.

Table A5.1: **Comparing the percentage of exposure in 18 cities of Iran based on seroprevalence data in late April/early June [1] to provincial estimates for population-level exposure based on excess mortality by the end of spring (20 June 2020) and summer (21 September 2020) 1399 in solar Hijri. (.) shows the 95% confidence intervals.**

Province	City	City-level exposure	Province-level exposure in spring	Province-level exposure in summer
Khuzestan	Ahvaz	9.4%(0.7-21.6)	16%(12-19)	47%(37-57)
Ardabil	Ardabil	20.0%(3.2-46.8)	13%(10-17)	37%(28-47)
Mazandaran	Babol*	22.4%(11.9-35.1)	21%(15-27)	35%(25-45)
Esfahan (Isfahan)	Esfahan	2.7%(0.0-9.8)	9%(6-11)	21%(15-27)
Golestan	Gorgan	43.9%(31.4-58.3)	31%(24-38)	56%(44-69)
Hamedan	Hamedan	8.3%(1.6-17.0)	6%(4-8)	24%(17-31)
Kerman	Kerman	8.2%(2.2-16.6)	3%(2-4)	21%(13-26)
Kermanshah	Kermanshah	17.3%(5.3-30.9)	9%(6-11)	22%(16-28)
Razavi Khorasan	Mashhad	14.8%(8.2-23.1)	7%(5-8)	29%(22-36)
Qom	Qom	58.5%(37.2-83.9)	36%(28-44)	57%(44-69)
Gilan (Guilan)	Rasht	72.6%(53.9-92.8)†	28%(19-36)	33%(22-44)
Kurdistan	Sanandaj	1.7%(0.0-6.0)	14%(11-18)	41%(31-51)
Mazandaran	Sari*	14.7%(7.8-22.4)	21%(15-27)	35%(25 - 45)
Fars	Shiraz	7.3%(1.2-14.5)	NA**	14% (10 - 17)
East Azerbaijan	Tabriz	5.6%(0.0-13.0)	9%(6-11)	25% (18 - 32)
Tehran	Tehran	16.3%(13.5-19.5)	8%(6-11)	22% (16 - 28)
West Azerbaijan	Urmia	10.0%(0.2-24.3)	6%(5-8)	24% (18 - 30)
Sistan-Baluchistan	Zahedan	12.1%(2.3-23.3)	NA**	24% (20 - 27)
Overall	Overall	17.0%(14.6-19.5)	10%(2-17)	27% (20 - 34)

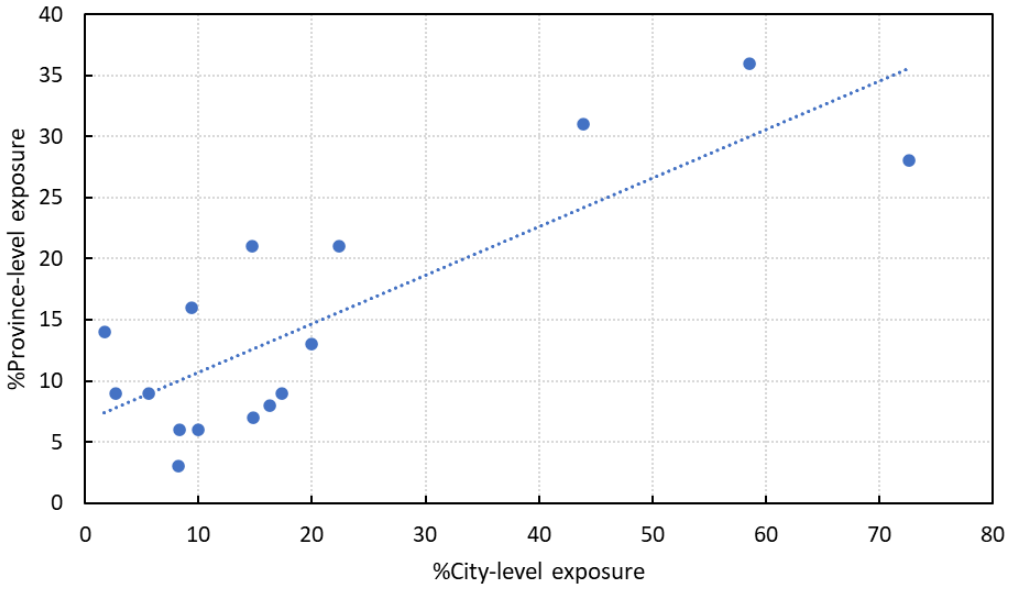


Figure A5.1: **A regression analysis based on the percentage of city-level seroprevalence estimates in late April/early June and percentage of provincial-level estimates by the end of spring (20 June 2020) according to seasonal excess mortality data.** The line of best fit (dashed line) corresponds to $R^2 = 0.67$. This analysis excludes Fars and Sistan-Baluchsitan provinces which did not show significant levels of excess mortality in spring [3]. If we include those two provinces with an assumed zero excess deaths in spring (and include the two cities, Shiraz and Zahedan, with their corresponding seroprevalence estimates), then the correlation drops to $R^2 = 0.63$. Also, if we remove the 72.6% exposure in Rasht based on seroprevalence as an outlier data, the correlation increases to $R^2 = 0.71$ (and $R^2 = 0.66$ if we assume the two provinces have zero excess deaths). In all these instances, the p-value is very significant ($p < .001$).

References

1. Ghafari, M. et al. Ongoing outbreak of COVID-19 in Iran: challenges and signs of concern with under-reporting of prevalence and deaths. *medRxiv* (2020).
2. Poustchi, H. et al. SARS-CoV-2 antibody seroprevalence in the general population and high-risk occupational groups across 18 cities in Iran: a population-based cross-sectional study. *Lancet Infectious Diseases* (2020).
3. Ghafari, M. et al. Excess deaths associated with the Iranian COVID-19 epidemic: a province-level analysis. *medRxiv* (2020).

Appendix 6: Supplementary material for chapter 7 part b

Table S1: **The overall attack rate (in percentage) by the end of May 2020 (wave 1), August 2020 (wave 2), January 2021 (wave 3), June 2021 (wave 4), and the week ending on 22 October 2021 (wave 5).**

Province	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
East Azerbaijan	5.44 (5.50 - 5.14)	11.68 (11.86 - 11.45)	39.85 (37.39 - 42.29)	60.36 (52.28 - 64.14)	84.92 (69.06 - 103.05)
West Azerbaijan	4.06 (3.85 - 4.34)	10.23 (8.87 - 10.50)	37.85 (34.37 - 42.99)	52.94 (45.84 - 60.46)	75.66 (60.25 - 98.53)
Ardabil	11.18 (9.90 - 11.52)	23.57 (20.61 - 26.34)	51.23 (45.54 - 54.60)	69.27 (60.02 - 77.94)	92.55 (76.19 - 112.14)
Isfahan	5.58 (4.92 - 5.28)	9.95 (8.66 - 9.62)	30.09 (26.10 - 30.96)	42.93 (37.36 - 45.99)	65.90 (51.80 - 81.46)
Alborz	5.32 (5.44 - 5.16)	10.50 (10.27 - 10.48)	27.44 (25.99 - 29.10)	45.49 (38.88 - 51.17)	64.09 (51.07 - 81.97)
Ilam	4.49 (4.13 - 4.23)	10.78 (10.17 - 10.55)	34.75 (31.70 - 36.22)	44.19 (40.23 - 47.35)	61.35 (51.01 - 73.74)
Bushehr	3.20 (2.53 - 3.24)	11.52 (8.40 - 11.39)	25.13 (20.63 - 25.25)	40.97 (32.96 - 45.54)	61.29 (44.63 - 82.10)
Tehran	7.69 (8.31 - 8.07)	12.67 (14.37 - 13.17)	32.42 (33.46 - 34.66)	51.17 (46.13 - 54.27)	77.06 (65.08 - 93.99)
Chahaar Mahal & Bakhtiari	7.06 (6.53 - 6.49)	13.46 (12.52 - 13.15)	40.07 (36.20 - 45.02)	56.42 (49.60 - 67.55)	76.28 (63.01 - 101.11)
South Khorasan	3.11 (2.94 - 3.49)	5.72 (5.26 - 6.53)	21.59 (19.20 - 23.92)	30.87 (27.58 - 33.24)	41.27 (34.65 - 48.50)
Razavi Khorasan	6.77 (6.27 - 6.66)	19.33 (17.53 - 21.33)	43.41 (38.66 - 45.49)	57.06 (49.79 - 64.07)	90.19 (71.76 - 113.32)
North Khorasan	6.46 (6.06 - 5.73)	14.52 (12.58 - 13.86)	36.11 (32.34 - 40.32)	48.61 (41.48 - 53.53)	77.73 (61.94 - 100.78)
Khuzestan	11.17 (10.51 - 13.65)	26.08 (23.70 - 29.49)	40.38 (36.19 - 42.13)	68.54 (58.79 - 76.98)	98.48 (79.82 - 120.59)
Zanjan	6.85 (7.59 - 7.15)	14.10 (14.18 - 14.59)	38.23 (35.99 - 41.37)	56.79 (51.31 - 63.59)	73.11 (60.85 - 87.84)
Semnan	4.12 (4.59 - 4.19)	8.27 (9.25 - 8.31)	30.87 (29.84 - 31.08)	42.82 (39.75 - 45.06)	57.26 (49.85 - 65.92)
Sistan & Baluchistan	1.24 (1.23 - 1.11)	5.02 (5.51 - 5.35)	14.02 (14.85 - 14.14)	26.46 (22.98 - 32.57)	46.60 (35.38 - 67.60)
Fars	1.64 (1.61 - 1.59)	6.51 (6.04 - 7.08)	25.69 (23.57 - 29.32)	38.77 (33.97 - 43.35)	56.63 (42.64 - 71.82)
Qazvin	12.92 (12.21 - 13.66)	19.79 (18.21 - 20.70)	46.85 (41.92 - 50.43)	64.95 (55.77 - 72.23)	85.97 (69.52 - 106.44)
Qom	14.87 (14.23 - 18.22)	20.10 (18.55 - 24.31)	40.20 (36.16 - 47.18)	51.89 (45.12 - 61.98)	84.74 (66.12 - 116.36)
Kurdistan	7.66 (6.86 - 7.83)	17.13 (16.40 - 18.52)	45.93 (43.35 - 55.61)	59.55 (54.04 - 69.37)	80.05 (66.82 - 103.86)
Kerman	2.10 (2.17 - 2.15)	8.11 (7.99 - 9.06)	23.45 (22.19 - 24.80)	32.84 (29.26 - 34.56)	50.58 (40.64 - 63.22)
Kermanshah	4.08 (4.39 - 4.31)	9.97 (9.86 - 9.13)	30.16 (28.50 - 32.97)	41.38 (37.50 - 46.08)	66.67 (54.07 - 86.87)
Kohgiluyeh & Boyer Ahmad	4.24 (3.88 - 3.79)	9.37 (8.61 - 8.72)	27.38 (24.45 - 27.12)	42.16 (36.62 - 46.16)	59.41 (48.25 - 74.93)
Golestan	20.27 (18.89 - 21.60)	33.20 (30.69 - 36.89)	55.32 (48.44 - 56.16)	71.67 (61.89 - 75.25)	105.43 (87.31 - 125.82)
Gilan	15.51 (15.13 - 17.90)	17.50 (17.26 - 20.12)	28.36 (27.51 - 31.21)	36.79 (34.53 - 40.22)	52.62 (45.21 - 67.60)
Lorestan	5.37 (5.64 - 5.41)	11.81 (12.20 - 12.09)	28.87 (27.28 - 30.13)	40.94 (36.45 - 43.29)	56.51 (47.10 - 68.36)
Mazandaran	11.19 (10.15 - 11.53)	18.16 (16.62 - 18.97)	31.36 (29.63 - 32.70)	40.39 (37.42 - 42.43)	58.36 (48.01 - 69.78)
Markazi	5.86 (5.45 - 5.22)	11.31 (10.48 - 10.60)	33.26 (31.29 - 33.89)	47.89 (42.32 - 51.57)	64.71 (53.01 - 75.59)
Hormozgan	3.51 (3.55 - 4.10)	12.58 (11.96 - 13.92)	21.11 (20.27 - 22.03)	32.48 (27.99 - 35.42)	48.52 (38.76 - 64.10)
Hamedan	3.74 (3.79 - 3.78)	10.69 (10.87 - 11.16)	28.74 (26.68 - 29.71)	45.23 (40.47 - 48.63)	63.97 (53.45 - 76.72)
Yazd	11.26 (12.05 - 11.49)	17.10 (18.44 - 17.86)	53.88 (51.44 - 59.97)	64.11 (59.99 - 72.87)	80.50 (71.43 - 93.68)

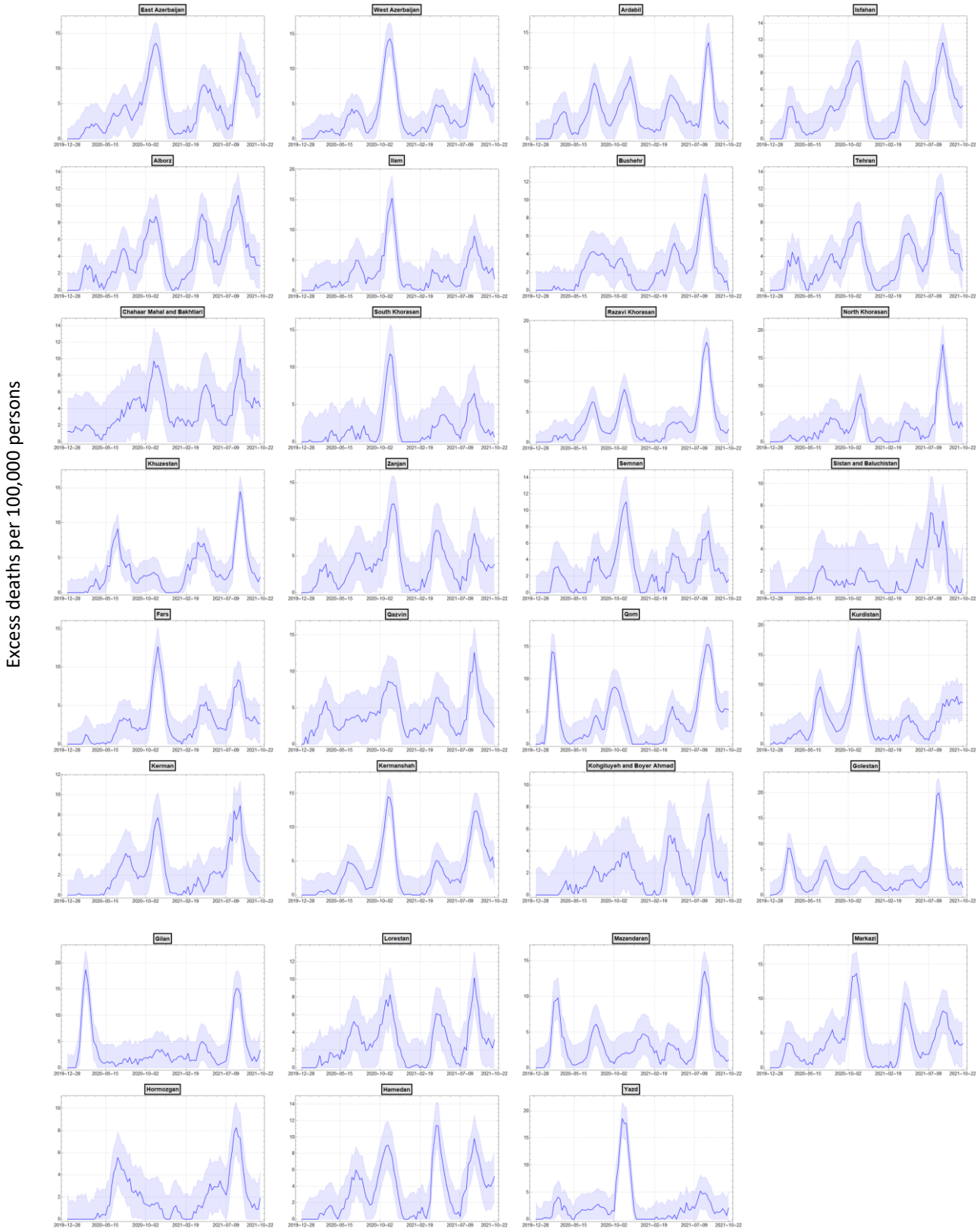


Figure S1: **One-month average of weekly excess mortality per 100,000 persons per province over time.** Shaded area shows the 95% forecast intervals. Note that the number of confirmed COVID-19 deaths are not reported at the province level in Iran.

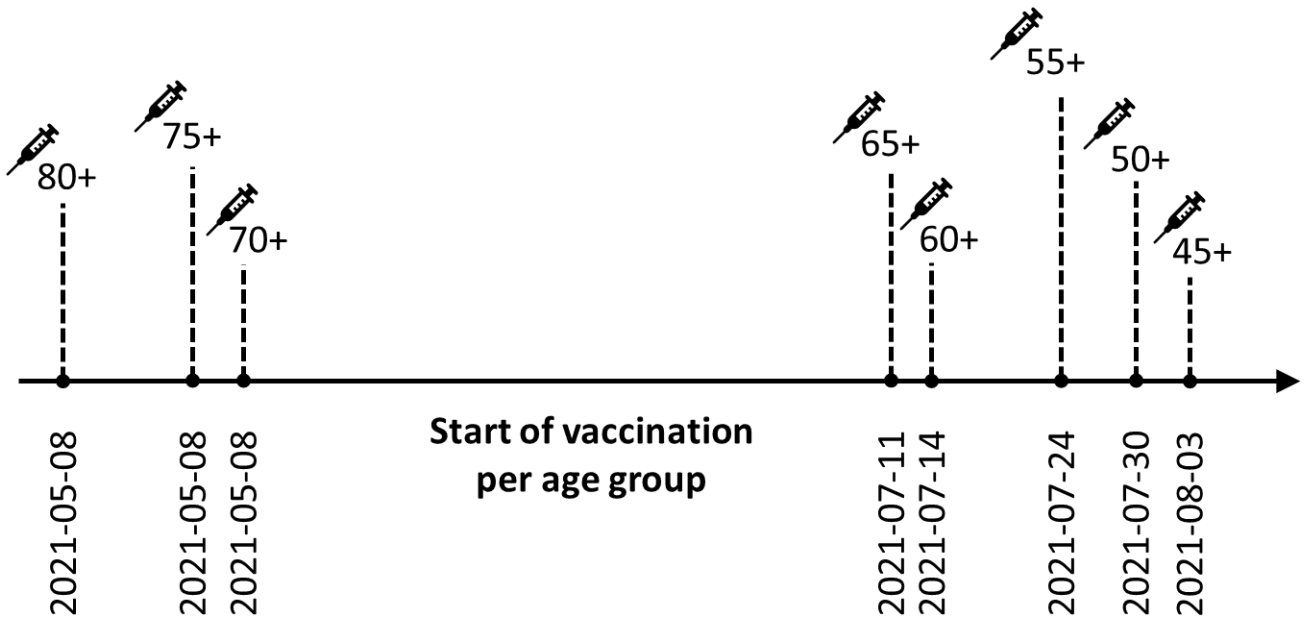


Figure S2: **Iran COVID-19 vaccination rollout dates per age group based on announcements made by MoHME (*behdasht.gov.ir*).**

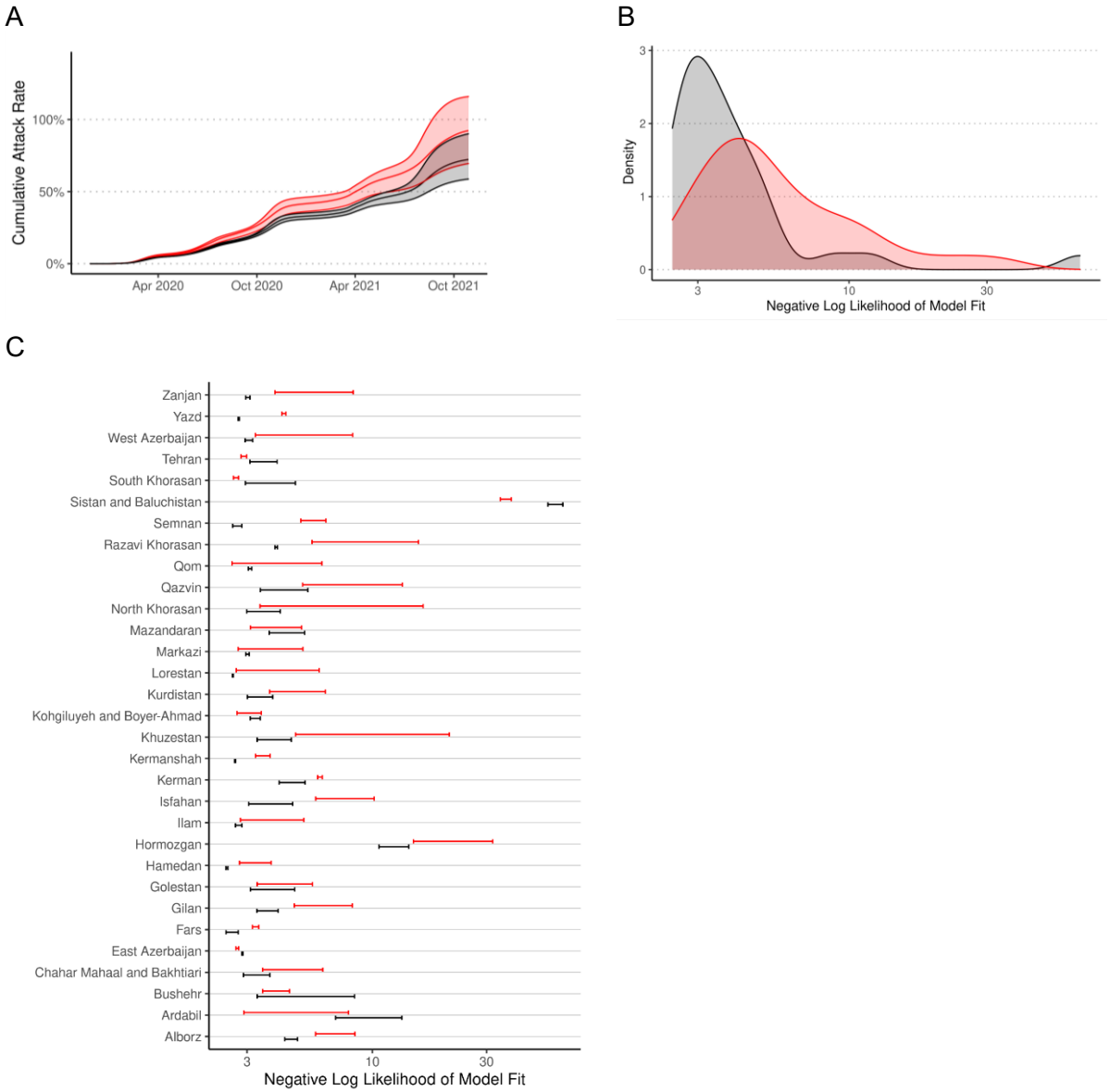


Figure S3: **(A)** Cumulative nationwide attack rates of SARS-CoV-2. Shaded area shows the variation in estimates from the three sets of model assumptions used for estimating attack rates. **(B, C)** Likelihood distribution of the modelled attack rates per province against seroprevalence data [1] using the O'Drsicoll et al. [2] (red) and Brazeau et al. [2] (black) age-stratified infection fatality rate estimates.

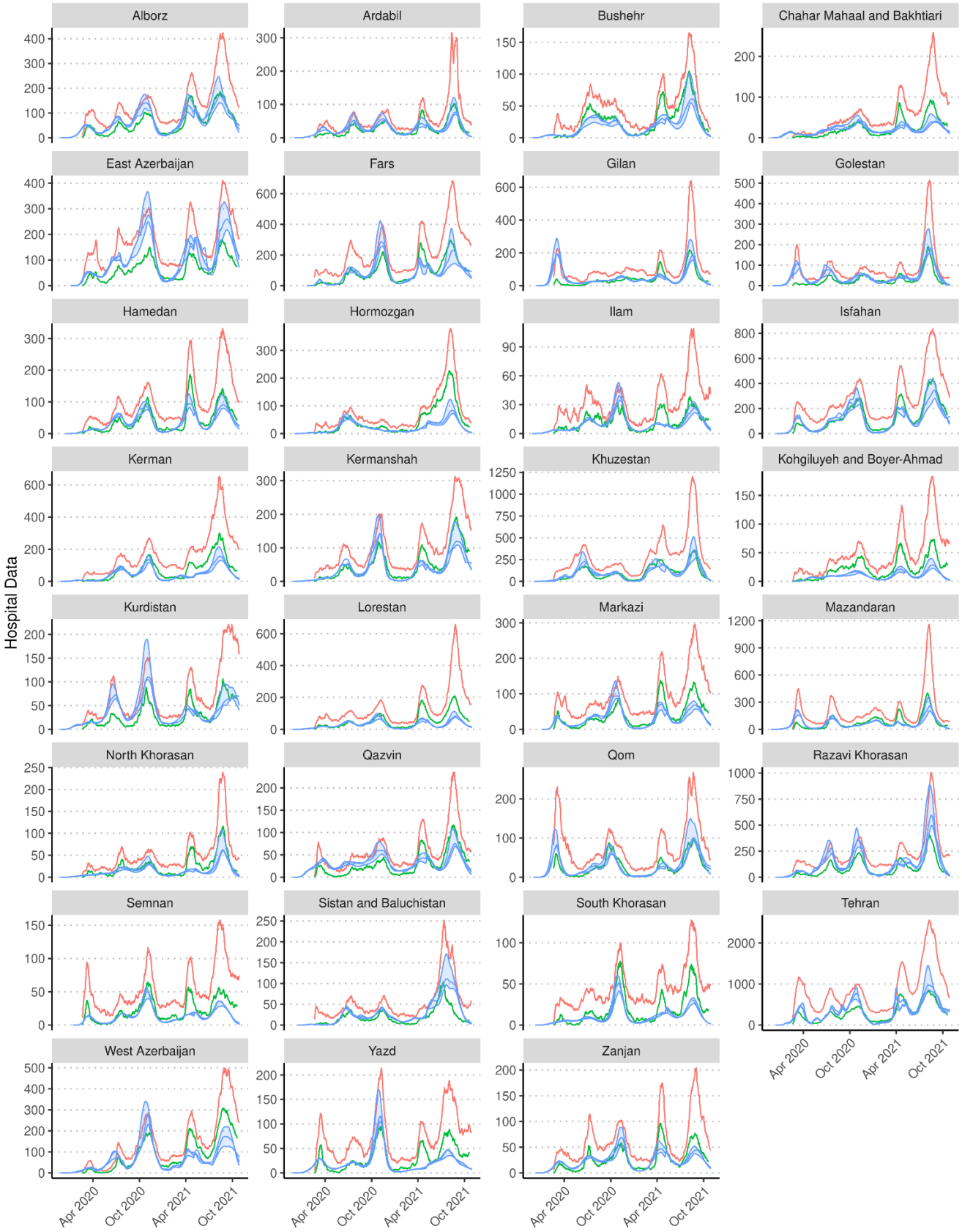


Figure S4: **Confirmed (green), suspected (red), and modelled (blue) daily hospital admissions per province over time.** Shaded area in blue shows the variation in modelled daily hospital admissions from the three sets of model assumptions used for estimating attack rates.

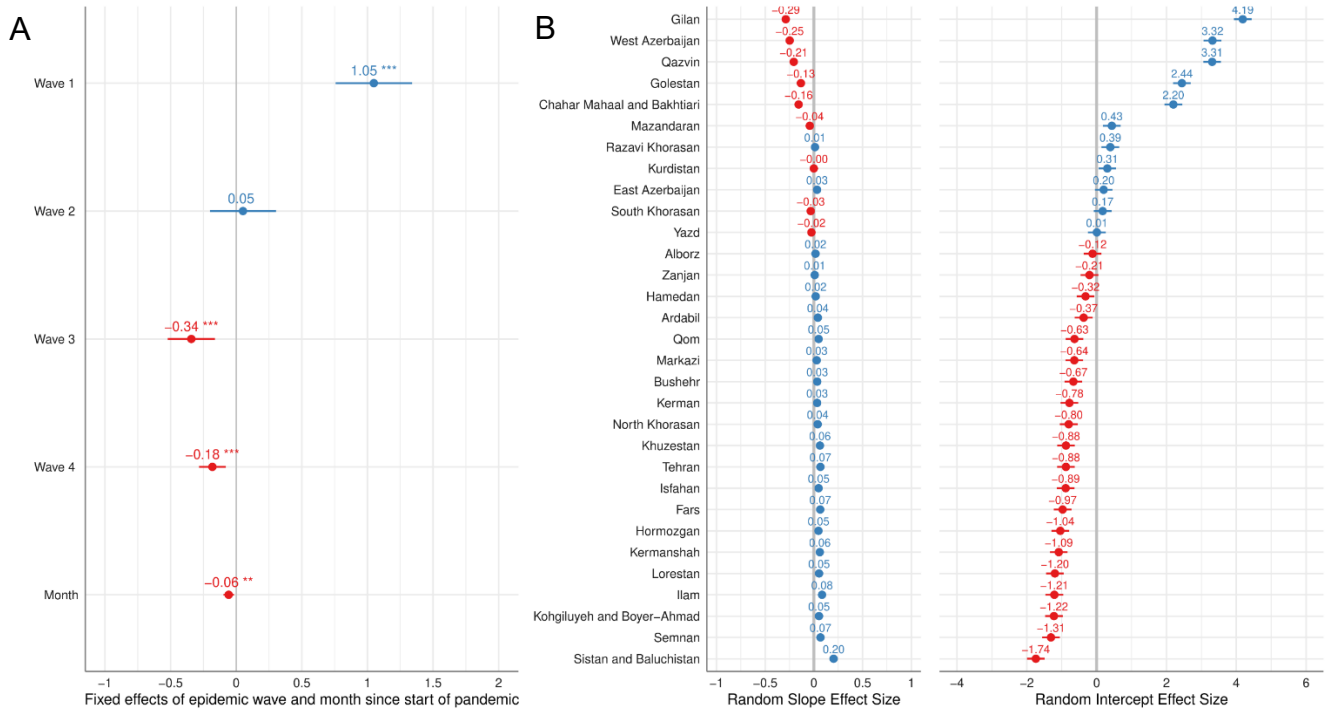


Figure S5: **Mixed effect model results investigating the relationship between model inferred hospitalisations and observed daily hospital admissions.** In **(A)** the fixed effect sizes are shown, with positive point estimates in blue and negative point estimates in red, with significant effects indicated with asterisks (*p<0.05, **p<0.01 and ***p<0.001%). In **(B)** the random effect sizes are shown for each province ranked by the size of the random intercept for each province.

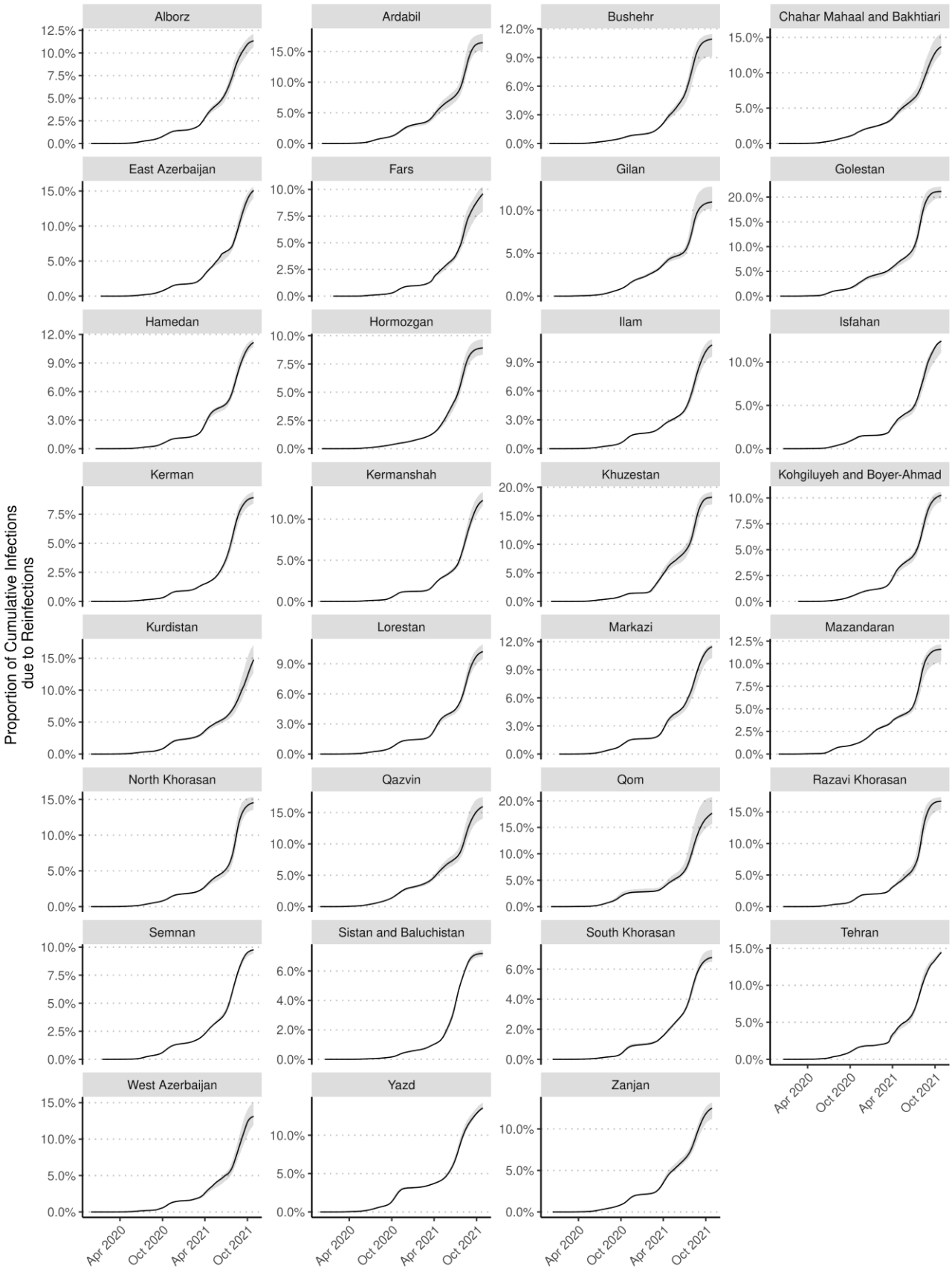
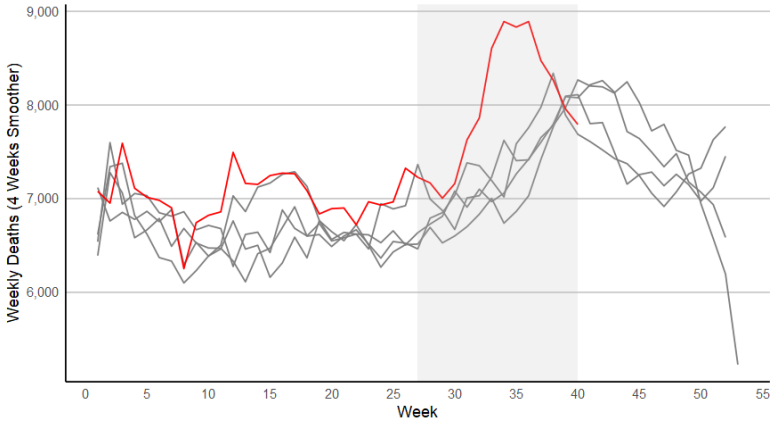


Figure S6: **Estimated proportion of infections due to reinfections per province over time.** Shaded area shows the variation in estimates from the three sets of model assumptions used for estimating attack rates.

A



B

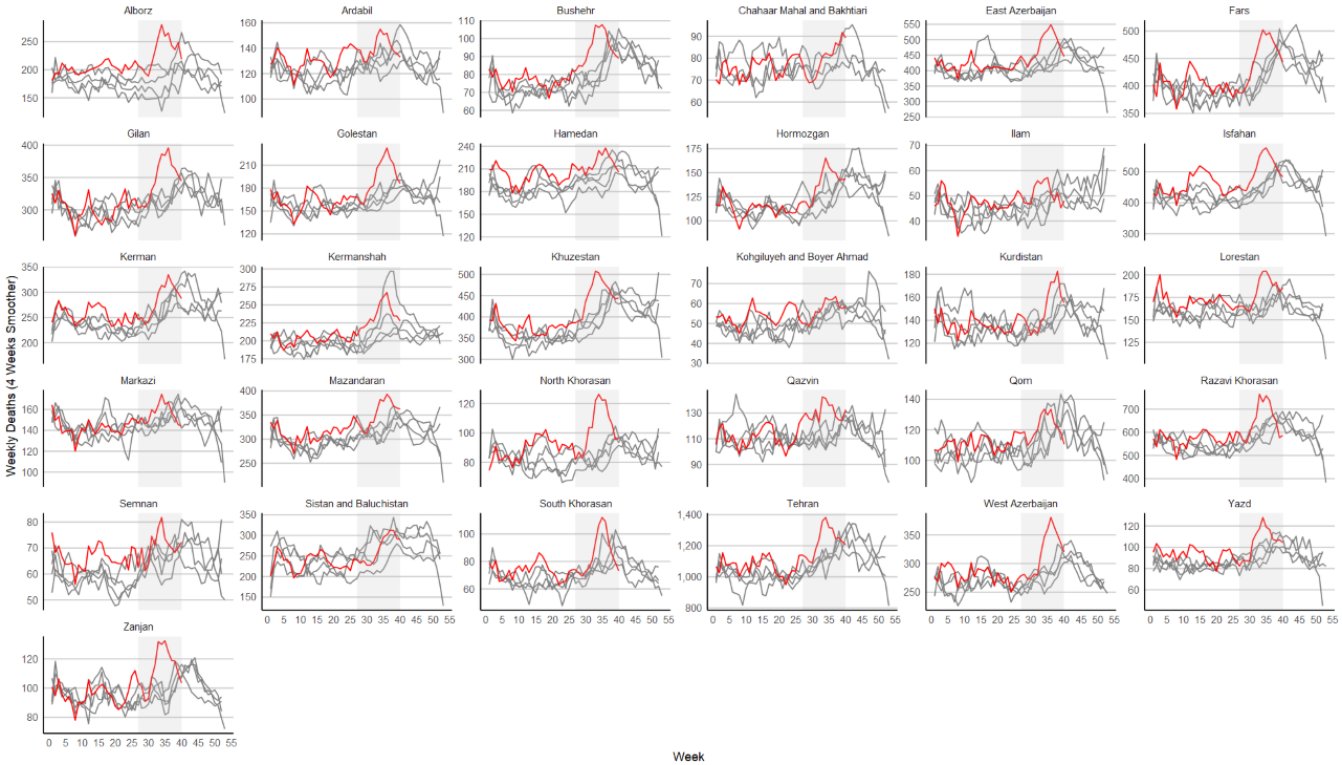
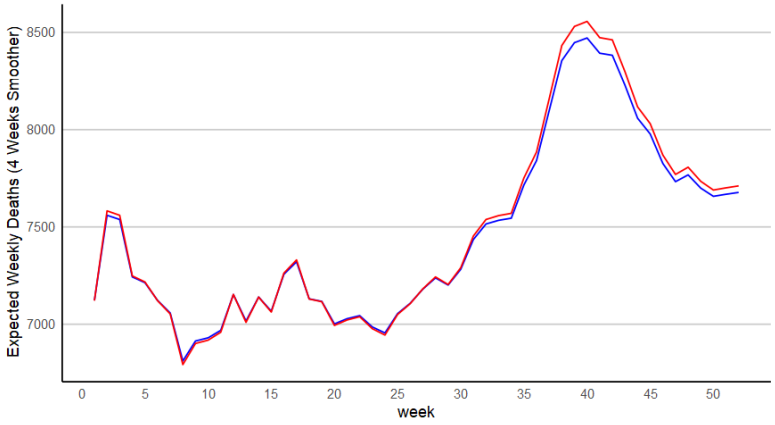


Figure S7: **One-month average weekly all-cause mortality data from 2015-03-21 to 2019-12-27** at the (**A**) national and (**B**) province level. All-cause mortality from 2015 to 2018 are shown gray and all-cause mortality during 2019 is shown in Red. Shaded area in gray highlights the weeks in autumn 2019 (2019-09-27 to 2019-12-27).

A



B



Figure S8: **Comparison between the expected one-month average of weekly excess mortality per 100,000 persons** (**A**) at the national level and (**B**) per-province over time using a linear regression (blue) and quasi-Poisson (red) model.