Dynamics and control of infectious disease epidemics: compartmental modelling of variations in infectiousness and symptoms during infection



William Hart

Keble College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary Term 2022

Acknowledgements

Firstly, a massive thank you to my supervisors, Robin Thompson and Philip Maini. Philip – thank you for your invaluable insights and advice, and for always finding the time to meet or to read draft manuscripts, often on very short notice. Robin – the vast amount of time you have dedicated to my academic development since we first met when I arrived in Oxford as an undergraduate, without which I would not be where I am today, is hugely appreciated. Here's to many more discussions about the state of English cricket.

This thesis would not have been possible without the collaborators who contributed to the publications from which the research chapters of this thesis are adapted: Leo Hochfilzer, Nik Cunniffe, Hyojung Lee and Hiroshi Nishiura (Chapter R1); Kit Yates (Chapter R2); and Seb Funk, Sam Abbott, Akira Endo, Joel Hellewell, Liz Miller and Nick Andrews (Chapter R4). Thanks also to Pauline Waight and other staff at Public Health England (now the UK Health Security agency) who contributed to data collection and management for the household study underlying Chapter R4, to members of the Wolfson Centre for Mathematical Biology for helpful discussions about my work, to members of the Juniper Consortium for useful comments on paper drafts, and to the anonymous reviewers whose suggestions helped to improve the work in each publication.

Thank you to the EPSRC for the funding that supported my DPhil, to the Keble Association for providing a travel grant that enabled me to travel to Canada for an intellectually stimulating workshop (it's a shame that COVID-19 dashed my hopes of attending more conferences and workshops in person), to Lady Margaret Hall for an ongoing lectureship, and to the Mathematical Institute and Worcester College for further teaching opportunities.

Finally, thank you to all my friends – particularly Allison, Amber, Ben, Beth, Brett, Casey, Ed, Freddie, Gilles, Hannah (who has had to put up with me as a flat-mate the last couple of years), Jack, James d-P, James W, Jim, John, Lewis, Oli, Phil, Sasha, Seymour, Will D-C and Will P – and family – particularly my siblings and grandparents, and of course my parents who continue to provide unwavering love and support. And to Meg the labrador, who never stopped wagging her tail to the end and is greatly missed.

Abstract

Through a series of case studies, this thesis explores how changes in infectiousness and symptoms during viral infections can be represented and inferred using compartmental models, in order to provide insights on epidemics in human populations.

First, we consider a compartmental model of Ebola virus disease outbreaks which relaxes the standard assumption made in epidemiological modelling studies that all symptomatic hosts are equally likely to be detected. We show that accounting for variations in symptoms during infection can lead to significantly different projections compared to a model in which symptoms are assumed to remain constant throughout infection, demonstrating the potential importance of accounting for individual host-scale dynamics in population-scale epidemic models.

We then develop an easy-to-use framework for informing population-scale epidemic models using host-scale data and models. Our approach involves using a multi-stage compartmental population-scale model with a large number of infected compartments in order to enable variations in infectiousness during infection (reflective of individual host-scale data) to be represented. Applying our approach to the case study of influenza A, we investigate the number of individuals from whom host-scale data need to be collected in order to ensure accurate population-scale epidemic projections.

We then turn our attention to SARS-CoV-2, and develop a novel approach for inferring changes in infectiousness during infection using data from infector-infectee transmission pairs. We show that this method, which is based on compartmental modelling, provides an improved fit to data collected in the early months of the COVID-19 pandemic compared to existing approaches. Our results indicate that many transmissions occur shortly before an infector develops symptoms, highlighting the benefits of contact tracing even if contacts from a short time window before symptom onset alone are traced.

Finally, we adapt our approach for inferring changes in infectiousness during SARS-CoV-2 infection in order to obtain updated estimates of the generation time distribution (which characterises the relative expected infectiousness profile of infected hosts at each time since infection) using data from UK households. We infer a shorter mean generation time in September–November 2020 compared to earlier months, which demonstrates the importance of continued data collection and analysis in order to inform future public health policy decisions.

Contents

Introduction **1**

I.1 Compartmental models of infectious disease epidemics 2
I.2 Variations in infectiousness during infection 5
I.3 Within-host modelling 10
I.4 Parameter inference 12
I.5 Outline of this thesis 15

R1 Accurate assessments of the effectiveness of interventions against Ebola may require models that account for variations in symptoms during infection **18**

R1.1 Introduction 18
R1.2 Methods 20
R1.2.1 Data 20
R1.2.2 Mathematical models 21
R1.2.3 Dependence of the detection rate on surveillance 23
R1.2.4 Parameters and model fitting 24
R1.2.5 Generating simulated data 26
R1.3 Results 27
R1.3.1 Simulation study 27
R1.3.2 Application to real-world Ebola data 33
R1.3.3 Model extensions 35
R1.3.4 Sensitivity analyses 37
R1.4 Discussion 42

R2 A theoretical framework for transitioning from host-scale to population-scale epidemiological dynamics: influenza A as a case study **47**

R2.1 Introduction 47
R2.2 Methods 49
R2.2.1 The SI_nR model 49

R2.2.2 The K&M model . . . 50
R2.2.3 Transitioning from host- to population-scale dynamics . . . 51
R2.2.4 Influenza case study . . . 54
R2.2.5 Errors in host- and population-scale dynamics . . . 56
R2.3 Results . . . 57
R2.3.1 Summary of multi-scale modelling framework . . . 57
R2.3.2 Equivalence between the compartmental and IDE methods in the large-compartment limit . . . 57
R2.3.3 Influenza case study . . . 61
R2.3.4 Analysis of sources of errors in population-scale projections 63
R2.3.5 Sensitivity analyses . . . 69
R2.4 Discussion . . . 76

R3 High infectiousness immediately before COVID-19 symptom onset highlights the importance of contact tracing 81
R3.1 Introduction . . . 81
R3.2 Methods . . . 83
R3.2.1 Notation and general details . . . 83
R3.2.2 Data . . . 84
R3.2.3 Incubation period . . . 85
R3.2.4 Independent transmission and symptoms model . . . 85
R3.2.5 Ferretti model . . . 86
R3.2.6 Mechanistic model . . . 88
R3.2.7 Likelihood . . . 92
R3.2.8 Parameter estimation . . . 94
R3.2.9 Maximum likelihood estimation and goodness of fit . . . 97
R3.2.10 Total proportion of non-symptomatic transmissions accounting for asymptomatic cases . . . 98
R3.2.11 Contact tracing and isolation . . . 99
R3.3 Results . . . 100
R3.3.1 Models and parameter estimation . . . 100
R3.3.2 Goodness of fit . . . 101
R3.3.3 Distributions of epidemiological time intervals . . . 103
R3.3.4 The contribution of non-symptomatic infectious individuals to transmission . . . 104
R3.3.5 Implications for isolation and contact tracing . . . 106
R3.3.6 Sensitivity analyses . . . 108
R3.4 Discussion . . . 112

R4 Inference of the SARS-CoV-2 generation time using UK household data **117**
R4.1 Introduction . . . 117
R4.2 Methods . . . 119
R4.2.1 Data . . . 119
R4.2.2 General infectiousness modelling framework . . . 120
R4.2.3 Independent transmission and symptoms model . . . 123
R4.2.4 Mechanistic model . . . 124
R4.2.5 Likelihood . . . 128
R4.2.6 Incubation period . . . 130
R4.2.7 Parameter estimation . . . 131
R4.3 Results . . . 133
R4.3.1 Models and parameter estimation . . . 133
R4.3.2 Comparison of posterior estimates . . . 134
R4.3.3 Estimation of realised generation times within households . 137
R4.3.4 Distributions of epidemiological time intervals . . . 138
R4.3.5 Temporal variation in the generation time distribution . . . 139
R4.3.6 Sensitivity analyses . . . 142
R4.4 Discussion . . . 148

Discussion **154**
D.1 Summary of findings . . . 154
D.2 Ongoing and future research . . . 164
D.3 Concluding remarks . . . 168

A1 Appendix to Chapter R1 **169**
A1.1 Model extensions . . . 169
A1.1.1 Model with variable infectiousness during infection . . . 169
A1.1.2 Model in which some hosts recover prior to the deterioration phase of an Ebola infection . . . 170
A1.2 Fitting models to non-cumulative data . . . 171
A1.3 Supplementary Figures . . . 172
A1.4 Supplementary Tables . . . 179

A2 Appendix to Chapter R2 **182**
A2.1 Explicitly including variability in the infectiousness time-courses of different infected hosts in the population-scale model . . . 182
A2.1.1 The $SI_{n,m}R$ model . . . 182

A2.1.2 Approximating the population-scale dynamics within the $SI_{n,m}R$ framework when there are multiple infection types . 183
A2.1.3 The limit of infinitely many compartments 184
A2.2 Alternative relationships between pathogen load and infectiousness 185
A2.3 Supplementary Figures 186

A3 Appendix to Chapter R3 **189**
A3.1 Discretisation of the serial interval distribution 189
A3.2 Supplementary Figures 190
A3.3 Supplementary Tables 192

A4 Appendix to Chapter R4 **196**
A4.1 Details of parameter estimation algorithm for the mechanistic model 196
A4.2 Relationship between the generation time, TOST and serial interval 196
A4.3 Extension of framework to account for co-primary cases 199
A4.4 Accounting for uncertainty in the incubation period distribution . . 201
A4.5 Supplementary Figures 202
A4.6 Supplementary Tables 209

Bibliography **212**

List of Figures

I.1 The SIR model. . . . 3
I.2 The SEIR model and the method of stages. . . . 5
I.3 The target cell-limited model. . . . 11

R1.1 Schematic showing how a model can be used to project the effect of changing control interventions on epidemic dynamics. . . . 19
R1.2 Flow diagram illustrating the different models that we considered. . . 22
R1.3 Identifiability of model parameters. . . . 28
R1.4 Using the constant symptoms model and variable symptoms model to project the impact of intensified surveillance on the simulated epidemic. . . . 29
R1.5 Effect of the choice of the detection probability in the constant symptoms model. . . . 31
R1.6 Using the constant symptoms model and variable symptoms model to project the effects of alternative interventions. . . . 34
R1.7 Results incorporating variable infectiousness during infection. . . . 36
R1.8 Results for a model in which some hosts recover prior to the deterioration phase of an Ebola infection. . . . 38
R1.9 Effect of the timing of surveillance intensification. . . . 39
R1.10 Effect of the surveillance frequency. . . . 40
R1.11 Using the constant symptoms model and variable symptoms model to project epidemic dynamics under intensified surveillance, when the models are fitted to non-cumulative incidence data. . . . 42

R2.1 Schematic demonstrating methods for transitioning from host- to population-scale epidemiological dynamics. . . . 58
R2.2 Real and synthetic host-scale data. . . . 62
R2.3 Transitioning from host- to population-scale influenza dynamics using the compartmental and IDE methods. . . . 62
R2.4 Number of compartments. . . . 64
R2.5 Number of hosts. . . . 65

R2.6 Extent of measurement error in host-scale data. 67
R2.7 Overall error in population-scale projections. 69
R2.8 Results assuming host-scale data are collected twice daily from each host. 71
R2.9 Trading off the number of hosts, d, and number of daily observations per host, f, when the total number of measurements, $M = 7fd$, is fixed. 72
R2.10 Effect of the relationship between pathogen load and infectiousness. 73
R2.11 Effect of the basic reproduction number, R_0. 75

R3.1 Schematic illustrating epidemiological time intervals in data from infector-infectee pairs. 82
R3.2 Schematic illustrating approaches for inference from transmission pair data. 101
R3.3 Posterior distributions of fitted parameters for the variable infectiousness model. 102
R3.4 Posterior distributions of fitted parameters for the independent transmission and symptoms model. 102
R3.5 Distributions of epidemiological time intervals estimated by fitting different models to data from 191 SARS-CoV-2 transmission pairs. . 103
R3.6 Discretised serial interval distributions. 105
R3.7 The contribution of non-symptomatic infectious individuals to transmission. 106
R3.8 Implications for isolation and contact tracing 107
R3.9 Distributions of epidemiological time intervals assuming an alternative incubation period distribution. 109
R3.10 The contribution of non-symptomatic infectious individuals to transmission assuming an alternative incubation period distribution. . . . 109
R3.11 Implications for isolation and contact tracing assuming an alternative incubation period distribution. 110
R3.12 Effect of the efficiency of isolation and contact tracing. 111

R4.1 Number of study households and household members by recruitment month and household size. 121
R4.2 Posterior distributions of fitted parameters for the independent transmission and symptoms model. 134
R4.3 Posterior distributions of fitted parameters for the mechanistic model. 135
R4.4 Comparison of posterior estimates between the two models. 136

R4.5 Estimation of realised generation times within study households. . . 137
R4.6 Generation time, TOST and serial interval distributions. 138
R4.7 Temporal changes in the generation time for the independent transmission and symptoms model. 140
R4.8 Temporal changes in the generation time for the mechanistic model. 141
R4.9 Monthly changes in the generation time from September–November 2020. 142
R4.10 Effect of accounting for the possibility of co-primary cases on estimates of the generation time. 143
R4.11 Temporal changes in the generation time for the independent transmission and symptoms model, accounting for the possibility of co-primary cases. 144
R4.12 Effect of accounting for uncertainty in the incubation period distribution on estimates of the generation time. 146
R4.13 Effect of the dependency of transmission on the household size on estimates of the generation time. 147

A1.1 Effect of the extent of measurement error in the simulated data. . . 172
A1.2 Robustness of our results to different values of parameters used to generate the synthetic data. 173
A1.3 Model fits for the model with variable infectiousness. 174
A1.4 Model fits for different surveillance frequencies. 175
A1.5 Effect of the mean detection probability. 176
A1.6 Effect of the extent of variation in symptom expression during infection in the variable symptoms model. 177
A1.7 Alternative dependence of the detection rate on surveillance. 178

A2.1 Comparison of errors at different scales. 186
A2.2 Results using the IDE method to transition from host-scale to population-scale dynamics. 187
A2.3 Effect of the extent of measurement error in population-scale data. . 187
A2.4 Effect of the extent of variability in host-scale parameter values. . . 188

A3.1 The contribution of asymptomatic cases to transmission. 190
A3.2 Posterior distributions of fitted parameters for the constant infectiousness model. 191
A3.3 Posterior distributions of fitted parameters for the Ferretti model. . 191

A4.1 Observed and model-predicted distributions of intervals between successive symptom onset dates within households. 202
A4.2 Temporal changes in the generation time, TOST and serial interval distributions. 203
A4.3 Temporal changes in fitted parameters for the mechanistic model. . 203
A4.4 Effect of the functional form of the generation time distribution for the independent transmission and symptoms model on estimates of the generation time. 204
A4.5 Effect of the functional form of the incubation period distribution for the independent transmission and symptoms model on estimates of the generation time. 205
A4.6 Further analysis of the effect of the dependency of transmission on the household size on estimates of the generation time. 206
A4.7 Effect of the relative infectiousness of entirely asymptomatic infected hosts on estimates of the generation time. 207
A4.8 Effect of the exclusion of individuals of unknown infection status on estimates of the generation time. 208

List of Tables

R2.1 Parameter values and initial conditions for the TCL within-host model. 54

R4.1 Previous SARS-CoV-2 generation time estimates. 119

A1.1 Assumed parameter values. 179
A1.2 Effect of surveillance. 180
A1.3 True and fitted parameter values in our analysis of simulated data. . 180
A1.4 Fitted model parameters in our analysis of data from the 2018–2020 DRC epidemic. 181
A1.5 Fitted model parameters in our analysis of data from the 2014–2016 epidemic in Liberia. 181

A3.1 Assumed parameter values. 192
A3.2 Fitted parameter values using data augmentation MCMC. 193
A3.3 Maximum likelihood parameter estimates and ΔAIC values. 194
A3.4 Mean and standard deviation of distributions of epidemiological time periods. 195

A4.1 Assumed parameter values. 209
A4.2 Fitted parameter values. 210
A4.3 Mean and standard deviation of distributions of epidemiological time periods. 211

Acronyms

AIC Akaike information criterion.

CrI Credible interval.

DRC Democratic Republic of the Congo.

IDE(s) Integro-differential equation(s).

K&M (model) Kermack and McKendrick (model).

MCMC Markov chain Monte Carlo.

ODE(s) Ordinary differential equation(s).

PCR Polymerase chain reaction.

PDE(s) Partial differential equation(s).

PHE Public Health England.

SEIR (model) Susceptible-exposed-infectious-removed (model).

SIR (model) Susceptible-infected-removed (model).

TCL (model) Target cell-limited (model).

TOST Time from onset of symptoms to transmission.

UKHSA United Kingdom Health Security Agency.

Introduction

Infectious disease outbreaks in human populations are responsible for severe public health consequences [59, 124, 172, 211, 220]. As a high-profile and topical example, the COVID-19 pandemic has had a devastating impact, with the UK alone having reported 160,038 deaths within 28 days of testing positive for SARS-CoV-2 (the virus that causes COVID-19) infection as of 16$^{\text{th}}$ February 2022 [230]. Governments worldwide imposed stringent non-pharmaceutical interventions, including national lockdowns, to combat the pandemic [22, 106]. Many restrictions have subsequently been removed or relaxed following the roll-out of COVID-19 vaccines [22], although novel SARS-CoV-2 variants remain a threat [79, 98, 223]. Mathematical modelling has helped to guide policy responses to the pandemic [28, 218], with models having been used, for example, to estimate transmissibility (characterised by the time-dependent reproduction number) [1, 12, 232] and to project future epidemic dynamics under different possible control interventions [70, 81, 141, 218].

Compartmental models are widely used to model the population-scale outbreak dynamics of SARS-CoV-2 [69, 70, 79, 130, 171, 201] and other pathogens, including Ebola [3, 10, 18, 54, 55, 239] and influenza [38, 53, 56, 59, 60, 89, 169, 215, 221]. In compartmental models, individuals are divided into compartments depending on their current infection status and sometimes other characteristics, such as age [131]. Usually, compartmental models involve an assumption that the infectiousness of each infected host is constant over the course of the infectious period [131]. However, infectiousness is, in reality, likely to vary, for example due to individual host-scale dynamics, such as changes in the pathogen load and the extent of symptoms during infection [111, 167]. Particularly for pathogens such as SARS-CoV-2 for which transmission can occur from presymptomatic infected individuals or those with mild initial symptoms [30], behavioural responses to symptoms (e.g. staying at home once symptoms develop or become more severe) may also reduce the relative risk of transmission at later times since infection [116, 160]. Finally, non-pharmaceutical interventions, such as case isolation and contact tracing, may affect the temporal profile of transmission risk [85, 90].

This thesis explores how compartmental models can be used to model variations in infectiousness and symptoms during infection. Specifically, in the first part of the thesis (Chapters R1 and R2), we demonstrate how population-scale compartmental models can be informed with information about how symptoms (Chapter R1) and infectiousness (Chapter R2) vary during infection in order improve the accuracy of epidemic projections. Then, in the second part of the thesis (Chapters R3 and R4), we show how an approach based on compartmental modelling can be used to predict how infectiousness varies during SARS-CoV-2 infection using data from infector-infectee transmission pairs (Chapter R3) and households (Chapter R4). While we aim to develop widely applicable modelling techniques, we also apply the approaches developed in this thesis to three viral pathogens – as well as focussing on SARS-CoV-2 in Chapters R3 and R4, we also consider Ebola and influenza as case studies in Chapter R1 and Chapter R2, respectively.

In the remainder of this chapter, we first introduce important epidemiological modelling (Sections I.1–I.3) and parameter inference (Section I.4) techniques that we use and adapt throughout this thesis. Due to length limitations, we do not attempt to provide a detailed full survey of the epidemiological modelling or parameter inference literature but, instead, focus on those methods that directly relate to the research in this thesis. We then conclude this introductory chapter by providing an outline of the main research chapters of this thesis.

I.1 Compartmental models of infectious disease epidemics

The use of compartmental models to model infectious disease epidemics can be traced back to the early 20^{th} century. In the seminal 1927 work of Kermack and McKendrick [133], the authors introduced the susceptible-infected-removed (SIR) compartmental model (Figure I.1) in which individuals are characterised as being susceptible to the pathogen (S), infected (I) or removed (either recovered and immune, or dead; R). In its most basic form [131], the SIR model is given by the system of ordinary differential equations (ODEs),

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI, \tag{I.1}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta SI - \mu I, \tag{I.2}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \mu I, \tag{I.3}$$

where $S(t)$, $I(t)$ and $R(t)$ are the numbers of susceptible, infected and recovered individuals respectively, t is the time since the start of the outbreak, β is the transmission rate parameter (i.e. transmissions occur at total rate βSI), and μ is the recovery rate, so that $1/\mu$ is the mean infectious period. Since demographics are neglected in the most simple form of the model, the total population size,

$$N = S(t) + I(t) + R(t), \tag{I.4}$$

remains constant, so $R(t)$ can be calculated as $R = N - (S + I)$, and equation (I.3) indeed decouples from equations (I.1) and (I.2). Note that the transmission term βSI is sometimes replaced by $\beta SI/N$ to reflect frequency-dependent, rather than density-dependent, transmission [131] (this distinction is important in Chapter R4 of this thesis, where we consider different possible dependencies of household SARS-CoV-2 transmission on the household size).

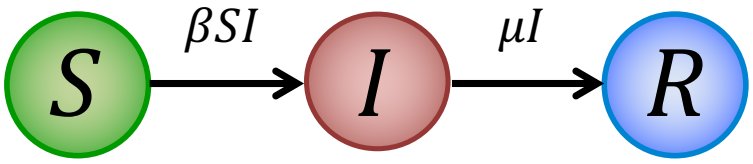


Figure I.1: The SIR model. Flow diagram illustrating the compartments of the SIR model, and the net rates at which individuals pass between the compartments.

Under the SIR model, it is straightforward to show that if a small number of infected individuals are introduced into an otherwise entirely susceptible population, then the pathogen will initially spread in the population if and only if

$$R_0 := \frac{\beta N}{\mu} > 1. \tag{I.5}$$

The quantity R_0 is called the *basic reproduction number*. In general, the basic reproduction number is defined to be the average number of secondary cases arising from a single infected individual in an otherwise entirely susceptible population [131].

Deterministic models, comprising systems of ODEs, are fast to solve numerically and are often analytically tractable. However, stochasticity may be important in real-world epidemiological systems, particularly when the number of infected individuals is small [131]. A stochastic analogue of the SIR model can be defined as a continuous-time Markov process with the following possible transitions [131]:

$$(S, I) \xrightarrow{\beta SI} (S - 1, I + 1) \qquad \text{(infection)}, \tag{I.6}$$

$$(S, I) \xrightarrow{\mu I} (S, I - 1) \qquad \text{(recovery)}, \tag{I.7}$$

i.e. infections occur at total rate βSI and recoveries at total rate μI (the variables and parameters have analogous definitions to those in the deterministic model). Such a model can be simulated using Gillespie's direct algorithm [97]. An important feature of stochastic epidemic models is that, unlike in deterministic models, an emerging outbreak may die out without a major epidemic occurring even if $R_0 > 1$ [11, 131, 221]. Whereas in Chapters R1 and R2 of this thesis we focus on deterministic epidemic models, in Chapters R3 and R4 we develop a likelihood-based framework for inference of important epidemiological quantities (such as the generation time – see Section I.2 below) using data from infector-infectee transmission pairs (Chapter R3) and households (Chapter R4) which is based on stochastic compartmental models.

Numerous extensions to the basic SIR model are possible. A common variant of the SIR model is the susceptible-exposed-infectious-removed (SEIR) model [131], in which there is assumed to be a latent (E) stage, during which an individual is infected but not yet infectious. In the simplest form of the SEIR model (Figure I.2A), infected individuals transition from the E stage to the I stage at a constant rate, γ, and similarly transition from the I stage to the R stage at a constant rate, μ. This corresponds to an assumption that the latent and infectious periods each follow an exponential distribution [237] (with means $1/\gamma$ and $1/\mu$, respectively). However, it has been shown that gamma distributions often better fit the actual distributions of these epidemiological periods [151, 237], and that incorrectly assuming exponential distributions may lead, for example, to erroneous estimates of the basic reproduction number [237]. Gamma-distributed latent and infectious periods may be accounted for within a compartmental model by sub-dividing the E and I classes into multiple compartments, through which each infected host progresses [11, 237]. This approach, sometimes referred to as the "method of stages" [152] or "linear chain trick" [23], leads to the $\mathrm{SE}_m\mathrm{I}_n\mathrm{R}$ model,

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta \sum_{i=1}^{n} SI_i, \tag{I.8}$$

$$\frac{\mathrm{d}E_1}{\mathrm{d}t} = \beta \sum_{i=1}^{n} SI_i - \gamma E_1, \tag{I.9}$$

$$\frac{\mathrm{d}E_i}{\mathrm{d}t} = \gamma E_{i-1} - \gamma E_i, \qquad i = 2, \ldots, m, \tag{I.10}$$

$$\frac{\mathrm{d}I_1}{\mathrm{d}t} = \gamma E_m - \mu I_1, \tag{I.11}$$

$$\frac{\mathrm{d}I_i}{\mathrm{d}t} = \mu I_{i-1} - \mu I_i, \qquad i = 2, \ldots, n, \tag{I.12}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \mu I_n, \tag{I.13}$$

Here, the latent and infectious periods each follow gamma distributions with means m/γ and n/μ, respectively, and shape parameters m and n [237]. A schematic illustrating the $\mathrm{SE}_m\mathrm{I}_n\mathrm{R}$ model with $m = 2$ and $n = 3$ is illustrated in Figure I.2B.

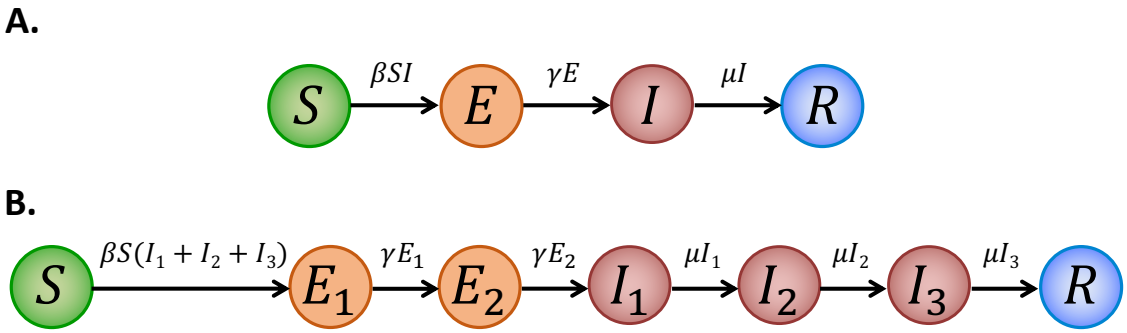


Figure I.2: The SEIR model and the method of stages. A. Flow diagram illustrating the SEIR model. B. Flow diagram illustrating an example of the method of stages, in which the E and I compartments in the SEIR model are divided into two and three sub-compartments, respectively (giving an $\mathrm{SE}_2\mathrm{I}_3\mathrm{R}$ model).

In equations (I.8)–(I.13) and Figure I.2B, all infectious hosts (i.e. individuals in any I_i stage) are assumed to transmit the pathogen (to each susceptible) at the same rate, β. While this constant infectiousness assumption is common in studies using the method of stages [23, 151, 152, 170, 237], some studies have considered variable infectiousness between different infectious stages [58, 63, 121] – for example, one study [121] considered a model in which different infectious compartments correspond to distinct phases of HIV infection. We consider compartmental models with gamma-distributed stage durations throughout this thesis, and variable infectiousness between different infectious stages is a major focus of Chapter R2.

Other extensions to the compartmental framework, which we do not consider in this thesis, include accounting for demography, spatial effects or age structure [131]. While we use compartmental models throughout this thesis (as well as focussing on modelling variations in infectiousness during infection – see Section I.2 below), alternative modelling approaches include individual-based [242] and/or network-based models [129].

I.2 Variations in infectiousness during infection

In population-scale compartmental models, an assumption is usually made that the transmission risk of an infectious individual (to each susceptible individual) either does not change during their infectious period [131], or changes at most a discrete

number of times [58, 121]. An alternative, potentially more realistic, approach is to instead model infectiousness as a function of time since infection [77].[1]

Alongside the SIR compartmental model, Kermack and McKendrick also introduced an integro-differential equation (IDE) epidemic model (sometimes referred to as a time since infection [196] or age of infection [26] model, or a distributed delay differential equation model [76]) in their 1927 work [133]. One form of this model (hereafter denoted the K&M model[2]) is given by

$$\frac{\mathrm{d}S}{\mathrm{d}t} = S(t)\int_0^\infty \beta(\tau;t)\frac{\mathrm{d}S(t-\tau)}{\mathrm{d}t}\mathrm{d}\tau, \tag{I.14}$$

where $S(t)$ is the number of susceptible individuals at time t since the start of the outbreak and $\beta(\tau;t)$ is the expected force of infection exerted onto each susceptible individual in the population by a single infected individual at time τ since infection (and at time t since the start of the outbreak). We refer to $\beta(\tau;t)$ as the (expected) *infectiousness profile* (with respect to the time since infection – we also consider the related infectiousness profile relative to the time since symptom onset in Chapters R3 and R4). To interpret this model, we note that the incidence rate of new infections at time t is

$$i(t) = -\frac{\mathrm{d}S}{\mathrm{d}t}, \tag{I.15}$$

and we have

$$i(t) = S(t)\int_0^\infty \beta(\tau;t)i(t-\tau)\mathrm{d}\tau, \tag{I.16}$$

i.e. the force of infection exerted onto each susceptible individual at time t is given by the integral over all times since infection, τ, of the incidence at time $t-\tau$ multiplied by the expected infectiousness at time t of individuals who were infected at time $t-\tau$. In this formulation of the K&M model, in order to solve the model (numerically) to calculate $S(t)$ for $t > 0$, the value of $S(t)$ must be prescribed for all $t \leq 0$ (note that we use a slightly different form of the K&M model in Chapter R2, explicitly including an extra model term for the initial conditions instead of requiring $S(t)$ to be known at negative times).

In fact, a range of compartmental models are obtained as special cases of the K&M IDE model [47, 63, 76]. For example, the deterministic SIR model (equations (I.1)–(I.3), replacing β in those equations with $\tilde{\beta}$ to distinguish this parameter from $\beta(\tau;t)$)

[1] Note that in this thesis, we use the terms "infectiousness" and "transmission risk" synonymously, i.e. accounting for both the rate at which an infected individual has contacts with susceptible individuals and the probability that each such contact leads to a transmission occurring.

[2] Note that some authors refer to the SIR model as the "Kermack and McKendrick model" (or similar), but we follow other studies [26, 63] by using this name to denote the more general IDE model introduced in [133].

is equivalent to equation (I.14) with expected infectiousness profile

$$\beta(\tau;t) = \tilde{\beta}\exp(-\mu\tau), \tag{I.17}$$

where $\tilde{\beta}$ is the transmission rate of an infectious individual and $\exp(-\mu\tau)$ gives the probability that an individual is still infectious at time τ since infection (i.e. the probability that the duration of the infectious period exceeds τ). Conversely, we show in Chapter R2 that the K&M model with general infectiousness profile $\beta(\tau)$ can be accurately approximated using a multi-stage compartmental model.

The basic reproduction number is given in terms of the expected infectiousness profile by [77, 88]

$$R_0 = N\int_0^\infty \beta(\tau;0)\mathrm{d}\tau, \tag{I.18}$$

(assuming a single newly infected individual is introduced into the population at time 0 and that there is no change in the infectiousness profile over the course of their infection), where N is the size of the population under consideration (which is assumed to be constant). More generally, the (instantaneous) *time-dependent reproduction number*, R_t (the average number of transmissions arising from an individual who is infected at time t, assuming no change in transmission conditions during the course of their infection), is given by [88]

$$R_t = S(t)\int_0^\infty \beta(\tau;t)\mathrm{d}\tau. \tag{I.19}$$

While we focus on the instantaneous reproduction number in this section, a case reproduction number, which accounts for changes in the availability of susceptibles and in the expected infectiousness profile during the course of infection, can also be defined [88].

The expected infectiousness profile can be decomposed as [88]

$$\beta(\tau;t) = \frac{R_t}{S(t)}f_{gen}(\tau;t), \tag{I.20}$$

where $f_{gen}(\tau;t)$ has total integral one over all times since infection, τ (at each fixed value of t). The function $f_{gen}(\tau;t)$ gives the probability density function of the distribution of the *generation time* – the interval between an individual being infected themselves and transmitting the pathogen [13, 77, 105, 210, 235] – at least assuming transmission conditions do not change during infection [88]. In Chapter R4 we demonstrate that realised generation times in households are in fact shorter than would be predicted by the normalised infectiousness profile, since the number of susceptible household members available for an infected individual to infect may decrease during infection. Nonetheless (as is standard), except where explicitly stated

otherwise, throughout this thesis we refer to the normalised expected infectiousness profile as the generation time distribution.

Throughout the majority of this thesis, we make the common assumption that the generation time distribution, $f_{gen}(\tau;t)$, is independent of the calendar time, t (except in Section R4.3.5 of Chapter R4, where we estimate the SARS-CoV-2 generation time separately for three different time periods during 2020). For simplicity we therefore make the same assumption in the remainder of this section, hereafter denoting the generation time distribution by $f_{gen}(\tau)$. Similarly, since we do not consider temporal changes in $\beta(\tau;t)$ in Chapter R2, and are mainly interested in the shape of the infectiousness profile (i.e. the generation time distribution) in Chapters R3 and R4, we suppress the t-dependence of $\beta(\tau;t)$ in those chapters.

Using the decomposition of $\beta(\tau;t)$ in equation (I.20), the K&M model can be written as the renewal equation,

$$i(t) = R_t \int_0^\infty f_{gen}(\tau) i(t-\tau) \mathrm{d}\tau. \tag{I.21}$$

If the generation time distribution is known, this equation can be used to estimate the time-dependent reproduction number using incidence data, since we have [88]

$$R_t = \frac{i(t)}{\int_0^\infty f_{gen}(\tau) i(t-\tau) \mathrm{d}\tau}. \tag{I.22}$$

An important special case of this equation occurs if the epidemic initially grows exponentially at rate r (i.e. $i(t) \propto \exp(rt)$), in which case the basic reproduction number and exponential growth rate are linked by the equation [235]

$$R_0 = \frac{1}{\int_0^\infty f_{gen}(\tau) \exp(-r\tau) \mathrm{d}\tau}. \tag{I.23}$$

Because real-world incidence data are reported at discrete time points, it is convenient to discretise equation (I.22) to obtain [88]

$$R_t = \frac{i_t}{\sum_{s=0}^\infty p_{gen,s} i_{t-s}}. \tag{I.24}$$

Here, i_t is the number of new infections reported at time point t, and $p_{gen,s}$ is the probability mass function of the discretised generation time distribution (t and s are often in units of days, but other units could be used if incidence data are not reported daily).

Renewal equation-based approaches are commonly used to estimate the time-dependent reproduction number (see, for example, [1, 62, 101, 214]). Real-world incidence data may not exactly correspond to the (daily) number of new infections –

e.g. for SARS-CoV-2, the UK reports case numbers by the date on which a (positive) test is taken; such data reflect the numbers of new infections at earlier time points because of delays being infected and taking a test (for example, an infected individual may not take a test until they develop symptoms). Therefore, some approaches for estimating the time-dependent reproduction number explicitly correct for delays between infection and reporting [1]. Other studies have considered a similar renewal framework to that described above but instead link the time-dependent reproduction number, the incidence of new symptomatic cases, and the distribution of the serial interval (the difference between symptom onset times in an infector-infectee pair) [62, 214].

While we have derived a simple version of the renewal equation here (equations (I.22) and (I.24)), related approaches for estimating the time-dependent reproduction number, R_t, often account for other factors. For example, two common extensions are using Bayesian inference (see Section I.4 below) to account for randomness in incidence data (due to stochasticity in underlying epidemic dynamics and/or reporting errors) and uncertainty in the estimated value of R_t [1, 62, 101, 214], and estimating a single value of the reproduction number over a sliding time window of length τ (i.e. when estimating R_t at time t, an assumption is made that the reproduction number was constant between times $t-\tau$ and t) in order to smooth out random variations in incidence that are not due to changes in R_t [1, 62, 101, 188, 214].

As indicated by equations (I.22) and (I.24), estimates of the time-dependent reproduction number often rely on estimates of the generation time distribution. In addition, since the generation time distribution represents the normalised infectiousness profile of infected individuals, this distribution is important in mathematical analyses of interventions such as case isolation and contact tracing (e.g. [85, 90] and Chapter R3 of this thesis). Models incorporating variations in infectiousness during infection, such as the K&M model, can also be used to project population-scale epidemic dynamics (see Chapter R2 of this thesis). Estimation of the generation time distribution is a major focus of this thesis. Two contrasting approaches exist for estimating the generation time distribution:

i. First, data comprising symptom onset dates in infector-infectee transmission pairs (see e.g. [85, 93] and Chapter R3 of this thesis) or households (see e.g. [40, 82] and Chapter R4 of this thesis) can be used. Usually, the exact infection times of each infector-infectee pair are unobserved. Therefore, because the generation time and serial interval distributions may differ [145] (note that in data where infector-infectee pairs are known exactly, the serial interval of each pair can

be calculated directly), estimation of the generation time distribution using such data typically requires an assumption regarding the relationship between infectiousness and symptoms, with the most straightforward assumption being that the transmission risk is independent of exactly when symptoms develop [85, 93] (see Chapter R3, where we analyse different possible assumptions about the relationship between SARS-CoV-2 transmission and symptoms).

ii. Alternatively, within-host data can be used to estimate the infectiousness profile (see Chapter R2 of this thesis). For a viral infection, such data may comprise longitudinal observations of the viral load over the course of infection for a cohort of infected individuals. By fitting a within-host model to viral load data, the viral load (of each individual in the cohort) can be estimated as a continuous curve relative to the time since infection (see the introduction to within-host modelling in Section I.3 below). The infectiousness profile can then be estimated by assuming a functional relationship between viral load and infectiousness – for example, linear [35, 49, 108, 111, 174] and log-linear [49, 108] relationships between viral load and infectiousness have been considered for influenza, while a log-linear relationship has been suggested for SARS-CoV-2 [45, 125, 128, 161].

I.3 Within-host modelling

One simple model of within-host viral dynamics is the target cell-limited (TCL) model [36] (Figure I.3). This model was first developed to describe HIV infection [194], but has since been applied to other viral infections including hepatitis C [178] and influenza [17]. The model is given by the system of ODEs

$$\frac{\mathrm{d}T}{\mathrm{d}\tau} = -\beta TV, \tag{I.25}$$

$$\frac{\mathrm{d}I}{\mathrm{d}\tau} = \beta TV - \delta I, \tag{I.26}$$

$$\frac{\mathrm{d}V}{\mathrm{d}\tau} = pI - cV, \tag{I.27}$$

where $T(\tau)$ is the number of susceptible target cells, $I(\tau)$ is the number of infected target cells, $V(\tau)$ is the quantity of free virus, and τ is the time since infection (note that the variables and parameters in this model are not the same as those denoted by the same symbols in the population-scale models considered in Section I.1). In this model, susceptible target cells are infected by free virus at total rate βTV per day, where β is the infection rate parameter. Infected target cells shed virus at rate

p and die at rate δ, while free virus is cleared at rate c. An immune response is not explicitly described in the model, but is implicitly included in the parameters δ and c [17]. We consider this model in the context of influenza in Chapter R2.

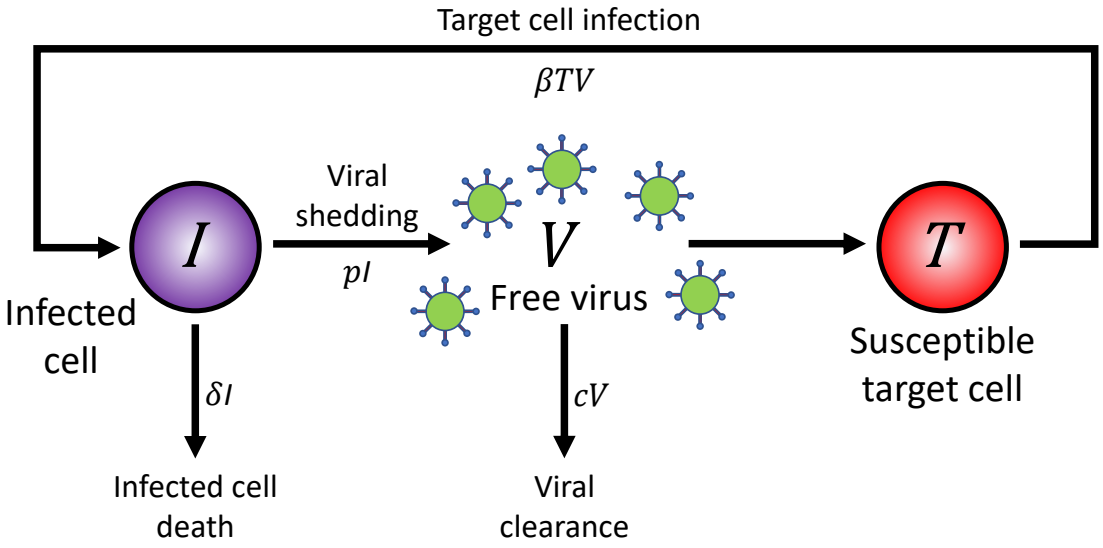


Figure I.3: The target cell-limited model. Flow diagram illustrating the TCL within-host model.

An important extension to the basic TCL model is the inclusion of an eclipse phase, during which a recently infected cell has not yet begun to shed the virus – this extension is common in within-host influenza models [17, 19, 207, 233]. Other extensions to the basic TCL model include accounting for the proliferation of both susceptible and infected cells [67], while more detailed models of within-host dynamics explicitly account for the effect of the innate [17, 35, 48, 107, 192] and/or adaptive [48, 110, 120] immune response.

In addition to their use in improving understanding of infection dynamics, and for estimating biologically meaningful parameter values that cannot be determined experimentally [19], within-host models have also been used to predict and test the effects of antiviral drugs [36, 207]. The emergence of resistance to antiviral drugs has also been considered in modelling studies [109]. Finally, as described above, within-host models fitted to data can be used to estimate the infectiousness profile of infected hosts, which can, in turn, be used (for example) to project population-scale epidemic dynamics [167, 174] – this "multi-scale" epidemic modelling approach is the focus of Chapter R2 of this thesis.

I.4 Parameter inference

Throughout this thesis, we estimate parameter values of different mathematical models by fitting the models to a variety of data types using three techniques: least-squares estimation, maximum likelihood estimation and Bayesian inference using Markov chain Monte Carlo (MCMC). A brief introduction to each of these three methods is given below.

In Chapters R1 and R2 we fit deterministic models (a population-scale model in Chapter R1 and a within-host model in Chapter R2) to data using (non-linear) least-squares estimation [208]. Generally, this approach involves finding the value of the vector of model parameters, denoted θ, that minimises the sum of squares residual [208],

$$SS_{res}(\theta) = \sum_{i=1}^{n} \left(f\left(t_i;\theta\right) - y_i\right)^2. \tag{I.28}$$

Here, y_i denotes the recorded value of an observable (the cumulative number of Ebola cases in Chapter R1 and the viral load of an infected individual in Chapter R2) at time-point t_i (the time since the start of the outbreak in Chapter R1 and the time since infection in Chapter R2), $f(t_i;\theta)$ denotes the corresponding predicted value using the model under consideration, and n is the total number of observations.

Maximum likelihood estimation is a parameter inference method for stochastic models [72] that we use in Chapter R3. In general, if $p\left(\boldsymbol{x} \mid \theta\right)$ is the probability of observing data $\boldsymbol{x}$ (where in Chapter R3, $\boldsymbol{x}$ is the vector of symptom onset times in infector-infectee transmission pairs) given the parameter values, θ, of the model under consideration, then the likelihood function is defined as

$$L(\theta) = p\left(\boldsymbol{x} \mid \theta\right), \tag{I.29}$$

(where we suppress the dependence of $L(\theta)$ on the observed data) [72]. Maximum likelihood estimation involves finding the parameter values, θ, that maximise the likelihood, $L(\theta)$ [72].

One special case of maximum likelihood estimation occurs when time-series data are assumed to be generated by a deterministic model, but with a normally distributed random variate of mean zero and standard deviation σ (representing, for example, a measurement error) added to each observation (independently for different observations). In this case (using the same notation as above), the likelihood is given by

$$L(\theta) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(f\left(t_i;\theta\right) - y_i\right)^2\right). \tag{I.30}$$

It follows that least-squares estimation and maximum likelihood estimation are equivalent in this special case, which occurs in Chapters R1 and R2 when we generate synthetic data assuming normally distributed measurement errors (note that we conduct separate analyses using simulated and real data in Chapter R1). This equivalence is true whether or not the exact value of σ is known (at least provided f does not depend on σ), but if σ is unknown then an additional equation can be derived in order to estimate it alongside the model parameters, θ.

Finally, in Chapters R3 and R4 we use MCMC [96] to fit stochastic models to data from transmission pairs (Chapter R3) and households (Chapter R4). When using MCMC, we take a Bayesian approach to parameter inference that involves calculating a distributional estimate of possible values of the model parameters, θ, given the observed data, $\boldsymbol{x}$ (this distributional estimate, denoted $p(\theta \mid \boldsymbol{x})$, is called the posterior distribution) [72]. In Bayesian inference, a prior distribution, $\pi(\theta)$, is specified, representing one's beliefs about the distribution of possible values of the model parameters, θ, before any data are observed [72]. It then follows from Bayes' theorem that the posterior distribution, $p(\theta \mid \boldsymbol{x})$, satisfies [72]

$$p(\theta \mid \boldsymbol{x}) \propto p\left(\boldsymbol{x} \mid \theta\right) \pi(\theta) = L(\theta)\pi(\theta). \tag{I.31}$$

MCMC methods provide a family of approaches for obtaining a sample from the posterior distribution, $p(\theta \mid \boldsymbol{x})$ [96]. One of the simplest MCMC algorithms, the Metropolis-Hastings algorithm [96, 117], is described in Algorithm I.1 (we describe this algorithm in relation to Bayesian inference, but in principle it can be used to sample from any probability distribution). Briefly, this algorithm generates a sequence of values of the model parameters, $\theta = \theta_n$ (where $n = 0, \ldots, N$ for some N), that converges to a sequence of samples from the posterior distribution for large n. In each step, a candidate value of $\theta = \theta_{prop}$ is sampled from some proposal distribution $Q\left(\theta_{prop} \mid \theta_{n-1}\right)$, where in Algorithm I.1 the proposal distribution can be any distribution that is symmetric in the sense of satisfying $Q(\tilde{\theta} \mid \theta) = Q(\theta \mid \tilde{\theta})$, although this algorithm can also be adapted for non-symmetric proposal distributions. The proposed value of θ is accepted (i.e. we set $\theta_n = \theta_{prop}$) with a probability that depends on the relative posterior densities at θ_{prop} and θ_{n-1} (where the proposal is always accepted if the posterior density is higher at θ_{prop} than at θ_{n-1}), and is otherwise rejected (i.e. we set $\theta_n = \theta_{n-1}$). For a proof that the sequence θ_n converges to a set of samples from the posterior distribution, see (for example) [96].

While the chain in Algorithm I.1 eventually converges to the posterior distribution of θ, initial estimates (i.e. θ_n for small n) may not follow this distribution. Therefore, the first M estimates are usually discarded for some $M < N$ (the "burn-in" period)

Algorithm I.1: The Metropolis-Hastings algorithm.

Initialise $\theta = \theta_0$;

for $n = 1, \ldots, N$ **do**

  Sample θ_{prop} from $Q(\theta_{prop} \mid \theta_{n-1})$;

  Calculate $\alpha_n = L(\theta_{prop})\pi(\theta_{prop})/(L(\theta_{n-1})\pi(\theta_{n-1}))$;

  Generate a random number, r_n, uniformly distributed between 0 and 1;

  if $r_n \leq \alpha_n$ **then**

    Set $\theta_n = \theta_{prop}$;

  else

    Set $\theta_n = \theta_{n-1}$;

  end

end

when this algorithm is used to obtain a sample from the posterior [96]. Also, since the θ_n are likely to be auto-correlated (i.e. after the burn-in period, the θ_n comprise an approximate sample from the posterior but not necessarily an independent sample), often only every j samples are kept after the burn-in (for some j) in order to reduce this auto-correlation ("thinning") [96].

In both Chapter R3 and Chapter R4, the likelihood given the observed data is difficult to calculate. This is because the data in those chapters comprise symptom onset dates in infector-infectee pairs (Chapter R3) and households (Chapter R4), sometimes supplemented with intervals of possible infection dates. Therefore, while relatively simple expressions exist for the likelihood given exact times of both symptom onset (within recorded symptom onset dates) and infection, these event times are unknown in those two chapters. An approximation to the likelihood in Chapter R3 can be obtained by assuming that symptom onset events can only occur at a single time each day (e.g. in the middle of the day), and integrating over possible infection times (this approximation was used when calculating maximum likelihood estimates, which were useful for model comparison). However, such an approach was not feasible in Chapter R4, since uncertainty in who-infected-whom in households leads to a more complex likelihood function than in Chapter R3.

One approach for accounting for uncertainty in precise event times is data augmentation MCMC, where data that are not directly observed are imputed and updated as part of the MCMC algorithm [142]. We use this approach in Chapters R3 and R4, augmenting the observed data with exact symptom onset times in Chapter R3 (in this case integrating over possible infection times) and with both infection and

symptom onset times in Chapter R4. One way to interpret data augmentation MCMC is to think of the imputed data as additional model parameters that are estimated as part of the MCMC procedure [43]. The MCMC algorithms used in Chapters R3 and R4 are variants of the Metropolis-Hastings algorithm in which, in alternating steps of the chain, either the model parameters are updated or the augmented data are updated (a form of Gibbs sampling [96]).

I.5 Outline of this thesis

In the remainder of this thesis, we present four research chapters (Chapters R1–R4) before concluding with an overall discussion. We have published the research underlying each of Chapters R1–R4 in peer-reviewed journals [112, 113, 115, 116]. Below, we summarise the four main research chapters.

In Chapter R1 we relax a standard assumption that is implicit in epidemiological models of Ebola outbreaks which are commonly used to analyse interventions targeting symptomatic hosts (such as isolation or treatment) [52, 140, 163, 165]: that all symptomatic hosts are equally likely to be detected. In other words, following an incubation period, the level of symptoms displayed by an individual host (who has not yet been detected) is usually assumed to remain constant throughout infection. However, individuals with Ebola infections in reality progress from initial non-specific symptoms through to more severe phases of infection [20]. In this chapter, we therefore compare projections of two compartmental models of Ebola interventions: one in which a constant symptoms level is assumed, and a more epidemiologically realistic model that accounts for varying symptoms during infection. Both models can reproduce observed epidemic data, as we show by fitting the models to data from the 2018–2020 epidemic in the Democratic Republic of the Congo (DRC) and the 2014–2016 epidemic in Liberia. However, for both of these epidemics, when interventions are altered identically in the two models, projections from the models differ. This result demonstrates the potential importance of including individual host-scale dynamics, such as varying symptoms, in models used to plan interventions against outbreaks of Ebola and other pathogens.

In Chapter R2 we turn our attention to variations in infectiousness (rather than symptoms) during infection. Specifically, we focus on multi-scale epidemic models [9, 25, 51, 61, 63, 73, 92, 95, 108, 144, 154, 155, 181, 249], in which a within-host model (which may be parameterised using infection data collected in longitudinal cohort studies) is used to predict the expected infectiousness profile of infected hosts, which is, in turn, used to inform population-scale projections. Until now, most multi-scale

models have been complex, requiring significant modelling expertise to run. In this chapter, we formulate an alternative, easy-to-use, multi-scale modelling framework. Our approach, which uses a multi-stage compartmental population-scale model with a large number of infected stages, generates identical results compared to previous more mathematically complex approaches. We apply our framework to the case study of influenza A in humans. By using a within-host viral dynamics model to generate synthetic host-scale data, we explore the effects of limited and inaccurate host-scale data on the accuracy of population-scale projections. If infection data are collected daily, we find that a cohort of at least 40 hosts is required for an expected population-scale projection error of no more than 10%. Errors in population-scale projections may be reduced by including more hosts in future cohort studies or by increasing the frequency of observations for each host. This work, therefore, provides not only an accessible epidemiological modelling framework but also an insight into the data required for accurate projections using multi-scale models.

In Chapter R3 we focus on inferring the infectiousness profile of SARS-CoV-2 infections (with respect to both the time since infection and the time since symptom onset) – which is critical to assess the effectiveness of public health measures such as isolation and contact tracing – using data describing symptom onset dates in infector-infectee transmission pairs (rather than host-scale data as in Chapter R2). Most earlier estimates of the SARS-CoV-2 infectiousness profile with respect to the time since infection (i.e. the generation time distribution) have involved an assumption that the infectiousness of each infected host is independent of exactly when they develop symptoms [74, 85, 93, 137, 145], but this assumption is not epidemiologically justified [16, 145]. In this chapter, we develop an alternative mechanistic approach, which is based on compartmental modelling, for inferring changes in infectiousness during SARS-CoV-2 infection using transmission pair data. We compare estimates of key epidemiological quantities generated using our mechanistic method with analogous estimates generated using previous approaches. The mechanistic method provides an improved fit to data from SARS-CoV-2 infector–infectee pairs compared to previously used methods. Our best-fitting model indicates a high proportion of presymptomatic transmissions, with many transmissions occurring shortly before the infector develops symptoms. High infectiousness immediately prior to symptom onset highlights the value of contact tracing, even if contacts from a short time window before symptom onset alone are traced.

Finally, in Chapter R4 we update the results, and further develop the methodology, of Chapter R3. While numerous estimates of the SARS-CoV-2 generation time distribution exist using data from the earliest months of the COVID-19 pandemic

(for example, we used data from December 2019–March 2020 in Chapter R3), very few estimates existed using more recent data at the time of conducting the analyses in Chapter R4. In this chapter, we use data from a household study conducted from March–November 2020 in the UK to obtain updated estimates of the generation time. We consider both the commonly used approach in which infectiousness is assumed to be independent of when symptoms develop, and the mechanistic model developed in Chapter R3 in which transmission and symptoms are linked explicitly. In order to estimate the generation time using data from households rather than transmission pairs (as in Chapter R3), we update our methodology, accounting for susceptible depletion and uncertainty regarding who-infected-whom within households. Assuming independent transmission and symptoms, we estimate a mean generation time similar to previous estimates from other countries, but with a higher standard deviation. Using our mechanistic approach, we estimate a longer mean generation time than suggested by the independent transmission and symptoms model, and a similar standard deviation. Both models suggest a shorter mean generation time in September-November 2020 compared to earlier months. Since the SARS-CoV-2 generation time appears to be changing, continued data collection and analysis is necessary to inform future public health policy decisions.

Taken together, this research shows how compartmental epidemic models can be adapted both to inform estimates of population-scale dynamics with details of individual host-scale dynamics (such as variations in infectiousness and symptoms during each infection), and to estimate how infectiousness varies during infection. In addition, in Chapters R1 and R3 we demonstrate how assessments of the possible impacts of control interventions may differ when using the methods we develop compared to commonly used approaches.

Chapter R1

Accurate assessments of the effectiveness of interventions against Ebola may require models that account for variations in symptoms during infection

We have published the research underlying this chapter in the journal *Epidemics* [112].

R1.1 Introduction

The 2018–2020 Ebola epidemic in the DRC was the second largest in history, with 3470 cases (3317 confirmed and 153 probable) recorded [245]. Mathematical models are increasingly being used for exploring the effects of different possible control interventions during Ebola epidemics [33, 163, 241]. Typically, the values of model parameters are chosen so that the model output matches observed epidemic data (model fitting; Figure R1.1A), and then interventions are introduced in the fitted model to project how the course of the epidemic is altered by different possible control strategies (intervention testing; Figure R1.1B).

A commonly used model for characterising epidemics of diseases including Ebola is the SEIR model [10, 54, 55], and extensions to this basic model include explicit incorporation of transmission from Ebola deceased hosts [3, 18, 239] or accounting for mismatches between symptoms and infectiousness [215, 221]. Possible interventions include isolation of symptomatic hosts, which can be included in the SEIR model by removing individuals from the infectious compartment. All individuals in the infectious compartment are usually assumed to be symptomatic, with the level of

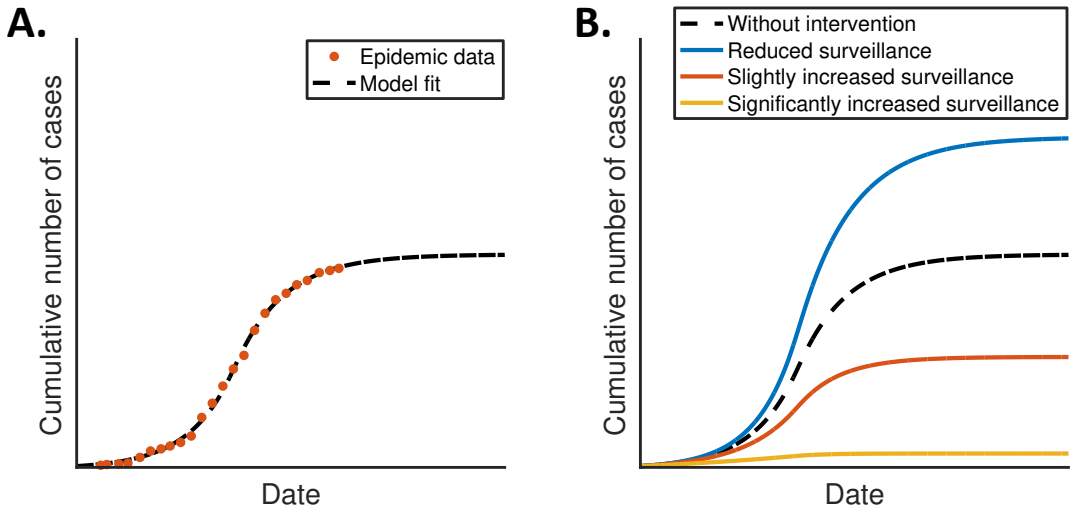


Figure R1.1: Schematic showing how a model can be used to project the effect of changing control interventions on epidemic dynamics. A. Model fitting. Model parameters are chosen so that the model output (black dashed) approximates epidemic data (red dots). B. Intervention testing. Alternative control interventions are introduced into the fitted model, and projected dynamics under these new control interventions can be calculated – we show projections of the effects of reduced surveillance (blue), slightly intensified surveillance (red) and significantly intensified surveillance (yellow) compared to a baseline level that was assumed when fitting the model to data. Here, an intensification of surveillance is assumed to lead to improved detection and control of infectious hosts, thereby reducing the total number of cases.

symptoms being assumed constant and therefore independent of the stage of infection. For example, symptomatic individuals were assumed to be isolated at a constant rate in [52] and hospitalised at a constant rate in [163].

However, in reality, it is not the case that all symptomatic hosts are equally symptomatic. During an Ebola infection, an infected host progresses through different stages [20] – from initial non-specific symptoms (fever, headache and myalgia) to a gastrointestinal phase (diarrhoea, vomiting, abdominal symptoms and dehydration), and then either to a deterioration phase (collapse, neurological manifestations and bleeding) or recovery. Individuals with non-specific symptoms are less likely to be observed and treated/isolated than individuals who have progressed further through infection and have developed more specific and more serious symptoms.

In this research chapter, we investigate whether explicitly accounting for variations in symptom expression during the course of an Ebola infection leads to different epidemic dynamics compared to assuming a constant level of symptoms. To do this, we compare projections derived from a model in which infectious hosts have a constant level of symptoms throughout infection (the "constant symptoms model") with those from a model in which variable symptoms during infection are accounted

for (the "variable symptoms model"). We first fit the two models to epidemic data, considering (in two separate analyses) data from two Ebola epidemics: the 2018–2020 Ebola epidemic in the DRC and the 2014–2016 epidemic in West Africa (the largest Ebola epidemic in history [246]). Then, to explore the effects of intensifying surveillance and control, we compare projections generated by the two models when control interventions in the models are altered.

These analyses demonstrate that models with or without variable symptoms can reproduce observed disease incidence time series, but that projections from the models are different when interventions are altered, even when the change in interventions is identical in both models. Our results highlight the need to consider whether variations in symptom expression during infection should be included in models of Ebola epidemics. Without accounting for variable symptoms, projections of the effects of interventions may be incorrect.

R1.2 Methods

R1.2.1 Data

To show that our results are not conditioned on particular properties of data from a single epidemic, we conducted an analysis of simulated data (see Section R1.2.5 below) as well as two separate analyses in which we considered data from two different Ebola epidemics.

In the first analysis of real epidemic data, we used data on the numbers of cases reported approximately weekly during the 2018–2020 Ebola epidemic in the DRC. In January 2019, Dr Peter Salama, then Deputy Director-General of Emergency Preparedness and Response at the World Health Organization, suggested that this epidemic comprised several distinct outbreaks in different affected areas [195]. Indeed, disease incidence time series display distinct phases (large numbers of cases at the end of July and beginning of August 2018, followed by low numbers of cases in September, and then larger numbers of cases again thereafter), likely due to spatial effects of spread of the virus which are not captured by standard non-spatial compartmental models [131]. For this reason, we focussed on data from the health zone of Beni, a city in the north-east of the DRC, and the neighbouring health zone of Kalunguta. This region was severely impacted by the 2018–2020 epidemic. Data were obtained from World Health Organization disease outbreak news reports from 4$^{\text{th}}$ August 2018 to 10$^{\text{th}}$ January 2019 [244].

In the second analysis of real epidemic data, we considered data consisting of the numbers of cases reported approximately weekly in Liberia during the 2014–2016

Ebola epidemic, which were obtained from the World Health Organization [243]. Data were taken from 23rd March 2014 to 3rd May 2015.

The two epidemic datasets are available alongside the publication from which this research chapter was adapted [112].

R1.2.2 Mathematical models

We extended the commonly used SEIR model to account explicitly for detection followed by isolation of infectious individuals. Specifically, we added a "controlled" (C) compartment representing the number of individuals that have ever been controlled (detected and isolated) up to the current time. In addition, we divided the infected (I) compartment into three sub-compartments, corresponding to the distinct stages of an Ebola infection described in Section R1.1. As mentioned in Section I.1, this results in the infectious period being gamma-distributed (in the absence of control), rather than exponentially distributed – and gamma distributions have been found to characterise epidemiological periods accurately in a range of systems [63, 216, 237]. When control is added, this gave rise to the SEI_3RC model,

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta\left(t\right) S\left(I_1 + I_2 + I_3\right), \tag{R1.1}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \beta\left(t\right) S\left(I_1 + I_2 + I_3\right) - \gamma E, \tag{R1.2}$$

$$\frac{\mathrm{d}I_1}{\mathrm{d}t} = \gamma E - \mu I_1 - \delta_1 I_1, \tag{R1.3}$$

$$\frac{\mathrm{d}I_2}{\mathrm{d}t} = \mu I_1 - \mu I_2 - \delta_2 I_2, \tag{R1.4}$$

$$\frac{\mathrm{d}I_3}{\mathrm{d}t} = \mu I_2 - \mu I_3 - \delta_3 I_3, \tag{R1.5}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \mu I_3, \tag{R1.6}$$

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \delta_1 I_1 + \delta_2 I_2 + \delta_3 I_3. \tag{R1.7}$$

Here, $\beta(t)$ is the transmission rate parameter (which was assumed to vary temporally – see the next paragraph), $1/\gamma$ and $3/\mu$ give the mean latent and infectious periods (in the absence of control), respectively, and δ_i is the detection (and control) rate of each individual in infectious stage I_i. To determine the impact on projections of variable symptoms during infection alone, in the baseline version of the model we assumed that all infectious hosts are equally infectious (at any given calendar time, t; this assumption is relaxed in Figure R1.7 and Appendix A1.1.1).

In our analyses, similarly to other Ebola models [54, 163], we made an assumption that the transmission rate parameter in the model is temporally varying, to reflect

changes in transmissibility during the epidemic. This could, for example, be due to changes in behavioural responses or alterations to interventions (aside from detection and isolation, which we model explicitly). In particular, we assumed that

$$\beta(t) = \begin{cases} \beta_0, & \text{for } t \leq T, \\ \beta_1, & \text{for } t > T. \end{cases} \tag{R1.8}$$

We then considered two alternative versions of the model. In the first (the constant symptoms model – Figure R1.2A), we assumed that all infectious individuals are successfully detected and isolated at the same average rate per day, so that $\delta_1 = \delta_2 = \delta_3 = \delta$, say. This assumption is common to any epidemiological model that includes interventions aimed at symptomatic hosts, unless differences in symptom expression are accounted for explicitly. The constant symptoms model is therefore (in this sense) similar to most epidemiological models that have been used to represent Ebola epidemics previously (e.g. [52, 163]).

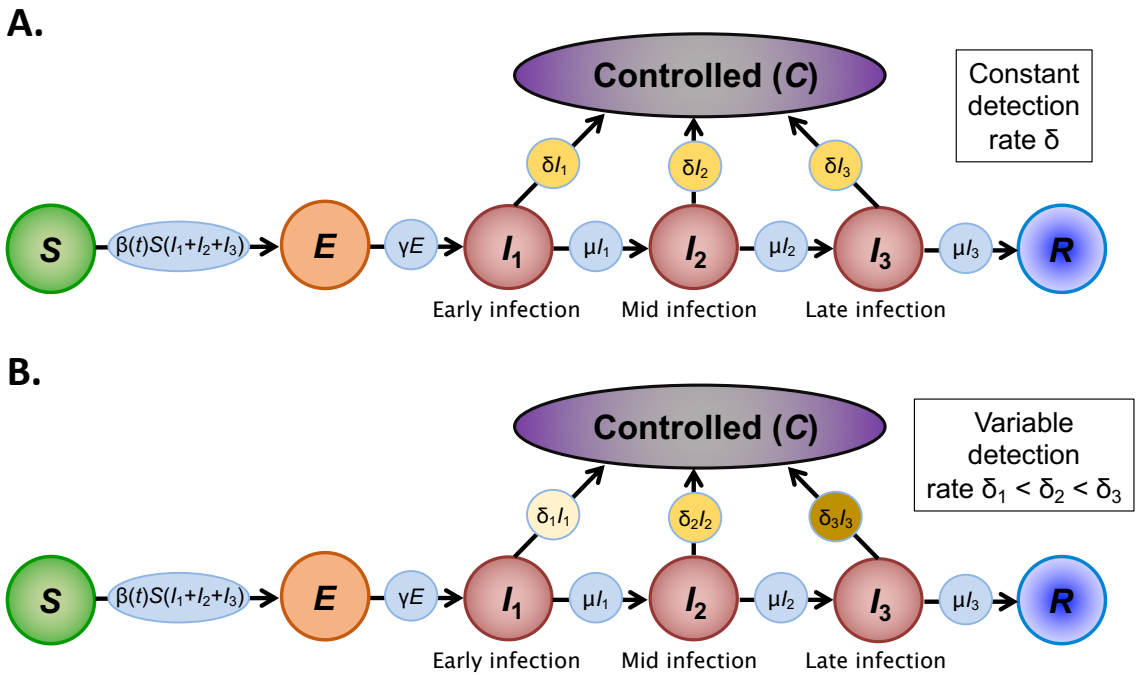


Figure R1.2: Flow diagram illustrating the different models that we considered. A. Constant symptoms model, in which individuals in each stage of infection are equally likely to be detected and isolated (so that the detection rate, δ, is equal for all three infectious stages); B. Variable symptoms model, in which symptoms are assumed to worsen during an Ebola infection (so that the detection rate is smaller for individuals in earlier infection compared to later infection, i.e. $\delta_1 < \delta_2 < \delta_3$). We show how additional epidemiological complexity can be added to these models in Appendix A1.1.

In addition, we considered the more realistic case in which symptoms become more severe as infection progresses, so that $\delta_1 < \delta_2 < \delta_3$. We refer to the resulting

model as the variable symptoms model (Figure R1.2B). This model reflects the fact that, in reality, individuals with initial mild symptoms are less likely to be detected and isolated to prevent further transmission, compared to individuals with more developed symptoms who are in the gastrointestinal or deterioration phases.

We considered the numerical solutions of the models described above in a host population of (constant) size

$$N = S + E + I_1 + I_2 + I_3 + R + C \tag{R1.9}$$

individuals. The basic reproduction number at the beginning of the epidemic and in the absence of surveillance is given by

$$R_0 = \frac{3\beta_0 N}{\mu}, \tag{R1.10}$$

while the effective reproduction number at the beginning of the epidemic but accounting for surveillance is

$$R_{0,eff} = \beta_0 N \left(\frac{1}{\mu + \delta_1} + \frac{\mu}{(\mu + \delta_1)(\mu + \delta_2)} + \frac{\mu^2}{(\mu + \delta_1)(\mu + \delta_2)(\mu + \delta_3)} \right), \tag{R1.11}$$

where the three terms inside the sum represent the expected duration of the I_1 stage, the expected duration of the I_2 stage multiplied by the probability of an infected host reaching the I_2 stage before being detected (and isolated), and the expected duration of the I_3 stage multiplied by the probability of an infected host reaching the I_3 stage before being detected, respectively.

The proportion of infected individuals who are ever detected and isolated during the course of infection is

$$q_C = 1 - \frac{\mu^3}{(\mu + \delta_1)(\mu + \delta_2)(\mu + \delta_3)}, \tag{R1.12}$$

while the proportion of transmissions prevented from each infected host by detection and isolation (relative to the number of transmissions that would occur in the absence of surveillance) is

$$\varepsilon_C = 1 - \frac{R_{0,eff}}{R_0} \tag{R1.13}$$

$$= 1 - \frac{1}{3} \left(\frac{\mu}{\mu + \delta_1} + \frac{\mu^2}{(\mu + \delta_1)(\mu + \delta_2)} + \frac{\mu^3}{(\mu + \delta_1)(\mu + \delta_2)(\mu + \delta_3)} \right). \tag{R1.14}$$

R1.2.3 Dependence of the detection rate on surveillance

The values of the parameters characterising the rate of Ebola detection and isolation, i.e. δ_1, δ_2 and δ_3, depend on the level of surveillance, which includes various passive

and active case finding strategies. We did not model explicitly the wide range of different surveillance activities that take place during an Ebola response (see Section R1.4). Instead, to provide a concrete setting in which to illustrate that projections can be different under the constant symptoms and variable symptoms models, we considered a simplified scenario in which each host is checked for infection exactly every Δ days (for a similar approach, see [65]).

Under this scenario, each time monitoring occurs there is a detection probability of p_i for individuals in compartment I_i (for $i = 1, 2, 3$). Now, if a host enters compartment I_i (for $i = 1$, 2 or 3) at a time that is uniformly distributed between surveillance times, then the expected period until the next surveillance time is $\Delta/2$. Ignoring progression through the infectious stages or to the R compartment, the expected number of surveys until the host is detected is given by $1/p_i$, so that the expected time after the first surveillance time until detection is $\Delta(1/p_i - 1)$. Summing these times gives

$$\text{Expected time in } I_i \text{ until detection} = \Delta/2 + \Delta(1/p_i - 1) \tag{R1.15}$$

$$= \Delta(1/p_i - 1/2). \tag{R1.16}$$

This surveillance regime can therefore be approximated in our SEI_3RC model by taking the rate of detection of a host in compartment I_i to be the multiplicative inverse of this expression, i.e.

$$\delta_i = \frac{1}{\Delta\left(\frac{1}{p_i} - \frac{1}{2}\right)}. \tag{R1.17}$$

An alternative possibility in which the interval between surveillance times is instead assumed to be exponentially distributed is considered in Figure A1.7. In this case, the expected period from entering compartment I_i until the next surveillance time is Δ, so the expected time in I_i until detection is Δ/p_i. Therefore, we instead took $\delta_i = p_i/\Delta$ in this supplementary analysis.

R1.2.4 Parameters and model fitting

Values of model parameters that were not estimated by fitting the models to epidemic data are given in Table A1.1. In most of our analyses (except where stated), we considered two different surveillance regimes. When the models were fitted to epidemic data, under weak surveillance, we assumed that the baseline surveillance period was $\Delta = 21$ days. When the fitted models were then used to project the impact of intensified surveillance, the models were simulated with the surveillance

period changed to $\Delta = 14$ days. We assumed that the detection probability in the constant symptoms model was $p_1 = p_2 = p_3 = 0.6$. When we accounted for the possibility that symptoms change as hosts progress through infection, we instead used default values of $p_1 = 0.1$, $p_2 = 0.8$ and $p_3 = 0.9$ so that the mean value of p_1, p_2 and p_3 was equal to the value of these parameters in the constant symptoms model.

The proportion of isolated hosts, q_C, and proportion of transmissions prevented by isolation, ε_C, under the above choices of surveillance frequency and detection probabilities, are given in Table A1.2. Alternative scenarios in which we instead chose the detection probability in the constant symptoms model so that either q_C or ε_C takes the same value in the two models (under baseline surveillance) are considered in Figure R1.5. Different surveillance periods, Δ, are considered in Figures R1.10 and A1.4, and different detection probabilities, p_i, in Figures A1.5 and A1.6.

The values of the transmission rates, β_0 and β_1, as well as the date on which the transmission rate changes, T, were obtained by fitting the outputs of the models to the epidemic data (separately for each model and for each epidemic dataset). The start date of the epidemic, T_0, was also estimated during the fitting procedure (where we generated numerical solutions starting with a single host in the E compartment at the start time of the epidemic, T_0, with all other individuals susceptible).

Model fitting was performed using least-squares estimation. Specifically, we calculated the parameter values which minimised the sum of squares residual,

$$SS_{res} = \sum_{i=1}^{n} \left(C\left(t_i\right) + R\left(t_i\right) - y_i \right)^2, \tag{R1.18}$$

where $C(t_i) + R(t_i)$ is the cumulative number of detected or removed hosts in the model under consideration, and y_i is the cumulative number of cases in the data, at each time of observation, t_i, and n is the total number of observations. Note that we assumed that individuals who had not previously been detected (and therefore never accessed treatment) were detected upon death, although this assumption is relaxed in Appendix A1.1.2 and Figure R1.8. We also considered fitting to non-cumulative data in Appendix A1.2 and Figure R1.11.

To compare the goodness of fit between the constant symptoms and variable symptoms models (for a given epidemic dataset), we calculated the coefficient of determination (R^2 value) for the two models, using the formula [75]

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}. \tag{R1.19}$$

Here,

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{R1.20}$$

is the total sum of squares, where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{R1.21}$$

is the mean of the data. A model with an R^2 value closer to one (i.e. a higher value) provides a better fit to the data than a model with a smaller R^2 value.

R1.2.5 Generating simulated data

In addition to analysing data from real Ebola epidemics, we carried out an analysis in which we used simulated data. We generated a hypothetical epidemic using the variable symptoms model with the "true" parameters shown in Tables A1.1 and A1.3. These parameters correspond to an initial basic reproduction number of $R_0 = (3\beta_0 N)/\mu = 2$ in the absence of interventions. We also considered different true values of fitted parameters (Figure A1.2).

We assumed that the cumulative number of cases was observed weekly for 52 weeks. In our analyses of real epidemic data, the times T and T_0 represent specific dates. In analyses using synthetic data (since we were not considering an epidemic taking place on specific dates), we instead defined T and T_0 as times relative to the beginning of the week in which the first observed case was reported (which was defined to be time zero), so that (for example) the first datapoint with observed cases is recorded at time 7 days.

Observations of the cumulative number of detected cases ($C + R$) were assumed to be recorded imperfectly. This was characterised by adding a normally distributed error with mean zero and standard deviation $\sigma = 4$ (different values of σ are considered in Figure A1.1) to the recorded values of the cumulative numbers of cases each week. The cumulative numbers of cases each week were then rounded to the nearest whole number (with a minimum possible value of zero). Adding noise in this way leads to the possibility that the cumulative number of recorded cases is not monotonic increasing. This can occur in reality, since the number of recorded cases is often corrected after the fact, so that (for example) taking data on the cumulative number of cases from World Health Organization case reports may generate non-monotonic increasing data.

R1.3 Results

As described in Section R1.1, fitted models are often used to test potential control interventions (Figure R1.1). We considered estimating the values of parameters of two models – the constant symptoms and variable symptoms models – using different epidemic datasets, and then using these models to project the impacts of different interventions. Initially, we analysed simulated data to verify that our model fitting procedure could identify the values of the parameters used to generate the data. We then conducted a further investigation in which we fitted the models to data from two Ebola epidemics – first, the 2018–2020 Ebola epidemic in the DRC, and second the Ebola epidemic in West Africa in 2014–2016.

R1.3.1 Simulation study

First, we generated synthetic data on the cumulative numbers of detected hosts during a simulated Ebola epidemic. To do this, we ran the more realistic variable symptoms model, and then supposed that the value of $C + R$ was observed every week for 52 weeks, assuming (in most of our analyses of simulated data) a normally distributed error on the recorded values. We then estimated the values of the parameters β_0, β_1, T and T_0 in the variable symptoms and constant symptoms models using least-squares estimation, assuming the values of other parameters to be known exactly. The estimated parameter values for the variable symptoms model closely matched those used to simulate the data (estimated parameter values are given in Table A1.3).

To investigate whether or not parameter values could be identified unambiguously in each model, we plotted the sum of squares residual for different fixed (T, T_0) pairs (fitting the remaining parameters, β_0 and β_1, at each (T, T_0) pair; Figure R1.3A,C). Similarly, we then plotted the sum of squares residual at different pairs of values of

$$(R_0, R_1) = \left(\frac{3\beta_0 N}{\mu}, \frac{3\beta_1 N}{\mu} \right), \tag{R1.22}$$

(where R_0 represents the basic reproduction number at the beginning of the epidemic and in the absence of surveillance and R_1 is defined for convenience, and where the values of N and μ given in Table A1.1 were assumed to be known), each time fitting T and T_0 (Figure R1.3B,D). In each panel of Figure R1.3, there was a clear unique parameter pair that minimised the sum of squares residual, confirming the identifiability of fitted parameters in both models.

Comparing the fitted models to the simulated data, we found that the two models could fit the data equally well (Figure R1.4A; the R^2 value was 0.99968 for both

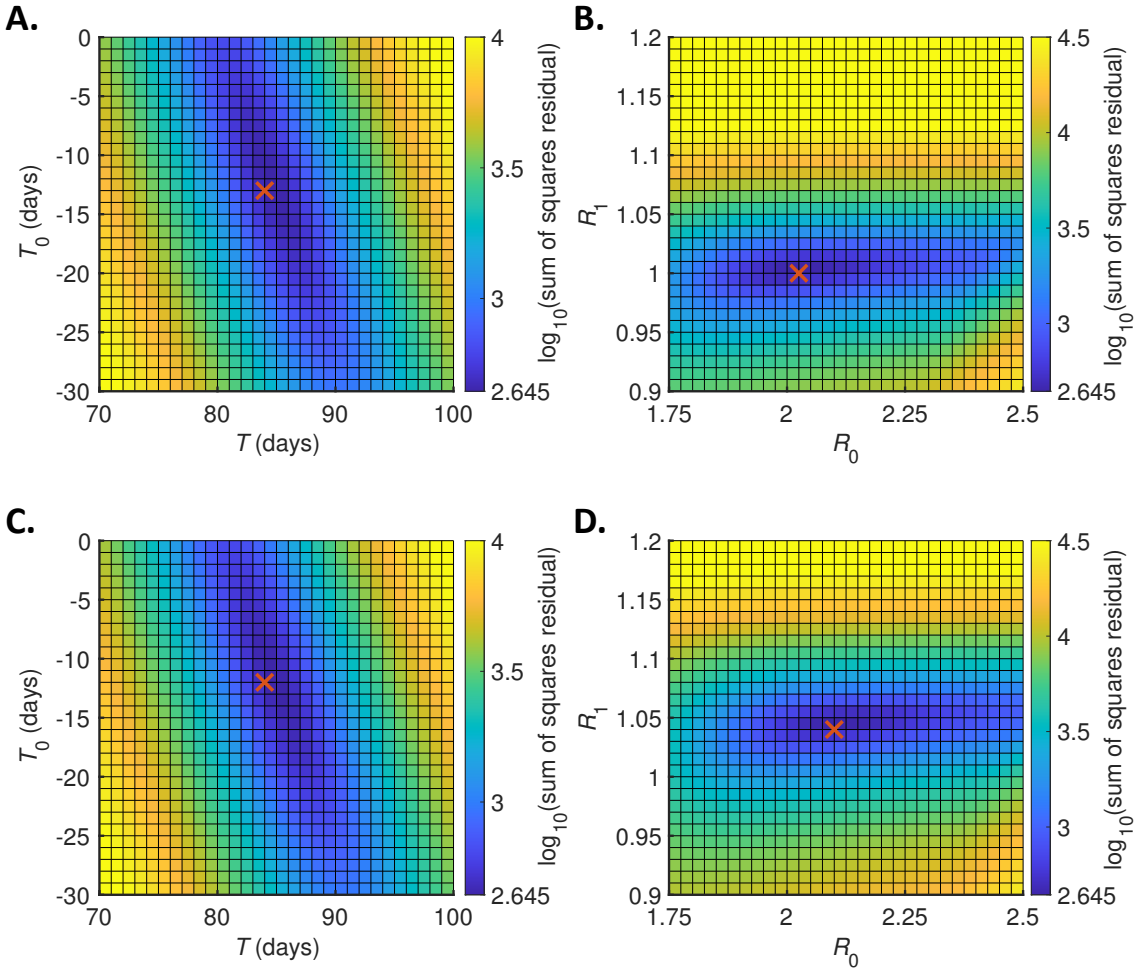


Figure R1.3: Identifiability of model parameters. A. The logarithm of the sum of squares residual for the variable symptoms model, when T and T_0 are fixed at each pair of values shown on the axes, and β_0 and β_1 are fitted to the data. The (T, T_0) pair corresponding to the minimum residual is indicated by a red cross. Errors larger than 10^4 are plotted in yellow. The "true" values of T and T_0 (i.e. the parameter values that were used to generate the synthetic data using the variable symptoms model) are $(T, T_0) = (84 \text{ days}, -14 \text{ days})$. B. The logarithm of the sum of squares residual for the variable symptoms model, when $R_0 = 3\beta_0 N/\mu$ and $R_1 = 3\beta_1 N/\mu$ are fixed at each pair of values shown on the axes, and T and T_0 are fitted to the data. The (R_0, R_1) pair corresponding to the minimum residual is indicated by a red cross. Errors larger than $10^{4.5}$ are plotted in yellow. The "true" values of these parameters are $(R_0, R_1) = (2, 1)$. C–D. Equivalent figures to A–B, for the constant symptoms model.

models). We then used the models to obtain projections of epidemic dynamics under an increased surveillance frequency of once every 14 days, compared to a frequency of once every 21 days assumed when fitting the models to data (Figure R1.4B). Since we modelled surveillance as an active process, where detected individuals are assumed to be isolated and unable to transmit the pathogen, both models projected intensified surveillance to lead to improved epidemic control and therefore fewer

detected cases overall.

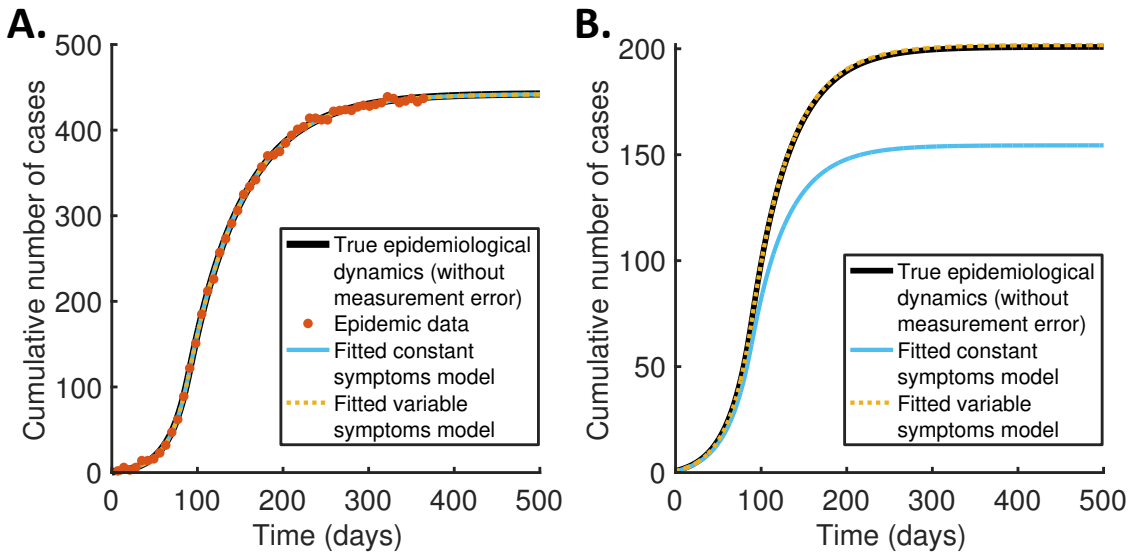


Figure R1.4: Using the constant symptoms model and variable symptoms model to project the impact of intensified surveillance on the simulated epidemic. A. The cumulative number of cases under: the variable symptoms model with the "true" parameter values in Table A1.3 and with no measurement error (black); the synthetic data comprised of weekly numbers of cases with normally distributed error (red dots); the constant symptoms model fitted to the synthetic data (blue), and; the variable symptoms model fitted to the synthetic data (yellow dotted). B. Projections of the effect of intensified surveillance according to the models shown in A. Note the different y-axis limits in A and B.

Even though the two models provided a similar fit to the simulated data under baseline surveillance (Figure R1.4A), projections from the two models when surveillance was intensified were entirely different (Figure R1.4B), with fewer estimated cases under the fitted constant symptoms model than the variable symptoms model (projections under the fitted variable symptoms model closely matched the "true" dynamics that were obtained by intensifying surveillance in the variable symptoms model that was used to generate the synthetic data). This is likely because, under the variable symptoms model, the detection probability of infected individuals (in a given surveillance round) increases during infection, so individuals are typically isolated earlier in infection under the constant symptoms model than the variable symptoms model. As a result, even though the proportion of infected individuals who are ever isolated is in fact lower (under either baseline or intensified surveillance) under the constant symptoms model than the variable symptoms model, a higher proportion of transmissions is prevented by isolation under the constant symptoms model (see Table A1.2). Therefore, while increasing the surveillance frequency leads

to reduced transmission through improved case detection and isolation in both models, intensifying surveillance has a greater impact on transmission under the constant symptoms model. Specifically, the effective reproduction number at the start of the epidemic but accounting for surveillance, $R_{0,eff}$, is 10% lower under intensified than baseline surveillance for the constant symptoms model, but only 8% lower for the variable symptoms model.

R1.3.1.1 Effect of the choice of the detection probability in the constant symptoms model

In Figure R1.4 (and throughout this chapter, except where explicitly stated otherwise), we chose the detection probabilities (in a given surveillance round), p_i, in the constant symptoms and variable symptoms models such that the mean value of p_1, p_2 and p_3 was equal in the two models. This led to the proportion of infected hosts who are ever detected and isolated, q_C, being higher in the variable symptoms model than in the constant symptoms model (Table A1.2). An alternative possibility in which we still used the same detection probabilities in the variable symptoms model ($p_1 = 0.1$, $p_2 = 0.8$ and $p_3 = 0.9$), but instead chose $p_1 = p_2 = p_3 = 0.66$ in the constant symptoms models in order to ensure an equal value of $q_C = 35\%$ in the two models (under the baseline surveillance regime), is considered in Figure R1.5A–B.

The constant symptoms and variable symptoms models again provided very similar fits to the data (Figure R1.5A) under this alternative choice of p_i in the constant symptoms model. However, the higher choice of p_i in the constant symptoms model (0.66 in Figure R1.5A–B compared to 0.6 in Figure R1.4) leads to a higher proportion of transmissions (which would otherwise occur) being prevented by detection and isolation of infected hosts. As a result, the constant symptoms model projected fewer cases under intensified surveillance in Figure R1.5B than in Figure R1.4B (since intensifying surveillance has a greater effect on transmission in Figure R1.5), leading to a greater difference in projections between the constant symptoms and variable symptoms models (note that the variable symptoms model is identical in the two figures).

We note that q_C is identical for the two models under baseline surveillance but differs slightly under intensified surveillance in this analysis (47% in the constant symptoms model and 46% in the variable symptoms model). Therefore, to demonstrate that projections from the two models can differ even when the probability of ever being isolated, q_C, is identical for each model (since a greater proportion of transmissions is still prevented through isolation in the constant symptoms model than in the variable symptoms model), results are also shown in Figure R1.5B for

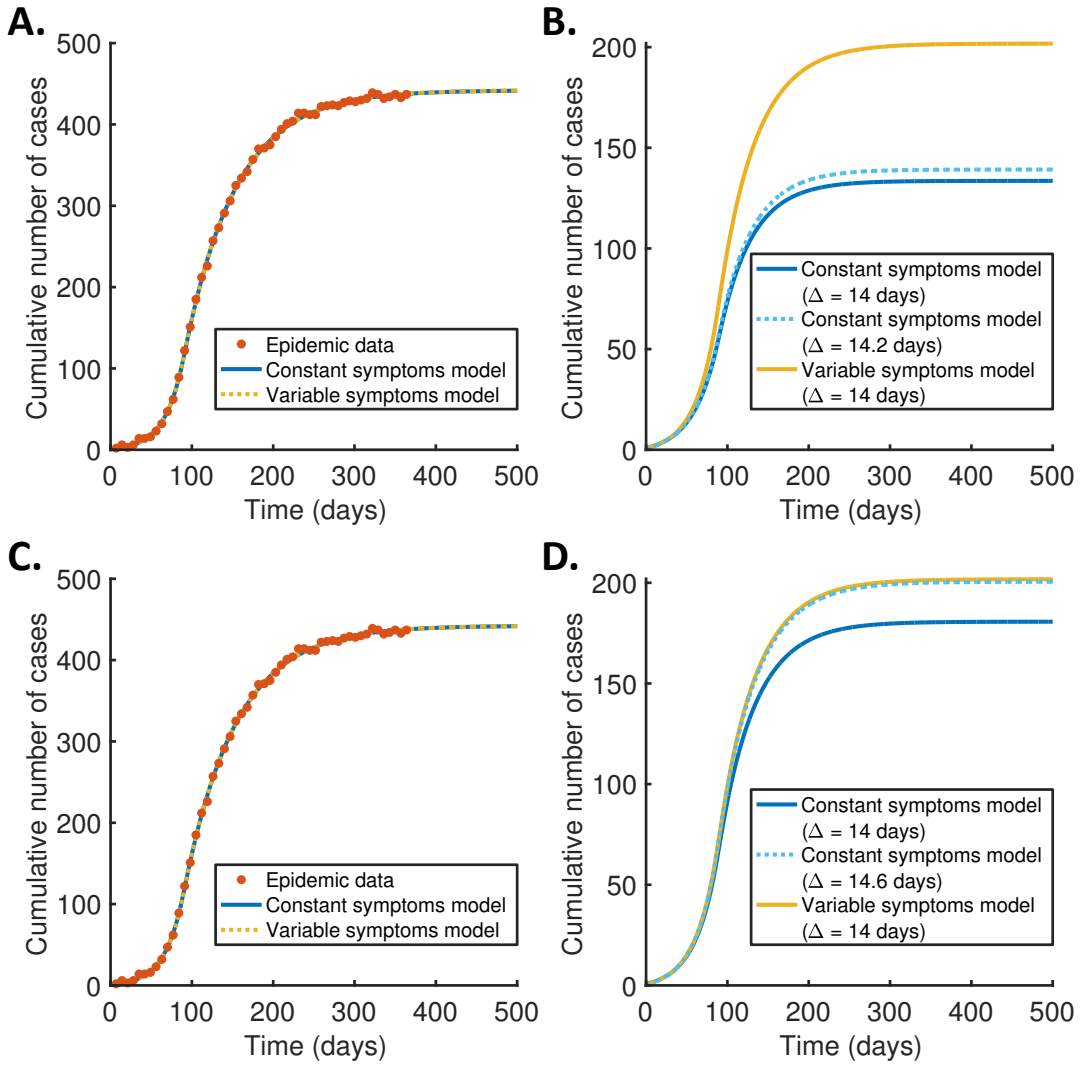


Figure R1.5: Effect of the choice of the detection probability in the constant symptoms model. A. Model fits to the simulated data (red dots), using the constant symptoms model (blue) and variable symptoms model (yellow dotted), assuming $p_1 = p_2 = p_3 = 0.66$ in the constant symptoms model in order to ensure an equal probability (under baseline surveillance) of ever being detected and isolated during infection, q_C, in the two models (see the text in Section R1.3.1.1; all other assumed parameter values are as in Figure R1.4). B. Corresponding projections of the effect of intensified surveillance, for the constant symptoms model (blue) and variable symptoms model (yellow) assuming surveillance once every $\Delta = 14$ days, and for the constant symptoms model assuming $\Delta = 14.2$ days (light blue dotted; this gives the same value of q_C as taking $\Delta = 14$ days in the variable symptoms model). C. Equivalent fits to the synthetic data, but instead assuming $p_1 = p_2 = p_3 = 0.52$ in the constant symptoms model in order to ensure an equal proportion of transmissions prevented through detection and isolation, ε_C, in the two models. D. Corresponding projections of the effect of intensified surveillance, for the constant symptoms model (blue) and variable symptoms model (yellow) assuming surveillance once every $\Delta = 14$ days, and for the constant symptoms model assuming $\Delta = 14.6$ days (light blue dotted; this gives the same value of ε_C as taking $\Delta = 14$ days in the variable symptoms model).

the constant symptoms model with the surveillance frequency altered to once every $\Delta = 14.2$ days (which gives the same value of q_C as the variable symptoms model with $\Delta = 14$ days).

As described in Section R1.3.1, we attributed the difference in projections between the two models in Figure R1.4B to the higher value of the proportion of transmissions prevented through detection and isolation, ε_C, in the constant symptoms model than in the variable symptoms model (ε_C is higher in the constant symptoms model because isolation typically occurs earlier in infection). In order to verify this is indeed the most important difference between the two models, we also conducted an analysis in which we set $p_1 = p_2 = p_3 = 0.52$ in the constant symptoms model in order to obtain an equal value of $\varepsilon_C = 18\%$ under baseline surveillance in each model (Figure R1.5C–D). In other words, in this analysis the effect of earlier detection in the constant symptoms model than in the variable symptoms model is offset by a lower mean detection probability per surveillance round.

We found a much smaller difference in projections between the two models when the surveillance frequency was increased to once every $\Delta = 14$ days in Figure R1.5D (note that ε_C still differs slightly between the two models under intensified surveillance in this case – $\varepsilon_C = 26\%$ in the constant symptoms model and $\varepsilon_C = 25\%$ in the variable symptoms model) than in Figure R1.4B. When we set $\Delta = 14.6$ days in the constant symptoms model and $\Delta = 14$ days in the variable symptoms model to ensure equal values of ε_C in the two models, projections from the two models were almost identical (Figure R1.5D). This result confirmed that the difference in ε_C between the two models is responsible for the difference in projections in Figure R1.4B.

R1.3.1.2 Effect of the extent of measurement error

To explore whether or not the error in projections due to incorrect model choice was significant compared to others sources of error, we explored the effect of the extent of the measurement error in the simulated data, σ, on our findings. Specifically, we calculated the extent to which the fitted constant symptoms and variable symptoms models underestimated the "true" total number of observed cases (i.e. the limiting value of $C + R$ at large times projected by the variable symptoms model using the "true" parameter values used to generate the simulated data) for a range of values of σ, under both the baseline and intensified surveillance regimes. We repeated this process over 1,000 simulated datasets for each value of σ to estimate the distribution of possible underestimates (Figure A1.1).

Before surveillance was intensified, the constant and variable symptoms models gave similar results at all values of σ, showing that either could approximate the true underlying epidemiological dynamics equally well (Figure A1.1A). However, following an intensification of surveillance, using the constant symptoms model led to systematic underestimation of the number of cases (Figure A1.1B). Even when the level of measurement error was large ($\sigma = 8$), the constant symptoms model led to very different projections of the effects of intensified surveillance compared to the variable symptoms model.

R1.3.1.3 Effect of the parameter values used to generate the synthetic data

We also considered the robustness of our results to the values of the parameters β_0, β_1 and T used to generate the synthetic data. Varying the parameters one-by-one, for each set of parameters we generated synthetic data using the variable symptoms model (for simplicity taking the measurement error to be zero in this analysis) and fitted the constant symptoms model to the data. In Figure A1.2 we plotted the underestimate in the total number of observed cases when the constant symptoms model was used to project the effect of intensified surveillance (compared to the corresponding projection using the variable symptoms model). For all parameter values considered, the constant symptoms model led to an underestimation of the true number of cases under intensified surveillance.

R1.3.2 Application to real-world Ebola data

To show that our main message – that accurate projections of the effects of interventions may require models that account for variations in symptoms during infection – is important for real world epidemics, we then repeated our analyses using data from two Ebola epidemics rather than using synthetic data.

We first considered data from the 2018–2020 Ebola epidemic in the DRC (Figure R1.6A). We again fitted the constant symptoms model and variable symptoms model to this dataset in turn, and found that both of these models could replicate the observed dynamics of the epidemic (Figure R1.6B; the R^2 value was 0.99862 for the constant symptoms model and 0.99861 for the variable symptoms model). Estimated parameter values for the model fits to the data from the DRC are given in Table A1.4. The estimated value of the basic reproduction number in the absence of interventions at the beginning of the outbreak ($R_0 = 3\beta_0/\mu$) was slightly smaller for the variable symptoms model ($R_0 = 1.8$) than for the constant symptoms model

($R_0 = 1.9$), demonstrating the well-known fact that the choice of model structure affects estimates of important epidemiological quantities [150, 151, 237, 239] (although we note that the reproduction number at the start of the epidemic, but accounting for surveillance, was very similar for the two models ($R_{0,eff} = 1.5$ for both models), with the difference in R_0 being attributable to the difference in the values of the proportion of transmissions prevented by isolation, ε_C, shown in Table A1.2).

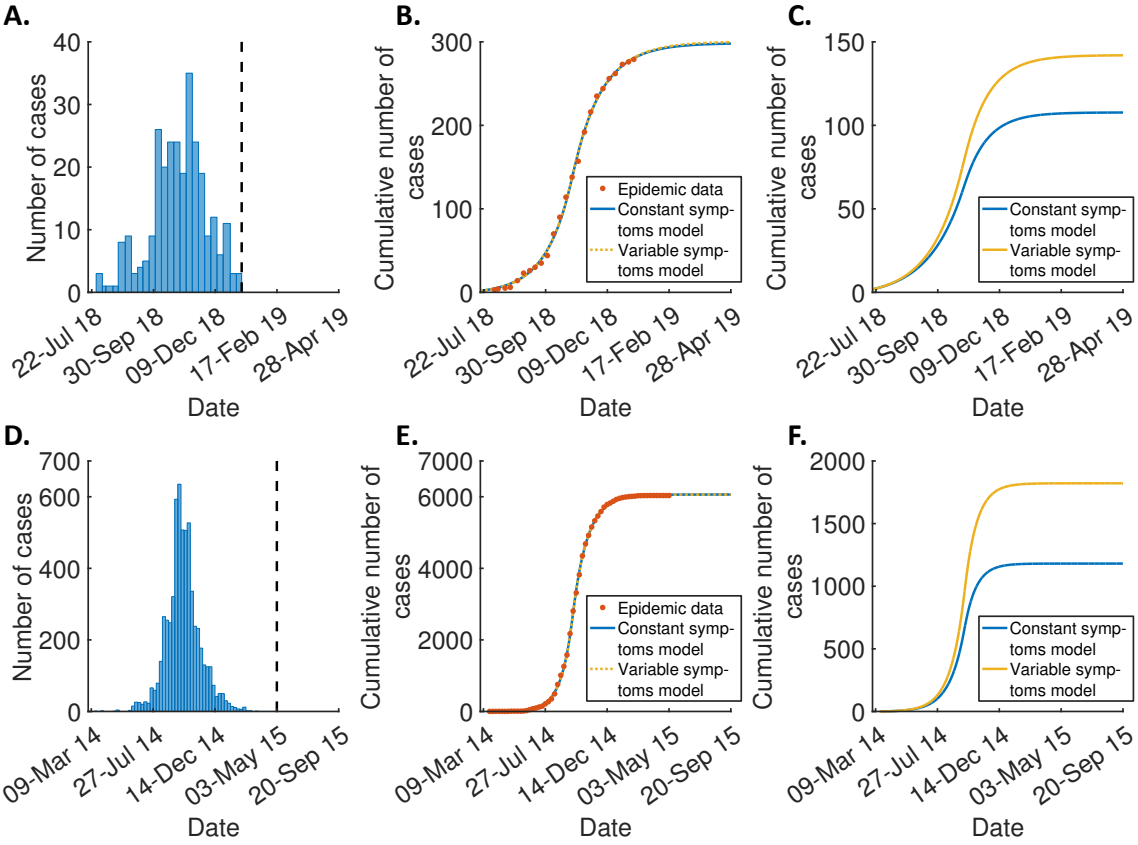


Figure R1.6: Using the constant symptoms model and variable symptoms model to project the effects of alternative interventions. A. The number of cases recorded on approximately weekly time intervals in the Beni and Kalunguta health zones in the DRC. The dashed black vertical line represents the time up to which data were available at the time of conducting our analyses. B. Model fits to the data (red dots), using the constant symptoms model (blue) and variable symptoms model (yellow dotted). C. Projections of the effect of intensified surveillance, using the constant symptoms model (blue) and variable symptoms model (yellow). D–F. Equivalent figures to A–C, using the data from the 2014–2016 Ebola epidemic in Liberia.

The fitted models were then used to project how the epidemic dynamics would have been altered under a different control intervention. In particular, we again increased the rate of detection in the fitted models, to represent projections under more intense surveillance and control efforts. When surveillance was intensified, projections using the constant symptoms model differed substantially from those using the variable

symptoms model (Figure R1.6C). In the case shown in Figure R1.6C, the constant symptoms model projected 24% fewer total cases than the more epidemiologically realistic variable symptoms model (108 cases in the constant symptoms model as opposed to 142 cases in the variable symptoms model). Consequently, even though the observed dynamics under the two models appear identical when fitted to data, the models produce different projections when control interventions are changed.

We then conducted the same analysis with data from the 2014–2016 Ebola epidemic in West Africa (Figure R1.6D–F; in this case, the R^2 value was 0.99989 for both models). Estimated parameter values for the model fits to the data from Liberia are given in Table A1.5. In this case, the constant symptoms model projected 35% fewer cases than the variable symptoms model under intensified surveillance (1181 cases in the constant symptoms model as opposed to 1822 cases in the variable symptoms model – a difference of 641 cases between the projections of the two models).

R1.3.3 Model extensions

To illustrate the principle that models with variable symptoms can generate different projections compared to models with constant symptoms, in most of our analyses we considered the simplest possible extension of the SEIR model in which the level of symptoms increases throughout infection. Here, to explore whether or not our findings still hold for more epidemiologically complex models, two extensions to our basic model are considered.

R1.3.3.1 Variable infectiousness during infection

In most of our analyses, we assumed that at any given time during the epidemic, all infectious and uncontrolled hosts transmitted the virus a constant rate (per susceptible individual). However, we also conducted an analysis in which the transmission rate varied throughout the course of an Ebola infection. Specifically, we considered scenarios in which infectiousness was either correlated with (Figure R1.7A) or correlated against (Figure R1.7D) the level of symptom expression (see Appendix A1.1.1 for details of the model with variable infectiousness). In both cases, we again fitted models assuming constant symptoms and variable symptoms to the epidemic datasets from the DRC and Liberia, obtaining similar fits (see Figure A1.3) to those shown in Figure R1.6B,E. However, when higher levels of symptoms were assumed to be associated with reduced infectiousness – for example due to a lower level of mixing in the population compared to hosts with less serious symptoms – our result that

projections are different between the constant symptoms and variable symptoms models was enhanced (Figure R1.7E,F).

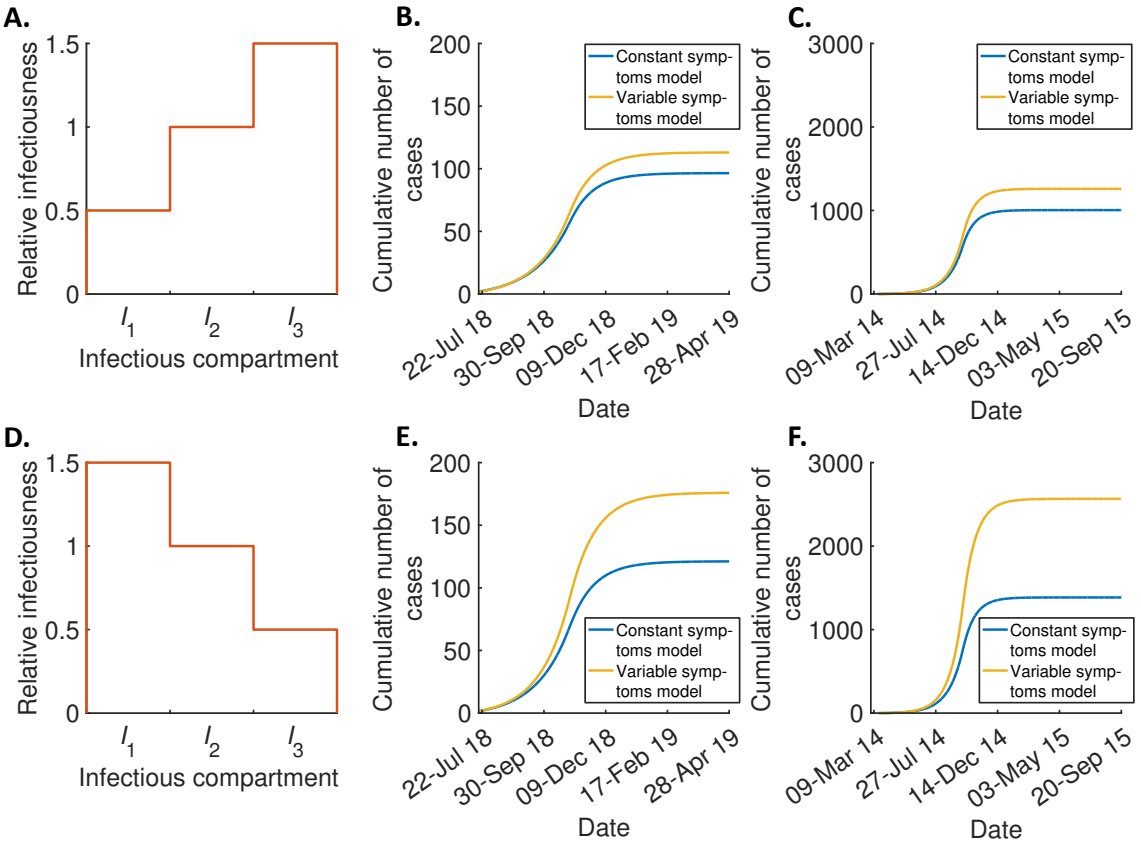


Figure R1.7: Results incorporating variable infectiousness during infection. A. The assumed relative infectiousness of infected individuals during each infectious stage, in a scenario where infectiousness increases during infection. B–C. Projections of the effect of intensified surveillance using the constant symptoms model (blue) and variable symptoms model (yellow), fitted to data from the 2018–2020 DRC epidemic (B) and the 2014–2016 Liberia epidemic (C), assuming the relative infectiousness values shown in A. Corresponding model fits are shown in Figure A1.3. D–F. Equivalent figures, but instead considering the possibility that infectiousness decreases during infection.

R1.3.3.2 Incorporating the possibility of recovery prior to the deterioration phase

We also considered a model extension in which, in the absence of surveillance, a proportion, q, of infected hosts recover during the gastrointestinal phase of Ebola infection – i.e. in the extended model, these individuals progress directly from the I_2 compartment to the R compartment, rather than progressing to the I_3 compartment (which represents the deterioration phase of infection). In the presence of surveillance, individuals who recover before detection were assumed to remain undetected (i.e. we assumed that such individuals do not contribute to recorded case counts). Details of

the extended model are given in Appendix A1.1.2. We considered values of q of 0.3, 0.5 and 0.7.

For each q value, the model fits to the two epidemic datasets were identical to those in Figure R1.6B,E when assessed by eye, and are therefore not shown. However, projected case numbers were always lower under intensified surveillance (Figure R1.8) than under the baseline surveillance level assumed when fitting the models to data, despite a higher proportion of infected hosts ever being detected under intensified surveillance than under baseline surveillance in the extended model (since when the surveillance frequency is increased, infected hosts are less likely to recover before being detected). In the case where the highest proportion of individuals recover rather than being detected ($q = 0.7$; Figure R1.8C,F), projections of the effect of intensified surveillance using the constant symptoms and variable symptoms models differ by more, compared to when a low proportion of individuals recover ($q = 0.3$; Figure R1.8A,D). Nonetheless, in every scenario that we considered, there was a substantial difference between the projected numbers of cases under the two models, supporting the finding that it may be necessary to include variable symptoms during infection in models of Ebola intervention strategies.

R1.3.4 Sensitivity analyses

R1.3.4.1 Effect of the timing of surveillance intensification

In the projections of the effect of intensified surveillance shown in Figure R1.6C,F, we assumed that the increased level of surveillance was in place throughout the epidemic (i.e. we explored what may have occurred under intensified surveillance during the whole outbreak). However, in outbreak response settings, an intensification of surveillance is likely to occur after the epidemic is underway. We therefore considered scenarios in which surveillance is instead intensified on different dates after the start of the epidemic (Figure R1.9).

If surveillance is intensified as soon as observed cases are first recorded (Figure R1.9A,D), then using the constant symptoms model rather than the more epidemiologically realistic variable symptoms model will lead to a larger error in projections, compared to a scenario of surveillance intensification later in the outbreak (Figure R1.9B,E). This is because early surveillance intensification allows more time for model projections to differ. The underestimate in the total number of observed cases when using the constant symptoms model (compared to the corresponding projection using the variable symptoms model) is plotted against the surveillance intensification date for the models fitted to the two epidemic datasets in Figure R1.9C,F.

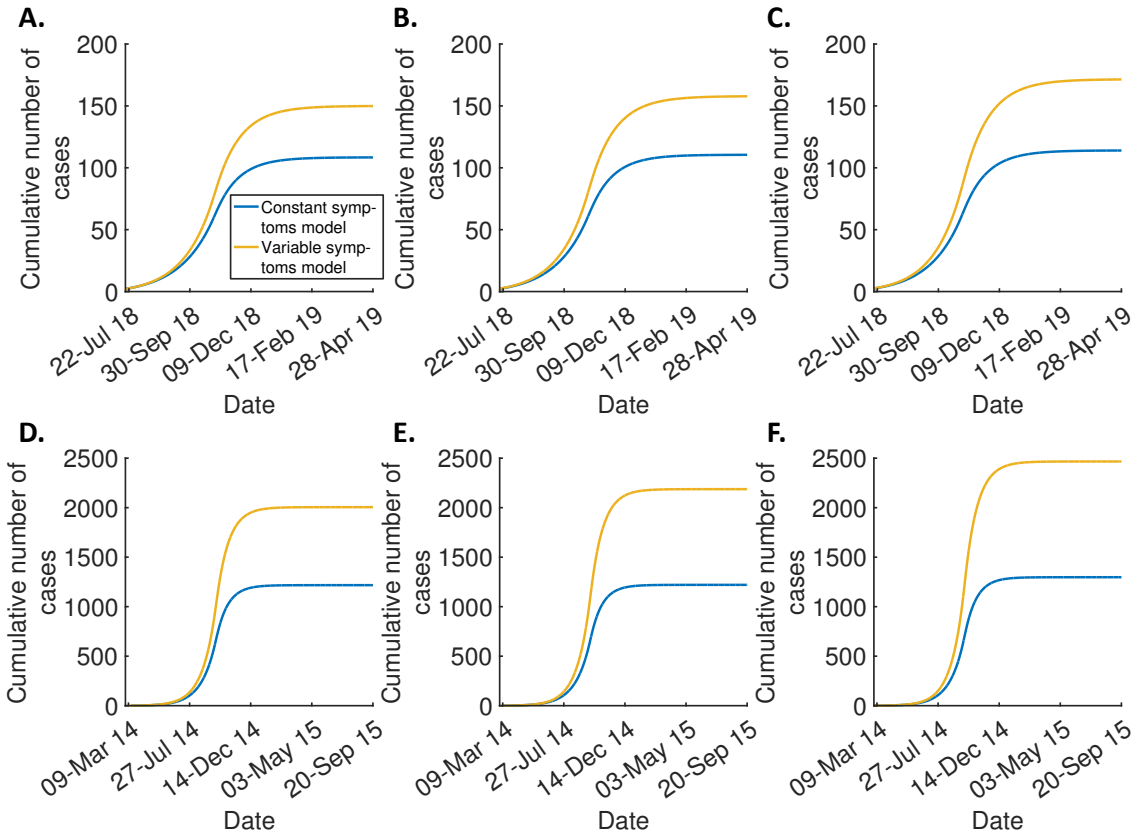


Figure R1.8: Results for a model in which some hosts recover prior to the deterioration phase of an Ebola infection. We considered a model in which a proportion q of infected hosts recover rather than passing into the I_3 compartment (see Appendix A1.1.2 for details of the extended model). A–C. The projected number of cases under intensified surveillance for $q = 0.3$ (A), $q = 0.5$ (B) and $q = 0.7$ (C), using the constant symptoms model (blue) and variable symptoms model (yellow) fitted to data from the 2018–2020 epidemic in the DRC (model fits to data are not shown). D–F. Equivalent figures to A–C, using the data from the 2014–2016 Ebola epidemic in Liberia.

R1.3.4.2 Effect of the surveillance frequency

We considered the robustness of our results to the level of surveillance assumed, both when fitting the models to epidemic data, and when altering the surveillance frequency in the fitted models to project the effects of changing interventions. The surveillance frequency is characterised by the parameter Δ, which represents the time interval between surveillance rounds. For ease of notation, here we denote the sampling interval assumed during model fitting as Δ_1 and the sampling interval under intervention testing as Δ_2.

Model fits to the data from the 2018–2020 Ebola epidemic in the DRC are shown for different values of Δ_1 in Figure A1.4A–C, and corresponding fits to the data from the 2014–2016 epidemic in Liberia are shown in Figure A1.4D–F – specifically, we considered values of $\Delta_1 = 7$ days, 14 days and 28 days. For each Δ_1 value, we found

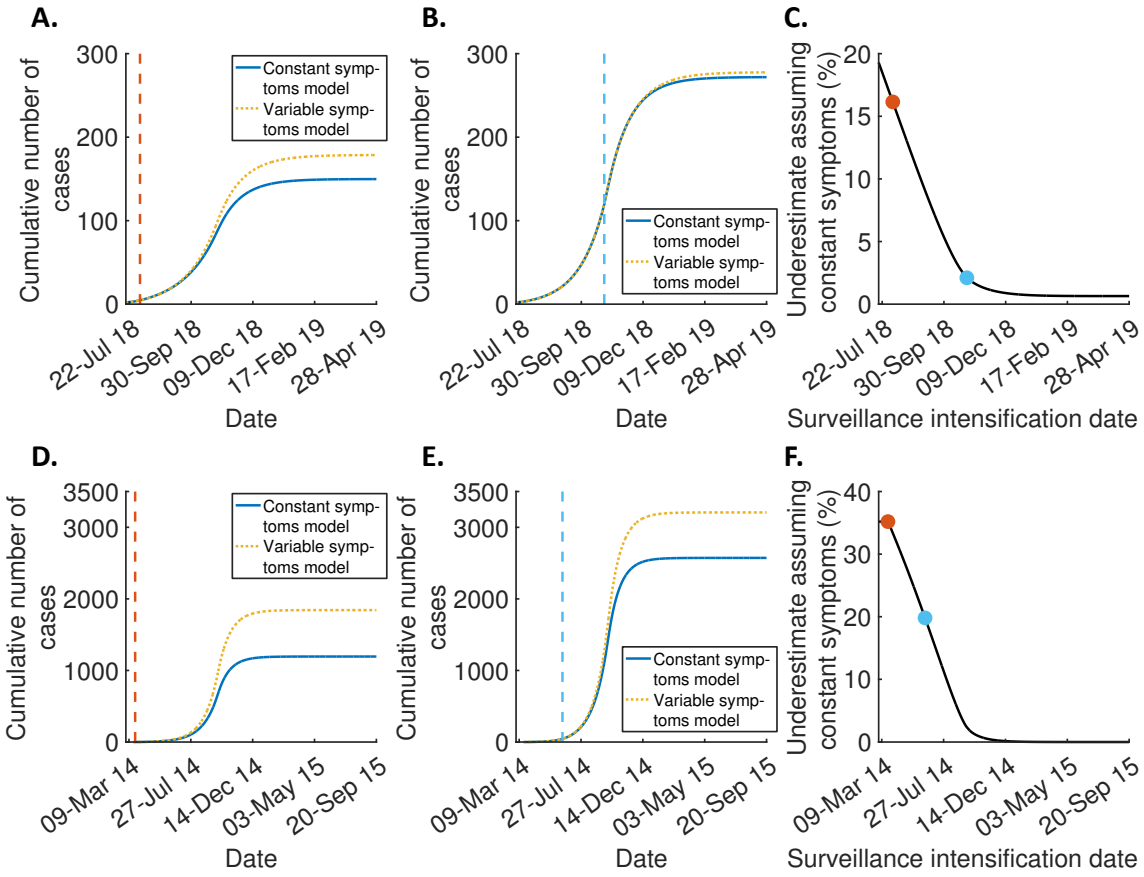


Figure R1.9: Effect of the timing of surveillance intensification. A–B. The projected number of cases, assuming surveillance intensification either on the date of the first case reports (A) or 12 weeks after this date (B), for the constant symptoms model (blue) and variable symptoms model (yellow dotted) fitted to data from the 2018–2020 epidemic in the DRC (note that surveillance was assumed to remain at its constant baseline level when fitting the models to data, so model fits are as shown in Figure R1.6B). The vertical dashed lines (red in A and light blue in B) represent the assumed dates of surveillance intensification. C. The underestimate in the total number of observed cases projected by the constant symptoms model compared to the variable symptoms model, expressed as a percentage of the number of cases projected by the variable symptoms model, for different times of surveillance intensification. The dots correspond to the times of surveillance intensification considered in A (red) and B (light blue). D–F. Equivalent panels to A–C, for the models fitted to data from the 2014–2016 Ebola epidemic in Liberia.

that both the constant symptoms and variable symptoms models could replicate the data closely.

In Figure R1.10 the relationship between the sampling frequencies (before and after the surveillance level is altered) and the difference in projections between the variable and constant symptoms models is summarised. For each value of Δ_1 considered, we calculated the percentage underestimate of the total number of observed cases under intensified surveillance by the constant symptoms model (compared to the corresponding estimate using the variable symptoms model) for

a range of values of Δ_2. When projections were made of the effect of intensifying surveillance (i.e. when $\Delta_2 < \Delta_1$), the constant symptoms model underestimated the total number of cases. Conversely, when surveillance was instead reduced (i.e. when $\Delta_2 > \Delta_1$) the constant symptoms model overestimated the total number of cases.

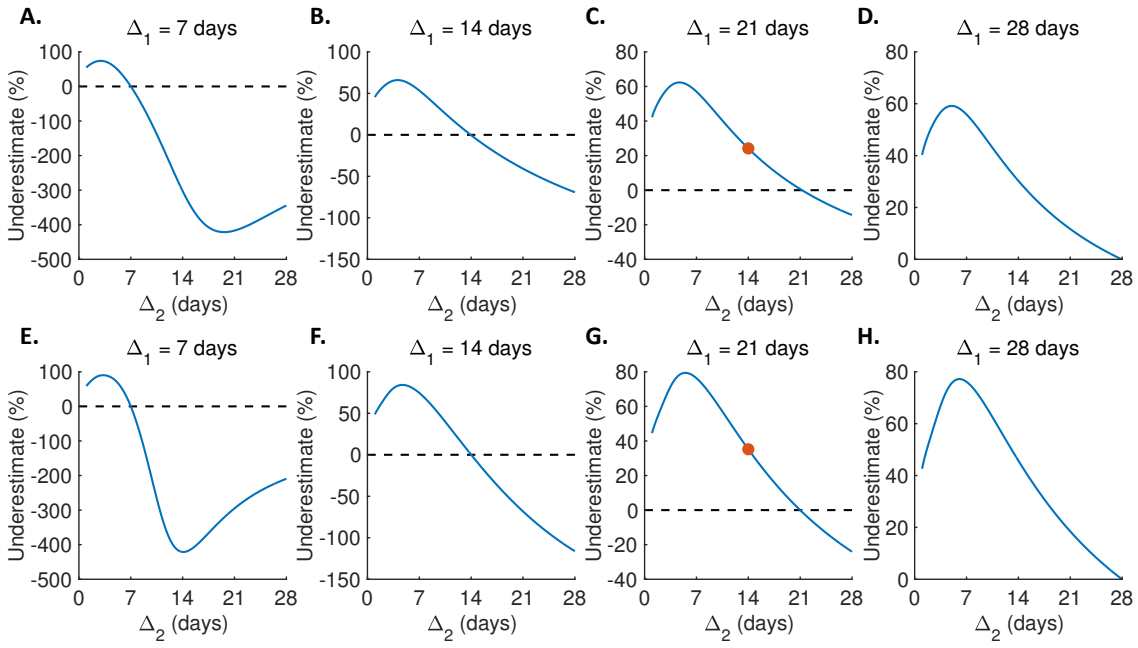


Figure R1.10: Effect of the surveillance frequency. A–D. The percentage underestimate in the projected total number of observed cases when the surveillance frequency is changed from Δ_1 (i.e. the value assumed when fitting the models to data) to Δ_2 if the constant symptoms model is used (compared to the projected number of cases using the variable symptoms model), for different values of Δ_2, when the two models were fitted to the data from the 2018–2020 epidemic in the DRC assuming $\Delta_1 = 7$ days (A), $\Delta_1 = 14$ days (B), $\Delta_1 = 21$ days (C) and $\Delta_1 = 28$ days (D). E–H. Equivalent panels showing underestimates when the models were fitted to the data from the 2014–2016 Ebola epidemic in Liberia. The red dots in C and G show the underestimates (24% and 35%, respectively) corresponding to Figure R1.6C,F.

R1.3.4.3 Effect of the detection probabilities

In addition, we investigated different assumptions regarding the level of symptoms in each stage of an Ebola infection by considering different values of the detection probability whenever surveillance occurs. In our main analyses, in the constant symptoms model we assumed the detection probabilities when an infected host is sampled to be $p_1 = p_2 = p_3 = 0.6$. In the variable symptoms model, we assumed the detection probabilities to be $p_1 = 0.1$, $p_2 = 0.8$ and $p_3 = 0.9$ in early, mid and late infection, respectively. The mean value of p_1, p_2 and p_3 was therefore the same in both models.

To explore the robustness of our results to the precise values of p_1, p_2 and p_3 used, we performed two separate analyses: first, we varied the value of the detection probability in the constant symptoms model (also varying p_2 in the variable symptoms model so that the mean detection probability in both models remained equal; Figure A1.5). Then, we altered the variation of p_1, p_2 and p_3 about the mean value in the variable symptoms model (Figure A1.6). Model fits for these analyses are not shown, since they are identical to those in Figure R1.6B,E when assessed by eye. In each case considered in Figures A1.5 and A1.6, the constant symptoms model led to an underestimation of the number of cases under intensified surveillance compared to the more realistic variable symptoms model.

R1.3.4.4 Alternative dependence of the detection rate on surveillance

In our main analyses, the detection rate, δ_i, of infected individuals in stage I_i, was assumed to depend on the interval between surveillance rounds, Δ, and the probability of detection in each surveillance round, p_i, according to the formula in equation (R1.17). This formula was derived under the assumption that surveillance occurs exactly every Δ days. In Figure A1.7, equivalent results to those in Figure R1.6B,C,E,F are shown under the alternative assumption of an exponentially distributed period between surveillance rounds (so that $\delta_i = p_i/\Delta$, as described in Section R1.2.3). Our main qualitative findings – that the constant and variable symptoms models provide very similar fits to epidemic data, but projections diverge when control is altered in the two fitted models – were not affected by the assumed dependence of the detection rate on surveillance.

R1.3.4.5 Model fits to non-cumulative data

In most of our analyses, we fitted the constant symptoms and variable symptoms models to data from two real-world Ebola epidemics. The data took the form of cumulative numbers of recorded cases at approximately weekly time intervals, and we performed the fitting using least-squares estimation. However, we also conducted an analysis to determine whether or not our main finding – that the constant and variable symptoms models project different effects of altered control interventions – is unchanged if we instead inferred the model parameter values by fitting to non-aggregated data (Figure R1.11). Details of how we fitted the models to non-cumulative data are given in Appendix A1.2.

The two models were able to fit the non-cumulative incidence data from the two outbreaks equally well (Figure R1.11A,D), but once again produced different projections of the effect of intensified surveillance (Figure R1.11C,F). For the models

fitted to data from the 2018–2020 outbreak in the DRC, the constant symptoms model projected 25% fewer total observed cases compared to the more realistic variable symptoms model when surveillance was intensified (Figure R1.11C), while for the models fitted to data from the 2014–2016 epidemic in Liberia, the constant symptoms model projected 35% fewer cases (Figure R1.11F).

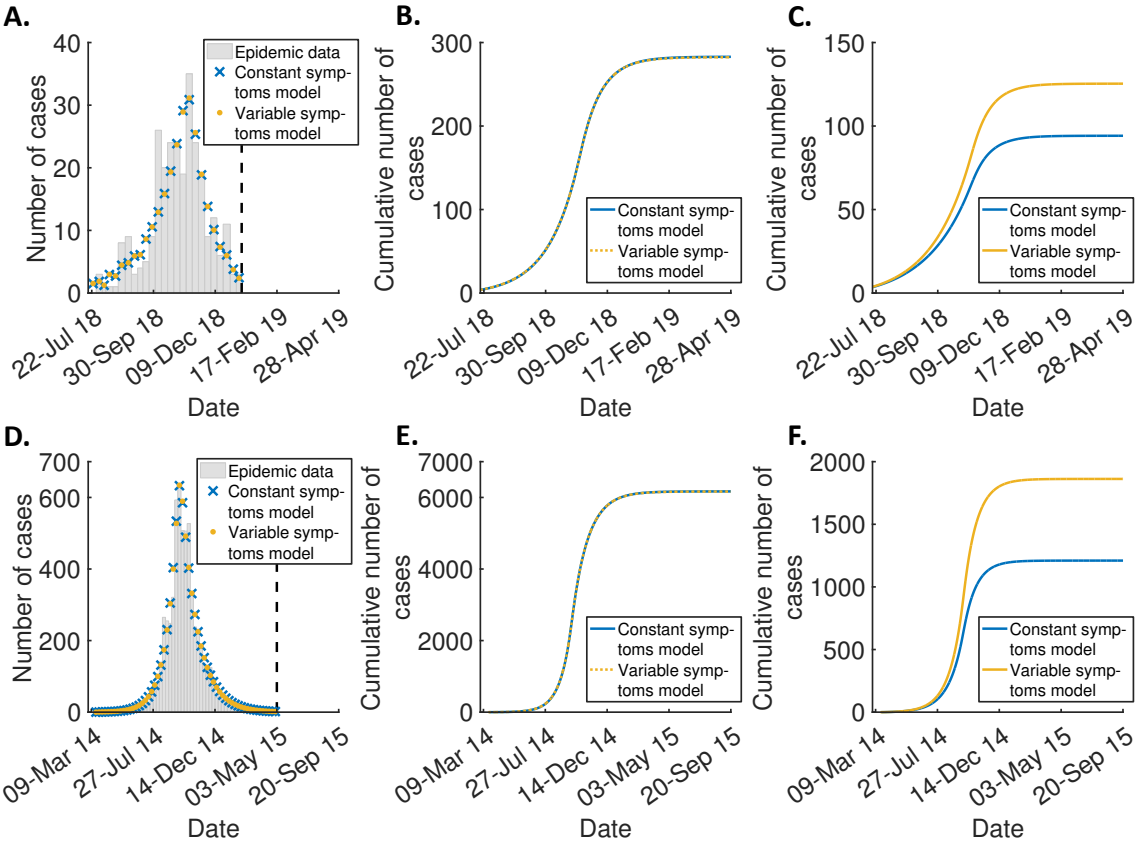


Figure R1.11: Using the constant symptoms model and variable symptoms model to project epidemic dynamics under intensified surveillance, when the models are fitted to non-cumulative incidence data. A. Non-cumulative data from the 2018–2020 Ebola outbreak in the DRC, consisting of the number of cases reported on approximately weekly intervals (grey bars), and model fits using the constant symptoms model (blue crosses) and variable symptoms model (yellow dots). Note that the fitted models project small numbers of cases before 22$^{\text{nd}}$ July 2018 (the lower x-axis limit in A). B. The cumulative number of cases under the fitted constant symptoms model (blue) and variable symptoms model (yellow). C. Projections of the effect of intensified surveillance using the constant symptoms model (blue) and variable symptoms model (yellow). D–F. Equivalent figures to A–C, using the data from the 2014–2016 Ebola epidemic in Liberia.

R1.4 Discussion

Epidemiological models that have been fitted to data are often used to project the dynamics of an epidemic under different control interventions [29, 64, 83, 140,

165]. However, most models considering interventions targeted at symptomatic hosts assume that those hosts display a constant level of symptoms that does not change during the course of infection. For many infectious diseases, however – including Ebola virus disease – there are different stages of an infection, and in each of these stages the level of symptoms is likely to be different. Individuals in early infection tend to have milder symptoms than those in later infection. As a result, symptomatic hosts in early infection are less likely to have appeared in the surveillance data that are routinely collected during an epidemic than those in later infection, who not only have had a longer period in which to be detected but are also likely to have developed more severe and recognisable symptoms.

Here, we have considered Ebola virus disease as a case study, and used two models to explore the dynamics of epidemics under different surveillance levels. We assumed that increased surveillance leads to improved detection and control of symptomatic hosts. We compared the output of a model in which, if an infected individual is surveyed, then the probability of detection is constant at each time during the infectious period (the constant symptoms model), to the equivalent projections from a model in which the probability of detection increases throughout infection (the variable symptoms model). We found that both these models can be fitted closely to data from Ebola epidemics (Figure R1.6B,E). However, when the level of surveillance in the models is increased, we found that the more epidemiologically realistic variable symptoms model projected a larger number of cases (e.g. Figure R1.6C,F). This is attributable to the fact that, while more frequent surveillance was always found to reduce transmission through improved case detection and isolation, detection typically occurs later in infection in the more realistic variable symptoms model, by which time fewer transmissions can be prevented through isolation. This finding is concerning, because it suggests that using the epidemiologically unrealistic constant symptoms model, which is of the type most commonly used in practice, may lead to an overestimation of the likely efficacy of additional interventions. Thus, it might be important to use the more realistic model to project the quantitative effects of interventions aimed at reducing the impacts of Ebola epidemics.

Variations in symptoms between different stages of infection, as well as the signature of such variable symptoms in those types of data that are collected during an epidemic, have to date received little attention. Until now, the impact of variable symptoms on projections of models used for testing Ebola interventions has never been assessed rigorously. However, our approach of splitting the infectious and symptomatic period into different sub-compartments was inspired by the so-called "method of stages" [152, 237], a technique most often used to model gamma-distributed

epidemiological periods [11, 157, 221]. Within that framework, varying infectiousness – rather than symptoms – over the course of infection has been considered previously [58, 63, 121]. In Chapter R2 we will show that if a large number of compartments is used in such an approach, then a model in which the (expected) infectiousness profile of infected hosts follows a user-specified continuous curve can be obtained. A similar modelling framework could be adopted here so that the level of symptoms can be specified as a continuous curve. We did not pursue such an approach in this chapter, since discrete changes in symptom expression in each symptomatic host are sufficient to make our underlying point that accurate projections of the effects of Ebola interventions may require models that account for variations in symptoms during infection. The general conclusion here that it may be necessary to account for individual host-scale dynamics (such as varying symptoms during infection) in population-scale models did, however, motivate us to develop the more detailed mathematical framework for connecting individual host-scale dynamics and population-scale epidemic dynamics in Chapter R2 (albeit with a focus on varying infectiousness rather than symptoms in that chapter).

To conduct our analyses, we sought to develop a simple model in which the level of symptoms increases during infection, and to compare the results from this model to those from the analogous model in which there is a constant level of symptoms during infection. Practical use of either model during an Ebola epidemic would require adjustment for the particular epidemic under consideration. For example, transmission in different settings could be included in a single model, such as spread in hospitals, community care centres, at funerals or in the wider community [33, 163]. We modelled detection and isolation of symptomatic hosts here, but other interventions such as vaccination could be modelled explicitly [87]. If an Ebola vaccine is not perfectly effective (this was contentious for the vaccine that was used in the 2018–2020 Ebola epidemic [119, 153, 166]), the possibility that vaccination might reduce symptoms while not completely stopping infectiousness could be included in our approach. A model that includes spatial spread of the pathogen or transmission through social contact networks might be required to replicate observed data [136, 139], or different geographical areas could be considered separately [163, 241]. To demonstrate that variable symptoms can affect projections of the effects of interventions, we assumed that all infected individuals pass through three stages of an Ebola infection (from non-specific symptoms to a gastrointestinal phase and then to a deterioration phase, represented in Figure R1.2 by early, middle and late infection), whereas in reality some hosts might recover rather than passing to the deterioration phase [20]. At the cost of an additional parameter value to be determined, it is

straightforward to include this additional detail in a compartmental epidemiological model. We demonstrated this extension to our approach in Figure R1.8, showing that our main finding – that accurate projections of the effectiveness of interventions against Ebola may require models that account for variations in symptoms during infection – is robust to the inclusion of this increased epidemiological complexity.

We modelled surveillance in a simple fashion, by assuming that infected hosts are surveyed on average at periodic intervals and that there is a particular probability of detection whenever a host is surveyed. For real-world epidemic projections, it may be necessary to consider the wide range of different surveillance approaches used in practice, including contact tracing from known cases [186] and rural village visitations to detect cases in locations where access to healthcare is limited [176], as well as disruptions to surveillance caused by factors including armed conflict [102].

We parameterised our models using the most straightforward possible approach – namely using least-squares estimation to fit the total numbers of detected or removed individuals in the relevant model compartments to data on the cumulative numbers of symptomatic cases – since the precise method of parameter inference was not central to our message. This type of approach has been widely used previously [54, 57, 202, 206, 238], although we note that fitting stochastic models to non-cumulative incidence data is usually preferable [104, 135, 156, 227], particularly when assessing uncertainty in parameter estimates and forward projections. However, we did not seek to quantify uncertainty in the estimated values of model parameters here. Nonetheless, our main finding – that it may be necessary to include variable symptoms in models of Ebola interventions – was unchanged in a sensitivity analysis in which model fitting was conducted using non-cumulative data (Figure R1.11). We expect this result to hold true whether deterministic or stochastic models are fitted to raw or cumulative data, and irrespective of the exact method of parameter inference used. However, extending our approach to a stochastic setting for practical applications, including quantification of uncertainty in parameter estimates and projections, is an area for future work. Different assumptions regarding the types and forms of noise in observations, driven by factors including stochasticity in the underlying epidemiological dynamics as well as errors in reporting, could also be considered. In our analyses using synthetic data, we added noise to the datasets by assuming that reporting errors followed a normal distribution, but alternative error structures could be explored in future studies.

One advantage of the models that we used is that the surveillance level is assumed to affect the epidemiological dynamics themselves, rather than simply the observed dynamics. This is not always the case in epidemiological models: a common method

for accounting for under-reporting is simply to scale the incidence data up [164], thereby assuming a fixed percentage of infectious cases are detected with no impact on the numbers of cases generated by those individuals. Another approach is to assume that some individuals in the infectious stage of a compartmental model are unobserved [225]. In reality, detected hosts have a lower probability of transmitting the pathogen than undetected hosts due to the higher chance that detected individuals are subject to interventions, and our models reflect this. As a result, we always projected a lower total number of detected cases under intensified surveillance than under the assumed baseline surveillance level, since more frequent surveillance led to improved epidemic control. This was true even when we accounted for the possibility of individuals recovering without being detected (so that in this sensitivity analysis, intensified surveillance increased the proportion of cases that were detected; Figure R1.8).

In this work, we considered a control strategy of detection and isolation under different surveillance levels. The effects of including variable symptoms in models of other intervention strategies should be tested, to determine whether or not it is always necessary to account for changing levels of symptoms throughout infection. We note that including additional epidemiological detail in models does not always improve projections [215]. Simple models are easier to parameterise and interpret than more complex models, and so modellers should consider carefully, in each study, whether or not including variable symptoms will change model projections. We also note that, in theory, it might be possible to deploy commonly used epidemiological models with altered parameter values as a proxy for explicit consideration of variable symptoms [90]. For example, if the chance of detection in early infection is low, then the period of early non-specific symptoms could be considered as part of the incubation period. In that case, care should be taken when "lifting" the values of model parameters directly from the clinical literature, to ensure that the definitions of parameters in the model match those in the original studies.

In summary, in this chapter we have shown that including different levels of symptoms at different stages of an Ebola infection in epidemiological models can alter projections of the effects of intervention strategies compared to assuming a fixed level of symptoms. If variations in symptoms during infection – and their impacts on detectability – can be characterised by epidemiologists and then included in mathematical models, then decision makers will be able to make more informed choices as to which particular intervention, or combination of interventions, they should pursue.

Chapter R2

A theoretical framework for transitioning from host-scale to population-scale epidemiological dynamics: influenza A as a case study

We have published the research underlying this chapter in the *Journal of the Royal Society Interface* [113].

R2.1 Introduction

Population-scale epidemiological models are increasingly used to project the possible future dynamics of outbreaks [89, 220, 241] and to plan interventions [32, 82, 83, 222], while within-host models are used to understand the spread of infection at the individual host-scale [17, 36, 193, 252]. Standard population-scale models assume that the infectiousness of each infected host is constant over the course of the infectious period [131], but in reality infectiousness will vary as the infection progresses through the host due to changing pathogen loads [111, 167] and other factors including behavioural responses to infection [111].

Multi-scale models have been used to connect epidemiological dynamics at the host-scale (within-host) to those at the population-scale (between-host) [9, 25, 51, 61, 63, 73, 92, 95, 108, 144, 154, 155, 181, 249]. These models (sometimes referred to as nested models in the evolutionary epidemiology literature [73, 167]) tend to assume a specified relationship between the level of infection within a host and the rate at which the host transmits the pathogen to susceptible individuals [111, 167, 174]. A within-host model, parameterised by fitting to host data, can then be used to

determine the parameters of a population-scale model incorporating time-dependent infectiousness [167, 174]. In addition to host-scale dynamics affecting population-scale transmission, there may be reciprocal feedback from the population-scale to the host-scale [167] – for example, if there are multiple co-circulating strains of the pathogen [155].

In 2019, a review concluded that, while numerous multi-scale epidemiological modelling studies exist, relatively few include substantial use of data [51]. While one reason for this is the lack of widely available datasets [7, 111], we contend that another contributing factor is that previous multi-scale modelling frameworks have been complex, making them challenging to implement other than by highly specialist mathematical modellers. Such frameworks have often employed IDEs [61, 73, 92, 108, 155, 158, 167, 174], although alternatives such as individual-based stochastic models [49, 82, 154, 181, 249] have also been considered. IDEs are challenging to solve, requiring bespoke numerical methods [63]. Some studies using IDEs have involved explicit simulation of the full multi-scale model [92, 155, 158]. However, other studies have either only used the multi-scale framework to derive quantities such as the basic reproduction number of the pathogen rather than projecting temporal epidemic dynamics [73, 108], or have made simplifying assumptions such as taking a within-host model to be in equilibrium [9, 25, 61, 95]. Although an assumption that the pathogen load in each infected host is not changing (or changes only a limited number of times) might be appropriate for chronic infections, it leads to an approximate population-scale model that does not explicitly account for time-dependent infectiousness or other potentially complex host-scale dynamics.

In most previous studies that have used IDEs to transition from host- to population-scale, the progression of infection through all hosts has been assumed to be identical [174]. Host-scale dynamics are therefore characterised by a within-host model in which the values of model parameters (describing factors such as pathogen replication as well as immune responses) are the same for all hosts. These parameters have either assumed values that have not been derived rigorously from data [51], or have been obtained by fitting the model to data collected in longitudinal cohort studies from a small number of hosts [61, 108]. Within-host model parameters are, in fact, likely to vary between individuals [17, 35, 233], for example due to differences in immune responses [179], while measurement error may lead to inaccurate parameter estimates particularly given limited numbers of observations [174, 180]. As we show here, if host-scale data are only available from a limited number of hosts, then projections of population-scale epidemic dynamics using a multi-scale modelling approach may be inaccurate.

In this research chapter, we introduce a novel framework for transitioning from host- to population-scale epidemiological dynamics straightforwardly. Our method involves using a compartmental population-scale model with a large number of infected compartments. Compartmental models, comprising systems of ODEs, can be solved easily using standard numerical routines and software packages [63, 91, 123, 131], are straightforward to adapt to include further biological detail [63, 112, 131], and are widely used for epidemic modelling [131, 203]. We show rigorously that our modelling framework is equivalent to a more complex IDE approach, in the large-compartment limit of our method. Since the number of compartments is simply a choice for the user to make, our easy-to-use method can generate results that are as accurate as those from more complex approaches.

To demonstrate our framework, we consider modelling an outbreak due to the influenza A virus. We use a previously parameterised within-host model [17] to generate a synthetic dataset representative of real host-scale data, incorporating variability in the viral load time series between hosts (which may arise in reality due to factors such as differences in immune responses). Since the magnitude of this variability has been chosen to match data from a previous cohort study [17, 175], our dataset is comparable to obtaining data from a cohort study, but with the advantage that we can test our approach using many different possible cohorts of any size (from small cohort sizes up to very large cohort sizes that generate idealised data). We explore the effects of both the number of hosts from which data are available, and the extent of measurement error in host-scale data, on population-scale projections. Our work therefore provides insight into the data required for accurate projections using multi-scale epidemic models, as well as an accessible modelling framework that can be used for obtaining projections during future epidemics of a range of infectious diseases.

R2.2 Methods

R2.2.1 The SI_nR model

The population-scale SI_nR compartmental model [8, 63] of pathogen transmission in a population of N individuals is given by

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -S\sum_{j=1}^{n}\beta_j I_j, \tag{R2.1}$$

$$\frac{\mathrm{d}I_1}{\mathrm{d}t} = S\sum_{j=1}^{n}\beta_j I_j - \mu_1 I_1, \tag{R2.2}$$

$$\frac{\mathrm{d}I_i}{\mathrm{d}t} = \mu_{i-1}I_{i-1} - \mu_i I_i, \quad \text{for } i = 2, \ldots, n, \tag{R2.3}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \mu_n I_n. \tag{R2.4}$$

In this model, infected individuals sequentially pass through n infected compartments before recovering (or dying). The variable $S(t)$ represents the number of susceptible individuals, $I_i(t)$ the number of individuals in the i^{th} infected compartment, and $R(t)$ the number of recovered individuals, where t is the time since the start of the outbreak. The transmission rate parameter of individuals in the i^{th} infected compartment is β_i, and the mean duration of the i^{th} infected stage is $1/\mu_i$. The basic reproduction number of this model is [8]

$$R_0 = N \sum_{i=1}^{n} \frac{\beta_i}{\mu_i}. \tag{R2.5}$$

The equation (R2.4) for

$$R = N - S - \sum_{i=1}^{n} I_i \tag{R2.6}$$

decouples from equations (R2.1)–(R2.3), and is ignored in the remainder of this chapter.

Note that in this model formulation, the possibility of "latent compartments" (i.e. stages of infection during which infected hosts are not yet infectious) is implicitly included (since any transmission rate, β_j, can be zero). We chose not to explicitly distinguish "latent" and "infectious" compartments here (as we did in Chapter R1) for notational convenience.

R2.2.2 The K&M model

The K&M model [133] is an IDE epidemic model, in which the infectiousness of each infected host varies continuously in time. We considered a form of the K&M model given by

$$\frac{\mathrm{d}S}{\mathrm{d}t} = S(t)\left[\int_0^t \beta(\tau)\frac{\mathrm{d}S(t-\tau)}{\mathrm{d}t}\mathrm{d}\tau - \int_t^\infty \beta(\tau)g(\tau - t)\mathrm{d}\tau\right], \tag{R2.7}$$

where $S(t)$ is the number of susceptible individuals at time t since the start of the outbreak, $\beta(\tau)$ is the expected infectiousness profile (the expected transmission rate of an infected individual at time τ since infection; note that in this chapter, we neglect any dependency of the expected infectiousness profile on t as well as τ), and $g(\tau)$ represents the initial density of individuals who have been infected for time τ. The basic reproduction number of this model is [77]

$$R_0 = N \int_0^\infty \beta(\tau)\mathrm{d}\tau, \tag{R2.8}$$

where N is the (constant) total population size. Latent periods, as well as any distribution of latent and infectious periods between hosts, can be incorporated into the expected infectiousness profile, $\beta(\tau)$.

If there is initially a single newly infected individual, with the remainder of the population susceptible, then $g(\tau) = \delta(\tau)$, where $\delta(\tau)$ is the Dirac delta function. In this case, the K&M model reduces to

$$\frac{\mathrm{d}S}{\mathrm{d}t} = S(t)\left[\int_0^t \beta(\tau)\frac{\mathrm{d}S(t-\tau)}{\mathrm{d}t}\mathrm{d}\tau - \beta(t)\right]. \tag{R2.9}$$

The $\mathrm{SI}_n\mathrm{R}$ model is obtained as a special case of the K&M model [63, 76]. The corresponding infectiousness profile, $\beta(\tau)$, is the expected transmission rate of an individual at time τ since infection in the analogous stochastic $\mathrm{SI}_n\mathrm{R}$ model. In this case, the expected infectiousness profile is non-zero for all times since infection $\tau > 0$, even if latent compartments are included in the $\mathrm{SI}_n\mathrm{R}$ model. This is because the time that a given host spends in each infected compartment is exponentially distributed, and so may be arbitrarily short or long.

To solve the K&M model numerically, we used a forward Euler method. Specifically, we chose a time-step δt, and approximated the solution of the K&M model (in the form given by equation (R2.9)) using the finite-difference scheme

$$S_{i+1} = S_i\left[1 + \delta t\left(\sum_{j=1}^{i}\beta_j(S_{i+1-j} - S_{i-j}) - \beta_i\right)\right], \tag{R2.10}$$

where i is a non-negative integer, $t_i = i\delta t$, $S_i = S(t_i)$ and $\beta_i = \beta(t_i)$. The time-step was chosen to be sufficiently small to ensure that the error in the numerical solution was negligible compared to other sources of error in projections of population-scale dynamics that we considered in our analyses.

R2.2.3 Transitioning from host- to population-scale dynamics

We considered both an existing IDE approach (steps A–C below) and a new compartmental framework (steps A–B and D–E below) to transition from host- to population-scale dynamics. The methodological approach we used involved the following steps (a schematic is shown in Figure R2.1):

A. Fit a within-host model to longitudinally sampled data on host-scale dynamics, to estimate the pathogen load of each individual host at every time since infection.

B. Estimate the expected infectiousness profile, $\beta(\tau)$, at each time since infection, τ, by assuming that the infectiousness of each host depends on the pathogen load according to a pre-specified relationship between these quantities.

Then either C:

C. Solve the K&M IDE model, with infectiousness profile $\beta(\tau)$, to calculate the population-scale dynamics.

Or D–E:

D. Parameterise the $\mathrm{SI}_n\mathrm{R}$ model: choose the number of infected compartments, n, where n is assumed to be large. Then choose T such that $\beta(\tau)$ is zero or very small for $\tau > T$, and choose the parameters in the $\mathrm{SI}_n\mathrm{R}$ model to be

$$\mu_i = \frac{n}{T}, \tag{R2.11}$$

$$\beta_i = \frac{n}{T}\int_{(i-1)T/n}^{iT/n} \beta(\tau)\,\mathrm{d}\tau, \tag{R2.12}$$

for $i = 1, \ldots, n$. Explanation of these parameter choices is given below.

E. Solve the $\mathrm{SI}_n\mathrm{R}$ model numerically to approximate the population-scale dynamics.

R2.2.3.1 Choice of parameter values in the $\mathrm{SI}_n\mathrm{R}$ model

Here, we provide intuitive justification for the choice of parameter values in the $\mathrm{SI}_n\mathrm{R}$ model in our compartmental framework for transitioning from host- to population-scale dynamics (equations (R2.11) and (R2.12)). A rigorous proof that, under these parameter choices, the compartmental method is identical to the IDE method in the large-compartment limit is given in Section R2.3.2.

We took the rates of progression through the infected compartments in the $\mathrm{SI}_n\mathrm{R}$ model, μ_i, to all have the same value (where this value is dependent on the number of compartments), so that

$$\mu_1 = \mu_2 = \cdots = \mu_n = \frac{1}{\lambda(n)}, \tag{R2.13}$$

where $\lambda(n)$ is the mean time that a given host spends in each infected compartment (for a given number of compartments, n). We chose $\lambda(n)$ so that $\lambda(n) \to 0$ as $n \to \infty$, in order to ensure that the infectiousness of each host varies continuously in the limit $n \to \infty$ (the explicit choice of $\lambda(n)$ that we used is given below).

To choose the transmission rates, β_i, we considered an individual who spends exactly the mean time, $\lambda(n)$, in each infected compartment. At time τ since infection, such an individual will have infectiousness given by

$$\beta(\tau;n) = \begin{cases} \beta_1, & \text{for } 0 \leq \tau < \lambda(n), \\ \beta_2, & \text{for } \lambda(n) \leq \tau < 2\lambda(n), \\ \vdots & \\ \beta_n, & \text{for } (n-1)\lambda(n) \leq \tau < n\lambda(n), \\ 0, & \text{for } \tau \geq n\lambda(n). \end{cases} \tag{R2.14}$$

We chose the values of the parameters β_i so that $\beta(\tau;n) \to \beta(\tau)$ as $n \to \infty$ with $\lambda(n) \to 0$. This ensured that the individual infectiousness profile of a host who spends exactly the mean time in each infected compartment (as given in equation (R2.14)) approaches $\beta(\tau)$ in the limit $n \to \infty$. Therefore, since infectiousness also varies continuously when $n \to \infty$, each infected host has an individual infectiousness profile given exactly by $\beta(\tau)$ in the limit. In general, $\beta(\tau)$ may have unbounded support, in which case $\lambda(n)$ must be chosen so that $n\lambda(n) \to \infty$ as $n \to \infty$ (for example, $\lambda(n) \propto n^{-1/2}$ is one possible choice). However, in practice we may expect there to be some time since infection, T, such that $\beta(\tau)$ is zero (or very small) for $\tau > T$, in which case it is possible to fix $n\lambda(n) = T$ (i.e. we can take $\lambda(n) = T/n$).

Specifically, we chose the transmission rates to be

$$\beta_i = \frac{1}{\lambda(n)} \int_{(i-1)\lambda(n)}^{i\lambda(n)} \beta(\tau)\mathrm{d}\tau, \qquad i = 1, \ldots, n-1, \tag{R2.15}$$

$$\beta_n = \frac{1}{\lambda(n)} \int_{(n-1)\lambda(n)}^{\infty} \beta(\tau)\mathrm{d}\tau, \tag{R2.16}$$

so that β_i is the average value of $\beta(\tau)$ between times $(i-1)\lambda(n)$ and $i\lambda(n)$ since infection, and the choice of β_n ensures that (for each n) the $\mathrm{SI}_n\mathrm{R}$ model has the same basic reproduction number as the K&M model with infectiousness profile $\beta(\tau)$, although other exact choices of transmission rates are also possible.

If there exists a T such that $\beta(\tau)$ is zero or very small for $\tau > T$, so that it is possible to take $\lambda(n) = T/n$, then the choices of values of μ_i (equation (R2.13)) and β_i (equations (R2.15) and (R2.16)) described here reduce to the expressions given in the summary of our compartmental framework above (equations (R2.11) and (R2.12)). Except in our proof of the equivalence of the compartmental and IDE multi-scale frameworks in the large-compartment limit of the compartmental approach in Section R2.3.2 (where, for generality, we accounted for the possibility that $\beta(\tau)$ has unbounded support), we assumed the existence of such a T in our analyses.

R2.2.4 Influenza case study

R2.2.4.1 Host-scale dynamics

The TCL model of viral dynamics, which has previously been used to model influenza infection [17, 36, 207], is given by

$$\frac{\mathrm{d}T}{\mathrm{d}\tau} = -\beta TV, \tag{R2.17}$$

$$\frac{\mathrm{d}I}{\mathrm{d}\tau} = \beta TV - \delta I, \tag{R2.18}$$

$$\frac{\mathrm{d}V}{\mathrm{d}\tau} = pI - cV, \tag{R2.19}$$

where $T(\tau)$ is the number of susceptible target cells, $I(\tau)$ is the number of infected target cells, $V(\tau)$ is the quantity of free virus (in units of TCID_{50}/ml), and τ is the time since infection. Interpretations of model parameters are given in Table R2.1. The model has previously been parameterised for influenza A infection in humans [17] (Table R2.1).

Parameter	**Interpretation**	**Assumed value**
β	Infection rate of susceptible cells by virus	2.7×10^{-5} $(\mathrm{TCID}_{50}/\mathrm{ml})^{-1}$ day^{-1}
δ	Death rate of infected cells	4.0 day^{-1}
p	Viral shedding rate by infected cells	1.2×10^{-2} $(\mathrm{TCID}_{50}/\mathrm{ml})$ day^{-1}
c	Clearance rate of free virus	3.0 day^{-1}
$T(0)$	Initial number of susceptible cells	4×10^8
$I(0)$	Initial number of infected cells	0
$V(0)$	Initial quantity of free virus	9.3×10^{-2} TCID_{50}/ml

Table R2.1: Parameter values and initial conditions for the TCL within-host model. These parameter values were previously estimated for influenza A in humans [17].

We used the TCL within-host model to generate synthetic data from different hosts. To incorporate variability between hosts, we assumed that the parameters δ and $V(0)$ in the TCL model vary between individuals (representing variation in the strength of the immune response and in the initial viral load). For each host, $\log_{10}(\delta)$ was sampled from a normal distribution with mean 0.60 $\log_{10}\left(\mathrm{day}^{-1}\right)$ and standard deviation 0.25 $\log_{10}\left(\mathrm{day}^{-1}\right)$, and $\log_{10}(V(0))$ was sampled from a normal distribution with mean -1.03 $\log_{10}(\mathrm{TCID}_{50}/\mathrm{ml})$ and standard deviation 1.12 $\log_{10}(\mathrm{TCID}_{50}/\mathrm{ml})$.

These values were chosen to match variability in previous individual parameter estimates [17], while the lognormal distribution was used to guarantee positivity. All other parameters were fixed at the values given in Table R2.1.

We first conducted analyses in which the viral load was assumed to be observed exactly and continuously throughout infection, before conducting analyses in which measurements of the viral load were recorded once daily for one week after infection (we considered different frequencies of measurement in Section R2.3.5.3). In the latter case, we incorporated measurement error by adding a normally distributed random variate with standard deviation σ to the logarithm of each viral load measurement. When we accounted for measurement error, we used a default measurement error level of $\sigma = 1 \log_{10}(\text{TCID}_{50}/\text{ml})$, although we tested our results for a range of values of σ in Figures R2.6 and A2.3.

In analyses where we accounted for both non-continuous observation and measurement error, we fitted the TCL model to the daily data from each host using least-squares estimation. Specifically, the values of the parameters δ and $V(0)$ (for an individual host) were chosen to minimise the sum of squares distance between the logarithm of the viral load in the model and in the data, while all other parameter values were assumed to be known exactly and were fixed at the values in Table R2.1. To avoid unrealistically large estimates of the initial viral load, we imposed $V(0) \leq 10^3$ TCID_{50}/ml when we fitted the parameters (note that when we generated the synthetic data, the randomly sampled "true" value of $V(0)$ used to generate the data for a given host never exceeded this maximum). An example of synthetic data generated for a single host, in addition to the fitted TCL model, is given in Figure R2.6A.

R2.2.4.2 Expected infectiousness profile

In most of our analyses, we made the common assumption of a linear relationship between the viral load and infectiousness of each influenza-infected host [35, 49, 108, 111, 174] (alternative possibilities are considered in Appendix A2.2 and Figure R2.10). In particular, we assumed that

$$\beta^{(i)}(\tau) = kV^{(i)}(\tau), \tag{R2.20}$$

for constant k, where i represents the particular host under consideration. Therefore, the expected infectiousness, $\beta(\tau)$, was given in terms of the expected viral load, $\bar{V}(\tau)$, of a host at time τ since infection (calculated over a large number of realisations of the within-host model), by

$$\beta(\tau) = k\bar{V}(\tau). \tag{R2.21}$$

We fixed the constant k by assuming that the basic reproduction number,

$$R_0 = N \int_0^\infty \beta(\tau)\, \mathrm{d}\tau, \tag{R2.22}$$

was known. In our main analyses, we fixed $R_0 = 1.5$, which is consistent with estimates for influenza A [89] (different values of R_0 are considered in Figure R2.11). The expected infectiousness could therefore be calculated using the formula

$$\beta(\tau) = \frac{R_0}{N \int_0^\infty \bar{V}(x)\, \mathrm{d}x} \bar{V}(\tau). \tag{R2.23}$$

To calculate the "true" expected infectiousness profile, $\beta(\tau)$, we computed the expected viral load over 10,000 realisations of the within-host model. We also conducted analyses in which data were only available from a smaller number of hosts, d. In such cases, we simulated the within-host model d times to calculate the exact host-scale dynamics corresponding to each host, and used the data to estimate first $\bar{V}(\tau)$ and then $\beta(\tau)$. In analyses where we also incorporated measurement error, we used the host-scale dynamics estimated by fitting the within-host model to daily observations of the viral load for each host, in order to estimate $\beta(\tau)$.

R2.2.4.3 Population-scale dynamics

Both the compartmental and IDE methods were used to project the population-scale dynamics with "true" or estimated expected infectiousness profile, $\beta(\tau)$. To parameterise the $\mathrm{SI}_n\mathrm{R}$ model, we took $T = 7$ days in equations (R2.11) and (R2.12), since the (true) expected infectiousness was found to be very small after a week since infection. We considered a population of size $N = 1{,}000$ individuals, and assumed that there was initially a single newly infected individual, with all others susceptible. These initial conditions were implemented in the $\mathrm{SI}_n\mathrm{R}$ model by taking $I_1(0) = 1$ and $S(0) = 999$, with all other compartments containing zero hosts initially.

R2.2.5 Errors in host- and population-scale dynamics

We defined error metrics in order to quantify the errors that arise in both the host- and population-scale dynamics. These were defined as proportional errors, so as to enable comparison between errors at the different scales.

First, we defined the within-host error, E_{wh}, to be the difference between the exact and approximate infectiousness profiles, integrated over the entire course of infection, as a proportion of the area of the exact infectiousness profile. Therefore,

$$E_{wh} = \frac{\int_0^\infty |\beta_{\mathrm{approx}}(\tau) - \beta_{\mathrm{exact}}(\tau)|\, \mathrm{d}\tau}{\int_0^\infty \beta_{\mathrm{exact}}(\tau)\, \mathrm{d}\tau}, \tag{R2.24}$$

where $\beta_{\text{exact}}(\tau)$ and $\beta_{\text{approx}}(\tau)$ are the exact and approximate infectiousness profiles, respectively.

Similarly, we defined the between-host error, E_{bh}, in terms of the rate of new cases per day throughout the epidemic, i.e.

$$E_{bh} = \frac{\int_0^\infty \left|\dot{S}_{\text{approx}}(t) - \dot{S}_{\text{exact}}(t)\right| \mathrm{d}t}{\int_0^\infty -\dot{S}_{\text{exact}}(t)\,\mathrm{d}t}, \tag{R2.25}$$

where $S_{\text{exact}}(\tau)$ and $S_{\text{approx}}(\tau)$ are the exact and approximate numbers of susceptible individuals at time t since the start of the epidemic, respectively, and the dot denotes differentiation with respect to time.

R2.3 Results

R2.3.1 Summary of multi-scale modelling framework

We have developed a new compartmental framework for transitioning from host- to population-scale epidemic dynamics. In our approach, a within-host model is fitted to data from individual hosts to estimate the pathogen load of each measured host at every time since infection (Figure R2.1A). As with other multi-scale epidemic models [111, 167, 174], by assuming a functional relationship between pathogen load and infectiousness, the expected infectiousness profile, $\beta(\tau)$, of an infected host is then estimated at each time since infection, τ (Figure R2.1B). In previous approaches, the K&M IDE model [133] has then been used to calculate the population-scale dynamics (Figure R2.1C). In our approach, the expected infectiousness profile is instead used to parameterise a multi-stage compartmental model with a large number of infected compartments (Figure R2.1D), which can also be used to project the population-scale dynamics (Figure R2.1E).

R2.3.2 Equivalence between the compartmental and IDE methods in the large-compartment limit

In this section, we give a proof that the compartmental and IDE multi-scale modelling frameworks (Figure R2.1) are equivalent in the large-compartment limit of the compartmental approach. Specifically, we show that when the SI_nR compartmental model (Figure R2.1D) is parameterised as described in Section R2.2.3.1, the K&M IDE model (with "target" expected infectiousness profile $\beta(\tau)$) is obtained in the limit as the number of infected compartments, n, in the SI_nR model tends to infinity.

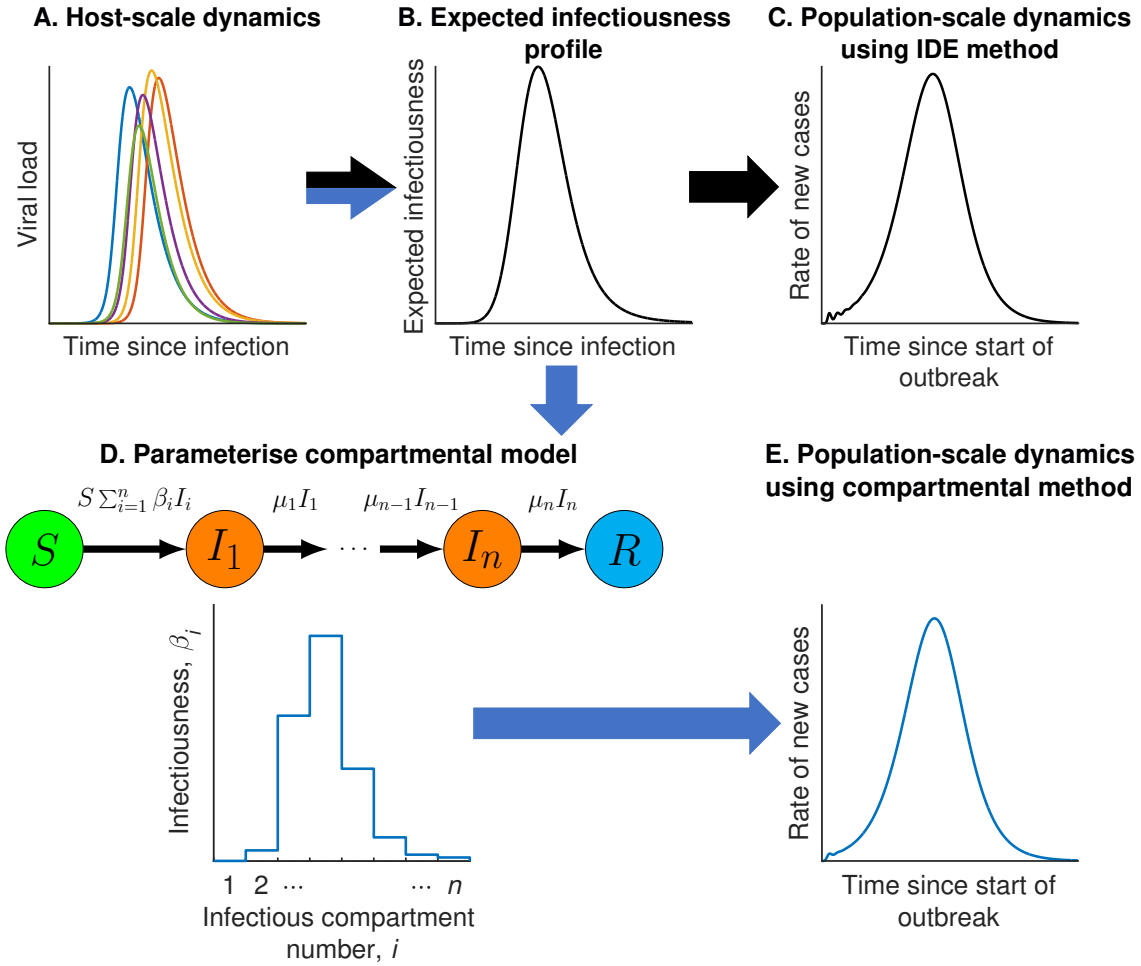


Figure R2.1: Schematic demonstrating methods for transitioning from host- to population-scale epidemiological dynamics. In standard approaches (A–C), measurements from within each host (A) are used to parameterise a within-host model, giving rise to an averaged infectiousness profile (B). These host-scale dynamics can then be nested in an IDE model (e.g. the K&M model [133]) used to project the population-scale dynamics (C). However, IDE models are challenging to solve. In contrast, in our approach (A,B,D,E), the expected infectiousness profile is instead used to parameterise a compartmental model (D) that can be used to project population-scale dynamics straightforwardly (E). Note that early-epidemic oscillations in C and E occur because the expected infectiousness of an infected host is close to zero in the first day of infection, leading to delays before successive generations of newly infected hosts begin to transmit the pathogen.

The system of ODEs which define the $\mathrm{SI}_n\mathrm{R}$ model (equations (R2.1)–(R2.4)) can be re-written as the system of equations,

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -S\int_0^{n\lambda(n)} \beta(\tau;n)I(t,\tau;n)\mathrm{d}\tau, \tag{R2.26}$$

$$\frac{\partial I(t,\tau;n)}{\partial t} = \begin{cases} -\dfrac{1}{\lambda(n)}\left(\dfrac{\mathrm{d}S}{\mathrm{d}t} + I(t,\tau;n)\right), & \text{for } 0 \leq \tau < \lambda(n), \\ \dfrac{1}{\lambda(n)}\left(I\left(t,\tau-\lambda(n);n\right) - I(t,\tau;n)\right), & \text{for } \lambda(n) \leq \tau < n\lambda(n). \end{cases} \tag{R2.27}$$

Here, $S(t)$ represents the number of susceptible individuals at time t since the start of the outbreak, and

$$I(t,\tau;n) \coloneqq \begin{cases} I_1(t)/\lambda(n), & \text{for } 0 \leq \tau < \lambda(n), \\ I_2(t)/\lambda(n), & \text{for } \lambda(n) \leq \tau < 2\lambda(n), \\ \vdots & \\ I_n(t)/\lambda(n), & \text{for } (n-1)\lambda(n) \leq \tau < n\lambda(n), \end{cases} \tag{R2.28}$$

provides an approximation to the density of infected individuals who have been infected for time τ, where $I_i(t)$ is the number of individuals in infected stage i (for $i = 1, \ldots, n$; the dependence of $I(t,\tau;n)$ on the number of infected compartments, n, is indicated explicitly for clarity). Finally, $\lambda(n) = 1/\mu_i$ is the mean duration of each infected stage, and $\beta(\tau;n)$ gives the transmission rate at time τ since infection of an infected individual who spends exactly time $\lambda(n)$ in each infected stage (the exact definition of $\beta(\tau;n)$ in terms of the transmission rates in each infected stage, β_i, is given in Section R2.2.3.1 – equation (R2.14)).

As described in Section R2.2.3.1, we chose $\lambda(n)$ such that $\lambda(n) \to 0$ as $n \to \infty$. If the target expected infectiousness profile, $\beta(\tau)$, has unbounded support, then it is also necessary to impose the condition $n\lambda(n) \to \infty$. However, if there exists some time since infection, T, such that $\beta(\tau) = 0$ for $t > T$, then it is possible to take $\lambda(n) = T/n$. In addition, we chose the transmission rates, β_i, such that $\beta(\tau;n) \to \beta(\tau)$ at each τ as $n \to \infty$ (equations (R2.15) and (R2.16)). Under these parameter choices, in the limit $n \to \infty$, equations (R2.26) and (R2.27) reduce to

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -S(t)\int_0^\infty \beta(\tau)I(t,\tau)\mathrm{d}\tau, \tag{R2.29}$$

$$\frac{\partial I(t,\tau)}{\partial t} + \frac{\partial I(t,\tau)}{\partial \tau} = 0, \quad \tau > 0, \tag{R2.30}$$

$$I(t,0) = -\frac{\mathrm{d}S}{\mathrm{d}t}, \tag{R2.31}$$

where $I(t,\tau) \coloneqq I(t,\tau;\infty)$. The partial differential equation (PDE) (R2.30), subject to boundary condition (R2.31) and a general initial condition $I(0,\tau) = g(\tau)$, has solution

$$I(t,\tau) = \begin{cases} g(\tau - t), & \text{for } t < \tau, \\ -\dfrac{\mathrm{d}S(t-\tau)}{\mathrm{d}t}, & \text{for } t > \tau. \end{cases} \tag{R2.32}$$

(This solution can be obtained using the method of characteristics). Substituting this solution into equation (R2.29), we find that

$$\frac{\mathrm{d}S}{\mathrm{d}t} = S(t)\left[\int_0^t \beta(\tau)\frac{\mathrm{d}S(t-\tau)}{\mathrm{d}t}\mathrm{d}\tau - \int_t^\infty \beta(\tau)g(\tau-t)\mathrm{d}\tau\right], \tag{R2.33}$$

which is the K&M model with expected infectiousness profile $\beta(\tau)$.

R2.3.2.1 Initial conditions

Whenever we solved the $\mathrm{SI}_n\mathrm{R}$ model numerically starting with a single newly infected host (e.g. the blue line in Figure R2.3B), we took the initial conditions $I_1(0) = 1$, $I_i(0) = 0$ for $i = 2, \ldots, n$, and $S(0) = N - 1$ (where N is the total population size). This choice of initial conditions gives

$$I(0, \tau; n) \coloneqq \begin{cases} \dfrac{1}{\lambda(n)}, & \text{for } 0 \leq \tau < \lambda(n), \\ 0, & \text{otherwise.} \end{cases} \tag{R2.34}$$

In the limit $n \to \infty$ with $\lambda(n) \to 0$, we have that $g(\tau) = I(0, \tau) = 0$ for $\tau > 0$, with $\int_0^\infty I(0, \tau)\mathrm{d}\tau = 1$. Therefore, $g(\tau) = \delta(\tau)$, where $\delta(\tau)$ is the Dirac delta function, which is the same initial condition that we used when solving the K&M model numerically (see Section R2.2.2).

R2.3.2.2 Error analysis

We also considered the errors in projections that can arise when a finite number of compartments is used in our compartmental framework. For simplicity, we focus here on the case where the infectiousness profile has bounded support (i.e. where $\beta(\tau) = 0$ for $\tau > T$), so that the mean duration of each infected compartment could be taken to be $\lambda(n) = T/n$.

When taking the limit $n \to \infty$ in equation (R2.27), the leading order error terms are $\mathcal{O}(\lambda(n)) = \mathcal{O}(1/n)$. In the case of an infectiousness profile with bounded support, we will also have

$$\beta(\tau; n) = \beta(\tau) + \mathcal{O}(1/n), \tag{R2.35}$$

for all $\tau > 0$, so that the error terms when taking the limit in equation (R2.26) will also be of $\mathcal{O}(1/n)$. This therefore suggests that the error in the population-scale dynamics (as defined by the error metric given in equation (R2.25)) when the compartmental method is used, relative to the dynamics projected using the IDE method, will be $\mathcal{O}(1/n)$ as the number of compartments, n, tends to infinity. This error estimate is tested in our case study of influenza below (Figure R2.4C).

R2.3.2.3 Note on the relative computational efficiency of the compartmental and IDE frameworks

Here, we consider the case where the expected infectiousness profile satisfies $\beta(\tau) = 0$ for $\tau > T$, and suppose that the SI_nR model with parameter choices as outlined in Section R2.2.3.1 is discretised using an Euler method with a time-step $\delta t = \lambda(n) = T/n$. It is straightforward to see that this numerical scheme is essentially equivalent to the finite-difference scheme given for the numerical solution of the K&M model in equation (R2.10), up to slight differences in the transmission rates. This therefore suggests that there is no innate difference in computational efficiency between the compartmental and IDE approaches (rather, this will depend on the precise numerical methods used to solve each model). The major advantage of our compartmental approach is its ease of use, while remaining similarly computationally efficient to the IDE approach.

R2.3.3 Influenza case study

To illustrate our multi-scale framework in a concrete setting, we considered the case study of influenza A infection in humans. We used the TCL within-host model, which has previously been fitted to data from a cohort study of influenza infection [17], to generate synthetic data from a large number of hosts (see Section R2.2.4.1). To demonstrate that the synthetic data that we generated were comparable to real data, we plotted real data from six hosts collected in a cohort study of influenza infection [175] (Figure R2.2A–F; note that the TCL model was previously fitted to the data shown in these panels to obtain the parameter estimates reproduced in Table R2.1 [17]) alongside synthetic data for six hosts (Figure R2.2G–L).

The synthetic viral load data were used to calculate the expected infectiousness profile (Figure R2.3A) under the assumptions of a linear relationship between viral load and infectiousness [35, 49, 108, 111, 174] and a basic reproduction number of 1.5 [89] (although other assumptions and values of the reproduction number are considered later), initially assuming perfectly characterised host-scale dynamics (i.e. assuming exact and continuous data were available from a large number of hosts). Both our compartmental approach, using a large number of infected compartments (n = 1,000), and the previously used IDE method (i.e. the K&M model) were then used to project the population-scale dynamics of an outbreak (Figure R2.3B). We considered a population of size N = 1,000 and assumed a single newly infected individual was introduced into an otherwise entirely susceptible population. The two approaches produced almost identical results – the error in the projected population-scale dynamics when the compartmental method was used, calculated as

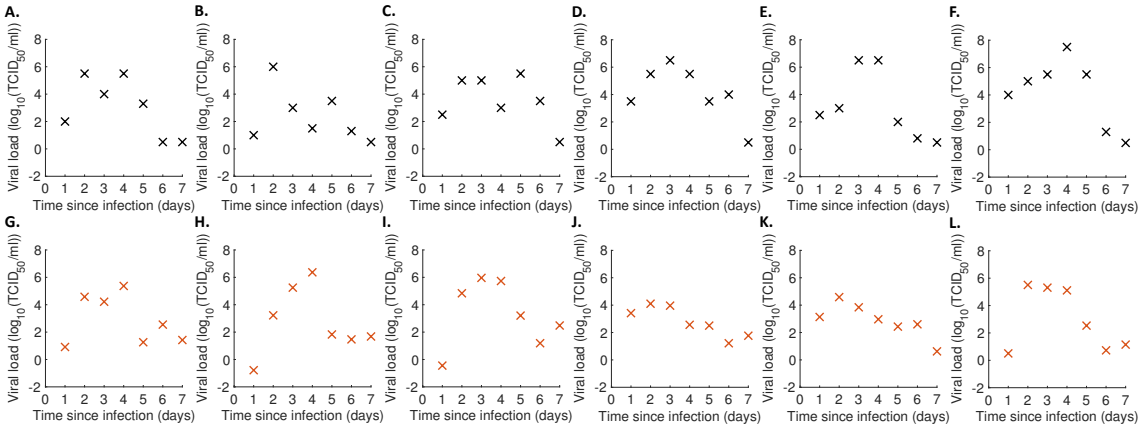


Figure R2.2: Real and synthetic host-scale data. A–F. Viral load data against the time since infection (black crosses) for six hosts, taken from a cohort study of influenza infection and previously used to parameterise the TCL within-host model [17, 175]. G–L. Example synthetic viral load data against the time since infection for six hosts (red crosses), generated using the TCL model (see Section R2.2.4.1) and assuming a normally distributed measurement error of standard deviation $\sigma = 1$ $\log_{10}$(TCID$_{50}$/ml) in each observation of the logarithm of the viral load.

a proportional error relative to the dynamics projected using the IDE method (see Section R2.2.5), was only 0.2%.

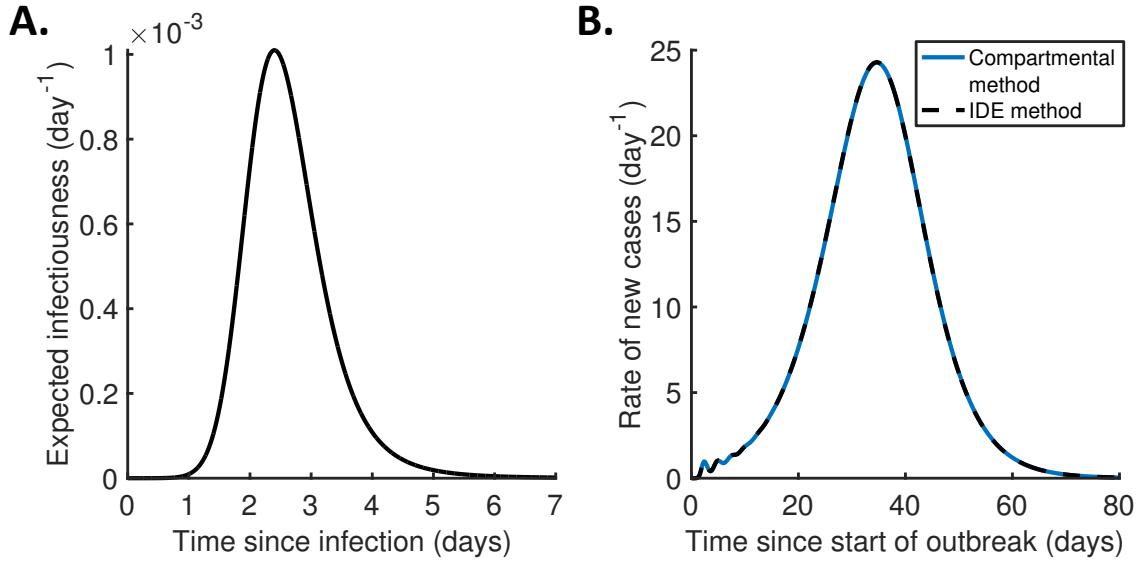


Figure R2.3: Transitioning from host- to population-scale influenza dynamics using the compartmental and IDE methods. A. The expected infectiousness profile, $\beta(\tau)$, when the host-scale dynamics are perfectly characterised. B. The population-scale dynamics, using our compartmental approach with $n = 1{,}000$ infected compartments (blue), and using the IDE method (black dashed), for the infectiousness profile shown in A.

R2.3.4 Analysis of sources of errors in population-scale projections

We considered the effect of three quantities on population-scale epidemic projections:

i. The number of compartments used in our framework.

ii. The number of hosts from which individual host-scale data are available.

iii. The extent of measurement error in host-scale data.

Note that the first of these three quantities is simply a choice for the user to make (i.e. a very large number of compartments can always be used in our framework to minimise this source of error), whereas the other two quantities characterise the availability and accuracy of host-scale data.

Initially, we considered these three factors in isolation, before testing their combined effects. We defined error metrics to quantify the errors that arose in the host-scale dynamics (the within-host error, E_{wh}) and in the population-scale dynamics (the between-host error, E_{bh}), using proportional errors in order to enable comparison between errors at the two scales (see Section R2.2.5).

R2.3.4.1 Number of compartments

Our error analysis in Section R2.3.2.2 suggested that in general, when our compartmental framework is used to estimate population-scale epidemic dynamics, the error in projections (compared to the corresponding estimates using the IDE approach) scales with $1/n$ as the number of compartments, n, becomes large. In order to test this error estimate, and also to investigate how many compartments are required for sufficiently accurate population-scale projections in our case study of influenza A, we assumed that the host-scale dynamics were perfectly characterised so that the true expected infectiousness profile (Figure R2.3A) was known exactly. We then generated the population-scale dynamics for different values of the number of infected compartments, n, and compared these to the dynamics projected using the IDE method (Figure R2.4A).

To quantify the improvement of the approximation as the number of compartments is increased, we plotted the error in the population-scale dynamics against the number of compartments, n (Figure R2.4B). In this case, $n = 24$ compartments were required for an error of 10% or below, while 47 compartments were needed for a 5% error (Figure R2.4B). We also plotted the error on a log-log scale for a wider range of n, in order to confirm that the error scales with $1/n$ as the number of compartments, n, becomes large (Figure R2.4C).

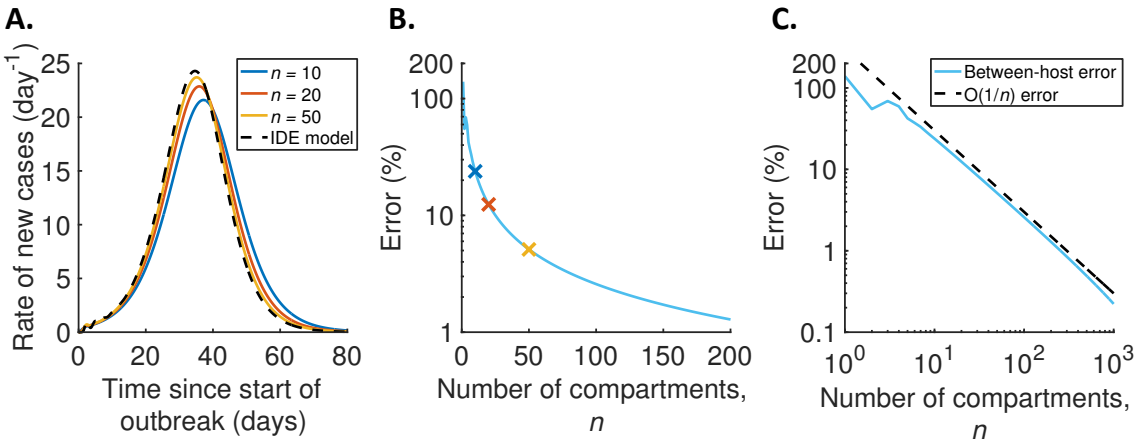


Figure R2.4: Number of compartments. A. The population-scale dynamics, using our compartmental approach with n =10 (blue), 20 (red) and 50 (yellow) infected compartments, and using the IDE method (black dashed), when the infectiousness profile is as in Figure R2.3A. B. The error in the population-scale dynamics against the number of compartments used in our framework, n, plotted using a log scale on the y-axis (light blue line). The crosses represent the errors corresponding to the curves of the same colour in A (these are at values of 24%, 12% and 5% error). C. The error in the population-scale dynamics for a larger range of n (light blue), plotted on a log-log scale and compared to an error proportional to $1/n$ (black dashed).

R2.3.4.2 Number of hosts

In most cohort studies used to inform multi-scale models, data are only available from a small number of hosts [61, 108], and within-host model parameters may vary significantly between hosts [17, 35, 233]. To investigate the error in population-scale projections that inadequate data may generate, we supposed that data were only available from d randomly chosen hosts. To isolate the effect of variability between hosts rather than measurement error, the exact viral load of each host was initially assumed to be known at every time since infection. We used the available data to estimate the expected infectiousness profile (Figure R2.5A) and calculated the approximate population-scale dynamics using our compartmental framework with $n = 1{,}000$ infected compartments (Figure R2.5B). For a fixed cohort size, d, significantly different projections of population-scale dynamics are possible, depending on which hosts are included in the study (Figure R2.5C). Therefore, we calculated the distributions of errors in both the host- and population-scale dynamics, relative to the case in which the host-scale dynamics were perfectly characterised, over 5,000 repeats for each of a range of cohort sizes, d (Figure R2.5D).

As the number of hosts increases, the errors at host- and population-scale both decrease in general (depending on precisely which hosts are included in the study cohort; Figure R2.5D). In this case – when exact and continuous data are available

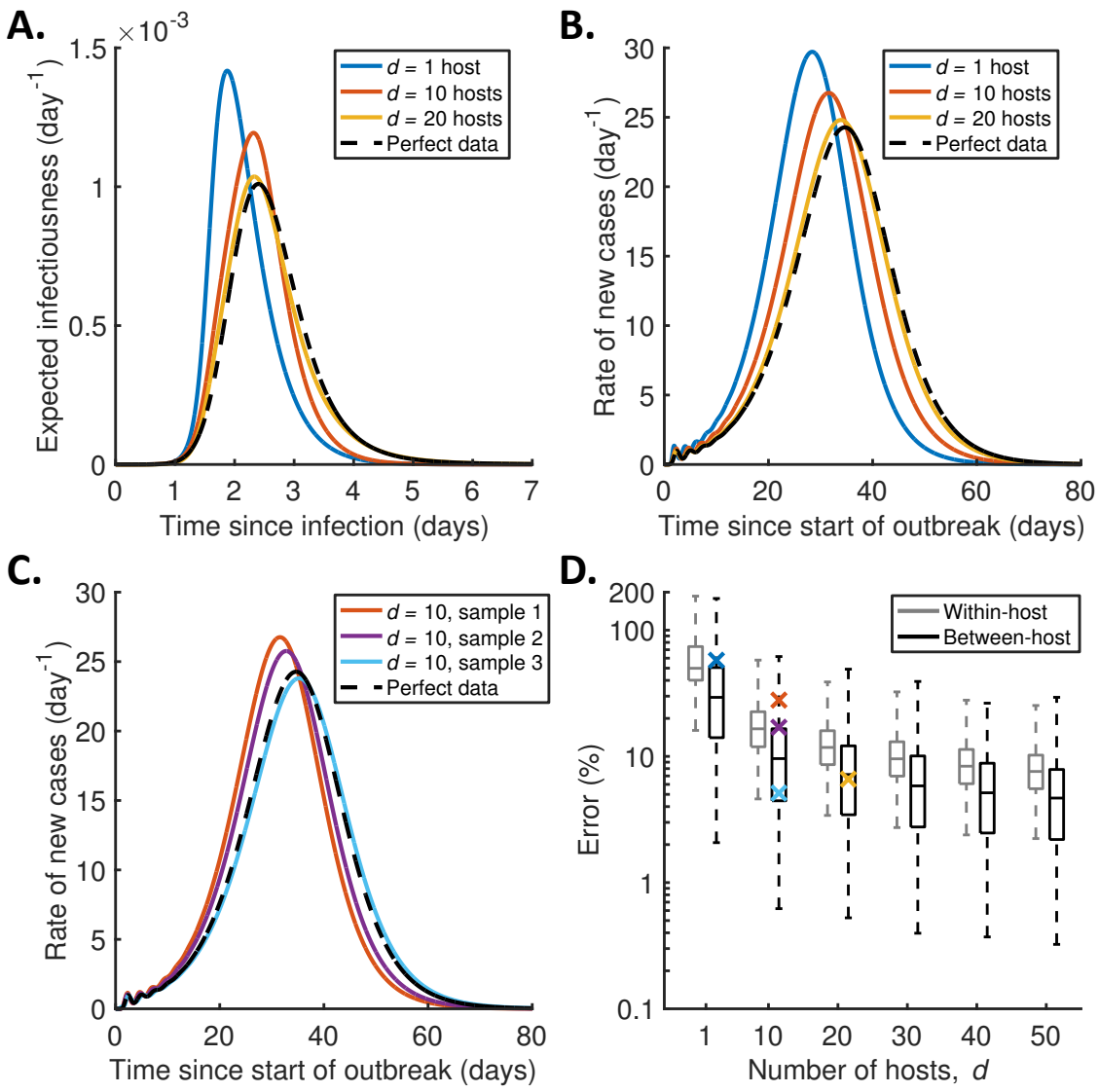


Figure R2.5: Number of hosts. A. Examples of estimates of the expected infectiousness profiles when exact and continuous data are available from $d = 1$ (blue), $d = 10$ (red) or $d = 20$ (yellow) randomly chosen hosts, and the expected infectiousness profile when the host-scale dynamics are perfectly characterised (black dashed). B. The projected population-scale dynamics obtained using each infectiousness profile in A, using $n = 1{,}000$ compartments in our framework. C. Three examples of possible approximate population-scale dynamics when data are available from $d = 10$ hosts. D. Box-and-whisker plots indicating the distributions of within-host (grey) and between-host (black) errors for different groups of hosts randomly chosen in the study cohort, for a range of values of the number of hosts, d, assuming host-scale data are exact and continuous. The boxes indicate the lower quartile, median and upper quartile, and the maximum length of each whisker is 1.5 times the interquartile range. The crosses represent the between-host errors corresponding to the curves of the same colour in B and C.

from each host – a cohort size of $d = 20$ hosts is sufficient for the between-host error to be no more than 10% on average (Figure R2.5D). However, since the errors are affected by the precise hosts included in the cohort, more hosts are required for a greater certainty of a small between-host error. For example, $d = 30$ hosts are required to ensure that the upper quartile of between-host errors for cohorts of that size is 10% or below.

Scatter plots comparing the errors at host- and population-scale for each cohort size considered in Figure R2.5D are shown in Figure A2.1. Although a large range of population-scale errors are possible for a given host-scale error, the magnitude of the population-scale error is usually smaller than that of the host-scale error (note that we used proportional errors to enable comparison of errors at different scales – see Section R2.2.5). For example, for a cohort size of $d = 20$ hosts, we found a smaller population- than host-scale error for 87% of possible cohorts (Figure A2.1C). Therefore, limited host-scale data do not necessarily preclude accurate population-scale projections, despite variability in viral dynamics between different hosts.

R2.3.4.3 Extent of measurement error in host-scale data

In longitudinal cohort studies, data are only collected from each host at a limited number of time points. For studies of influenza infection, data may be collected daily over the course of infection (for example [17, 35]), which lasts approximately one week [37]. However, there can be significant measurement error whenever the viral load is recorded [174, 180].

We considered viral load values recorded daily for each host for a week after infection (the effect of more frequent observations is explored in Section R2.3.5.3). To incorporate measurement error in the synthetic data, a normally distributed error with standard deviation σ was applied to the logarithm of each measurement. We then fitted the TCL model to the data for each host to estimate the host-scale dynamics (Figure R2.6A).

To demonstrate the effect of measurement error on population-scale projections, we assumed that data were available from $d = 10$ randomly chosen hosts, and compared estimates of the expected infectiousness profile, first under the assumption that the viral load of each host was known exactly at all times during infection, and second when there was measurement error in daily recordings of the viral load (Figure R2.6B). Our compartmental framework with $n = 1{,}000$ compartments was then used to project the population-scale dynamics in both cases (Figure R2.6C). We calculated the within-host and between-host errors that arose directly due to

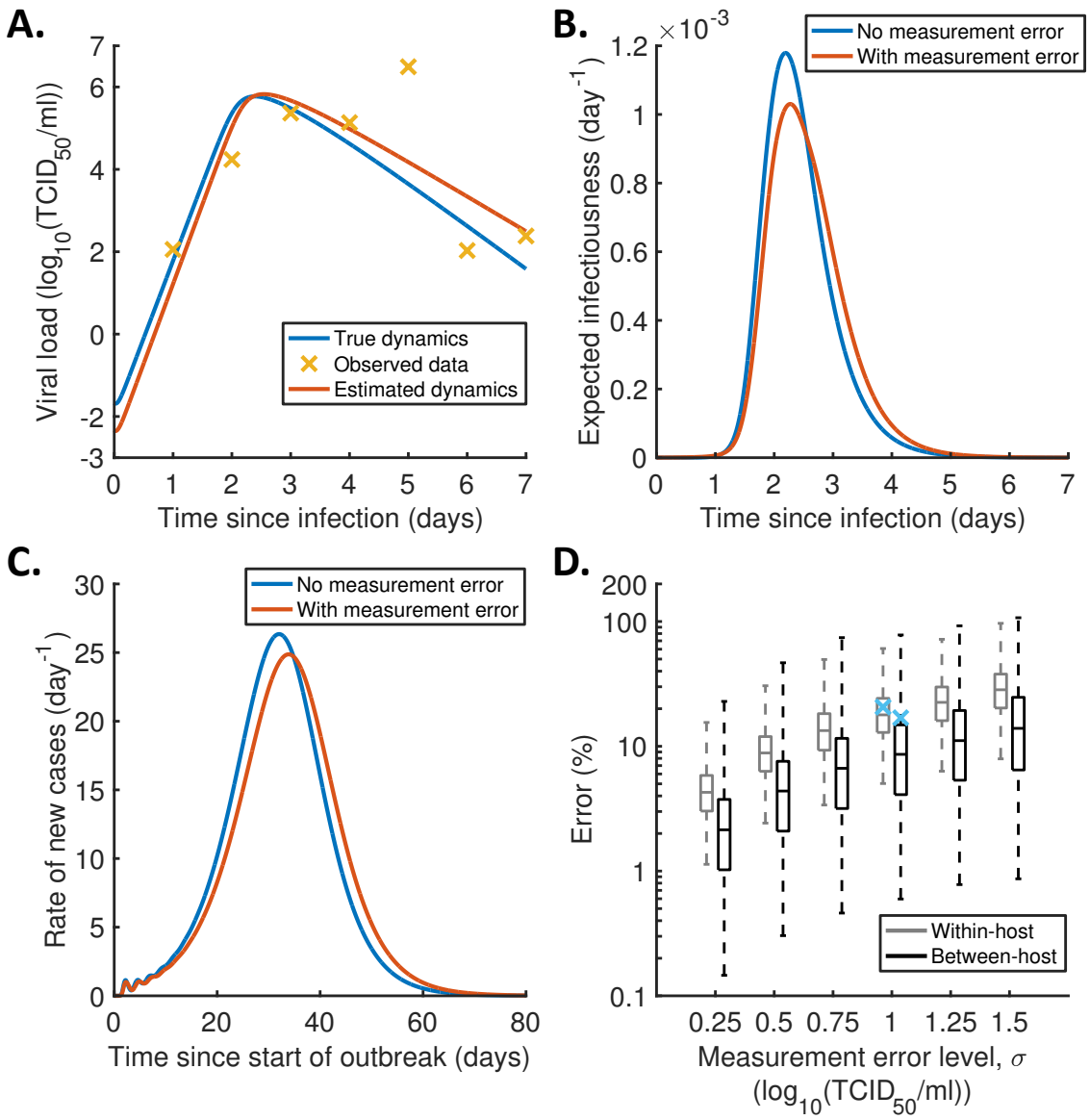


Figure R2.6: Extent of measurement error in host-scale data. A. Example of synthetic data for a single host: the true viral load of the host against the time since infection (plotted using a log scale on the y-axis; blue), daily synthetic data with measurement error (in the logarithm of the viral load) of standard deviation $\sigma = 1$ $\log_{10}(\text{TCID}_{50}/\text{ml})$ (yellow crosses), and the estimated viral load against the time since infection when the TCL model is fitted to the data (red). B. Examples of estimated expected infectiousness profiles, without measurement error (blue) and with measurement error (red), for $d = 10$ hosts. C. The population-scale dynamics for each infectiousness profile in B, using $n = 1{,}000$ compartments in our framework. D. Box-and-whisker plots indicating the distributions of within-host (grey) and between-host (black) errors arising directly due to measurement error, for a range of values of the extent of measurement error, σ. The light blue crosses represent the within-host error corresponding to B and the between-host error corresponding to C (these are at values of 21% and 17% error, respectively).

measurement error by taking the "true" dynamics to be those when exact and continuous data were available from the same 10 individuals. The distributions of these errors, each time calculated over 5,000 repeats for a range of values of σ, are shown in Figure R2.6D.

The errors at host-scale and at population-scale both increase with the measurement error level, σ. For values of σ of 1 $\log_{10}(\text{TCID}_{50}/\text{ml})$ or higher, the mean population-scale error exceeds 10%. In that case, when a cohort of only $d = 10$ hosts is used, measurement error alone may prevent accurate population-scale projections, even if there is no additional error contribution due to within-host model parameter variability.

R2.3.4.4 Overall error

So far, we have described analyses of the separate effects of host cohort size and measurement error on the characterisation of host-scale viral load time series, as well as the resulting impact on population-scale outbreak projections. However, in reality, both these sources of error would be present simultaneously. Errors would also occur if a small number of compartments is used in our multi-scale modelling approach, although this can be avoided by simply choosing a large number of compartments in the compartmental framework. Nonetheless, we also conducted an analysis in which all three potential sources of error were included (number of hosts, extent of measurement error, and number of compartments).

When we investigated the combined effect of these potential sources of error, we considered a measurement error level of $\sigma = 1$ $\log_{10}(\text{TCID}_{50}/\text{ml})$ in once daily measurements of the viral load, which generated synthetic data comparable to data recorded in cohort studies (Figure R2.2; different error levels are considered in Figure R2.6). Assuming that data were available from d randomly sampled hosts, our compartmental framework with n infected compartments was used to estimate the population-scale dynamics. We repeated this analysis 10,000 times each for different pairs of values of d and n, each time calculating the within- and between-host errors, relative to the case in which the host-scale dynamics were perfectly characterised and the IDE method was used (equivalent to using infinitely many compartments in our compartmental framework).

The distributions of within- and between-host errors, when using a large number of compartments ($n = 1{,}000$) in our framework, are plotted for different numbers of hosts (d) in Figure R2.7A. We found that data from $d = 40$ hosts are required for an average between-host error of 10% or below (compared to 20 hosts if data are recorded exactly, i.e. with no measurement error – see Figure R2.5D).

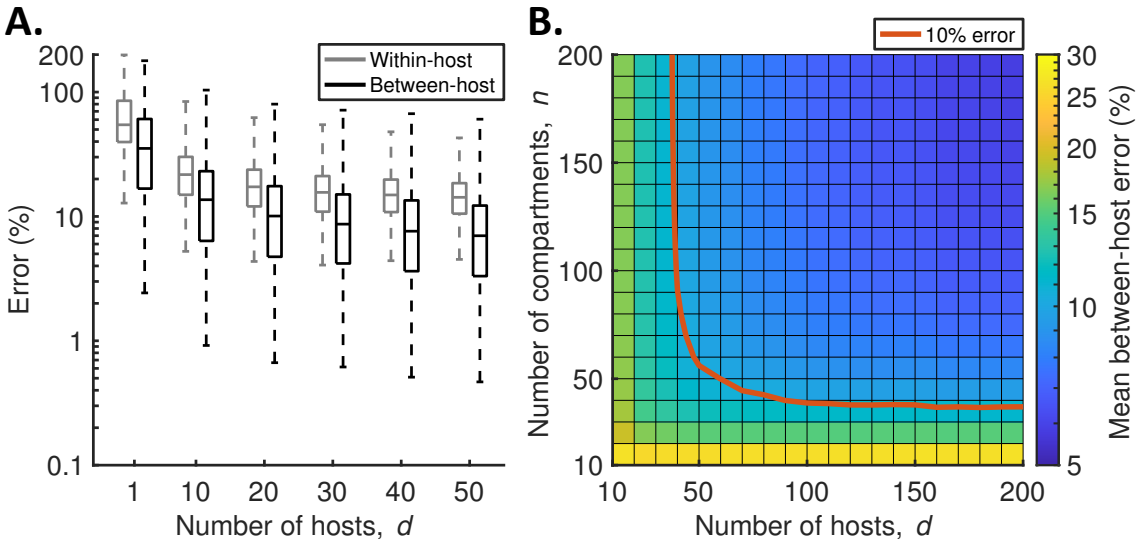


Figure R2.7: Overall error in population-scale projections. A. Box-and-whisker plots indicating the distributions of within-host (grey) and between-host (black) errors for different hosts chosen in the study cohort when $n = 1{,}000$ compartments are used in our framework, for a range of values of the number of hosts, d, assuming daily observations of the logarithm of the viral load with a normally distributed measurement error of standard deviation $\sigma = 1\ \log_{10}(\text{TCID}_{50}/\text{ml})$. B. The expected error in the population-scale dynamics, for different values of the number of compartments, n, and the number of hosts, d. The red line indicates where the error is 10%.

The mean error in the population-scale dynamics, for different numbers of hosts (d) and numbers of compartments (n), is shown in Figure R2.7B. In the case of $d = 40$ hosts (the minimum cohort size required to ensure a mean between-host error of no more than 10% when using a very large number of compartments), the user should choose at least $n = 60$ compartments in our approach to ensure a mean error of 10% or less. As few as 40 compartments are required if the cohort size is large (at least 70 hosts). Nonetheless, since the number of compartments is simply a choice for the user, rather than requiring any more data to be collected, we suggest that any user of our framework simply chooses a very large number of compartments. In the example considered here, we note that the benefit of using more compartments may become negligible when more than $n = 100$ compartments are included in our approach (Figure R2.7B).

R2.3.5 Sensitivity analyses

R2.3.5.1 Effect of explicitly including variability between different infected hosts in the population-scale model

In both the compartmental and IDE multi-scale frameworks we have considered in this research chapter, a population-scale model is used in which the population is assumed

to be homogeneous and well-mixed, with all infected hosts effectively assumed to follow the same infectiousness profile, $\beta(\tau)$. However, when we analysed the errors in population-scale epidemic projections that may arise when such projections are informed using host-scale data, we did account for variability in the individual infectiousness profiles of different infected hosts (since if host-scale data are only available from a small cohort of hosts, variability in host-scale dynamics can lead to errors in estimates of the expected infectiousness profile, $\beta(\tau)$).

In Appendix A2.1 we considered a more detailed population-scale compartmental model that explicitly accounts for variability in infectiousness progression between different infected hosts. As was the case for the SI_nR model used in most of our analyses, we found that in the limit in which each infected host passes through infinitely many (infinitesimally short) stages of infection, the K&M IDE model is attained (with an expected infectiousness profile, $\beta(\tau)$, which averages over all possible progression pathways). This result provides mathematical justification for using population-scale models which average over the host-scale dynamics of different infected individuals (see also [88]).

We note that in the extended model considered in Appendix A2.1, variability between different infected individuals is assumed to arise at random. If variation is not random, then a more complex population-scale model may be required – for example, an age-structured population-scale model may be needed if age affects host-scale dynamics [120]. Furthermore, even if variability between hosts is random, a stochastic population-scale model which explicitly accounts for this variability (such as the stochastic version of the model used in Appendix A2.1) may be needed in settings where a deterministic population-scale model is not appropriate – for example, stochasticity may be important at the very start of an outbreak, when a major epidemic is not certain to occur [8, 11, 213, 221].

R2.3.5.2 Results using the IDE method to transition from host- to population-scale dynamics

We repeated the analyses in Figure R2.5D, Figure R2.6D and Figure R2.7A, but using the IDE method rather than our compartmental framework to transition from host- to population-scale dynamics (Figure A2.2). In each case, we found that the results we obtained using the two approaches were almost identical.

R2.3.5.3 Effect of the frequency of data collection

In Figures R2.6 and R2.7 we assumed viral load data were collected once daily from each host (in the cohort of hosts from which such data were available), for a week

following infection. However, we also conducted supplementary analyses considering the possibility of more frequent data collection. If viral load data are collected from each host twice daily, rather than once daily, then estimates of individual host-scale dynamics may be more accurate (Figure R2.8A). We repeated the analyses in Figure R2.7, but assuming twice daily data collection from each host (Figure R2.8B–C). In this case, we found that data are only required from a cohort of 20 hosts for a 10% mean error.

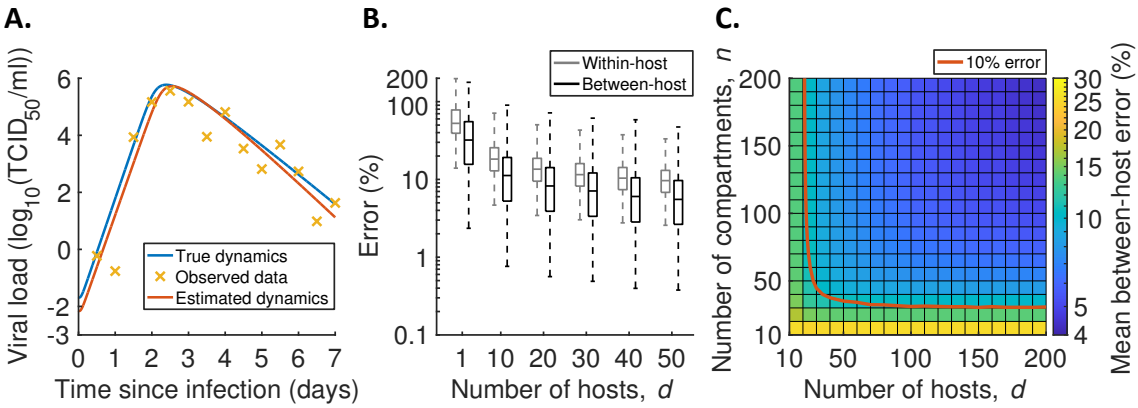


Figure R2.8: Results assuming host-scale data are collected twice daily from each host. A. Example of synthetic data for a single host: the true viral load of the host against the time since infection (blue), twice daily data with measurement error level $\sigma = 1$ $\log_{10}(\text{TCID}_{50}/\text{ml})$ (yellow crosses), and the viral load against the time since infection when the TCL model is fitted to the data (red). B. Box-and-whisker plots indicating the distributions of within-host (grey) and between-host (black) errors for different hosts chosen in the study cohort when $n = 1{,}000$ compartments are used in our framework, for a range of values of the number of hosts, d. C. The expected error in the population-scale dynamics when the compartmental method is used, against the number of compartments, n, and the number of hosts, d. The red line indicates where the error is 10%.

If, instead, the total number of measurements that can be taken is fixed, then it might be necessary to choose between sampling a large number of hosts infrequently, or a small number of hosts frequently. To explore this possibility, we compared different numbers of daily observations per host, f (assuming data collection for one week following infection, so that $7f$ measurements in total were collected from each host), for different fixed values of the total number of measurements, M, so that each time the (maximum possible) cohort size was $d = M/(7f)$. We considered values of M that were multiples of 84, so that data acquisition was possible at $f = 1, 2, 3$ and 4 daily measurements from each host.

In Figure R2.9 the expected error in the population-scale dynamics (each time calculated over 10,000 repeats, using $n = 1{,}000$ infected compartments in our framework) is plotted against M for $f = 1, 2, 3$ and 4 (with $d = M/(7f)$ in each

case). At larger values of the total number of measurements, M, the expected errors were found to be similar when data were collected either once daily from $d = d_1$ hosts or twice daily from $d = d_2 = 0.5d_1$ hosts (for example, when $d_1 = 40$, the respective errors are both 10%). On the other hand, data collection 3 or 4 times daily (from a smaller number of hosts) led to larger errors for a given fixed value of M compared to collection once or twice daily (from a larger number of hosts).

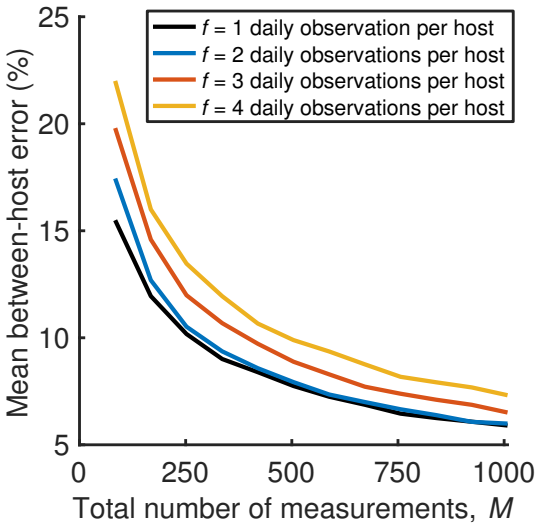


Figure R2.9: Trading off the number of hosts, d, and number of daily observations per host, f, when the total number of measurements, $M = 7fd$, is fixed. The expected error in the population scale dynamics is plotted against M for $f = 1$ (black), 2 (blue), 3 (red) and 4 (yellow), with $d = M/(7f)$ in each case.

R2.3.5.4 Effect of the relationship between pathogen load and infectiousness

In most of our analyses, we made the common assumption that infectiousness is directly proportional to the viral load [35, 49, 108, 111, 174]. However, we also considered two alternative scenarios, in which the infectiousness of each host either scales with the logarithm of their viral load [49, 108] (Figure R2.10A–C) or saturates at high viral loads [9, 58] (Figure R2.10D–F). These alternative relationships between pathogen load and infectiousness are detailed in Appendix A2.2. For both possibilities, assuming perfectly characterised host-scale dynamics so that the infectiousness profile was known exactly (Figure R2.10A,D), we used both the compartmental and IDE methods to transition to population-scale dynamics (Figure R2.10B,E). When our compartmental approach was used, we assumed that the expected infectiousness was negligible at times since infection greater than $T = 8$ days.

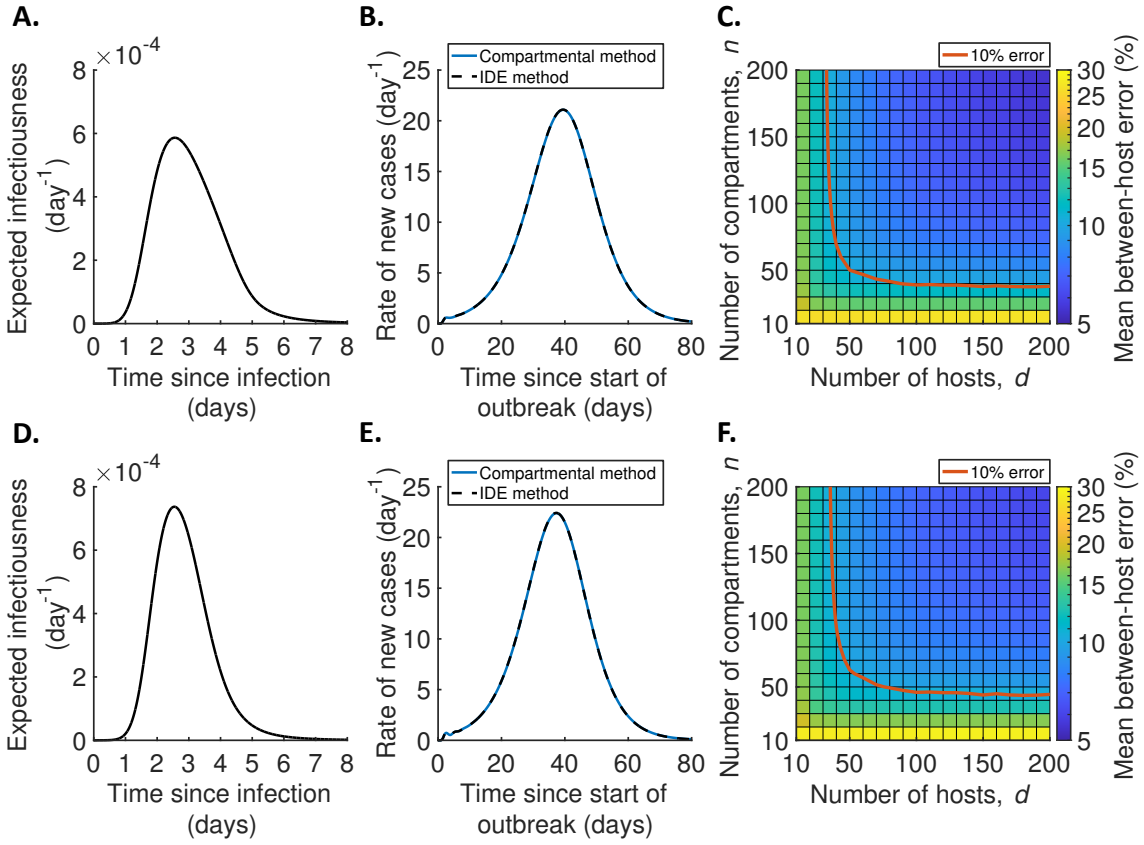


Figure R2.10: Effect of the relationship between pathogen load and infectiousness. A–C. Results assuming infectiousness scales with the logarithm of the viral load (see Appendix A2.2). A. The expected infectiousness profile, $\beta(\tau)$, when the host-scale dynamics are perfectly characterised. B. The population-scale dynamics, using our compartmental approach with $n = 1{,}000$ infected compartments (blue), and using the IDE method (black dashed), when the infectiousness profile is as in A. C. The expected error in the population-scale dynamics when the compartmental method is used, against the number of compartments, n, and the number of hosts, d. The red line indicates where the error is 10%. D–F. Equivalent figures to A–C, but instead assuming a relationship where infectiousness saturates at high viral loads (see Appendix A2.2).

We then repeated the analysis in Figure R2.7B but under the log-linear (Figure R2.10C) and saturation relationships (Figure R2.10F), in both cases obtaining similar results to those shown in Figure R2.7B – for example, when there were $d = 40$ hosts and we used $n = 200$ compartments in our framework, the mean population-scale errors were 9.7% (Figure R2.7B), 9.2% (Figure R2.10C) and 9.5% (Figure R2.10F) under the assumptions of a linear, log-linear or saturation relationship, respectively.

R2.3.5.5 Effect of the basic reproduction number, R_0

We then explored the effect of the assumed value of the basic reproduction number, R_0, on our results (Figure R2.11), considering values of R_0 both lower and higher than 1.5 (the value used in most of our analyses). The number of hosts required for a mean population-scale error of 10% or below was found to increase with R_0, with $d = 10$ hosts required when $R_0 = 1.1$ (Figure R2.11C) and $d = 70$ hosts required when $R_0 = 3$ (Figure R2.11I). On the other hand, for values of R_0 both lower and higher than 1.5, fewer compartments were required in our framework to ensure a mean error of 10% or below than in Figure R2.7B (where we found that, for $R_0 = 1.5$, a minimum of $n = 40$ compartments were required when the cohort size was large).

R2.3.5.6 Effect of the extent of measurement error in population-scale data.

We also repeated the analysis in Figure R2.7B for different values of the standard deviation, σ, of the measurement error applied to observations of the logarithm of the viral load, finding that both the number of hosts, d, and number of compartments, n, required for a population-scale error of 10% or below increased when there was more measurement error (Figure A2.3).

R2.3.5.7 Effect of the extent of variability in host-scale parameter values

When we generated host-scale data, we sampled the logarithms of the parameters δ and $V(0)$ in the TCL model from normal distributions with standard deviations 0.25 $\log_{10}(\text{day}^{-1})$ and 1.12 $\log_{10}(\text{TCID}_{50}/\text{ml})$, respectively, for each host. These values were chosen to match variability in previous individual parameter estimates for six hosts [17], but the level of variability was not known to high accuracy due to the small sample size in that study [17]. Therefore, we also considered changing the standard deviations of the distributions from which we sampled $\log_{10}(\delta)$ and $\log_{10}(V(0))$ to 0.25α $\log_{10}(\text{day}^{-1})$ and 1.12α $\log_{10}(\text{TCID}_{50}/\text{ml})$, respectively, and repeated our analyses in Figure R2.7B for different values of α (Figure A2.4). The number of hosts required for a population-scale error of no more than 10% increased with α (Figure A2.4).

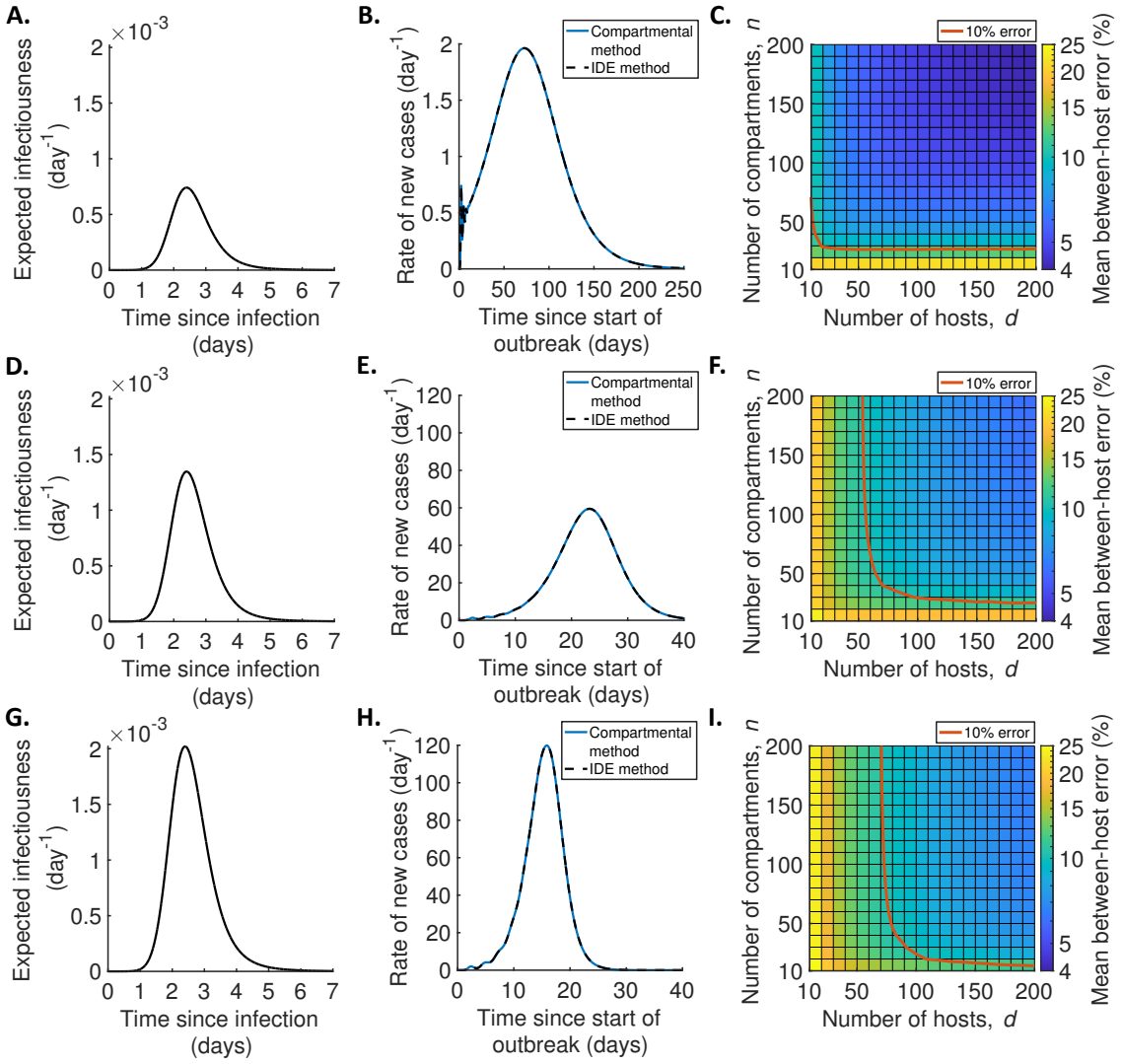


Figure R2.11: Effect of the basic reproduction number, R_0. A. The expected infectiousness profile, $\beta(\tau)$, when the host-scale dynamics are perfectly characterised, and assuming a basic reproduction number of $R_0 = 1.1$. B. The population-scale dynamics, using our compartmental approach with $n = 1{,}000$ infected compartments (blue), and using the IDE method (black dashed), when the infectiousness profile is as in A. C. The expected error in the population-scale dynamics when the compartmental method is used, against the number of compartments, n, and the number of hosts, d. The red line indicates where the error is 10%. D–F. Equivalent figures to A–C, for $R_0 = 2$. G–I. Equivalent figures to A–C, for $R_0 = 3$.

R2.4 Discussion

In this research chapter, we have introduced a novel, easy-to-use, compartmental framework for nesting host-scale data in population-scale epidemic models. In the large-compartment limit, our method is mathematically equivalent to more complicated approaches that involve IDEs. However, our method has the advantage that it can be used straightforwardly, allowing it to be applied widely in future. We provided adaptable computing code alongside the publication from which this research chapter is adapted [113] in order to facilitate future use of our approach.

To illustrate our method, we considered the example of influenza A infection in humans. A viral dynamics model [17] was used to generate a synthetic dataset describing changing viral loads in a cohort of hosts, which is representative of real host-scale data (Figure R2.2). We showed how our compartmental framework can be used to project the population-scale epidemic dynamics, and compared our projections to corresponding estimates obtained using the more complicated K&M IDE model. The population-scale projections from our framework closely matched those obtained using the IDE model, provided that a sufficient number of compartments was employed in our approach (Figure R2.3B and Figure R2.4).

The amount of data used in modelling studies of host-scale influenza dynamics has varied widely, with some studies using data from fewer than 10 hosts [17] but others from more than 40 hosts [35]. While multi-scale epidemic models have often been parameterised using either no or limited data [51], drawing robust population-scale conclusions from cohort studies involving a small number of hosts is likely to be challenging, since host-scale dynamics display significant variability between different individuals [17, 35, 233]. We therefore assessed the errors that arise in projected population-scale dynamics as a result of limited host-scale data, as well as considering measurement errors that can beset parameter inference from host-scale data [180]. We first investigated these effects separately (Figures R2.5 and R2.6), before considering both these effects in a single combined analysis (Figure R2.7). When host-scale data were collected once daily, we found that data from at least 40 hosts were required for a mean population-scale error of 10% or smaller when either our compartmental approach or the IDE method was used (Figure R2.7A and Figure A2.2C). However, since the precise value of the population-scale error depended on the exact subset of hosts that was included in the study, the error could be either larger or smaller than 10% even when data were available from 40 hosts (Figure R2.7A). As a result, larger numbers of hosts can increase the confidence that the error is below a pre-specified threshold value (Figure R2.7A). We considered daily measurements of pathogen load, since this frequency of data acquisition is

common to a number of previous longitudinal studies of influenza infection (e.g. [17, 35]). However, the accuracy of population-scale projections depends on the frequency with which data are collected (see Section R2.3.5.3), so regular data collection from each host in future cohort studies is important for ensuring that population-scale projections are as accurate as possible.

Our approach was motivated by earlier studies in which compartmental models with multiple latent and infectious stages were employed so that the standard assumption of exponentially distributed latent and infectious periods was relaxed (i.e. the method of stages) [23, 151, 152, 170, 237]. The use of multiple stages allows for gamma-distributed latent and infectious periods, and gamma distributions have been shown to characterise epidemiological periods accurately [151, 237]. However, in those studies [23, 151, 152, 170, 237], the level of infectiousness was assumed to be constant throughout the infectious period. We were therefore also inspired by previous research in which time-dependent infectiousness was incorporated into multi-stage compartmental models, in cases where the compartments correspond to clearly distinct phases of infection (for example, Figure R1.7 of our study of Ebola in Chapter R1, as well as a study of HIV [121]) or convenient time periods [58]. Our approach is more similar to a method used to include experimental data in models of plant disease [63], but differs from previous literature [58, 63, 112, 121] due to our use of a large number of infected compartments corresponding to different transmission rates in order to provide an easy method to transition from host- to population-scale that is accurate for any host-scale infection dynamics.

We were also motivated to develop a general framework for transitioning from host- to population-scale epidemic dynamics by our results in Chapter R1, where we demonstrated that population-scale Ebola models may need to account for details of host-scale dynamics to ensure accurate assessments of the effectiveness of interventions. While the focus of Chapter R1 was on variations in symptom expression during infection, rather than infectiousness as is the case here, combining the approaches of Chapters R1 and R2 by extending the multi-scale framework developed here to also account for variable symptoms would be an interesting area of future work. One possibility is to fit a within-host model, in which the level of symptom expression is included as a variable, to longitudinal host-scale data comprising both viral loads and symptom scores [35, 154]. Such an extension to our approach would enable the potential dependency of transmissibility on behavioural factors in addition to pathogen load (e.g. infected individuals staying at home once they develop symptoms or once symptom severity reaches a certain threshold) to be accounted for – see Chapter R3, where our best model indicates a decrease in the

transmission risk from individuals with SARS-CoV-2 infections immediately after developing symptoms.

While we sought to develop our framework for transitioning from host- to population-scale using the simplest possible population-scale model, our compartmental approach could be generalised, for example, to models incorporating age structure, spatial effects, social contact networks or stochasticity [131]. In an age-structured model, different within-host model parameter values (or even different models) could be used to describe host-scale dynamics in the different age groups, since there may be substantial differences in host-scale dynamics between individuals of different ages [120].

As well as developing a general multi-scale modelling framework, we chose to consider the case study of influenza A in humans because compartmental models are frequently used to model both host- and population-scale influenza dynamics, while there has also been significant interest in developing models linking the dynamics at the two scales [174]. In principle, however, our approach could be extended to model outbreaks of other pathogens for which host-scale dynamics are well-characterised. This would require careful consideration of the functional relationship between pathogen load and infectiousness, since this is likely to differ between pathogens [111]. In particular, the mode of transmission may be an important factor in determining suitable relationships for different pathogens [99]. Alternatively, if host-scale infection dynamics are not well-characterised, then an expected infectiousness profile that is estimated from transmission data [82, 182], rather than host-scale data, could also be embedded within our framework. In Chapters R3 and R4 of this thesis we focus on estimation of the infectiousness profile for SARS-CoV-2 using transmission data from infector-infectee pairs (Chapter R3) and households (Chapter R4). Combining the estimates in those two chapters with the modelling framework we have developed here is an avenue for future work.

To describe individual host-scale influenza dynamics we used the simple TCL within-host model. More detailed within-host models exist, and involve features including a delay before target cells begin to shed virus (an eclipse phase) [17] or explicit modelling of immune responses [19]. While the TCL model was sufficient to demonstrate our approach here, the expected infectiousness profile in our framework could be generated using a within-host model with any level of complexity. When we used the TCL model to generate synthetic host-scale data, we assumed that two within-host model parameters varied between hosts, using previous parameter estimates to determine the extent of parameter variability [17]. We incorporated

measurement error by adding a normally distributed random variate to daily observations of the logarithm of the viral load (although we also considered other frequencies of data collection in Section R2.3.5.3). Differences in both the extent of measurement error, and the extent of parameter variability between hosts, can lead to substantial differences in population-scale projection errors (Figures A2.3 and A2.4). Therefore, if our modelling framework is used to determine how many hosts should be included in future cohort studies, careful consideration of the measurement error and the variability in pathogen load time series between hosts is important.

We fitted the TCL model to the daily data from each host using a basic least-squares estimation approach, since the precise method of within-host model parameter inference is not central to our multi-scale modelling framework. However, it would be straightforward to extend our approach to consider different error structures and methods for fitting models to host-scale data. In particular, a non-linear mixed effects modelling approach – amounting to a partial pooling of the data between individuals – could be used. This would enable robust parameter estimation in a real dataset, particularly in settings in which the numbers of data points per host are small, and both the frequency and timing of data collection may vary between hosts [21, 217]. One interesting area of future observation is to use such a method to explore further whether or not there is an optimal balance between the number of hosts and the frequency of measurements per host, if total resources are limited (see the preliminary analysis exploring this potential trade-off in Figure R2.9).

In our main analyses, we made the common assumption that the infectiousness of an influenza-infected host is proportional to their viral load [35, 49, 108, 111, 174]. However, more complex possibilities could easily be incorporated into our framework. To demonstrate this, we analysed two alternative possibilities where infectiousness either scales with the logarithm of the viral load [49, 108] or saturates at high pathogen loads [9, 58] (Appendix A2.2 and Figure R2.10), in both cases obtaining similar results to those shown in Figure R2.7.

In summary, in this chapter we have developed a novel compartmental framework for nesting host-scale data in population-scale epidemiological models. We have demonstrated our easy-to-use approach in the context of influenza. Not only can our modelling approach be used to inform population-scale projections with data from individual hosts, but it can also be used to design cohort studies by determining which data need to be collected. As a result, clear communication between clinical epidemiologists who conduct cohort studies and epidemiological modellers will allow for optimal study design. Including host-scale dynamics in population-scale epidemiological models as proposed here has the potential to improve epidemic projections;

we hope that the simplicity of our approach will facilitate its use in a wide range of future outbreaks.

Chapter R3

High infectiousness immediately before COVID-19 symptom onset highlights the importance of contact tracing

We have published the research underlying this chapter in the journal *eLife* [116].

R3.1 Introduction

Data from SARS-CoV-2 infector-infectee pairs are used to estimate important epidemiological time intervals [14, 31, 74, 78, 84, 85, 93, 118, 147, 183, 198, 205] and to assess the relative contributions of presymptomatic and symptomatic transmission [31, 84, 85, 149, 226, 251]. Such data generally comprise symptom onset dates for each infector and infectee (in some cases supplemented with partial information about infection times, consisting of a range of possible exposure dates for infectors and/or infectees [85]; Figure R3.1). The distribution of the serial interval (the difference between the symptom onset times of the infector and infectee) can then be estimated from the data [78, 84, 183, 198]. Statistical models can also be used to infer the distributions of two other important epidemiological time intervals – the generation time (the difference between the infection times of the infector and infectee) [74, 84, 85, 93] and the time from onset of symptoms to transmission (TOST) [14, 84, 118] – from transmission pair data, and to estimate the proportion of presymptomatic transmissions [84, 85, 149, 226, 251].

The generation time and TOST distributions indicate the normalised average infectiousness profile of an infected host at each time since infection and time since symptom onset, respectively [88, 118] (so the probability density function of the generation time distribution is proportional to the quantity denoted $\beta(\tau)$ in the previous

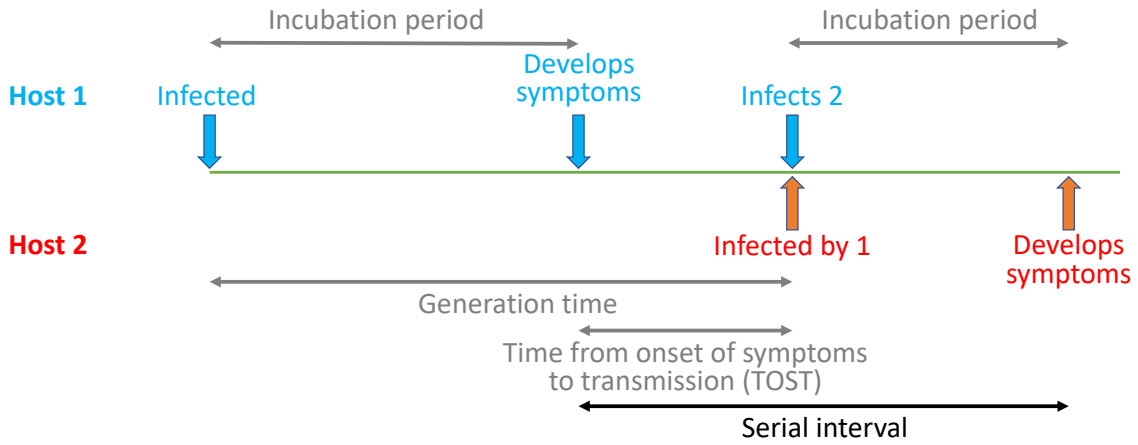


Figure R3.1: Schematic illustrating epidemiological time intervals in data from infector-infectee pairs. While the serial interval for each pair can be calculated directly from the data (with some uncertainty, given the unknown precise times of symptom appearance on the onset dates [219]), other time intervals, including the generation time and TOST, are unobserved (these are shown in grey).

research chapter, but scaled to have a total integral of one). These distributions are important for assessing the effectiveness of public health measures such as isolation [15, 240] and contact tracing [71, 85, 90]. Estimates of the SARS-CoV-2 generation time have typically involved an assumption that a host's infectiousness profile is independent of exactly when they develop symptoms [74, 85, 93, 137, 145]. However, such an assumption is unjustified [16, 145] and can lead to a poor fit to data [84].

In this research chapter, we develop a mechanistic approach for inferring key epidemiological quantities using data from infector-infectee pairs. This approach is motivated by compartmental epidemic models with gamma-distributed stage durations (such as the model we used in Chapter R1; see also, for example, [151, 237]) and changes in infectiousness during infection (such as our compartmental framework in Chapter R2; see also, for example, [5, 58, 94, 121]). As we show, our method provides an improved fit to data from SARS-CoV-2 transmission pairs compared to previous approaches, namely: (i) a model assuming that transmission and symptoms are independent [74, 85, 93, 137], and (ii) a previous statistical method in which this assumption is relaxed [84]. Under our best-fitting model, the proportion of presymptomatic transmissions is high, with many transmissions occurring in a short time window prior to symptom onset. We consider the implications of these results for contact tracing and isolation strategies.

R3.2 Methods

R3.2.1 Notation and general details

Here, we outline the notation used in this chapter when describing the different models that we considered. For a given transmission pair, we label the infector as 1 and the infectee as 2, and define:

$$t_{ik} = (\text{time of infection of host } k)\,, \quad k = 1, 2, \tag{R3.1}$$

$$t_{sk} = (\text{time of symptom onset of host } k)\,, \quad k = 1, 2, \tag{R3.2}$$

$$\tau_{inc,k} = (\text{incubation period of host } k)\,, \quad k = 1, 2, \tag{R3.3}$$

$$\tau_{gen} = (\text{generation time})\,, \tag{R3.4}$$

$$x_{tost} = (\text{TOST})\,, \tag{R3.5}$$

$$x_{ser} = (\text{serial interval})\,. \tag{R3.6}$$

In the above, t is used to denote calendar times, τ for time intervals relative to the time of infection, and x for time intervals relative to the time of symptom onset. We denote the probability density functions of the incubation period, generation time, TOST and serial interval as f_{inc}, f_{gen}, f_{tost} and f_{ser}, respectively, and use a capital F for the corresponding cumulative distribution functions.

In addition, we denote the expected infectiousness of an infected host at time τ since infection as $\beta(\tau)$, and the expected infectiousness at time x since symptom onset as $b(x)$. These infectiousness profiles are related to the generation time and TOST distributions, respectively, by

$$\beta(\tau) = \beta_0 f_{gen}(\tau)\,, \tag{R3.7}$$

$$b(x) = \beta_0 f_{tost}(x)\,. \tag{R3.8}$$

The exact value of the overall scaling factor, β_0, has no effect on our analyses in this chapter, since it simply adds a constant multiplicative factor to the likelihood function given below. We also let $\beta(\tau \mid \tau_{inc})$ and $b(x \mid \tau_{inc})$ be the expected infectiousness at time τ since infection and at time x since symptom onset, respectively, conditional on an incubation period of τ_{inc} (these are related by $\beta(\tau \mid \tau_{inc}) = b(\tau - \tau_{inc} \mid \tau_{inc})$ and $b(x \mid \tau_{inc}) = \beta(x + \tau_{inc} \mid \tau_{inc})$).

We considered several different models for infectiousness (details of individual models are given below). In each model, the conditional infectiousness, $\beta(\tau \mid \tau_{inc})$, or equivalently, $b(x \mid \tau_{inc})$, is specified. The distributions of the generation time and

TOST can be obtained from this conditional infectiousness by averaging over the incubation period distribution (which was assumed to be known):

$$\beta(\tau) = \beta_0 f_{gen}(\tau) = \int_0^\infty \beta(\tau \mid \tau_{inc}) f_{inc}(\tau_{inc}) \, \mathrm{d}\tau_{inc}, \tag{R3.9}$$

$$b(x) = \beta_0 f_{tost}(x) = \int_0^\infty b(x \mid \tau_{inc}) f_{inc}(\tau_{inc}) \, \mathrm{d}\tau_{inc}. \tag{R3.10}$$

Alternative (equivalent) expressions for the generation time and TOST distributions are available for some of the models considered (these are detailed below).

To obtain an expression for the serial interval distribution, note that the serial interval and TOST for a transmission pair are related by the equation

$$x_{ser} = x_{tost} + \tau_{inc,2}. \tag{R3.11}$$

We assumed throughout that x_{tost} and $\tau_{inc,2}$ are independent, so that the serial interval distribution is given by the convolution

$$f_{ser}(x_{ser}) = \int_0^\infty f_{tost}(x_{ser} - \tau_{inc}) f_{inc}(\tau_{inc}) \, \mathrm{d}\tau_{inc}. \tag{R3.12}$$

The proportion of presymptomatic transmissions (out of all transmissions generated by infected individuals who develop symptoms) can be calculated as

$$q_P = \int_{-\infty}^0 f_{tost}(x_{tost}) \, \mathrm{d}x_{tost}, \tag{R3.13}$$

although simpler equivalent expressions for individual models are also detailed later. An expression for the proportion of all transmissions that arise from individuals who are not displaying symptoms, accounting for both presymptomatic and asymptomatic transmissions, is derived in Section R3.2.10.

R3.2.2 Data

Following [84], we considered SARS-CoV-2 transmission pair data from five different studies [50, 85, 118, 248, 250], totalling 191 infector-infectee pairs (these data are provided alongside the publication from which this research chapter is adapted [116]). In all 191 transmission pairs, both the infector and the infectee developed symptoms, and the symptom onset date of each host was recorded. In four of the five studies [50, 85, 118, 248], intervals of exposure were available for either the infector or infectee (or both), whereas in the other study [250], only symptom onset dates were recorded.

R3.2.3 Incubation period

In our main analyses, the incubation period was assumed to follow a gamma distribution with shape parameter 5.807 and scale parameter 0.948 [143]. This corresponds to a mean incubation period of 5.5 days and a standard deviation of 2.3 days. However, to demonstrate that our main conclusions are robust to the exact incubation period distribution used, we also repeated our analyses using an alternative, more dispersed, gamma-distributed incubation period with a mean of 5.3 days and a standard deviation of 3.2 days [146] (Figures R3.9–R3.11).

R3.2.4 Independent transmission and symptoms model

R3.2.4.1 Model description

In this model, the infectiousness of each host at a given time since infection is assumed to be independent of their incubation period, so that

$$\beta\left(\tau \mid \tau_{inc}\right) = \beta\left(\tau\right) = \beta_0 f_{gen}\left(\tau\right), \tag{R3.14}$$

where the generation time distribution, f_{gen}, is prescribed. We assumed [85, 93] that

$$\tau_{gen} \sim \text{Gamma}\left(a, b\right), \tag{R3.15}$$

where a and b are shape and scale parameters, respectively, so that the mean generation time is $m_{gen} = ab$ and the standard deviation of generation times is $s_{gen} = a^{1/2}b$.

R3.2.4.2 TOST distribution

The TOST distribution for this model is given by

$$f_{tost}\left(x_{tost}\right) = \frac{1}{\beta_0}\int_0^\infty b\left(x_{tost} \mid \tau_{inc}\right) f_{inc}\left(\tau_{inc}\right) \mathrm{d}\tau_{inc} \tag{R3.16}$$

$$= \frac{1}{\beta_0}\int_0^\infty \beta\left(x_{tost} + \tau_{inc} \mid \tau_{inc}\right) f_{inc}\left(\tau_{inc}\right) \mathrm{d}\tau_{inc} \tag{R3.17}$$

$$= \int_0^\infty f_{gen}\left(x_{tost} + \tau_{inc}\right) f_{inc}\left(\tau_{inc}\right) \mathrm{d}\tau_{inc}. \tag{R3.18}$$

Alternatively, this formula can be derived by noting that

$$x_{tost} = \tau_{gen} - \tau_{inc,1}. \tag{R3.19}$$

In this model, τ_{gen} and $\tau_{inc,1}$ are assumed to be independent, so the TOST distribution is therefore given by the convolution of the distributions of τ_{gen} and $-\tau_{inc,1}$.

R3.2.4.3 Proportion of presymptomatic transmissions

The proportion of transmissions occurring before symptom onset (among infectors who eventually develop symptoms) is

$$q_P = \int_{-\infty}^{0} f_{tost}\left(x_{tost}\right) \mathrm{d}x_{tost} \tag{R3.20}$$

$$= \int_{-\infty}^{0} \int_{0}^{\infty} f_{gen}\left(x_{tost} + \tau_{inc}\right) f_{inc}\left(\tau_{inc}\right) \mathrm{d}\tau_{inc} \mathrm{d}x_{tost} \tag{R3.21}$$

$$= \int_{0}^{\infty} \int_{\tau_{gen}}^{\infty} f_{gen}\left(\tau_{gen}\right) f_{inc}\left(\tau_{inc}\right) d\tau_{inc} \mathrm{d}\tau_{gen} \tag{R3.22}$$

$$= \int_{0}^{\infty} f_{gen}\left(\tau_{gen}\right) \left(1 - F_{inc}\left(\tau_{gen}\right)\right) \mathrm{d}\tau_{gen}. \tag{R3.23}$$

R3.2.4.4 Prior distributions for estimated parameters

The vector of unknown model parameters, $\theta = (m_{gen}, s_{gen})$, was estimated when we fitted the model to the transmission pair data. When we used data augmentation MCMC to fit the model to the data (see Section R3.2.8), we assumed independent lognormal prior distributions for each parameter. The prior distributions for the mean, m_{gen}, and standard deviation, s_{gen}, of the generation time distribution had medians of 5 days and 2 days, respectively (these choices were informed by previous estimates of the SARS-CoV-2 generation time distribution [84, 85, 93]), and were chosen to ensure a prior probability of only 0.025 that these parameters exceeded the very high values of 10 days and 7 days, respectively. The exact prior distributions we used are detailed in Table A3.2.

R3.2.5 Ferretti model

R3.2.5.1 Model description

In this model, which was first proposed in [84], the conditional infectiousness is specified as the re-scaled skew-logistic distribution,

$$b\left(x \mid \tau_{inc}\right) = \begin{cases} \dfrac{C_F \beta_0 \exp\left[-\left(\dfrac{x m_{inc}}{\tau_{inc}} - \mu_F\right) / \sigma_F\right]}{\left(1 + \exp\left[-\left(\dfrac{x m_{inc}}{\tau_{inc}} - \mu_F\right) / \sigma_F\right]\right)^{\alpha_F + 1}}, & -\tau_{inc} \leq x < 0, \\ \dfrac{C_F \beta_0 \exp\left[-\left(x - \mu_F\right) / \sigma_F\right]}{\left(1 + \exp\left[-\left(x - \mu_F\right) / \sigma_F\right]\right)^{\alpha_F + 1}}, & x \geq 0. \end{cases} \tag{R3.24}$$

Here, m_{inc} is the mean incubation period, and μ_F, σ_F and α_F are model parameters that do not have straightforward epidemiological interpretations. The scaling factor,

C_F, in this expression is derived below. Note that for this model, straightforward expressions are not available for the TOST and generation time distributions, so we used the general formulae given in Section R3.2.1 (equations (R3.9) and (R3.10)) to calculate these distributions (numerically).

R3.2.5.2 Scaling factor in conditional infectiousness

To derive the correct scaling factor, C_F, in the conditional infectiousness, note that we require

$$\int_{-\infty}^{\infty} f_{tost}(x)\,\mathrm{d}x = \int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{1}{\beta_0} b(x \mid \tau_{inc})\, f_{inc}(\tau_{inc})\,\mathrm{d}\tau_{\mathrm{inc}}\mathrm{dx} = 1. \tag{R3.25}$$

Now,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{1}{\beta_0} b(x \mid \tau_{inc})\,\mathrm{d}x \\
&= \int_{-\tau_{inc}}^{0} \frac{C_F \exp\left[-\left(\frac{x m_{inc}}{\tau_{inc}} - \mu_F\right)/\sigma_F\right]}{\left(1+\exp\left[-\left(\frac{x m_{inc}}{\tau_{inc}} - \mu_F\right)/\sigma_F\right]\right)^{\alpha_F+1}}\mathrm{d}x \\
&\qquad + \int_{0}^{\infty} \frac{C_F \exp\left[-(x-\mu_F)/\sigma_F\right]}{\left(1+\exp\left[-(x-\mu_F)/\sigma_F\right]\right)^{\alpha_F+1}}\mathrm{d}x \qquad \text{(R3.26)} \\
&= \frac{C_F\sigma_F}{\alpha_F}\left[1-\left(1+e^{\mu_F/\sigma_F}\right)^{-\alpha_F}\right. \\
&\qquad \left. + \frac{\tau_{inc}}{m_{inc}}\left(\left(1+e^{\mu_F/\sigma_F}\right)^{-\alpha_F} - \left(1+e^{(m_{inc}+\mu_F)/\sigma_F}\right)^{-\alpha_F}\right)\right]. \qquad \text{(R3.27)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{1}{\beta_0} b(x \mid \tau_{inc})\, f_{inc}(\tau_{inc})\,\mathrm{d}\tau_{inc}\mathrm{d}x & \text{(R3.28)} \\
&= \int_{0}^{\infty}\left(\int_{-\infty}^{\infty} \frac{1}{\beta_0} b(x \mid \tau_{inc})\,\mathrm{d}x\right) f_{inc}(\tau_{inc})\,\mathrm{d}\tau_{inc} & \text{(R3.29)} \\
&= \frac{C_F\sigma_F}{\alpha_F}\left[1-\left(1+e^{(m_{inc}+\mu_F)/\sigma_F}\right)^{-\alpha_F}\right], & \text{(R3.30)}
\end{aligned}$$

so we have

$$C_F = \frac{\alpha_F}{\sigma_F\left(1-\left(1+e^{(m_{inc}+\mu_F)/\sigma_F}\right)^{-\alpha_F}\right)}. \tag{R3.31}$$

R3.2.5.3 Proportion of presymptomatic transmissions

The proportion of presymptomatic transmissions is given by

$$q_P = \int_{-\infty}^{0} f_{tost}(x)\,\mathrm{d}x \tag{R3.32}$$

$$= \int_0^\infty \int_{-\infty}^0 \frac{1}{\beta_0} b\left(x \mid \tau_{inc}\right) f_{inc}\left(\tau_{inc}\right) \mathrm{d}x \mathrm{d}\tau_{inc} \tag{R3.33}$$

$$= \int_0^\infty \frac{C_F \sigma_F \tau_{inc}}{\alpha_F m_{inc}} \left[\left(1 + e^{\mu_F/\sigma_F}\right)^{-\alpha_F} - \left(1 + e^{(m_{inc}+\mu_F)/\sigma_F}\right)^{-\alpha_F}\right] f_{inc}\left(\tau_{inc}\right) \mathrm{d}\tau_{inc} \tag{R3.34}$$

$$= \frac{\left(1 + e^{\mu_F/\sigma_F}\right)^{-\alpha_F} - \left(1 + e^{(m_{inc}+\mu_F)/\sigma_F}\right)^{-\alpha_F}}{1 - \left(1 + e^{(m_{inc}+\mu_F)/\sigma_F}\right)^{-\alpha_F}}. \tag{R3.35}$$

R3.2.5.4 Prior distributions for estimated parameters

The vector of unknown model parameters, $\theta = (\mu_F, \sigma_F, \alpha_F)$, was estimated when we fitted the model to the transmission pair data. We assumed independent (improper) uniform prior distributions for each model parameter when we used data augmentation MCMC to fit the model to the data (Section R3.2.8), imposing only that the values of σ_F and α_F were positive (whereas μ_F could take either positive or negative values).

R3.2.6 Mechanistic model

R3.2.6.1 Model description

In our mechanistic approach, each infection is divided into three stages: latent (E), presymptomatic infectious (P), and symptomatic infectious (I). The stage durations were assumed to be independent, and infectiousness was assumed to be constant over the duration of each stage. We denote the stage durations by $y_{E/P/I}$, their density and cumulative distribution functions by $f_{E/P/I}$ and $F_{E/P/I}$, and the transmission rates of hosts in the P and I stages by $\beta_{P/I}$, respectively. We define

$$\alpha_P = \beta_P / \beta_I \tag{R3.36}$$

to be the ratio of transmission rates in the P and I stages. In this model, the expected number of transmissions generated by each infected host is

$$\beta_0 = \beta_P m_P + \beta_I m_I, \tag{R3.37}$$

where $m_{P/I}$ are the respective mean durations of the P and I stages.

We further assumed that the durations of each stage follow gamma distributions, with

$$y_E \sim \text{Gamma}\left(k_E, \frac{1}{k_{inc}\gamma}\right), \tag{R3.38}$$

$$y_P \sim \text{Gamma}\left(k_P, \frac{1}{k_{inc}\gamma}\right), \tag{R3.39}$$

$$y_I \sim \text{Gamma}\left(k_I, \frac{1}{k_I \mu}\right), \tag{R3.40}$$

where

$$k_{inc} = k_E + k_P. \tag{R3.41}$$

In particular, the scale parameters of y_E and y_P were both assumed to be equal to $1/(k_{inc}\gamma)$, in order to ensure a gamma-distributed incubation period,

$$\tau_{inc} = y_E + y_P \sim \text{Gamma}\left(k_{inc}, \frac{1}{k_{inc}\gamma}\right). \tag{R3.42}$$

We fixed $k_{inc} = 5.807$ and $\gamma = 1/\left(5.807 \times 0.948\right)$, in order to obtain the specified incubation period distribution (see Section R3.2.3 above). When we fitted this model to transmission pair data, we assumed that $k_I = 1$, so that the symptomatic infectious period follows an exponential distribution. The (combinations of) parameters k_E/k_{inc} (representing the ratio of the mean latent (E) period and the mean incubation (combined E and P) period) and $1/\mu$ (representing the mean symptomatic infectious (I) period) were estimated when we fitted the model to data. We considered two versions of the model: one in which we assumed $\alpha_P = 1$ (the "constant infectiousness model"), and one in which α_P was also estimated (the "variable infectiousness model").

R3.2.6.2 Conditional infectiousness

Conditional on an incubation period of length τ_{inc}, the expected infectiousness of an infected host at time x since symptom onset is given by

$$b\left(x \mid \tau_{inc}\right) = \begin{cases} \beta_P \times p\left(Y_P \geq -x \mid Y_E + Y_P = \tau_{inc}\right), & -\tau_{inc} \leq x < 0, \\ \beta_I \times p\left(Y_I \geq x\right), & x \geq 0, \end{cases} \tag{R3.43}$$

where we here explicitly distinguish the random variables $Y_{E/P/I}$ from their observed values $y_{E/P/I}$ (i.e. the lengths of each stage of infection). Now,

$$\begin{aligned} &p\left(Y_P \geq -x \mid Y_E + Y_P = \tau_{inc}\right) \\ &\quad = \int_{-x}^{\infty} p\left(Y_P = y_P \mid Y_E + Y_P = \tau_{inc}\right) \mathrm{d}y_P && \text{(R3.44)} \\ &\quad = \int_{-x}^{\infty} \frac{p\left(Y_E + Y_P = \tau_{inc} \mid Y_P = y_P\right) p\left(Y_P = y_P\right)}{p\left(Y_E + Y_P = \tau_{inc}\right)} \mathrm{d}y_P && \text{(R3.45)} \\ &\quad = \int_{-x}^{\infty} \frac{f_E\left(\tau_{inc} - y_P\right) f_P(y_P)}{f_{inc}\left(\tau_{inc}\right)} \mathrm{d}y_P, && \text{(R3.46)} \end{aligned}$$

where we used Bayes' rule to obtain the second equality. For the special case of the gamma-distributed stage durations considered, we have that

$$\frac{f_E\left(\tau_{inc} - y_P\right) f_P(y_P)}{f_{inc}\left(\tau_{inc}\right)} = \frac{1}{\tau_{inc}} f_{Beta}\left(y_P/\tau_{inc}; k_P, k_E\right), \tag{R3.47}$$

where $f_{Beta}(x; a, b)$ is the probability density function of a beta-distributed random variable with shape parameters a and b (we denote the corresponding cumulative distribution function by $F_{Beta}(s; a, b)$). Therefore,

$$p(Y_P \geq -x \mid Y_E + Y_P = \tau_{inc}) = 1 - F_{Beta}(-x/\tau_{inc}; k_P, k_E), \tag{R3.48}$$

and so

$$b(x \mid \tau_{inc}) = \begin{cases} \beta_P (1 - F_{Beta}(-x/\tau_{inc}; k_P, k_E)), & -\tau_{inc} \leq x < 0, \\ \beta_I (1 - F_I(x)), & x \geq 0. \end{cases} \tag{R3.49}$$

$$= \begin{cases} \alpha_P C \beta_0 (1 - F_{Beta}(-x/\tau_{inc}; k_P, k_E)), & -\tau_{inc} \leq x < 0, \\ C\beta_0 (1 - F_I(x)), & x \geq 0, \end{cases} \tag{R3.50}$$

where

$$C = \frac{\beta_I}{\beta_0} = \frac{\beta_I}{\beta_P m_P + \beta_I m_I} = \frac{1}{\alpha_P m_P + m_I} \tag{R3.51}$$

$$= \frac{1}{\left(\frac{\alpha_P k_P}{k_{inc}\gamma} + \frac{1}{\mu}\right)} = \frac{k_{inc}\gamma\mu}{\alpha_P k_P \mu + k_{inc}\gamma}. \tag{R3.52}$$

R3.2.6.3 Generation time distribution

The generation time for an individual transmission can be written as

$$\tau_{gen} = y_E + y^*, \tag{R3.53}$$

where y^* is the time between the start of the P stage and the transmission occurring (we denote the density of this time interval by $f^*(y^*)$), and therefore the generation time distribution is given by the convolution

$$f_{gen}(\tau_{gen}) = \int_0^{\tau_{gen}} f^*(\tau_{gen} - y_E) f_E(y_E) \, \mathrm{d}y_E. \tag{R3.54}$$

To obtain an expression for $f^*(y^*)$, which can then be used to calculate the generation time distribution using the equation above, note that the expected infectiousness of an infected host at time y^* since the start of the P stage is equal to

$$b^*(y^*) = \beta_P \times p(Y_P \geq y^*) + \beta_I \times p(Y_P \leq y^*, Y_P + Y_I \geq y^*). \tag{R3.55}$$

The second probability can be evaluated by conditioning on the value of Y_P, to obtain

$$\begin{aligned}
&b^*(y^*) \\
&= \beta_P (1 - F_P(y^*)) \\
&\qquad + \beta_I \int_0^\infty p(Y_P \le y^*, Y_P + Y_I \ge y^* \mid Y_P = y_p) f_P(y_P) \, \mathrm{d}y_P \qquad \text{(R3.56)} \\
&= \beta_P (1 - F_P(y^*)) + \beta_I \int_0^{y^*} p(Y_I \ge y^* - y_P \mid Y_P = y_p) f_P(y_P) \, \mathrm{d}y_P \qquad \text{(R3.57)} \\
&= \beta_P (1 - F_P(y^*)) + \beta_I \int_0^{y^*} (1 - F_I(y^* - y_P)) f_P(y_P) \, \mathrm{d}y_P. \qquad \text{(R3.58)}
\end{aligned}$$

Therefore,

$$f^*(y^*) = C \left(\alpha_P (1 - F_P(y^*)) + \int_0^{y^*} (1 - F_I(y^* - y_P)) f_P(y_P) \, \mathrm{d}y_P \right). \qquad \text{(R3.59)}$$

R3.2.6.4 TOST distribution

The expected infectiousness of an infected host at time x since symptom onset is given by

$$b(x) = \begin{cases} \beta_P \times p(Y_P \ge -x), & x < 0, \\ \beta_I \times p(Y_I \ge x), & x \ge 0. \end{cases} \qquad \text{(R3.60)}$$

Therefore,

$$f_{tost}(x_{tost}) = \frac{1}{\beta_0} b(x_{tost}) = \begin{cases} \alpha_P C (1 - F_P(-x_{tost})), & x_{tost} < 0, \\ C (1 - F_I(x_{tost})), & x_{tost} \ge 0. \end{cases} \qquad \text{(R3.61)}$$

R3.2.6.5 Proportion of presymptomatic transmissions

The proportion of presymptomatic transmissions is

$$\begin{aligned}
q_P &= \frac{\beta_P m_P}{\beta_0} = \frac{\beta_P m_P}{\beta_P m_P + \beta_I m_I} = \frac{\alpha_P m_P}{\alpha_P m_P + m_I} \qquad \text{(R3.62)} \\
&= \frac{\left(\frac{\alpha_P k_P}{k_{inc} \gamma} \right)}{\left(\frac{\alpha_P k_P}{k_{inc} \gamma} + \frac{1}{\mu} \right)} = \frac{\alpha_P k_P \mu}{\alpha_P k_P \mu + k_{inc} \gamma}. \qquad \text{(R3.63)}
\end{aligned}$$

R3.2.6.6 Prior distributions for estimated parameters

The vector of unknown model parameters, $\theta = (k_E/k_{inc}, 1/\mu)$, was estimated when we fitted the constant infectiousness model to the transmission pair data, while the corresponding vector of estimated model parameters for the variable infectiousness

model was $\theta = (k_E/k_{inc}, 1/\mu, \alpha_P)$. In both cases, we assumed independent prior distributions for each entry of θ when we used data augmentation MCMC to fit the model to the data (Section R3.2.8). A lognormal prior distribution was used for $1/\mu$ and a beta prior distribution for k_E/k_{inc} (which was constrained to lie between zero and one). The prior distribution for k_E/k_{inc} was chosen to limit the prior probability of extreme values, with a prior 95% credible interval of $[0.1, 0.9]$. For the variable infectiousness model, since α_P represents the ratio between the transmission rates of hosts in the P and I stages, we assumed a lognormal prior distribution with median one for this parameter to ensure equal prior probabilities of values above and below one. This prior distribution was also chosen to limit the prior probability of extreme values, with a prior 95% credible interval of $[0.2, 4.8]$. The exact prior distributions we used are given in Table A3.2.

R3.2.7 Likelihood

Here, we derive the likelihood function that we used to fit each of the four models of infectiousness described above to the transmission pair data. For a single transmission pair (denoted n), conditional on the parameters, θ, of the model of infectiousness under consideration, the joint probability density that all of the following events occur:

i. host 1 (the infector) is infected in the time interval $[t_{i1,L}, t_{i1,R}]$ (where this interval may be infinitely wide);

ii. host 1 transmits the pathogen to host 2 (we write $1 \to 2$ to denote the occurrence of the transmission);

iii. the transmission from host 1 to host 2 occurs in the time interval $[t_{i2,L}, t_{i2,R}]$ (where again this interval may be infinitely wide);

iv. hosts 1 and 2 develop symptoms at times t_{s1} and t_{s2}, respectively (note that when we used data augmentation MCMC to fit the models to data, we augmented observed symptom onset dates with estimated precise times of symptom onset, as described in Section R3.2.8 below);

is given by

$$p\left(1 \to 2, t_{s1}, t_{s2},\ [t_{i1,L}, t_{i1,R}], [t_{i2,L}, t_{i2,R}] \mid \theta\right) \tag{R3.64}$$

$$= \int_{t_{i2,L}}^{t_{i2,R}} \int_{t_{i1,L}}^{t_{i1,R}} p\left(1 \to 2, t_{i1}, t_{s1}, t_{i2}, t_{s2} \mid \theta\right) \mathrm{d}t_{i1} \mathrm{d}t_{i2} \tag{R3.65}$$

$$= \int_{t_{i2,L}}^{t_{i2,R}} \int_{t_{i1,L}}^{t_{i1,R}} p\left(1 \to 2, t_{i2}, t_{s2} \mid t_{i1}, t_{s1}, \theta\right) p\left(t_{i1}, t_{s1} \mid \theta\right) \mathrm{d}t_{i1} \mathrm{d}t_{i2} \tag{R3.66}$$

$$= \int_{t_{i2,L}}^{t_{i2,R}} \int_{t_{i1,L}}^{t_{i1,R}} p\left(t_{s2} \mid 1 \to 2, t_{i1}, t_{s1}, t_{i2}, \theta\right) p\left(1 \to 2, t_{i2} \mid t_{i1}, t_{s1}, \theta\right) p\left(t_{i1}, t_{s1} \mid \theta\right) \mathrm{d}t_{i1} \mathrm{d}t_{i2} \tag{R3.67}$$

$$= \int_{t_{i2,L}}^{t_{i2,R}} \int_{t_{i1,L}}^{t_{i1,R}} p\left(1 \to 2, t_{i2} \mid t_{i1}, t_{s1}, \theta\right) p\left(t_{s1} \mid t_{i1}, \theta\right) p\left(t_{i1} \mid \theta\right) p\left(t_{s2} \mid t_{i2}, \theta\right) \mathrm{d}t_{i1} \mathrm{d}t_{i2}. \tag{R3.68}$$

We note that

$$p\left(1 \to 2, t_{i2} \mid t_{i1}, t_{s1}, \theta\right) \propto b\left(t_{i2} - t_{s1} \mid t_{s1} - t_{i1}, \theta\right), \tag{R3.69}$$

where the dependence of the conditional expected infectiousness, $b\left(x \mid \tau_{inc}, \theta\right)$, on the model parameters, θ, is indicated explicitly here. This is because the left-hand side gives the probability density of a transmission from 1 to 2 occurring at time t_{i2}, conditioned on the infection and onset times of 1, and is therefore proportional to the conditional infectiousness, $b\left(x_{tost} \mid \tau_{inc}, \theta\right)$. We also have that

$$p\left(t_{sk} \mid t_{ik}, \theta\right) = f_{inc}\left(t_{sk} - t_{ik}\right), \tag{R3.70}$$

for $k = 1, 2$. In an exponentially growing epidemic with growth rate r, the term $p\left(t_{i1} \mid \theta\right)$ will introduce a factor proportional to $e^{rt_{i1}}$ into the likelihood [85], although we neglect this correction here (note that our maximum likelihood estimates for the Ferretti model (Table A3.3) are similar to those obtained in [84], where the authors also fitted the Ferretti model to the same dataset but with this correction included).

When only transmission pair n is observed, the likelihood of the parameters, θ, of the model of infectiousness under consideration is given by

$$L^{(n)}\left(\theta\right) = \frac{1}{\beta_0} \int_{t_{i2,L}}^{t_{i2,R}} \int_{t_{i1,L}}^{t_{i1,R}} b\left(t_{i2} - t_{s1} \mid t_{s1} - t_{i1}, \theta\right) f_{inc}\left(t_{s1} - t_{i1}\right) f_{inc}\left(t_{s2} - t_{i2}\right) \mathrm{d}t_{i1} \mathrm{d}t_{i2}. \tag{R3.71}$$

Here, we added the constant factor $1/\beta_0$ for convenience (to cancel the factor of β_0 in the expression for the conditional infectiousness in each model), although note that, in general,

$$\frac{1}{\beta_0} \int_{-\infty}^{\infty} b\left(x_{tost} \mid \tau_{inc}, \theta\right) \mathrm{d}x_{tost} \tag{R3.72}$$

may not be equal to 1, since the expected number of transmissions generated by an infected host may depend on their incubation period.

Assuming each transmission pair in our dataset to be independent, the overall likelihood is therefore given by the product of the contributions, $L^{(n)}(\theta)$, from each individual transmission pair, i.e.

$$L(\theta) = \prod_{n=1}^{N} L^{(n)}(\theta), \tag{R3.73}$$

where N is the total number of transmission pairs.

R3.2.7.1 Calculation of the likelihood

We calculated the likelihood function described above by constraining the symptom onset time, t_s, of each host (which was imputed as part of the parameter estimation procedure described in Section R3.2.8 below) to lie on the grid

$$[t_{s,L} + \delta t, t_{s,L} + 2\delta t, \ldots, t_{s,L} + 1], \tag{R3.74}$$

where $t_{s,L}$ is the start of the day of onset for that host, and we took $\delta t = 0.125$ days. The contribution to the likelihood from each transmission pair, $L^{(n)}\left(\theta; t_{s1}^{(n)}, t_{s2}^{(n)}\right)$, was then calculated by discretising the integrals in equation (R3.71), with the infection time, t_i, of a given host constrained to the grid

$$\left[t_{i,L} + \frac{\delta t}{2}, \ldots, t_{i,R} - \frac{\delta t}{2}\right], \tag{R3.75}$$

where $t_{i,L}$ and $t_{i,R}$ are lower and upper bounds for the infection time of that host, respectively. Where no explicit lower bound for an individual's infection time was available in the data (note that the requirement for each individual to be infected before they developed symptoms always gave an upper bound, i.e. $t_{i,R} \leq t_{s,L}$), we assumed a maximum possible incubation period of 30 days to ensure a finite lower bound. Different discretisations were used for the infection and symptom onset times, both to avoid conditioning on an incubation period of zero days (since the conditional infectiousness may be undefined in this case) and to avoid the possibility of transmissions occurring at the exact time of symptom onset (since the infectiousness profile was allowed to be discontinuous at the onset time in our mechanistic model).

R3.2.8 Parameter estimation

R3.2.8.1 Overview

To account for uncertainty in the exact symptom onset times within the date of onset (and so avoid imparting bias by fitting continuous-time models to discrete-time

symptom onset data), we fitted the models to the data using data augmentation MCMC [31, 40, 41, 82, 204, 219]. In alternating steps of the chain, we updated either the vector of model parameters, θ, or the (estimated) precise symptom onset times of each infector and infectee. The chain was run for 2,500,000 steps, of which the first 500,000 were discarded as burn-in. Posterior distributions of parameters were obtained by recording only every 100 iterations of the chain. Point parameter estimates (posterior medians) were used to predict the distributions of the generation time, TOST and serial interval for each model.

R3.2.8.2 Details of data augmentation MCMC algorithm

We denote the vectors of symptom onset times for each infector and infectee by $\boldsymbol{t}_{s1}$ and $\boldsymbol{t}_{s2}$, and the corresponding likelihood by

$$L\left(\theta; \boldsymbol{t}_{s1}, \boldsymbol{t}_{s2}\right) = \prod_{n=1}^{N} L^{(n)}\left(\theta; t_{s1}^{(n)}, t_{s2}^{(n)}\right). \tag{R3.76}$$

In this expression, $L^{(n)}\left(\theta; t_{s1}^{(n)}, t_{s2}^{(n)}\right)$ is the contribution to the likelihood from transmission pair n, and $t_{s1}^{(n)}$ and $t_{s2}^{(n)}$ are the symptom onset times of the corresponding infector and infectee (i.e. the n^{th} entries of $\boldsymbol{t}_{s1}$ and $\boldsymbol{t}_{s2}$, respectively). We denote the prior distribution of θ by $\pi(\theta)$ (where the exact prior distributions we used for each model are described above and in Table A3.2).

We define the proposal distributions,

$$Q_1\left(\theta_{prop} \mid \theta\right), \qquad Q_2^{(n)}\left(t_{s1,prop}^{(n)}, t_{s2,prop}^{(n)} \mid t_{s1}^{(n)}, t_{s2}^{(n)}\right), \tag{R3.77}$$

which were taken to be symmetric, i.e.

$$Q_1\left(\theta_{prop} \mid \theta\right) = Q_1\left(\theta \mid \theta_{prop}\right), \tag{R3.78}$$

$$Q_2^{(n)}\left(t_{s1,prop}^{(n)}, t_{s2,prop}^{(n)} \mid t_{s1}^{(n)}, t_{s2}^{(n)}\right) = Q_2^{(n)}\left(t_{s1}^{(n)}, t_{s2}^{(n)} \mid t_{s1,prop}^{(n)}, t_{s2,prop}^{(n)}\right). \tag{R3.79}$$

The exact proposal distributions we used are detailed in Section R3.2.8.4. The data augmentation MCMC algorithm that we used is given in Algorithm R3.1.

R3.2.8.3 Initial values

For each model we considered, the initial parameter values, θ_0, were chosen arbitrarily. The initial symptom onset times, $t_{s1,0}$ and $t_{s2,0}$, were uniformly and independently sampled on the grid of possible symptom onset times for each host (see equation (R3.74)).

Algorithm R3.1: The data augmentation MCMC algorithm used to fit the models of infectiousness to the transmission pair data.

Initialise $\theta = \theta_0$, $\boldsymbol{t}_{s1} = \boldsymbol{t}_{s1,0}$ and $\boldsymbol{t}_{s2} = \boldsymbol{t}_{s2,0}$;
Calculate $L_0^{(n)} = L^{(n)}\left(\theta_0; t_{s1,0}^{(n)}, t_{s2,0}^{(n)}\right)$ for $n = 1, \ldots, N$;
Calculate $L_0 = \prod_{n=1}^{N} L_0^{(n)}$;
Calculate $\pi_0 = \pi(\theta_0)$;
for $m = 1, \ldots, M$ **do**
  if *m is odd* **then**
    Sample θ_{prop} from $Q_1\left(\theta_{prop} \mid \theta_{m-1}\right)$;
    Set $\boldsymbol{t}_{s1,m} = \boldsymbol{t}_{s1,(m-1)}$ and $\boldsymbol{t}_{s2,m} = \boldsymbol{t}_{s2,(m-1)}$;
    Calculate $L_{prop}^{(n)} = L^{(n)}\left(\theta_{prop}; t_{s1,m}^{(n)}, t_{s2,m}^{(n)}\right)$ for $n = 1, \ldots, N$;
    Calculate $L_{prop} = \prod_{n=1}^{N} L_{prop}^{(n)}$;
    Calculate $\pi_{prop} = \pi(\theta_{prop})$;
    Generate a random number, r, uniformly distributed between 0 and 1;
    if $r \leq \pi_{prop} L_{prop} / (\pi_{m-1} L_{m-1})$ **then**
      Set $\theta_m = \theta_{prop}$, $L_m^{(n)} = L_{prop}^{(n)}$ for each n, $L_m = L_{prop}$, and $\pi_m = \pi_{prop}$;
    else
      Set $\theta_m = \theta_{m-1}$, $L_m^{(n)} = L_{m-1}^{(n)}$ for each n, $L_m = L_{m-1}$, and $\pi_m = \pi_{m-1}$;
    end
  else
    Set $\theta_m = \theta_{m-1}$ and $\pi_m = \pi_{m-1}$;
    for $n = 1, \ldots, N$ **do**
      Sample $t_{s1,prop}^{(n)}$ and $t_{s2,prop}^{(n)}$ from
      $Q_2^{(n)}\left(t_{s1,prop}^{(n)}, t_{s2,prop}^{(n)} \mid t_{s1,(m-1)}^{(n)}, t_{s2,(m-1)}^{(n)}\right)$;
      Calculate $L_{prop}^{(n)} = L^{(n)}\left(\theta_m; t_{s1,prop}^{(n)}, t_{s2,prop}^{(n)}\right)$;
      Generate a random number, r, uniformly distributed between 0 and 1;
      if $r \leq L_{prop}^{(n)} / L_{m-1}^{(n)}$ **then**
        Set $t_{s1,m}^{(n)} = t_{s1,prop}^{(n)}$, $t_{s2,m}^{(n)} = t_{s2,prop}^{(n)}$, and $L_m^{(n)} = L_{prop}^{(n)}$;
      else
        Set $t_{s1,m}^{(n)} = t_{s1,(m-1)}^{(n)}$, $t_{s2,m}^{(n)} = t_{s2,(m-1)}^{(n)}$, and $L_m^{(n)} = L_{m-1}^{(n)}$;
      end
    end
    Calculate $L_m = \prod_{n=1}^{N} L_m^{(n)}$;
  end
end

R3.2.8.4 Proposal distributions

The proposal distribution $Q_1(\theta_{prop} \mid \theta)$ was taken to be multivariate normal with mean vector θ and covariance matrix Σ. The covariance matrix, Σ (note that Σ was fixed, i.e. this matrix did not depend on the values of θ and θ_{prop}), was tuned (by running the parameter fitting procedure for a smaller number of iterations than used in our final analyses, with different choices of Σ) as follows: the diagonal entries of Σ were chosen to ensure that the ratios between the standard deviations of proposal distributions for individual model parameters were approximately equal to the corresponding ratios between the standard deviations of posterior distributions. Each off-diagonal entry of Σ was only set to a non-zero value when the posterior distributions indicated a correlation of at least 0.5 in absolute value between the parameter pair corresponding to that entry of Σ, in which case the correlation in the proposal distributions was chosen to approximately match the correlation in the posterior distributions of the parameter pair – this was the case for the pairs $(k_E/k_{inc}, \alpha)$ and $(1/\mu, \alpha)$ in the variable infectiousness model, and for each parameter pair in the Ferretti model. The overall scaling of Σ was then chosen to ensure an overall acceptance rate of approximately 30% (over MCMC steps in which θ was updated).

We sampled the proposed symptom onset times for each host, $t^{(n)}_{s1,prop}$ and $t^{(n)}_{s2,prop}$, uniformly on the grid of possible onset times for that host (see equation (R3.74)), and independently of both the corresponding times in the previous step of the chain, and the onset times of all other hosts.

R3.2.9 Maximum likelihood estimation and goodness of fit

In order to provide a straightforward comparison of the goodness of fit between models, we also determined the parameter values, $\hat{\theta}$, that maximised the likelihood, $L(\theta)$, for each model, under the assumption that each infected host developed symptoms exactly in the middle of their recorded onset date (note that the prior distributions detailed for each model above and in Table A3.2 were not used when maximising the likelihood). The Akaike information criterion (AIC) for each model could then be calculated as [4]

$$\text{AIC} = 2 \times (\text{number of estimated parameters}) - 2 \log\left(L\left(\hat{\theta}\right)\right), \tag{R3.80}$$

where three parameters were estimated for the variable infectiousness model and the Ferretti model, and two parameters were estimated for the constant infectiousness model and the independent transmission and symptoms model. Lower AIC values

indicate a better fit to the data, but with a penalty for models with a higher number of fitted parameters. We calculated the ΔAIC value for each model as the difference between the AIC values for that model and for the model with the lowest AIC value. Since the maximum likelihood estimates, $\hat{\theta}$, did not account for uncertainty in exact symptom onset times, they were not used elsewhere in our analyses (however, the maximum likelihood estimates all lay within the credible intervals obtained in the more complex data augmentation MCMC parameter fitting procedure – see Tables A3.2 and A3.3).

R3.2.10 Total proportion of non-symptomatic transmissions accounting for asymptomatic cases

Here, we derive an expression for the total proportion of non-symptomatic transmissions once entirely asymptomatic cases are accounted for, and describe how this expression was used in our analyses. The (instantaneous) reproduction number, denoted R, can be decomposed as

$$R = p_A R_A + (1 - p_A)(R_P + R_I), \tag{R3.81}$$

where p_A is the proportion of completely asymptomatic cases, R_A is the expected number of secondary transmissions generated by each asymptomatic infected host, and $R_{P/I}$ are the expected numbers of transmissions generated before and after symptom onset, respectively, by a host who develops symptoms. The total proportion of non-symptomatic transmissions is then given by

$$\frac{p_A R_A + (1 - p_A) R_P}{R} = \frac{p_A R_A + (1 - p_A) R_P}{p_A R_A + (1 - p_A)(R_P + R_I)} \tag{R3.82}$$

$$= \frac{p_A \alpha_A + (1 - p_A) q_P}{p_A \alpha_A + (1 - p_A)}, \tag{R3.83}$$

where here

$$q_P = \frac{R_P}{R_P + R_I} \tag{R3.84}$$

is the proportion of transmissions generated prior to symptom onset by hosts who develop symptoms (expressions for q_P are given for the individual models above), and

$$\alpha_A = \frac{R_A}{R_P + R_I} \tag{R3.85}$$

is the ratio between the average number of secondary cases generated by an asymptomatic infected host and the number generated by a host who develops symptoms at some stage during infection.

For each model, we used the posterior parameter distributions that were obtained when we fitted the model to data to obtain a sample from the posterior distribution of q_P. In order to estimate the total proportion of non-symptomatic transmissions, we assumed the distributions

$$p_A \sim \text{Beta}\,(85, 186)\,, \quad \text{[median 0.31, 95\% credible interval (CrI) 0.26–0.37]}\,, \tag{R3.86}$$

$$\alpha_A \sim \text{Lognormal}\left(-1.04, 0.65^2\right), \quad \text{[median 0.35, 95\% CrI 0.10–1.26]}\,, \tag{R3.87}$$

which were chosen for consistency with confidence intervals obtained in a meta-analysis [30]. These distributions are shown in Figure A3.1. We then combined samples from the assumed distributions of p_A and α_A with the sample that we generated from the posterior distribution of q_P to obtain a distribution for the total proportion of non-symptomatic transmissions.

R3.2.11 Contact tracing and isolation

We compared assessments of the efficiency of isolation and contact tracing strategies between the different models. While mathematically complex methods for modelling contact tracing exist (see e.g. [85, 88]), we instead took a relatively straightforward approach to explore how the predicted effectiveness of three important aspects of isolation and contact tracing policies varied between models.

First, we considered the proportion of transmissions that can be prevented if a symptomatic host is isolated d_1 days after symptom onset. Assuming that a proportion ε_1 of infectious contacts that would otherwise occur are prevented during the isolation period (and assuming the individual is isolated until they are no longer infectious), the overall proportion of transmissions prevented through isolation is

$$\varepsilon_1 \left(1 - F_{tost}\left(d_1\right)\right). \tag{R3.88}$$

We then predicted the proportion of the presymptomatic infectious contacts (i.e. presymptomatic contacts in which a transmission occurred) of a symptomatic index case that will be identified, if contacts that occurred within a contact elicitation window of up to d_2 days before the index case developed symptoms are traced. In this scenario, assuming that it is possible to trace a fraction ε_2 of the host's presymptomatic contacts (from within the contact elicitation window), then the proportion of presymptomatic infectious contacts found is equal to

$$\frac{\varepsilon_2 \left(q_P - F_{tost}\left(-d_2\right)\right)}{q_P}. \tag{R3.89}$$

Finally, we considered the proportion of onward transmissions that can be prevented if an infected individual, who is identified through contact tracing, is isolated d_3 days after exposure. Assuming that a proportion ε_3 of infectious contacts that would otherwise occur are prevented during the isolation period, the overall proportion of onward transmissions prevented through isolation is

$$\varepsilon_3 \left(1 - F_{gen}\left(d_3\right)\right). \tag{R3.90}$$

In our main analysis of contact tracing and isolation (Figure R3.8), we assumed that $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = 1$ (i.e. perfect isolation of symptomatic hosts, contact identification, and isolation of infected contacts). Values of ε_1, ε_2 and ε_3 below 1 are considered in Figure R3.12.

R3.3 Results

R3.3.1 Models and parameter estimation

We considered four different models of infectiousness (see Sections R3.2.4–R3.2.6 for details):

i. **The "variable infectiousness model".** Our mechanistic approach (Figure R3.2B, solid line) with the relative infectiousness levels for presymptomatic (P) and symptomatic (I) infectious hosts estimated from the data.

ii. **The "constant infectiousness model".** Our mechanistic approach (Figure R3.2B, dashed line) with identical infectiousness levels for presymptomatic (P) and symptomatic (I) infectious hosts.

iii. **The "Ferretti model".** The best-fitting statistical model from [84], in which the presymptomatic portion of an individual's infectiousness profile is scaled (horizontally) depending on the duration of their incubation period.

iv. **The "independent transmission and symptoms model".** The standard approach [85, 93] (Figure R3.2A) in which infectiousness is assumed independent of symptoms.

We fitted each model to data from 191 SARS-CoV-2 transmission pairs [84, 116] obtained by combining data from five studies [50, 85, 118, 248, 250]. To account for uncertainty in the precise times of symptom appearance within the day of onset for the infector and infectee [219], we used data augmentation MCMC. Posterior distributions of all fitted model parameters are shown for the variable

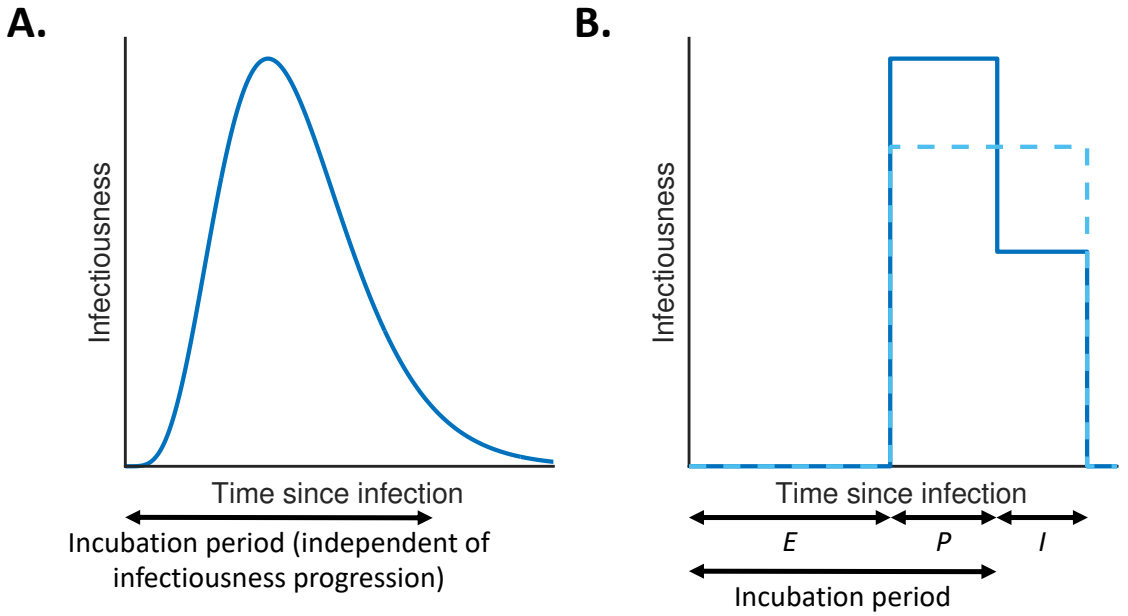


Figure R3.2: Schematic illustrating approaches for inference from transmission pair data. In standard approaches (A) for inferring infectiousness profiles from transmission pair data, the infectiousness of a host at a given time since infection is assumed to be independent of their incubation period. In our approach (B), we link a host's infectiousness with whether or not they have developed symptoms. We assume that individuals are not infectious during the latent (E) period, and that infectiousness may either vary between the presymptomatic infectious (P) and symptomatic infectious (I) periods (solid line – this corresponds to our "variable infectiousness" model), for example due to changing behaviour in response to symptoms [160], or be identical in these two time periods (dashed line – this corresponds to our "constant infectiousness" model).

infectiousness model (which was found to provide the best fit to the data, as described in Section R3.3.2) in Figure R3.3, for the commonly used independent transmission and symptoms model in Figure R3.4, and for the other two models in Appendix A3 (Figures A3.2 and A3.3). Point estimates (posterior medians) and 95% credible intervals are given in Table A3.2.

Note that the Ferretti model and the independent transmission and symptoms model were also fitted to the same data in [84] (the parameter estimates obtained in that study [84] lie within the credible intervals shown in Table A3.2), but estimates of epidemiological quantities obtained using those models were not compared directly in that study.

R3.3.2 Goodness of fit

To compare the goodness of fit of the four models to the transmission pair data, we calculated the AIC for each model. When calculating AIC values, we considered maximum likelihood parameter estimates under the assumption of each symptom

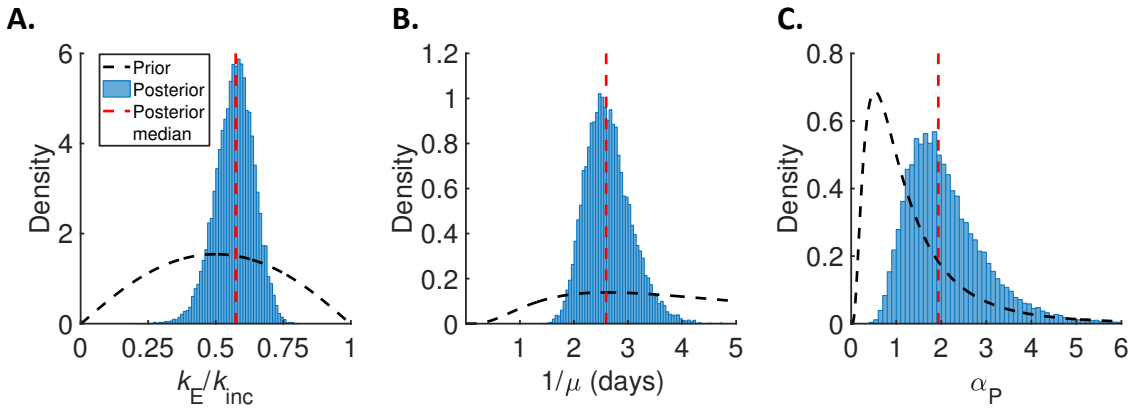


Figure R3.3: Posterior distributions of fitted parameters for the variable infectiousness model. Prior distributions (black dashed lines) of estimated parameters in the variable infectiousness model, and posterior distributions (blue bars) and posterior medians (vertical red dashed lines) obtained when we used data augmentation MCMC to fit this model to the transmission pair data. A. Ratio of mean durations of the latent (E) and incubation (combined E and P) periods, k_E/k_{inc}. B. Mean symptomatic infectious (I) period, $1/\mu$. C. Ratio of transmission rates in the presymptomatic infectious (P) and symptomatic infectious (I) stages of infection, α_P.

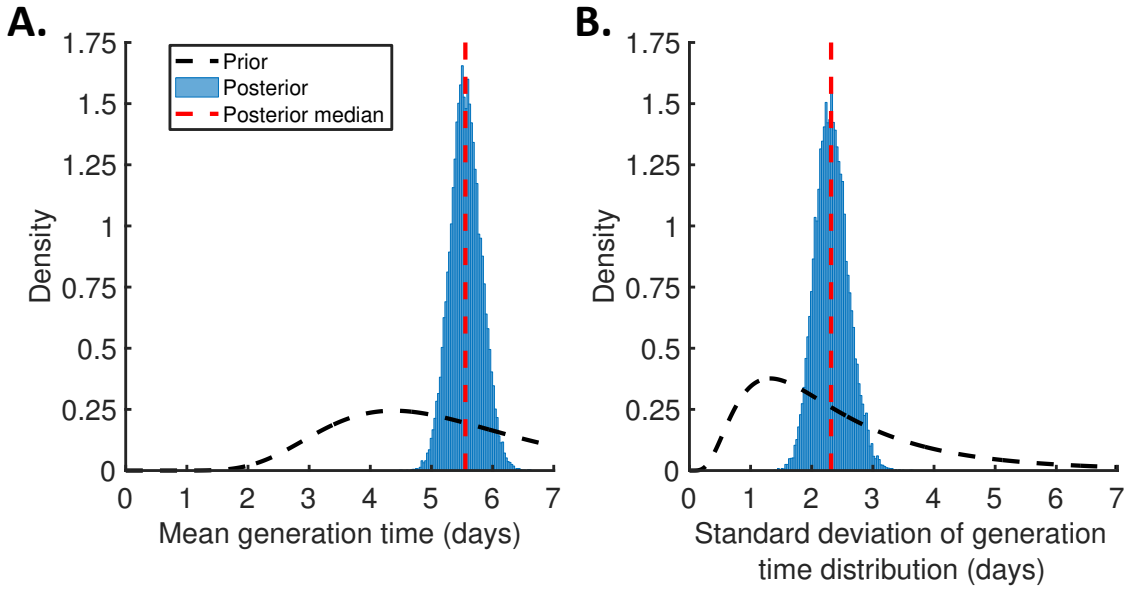


Figure R3.4: Posterior distributions of fitted parameters for the independent transmission and symptoms model. Prior distributions (black dashed lines) of estimated parameters in the independent transmission and symptoms model, and posterior distributions (blue bars) and posterior medians (vertical red dashed lines) obtained when we used data augmentation MCMC to fit this model to the transmission pair data. A. Mean generation time. B. Standard deviation of the generation time distribution.

onset occurring in the middle of the recorded onset date (Table A3.3; note that these maximum likelihood estimates were not used in any other analyses), to avoid the need to compare models based on likelihoods calculated using augmented data.

The best fit to the data was obtained using the variable infectiousness model (ΔAIC = 0). The constant infectiousness model gave the next best fit (ΔAIC = 1.3), followed by the Ferretti model (ΔAIC = 5.1). Finally, the model with the standard assumption of independent transmission and symptoms gave the worst fit (ΔAIC = 38.9).

R3.3.3 Distributions of epidemiological time intervals

For each model, we calculated the generation time (Figure R3.5A), TOST (Figure R3.5B) and serial interval (Figure R3.5C) distributions using point estimates of the fitted parameters (specifically, we used the medians of the posterior distributions obtained using data augmentation MCMC – these point parameter estimates are given in Table A3.2).

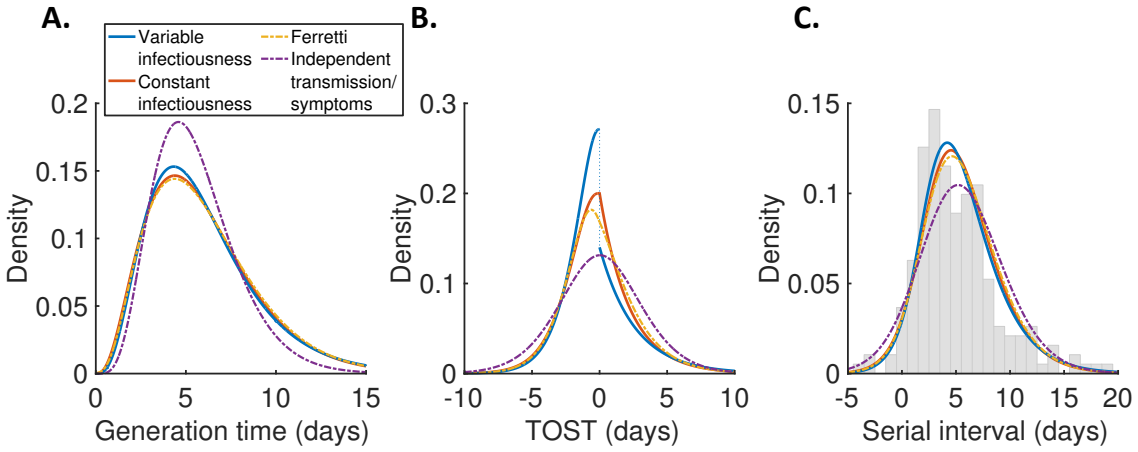


Figure R3.5: Distributions of epidemiological time intervals estimated by fitting different models to data from 191 SARS-CoV-2 transmission pairs. A. Generation time, indicating the relative expected infectiousness of a host at each time since infection. B. TOST, indicating the relative expected infectiousness of a host at each time since symptom onset. C. Serial interval, indicating the periods between infectors and infectees developing symptoms. The empirical serial interval distribution from the transmission pair data [84, 116] is shown as grey bars. The means and standard deviations of these distributions are given in Table A3.4. In all panels, lines represent: variable infectiousness model (blue), constant infectiousness model (red), Ferretti model (orange dashed), and independent transmission and symptoms model (purple dashed).

The predicted variability in the generation time between individuals was lower for the independent transmission and symptoms model compared to the other three

models (Figure R3.5A). On the other hand, the TOST distribution was most concentrated around the time of symptom onset for the best-fitting variable infectiousness model, and least concentrated for the independent transmission and symptoms model (Figure R3.5B). In the best-fitting model, a decrease in infectiousness was inferred following symptom onset, likely due to behavioural factors that reduce the transmission risk following symptom appearance [160].

The empirical serial interval distribution is also plotted alongside the model predictions in Figure R3.5C, to give an approximate visual confirmation of the relative goodness of fit of the different models to the data. However, this involved comparing continuous-time model-predicted distributions with discrete-time data. Therefore, we also calculated discretised versions of the serial interval distributions for each model, using the method proposed in [62] (Figure R3.6). This approach, which we describe in Appendix A3.1, properly accounts for the fact that, for example, if the exact (continuous) serial interval for a given transmission pair was 1.5 days, then (depending on the exact time of day at which the infector developed symptoms) the dates on which the infector and infectee developed symptoms may have differed by either one or two days.

We note that since the transmission pair data contained intervals of possible exposure times in addition to symptom onset dates, the comparisons shown in Figure R3.5C and Figure R3.6 only give a partial picture of the goodness of fit of the four models. On the other hand, the AIC values reported above account for these exposure intervals and therefore give a more complete comparison of the goodness of fit of the different models.

R3.3.4 The contribution of non-symptomatic infectious individuals to transmission

Using the full posterior distributions of model parameters obtained when fitting the models to data, we calculated posterior estimates of the proportion of transmissions occurring before symptom onset (for hosts who develop symptoms) for each model (Figure R3.7A). The median (95% CrI) proportion of presymptomatic transmissions was 64% (52–74%), 56% (50–62%), 56% (48–62%), and 49% (43–56%) under the variable infectiousness model, constant infectiousness model, Ferretti model, and independent transmission and symptoms model, respectively. The central estimate of 64% of transmissions occurring prior to symptom onset using the best-fitting variable infectiousness model is higher than estimated in most previous studies in which the generation time and/or TOST were estimated [14, 84, 85, 118]. In the wider literature, we note significant variation in estimates of the contribution of presymptomatic

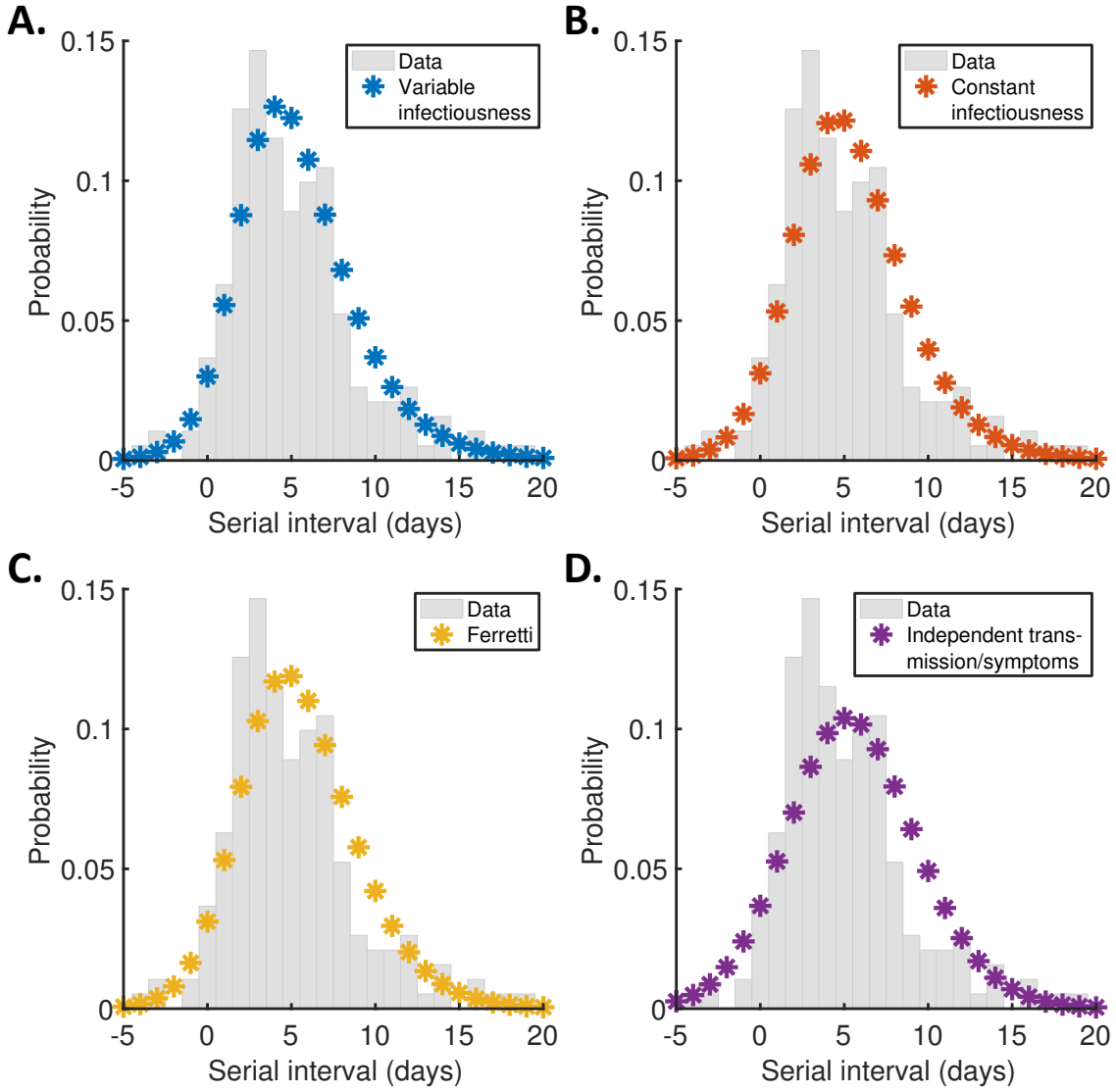


Figure R3.6: Discretised serial interval distributions. Discretised versions of the serial interval distributions shown in Figure R3.5C, calculated using the method proposed in [62] (see Appendix A3.1), plotted as stars alongside the empirical serial interval distribution from the transmission pair data (grey bars). A. Variable infectiousness model. B. Constant infectiousness model. C. Ferretti model. D. Independent transmission and symptoms model.

transmission (obtained under a range of different modelling assumptions), including estimates exceeding 64% [39, 93, 226].

We also combined the estimates in Figure R3.7A with the results of a previous study [30] in which the extent of asymptomatic transmission (i.e. transmissions from individuals who never display symptoms) was characterised (Figure A3.1), to obtain estimates for the total proportion of non-symptomatic (either presymptomatic or asymptomatic) transmissions for the different models (Figure R3.7B). The non-

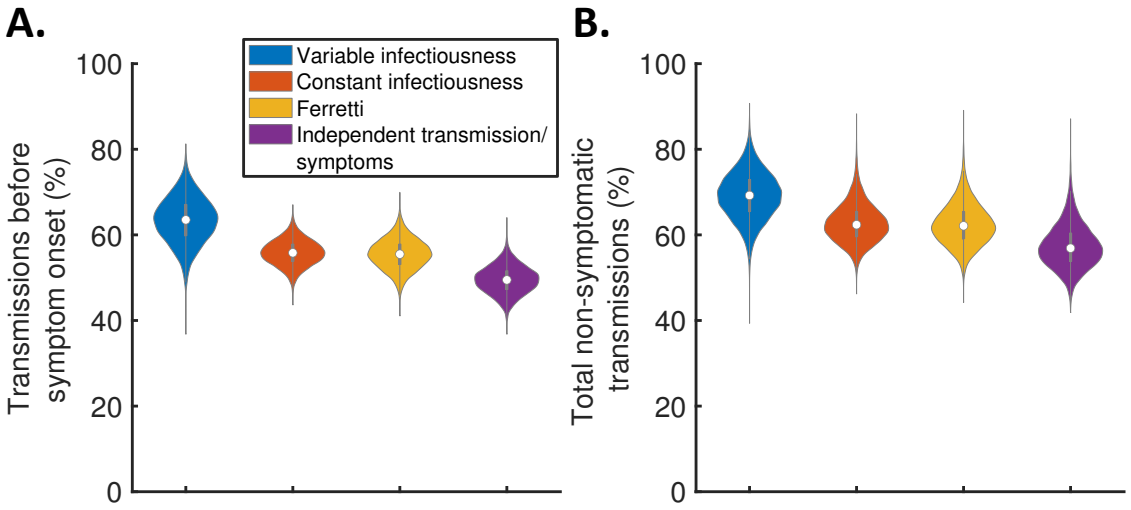


Figure R3.7: The contribution of non-symptomatic infectious individuals to transmission. A. Violin plots indicating posterior distributions of the proportion of transmissions occurring prior to symptom onset for infectors who develop symptoms (i.e. not considering transmissions from individuals who remain asymptomatic throughout infection) for the different models. B. Posterior distributions of the total proportion of non-symptomatic transmissions, accounting for transmissions from asymptomatic infectious individuals (Figure A3.1), for the different models. In both panels, violins represent: variable infectiousness model (blue), constant infectiousness model (red), Ferretti model (orange), and independent transmission and symptoms model (purple).

symptomatic proportion was highest for the variable infectiousness model and lowest for the independent transmission and symptoms model.

R3.3.5 Implications for isolation and contact tracing

We explored the implications of our results for isolation and contact tracing (Figure R3.8) under the simplifying assumptions of perfect isolation (i.e. isolation prevents transmission completely) and perfect contact tracing (i.e. all contacts are traced successfully during periods of contact tracing). Imperfect isolation and contact tracing are considered in Figure R3.12.

Considering a scenario in which an infected host (referred to in this section as the "index case") is detected and isolated following symptom onset, we first calculated how many transmissions from the index case are expected to be prevented for different time delays between the index case developing symptoms and being isolated (Figure R3.8A), compared to a scenario in which the index case is never isolated. We then considered tracing the contacts of that index case, inferring the proportion of infectious contacts (i.e. contacts in which a transmission happened) occurring before the index host developed symptoms that are identified for different contact elicitation windows (Figure R3.8B). As an example, a contact elicitation

window of two days before symptom onset means that all contacts of the index case that occurred in the two days prior to the index case developing symptoms are traced (in addition to contacts that occurred after the index case developed symptoms). Finally, we considered isolation of infected contacts of the index case. For different time periods between the index case transmitting the virus to the contact and the contact being isolated, we calculated the expected proportion of transmissions by the contact (that would be generated if the contact did not isolate) prevented through the contact's isolation (Figure R3.8C).

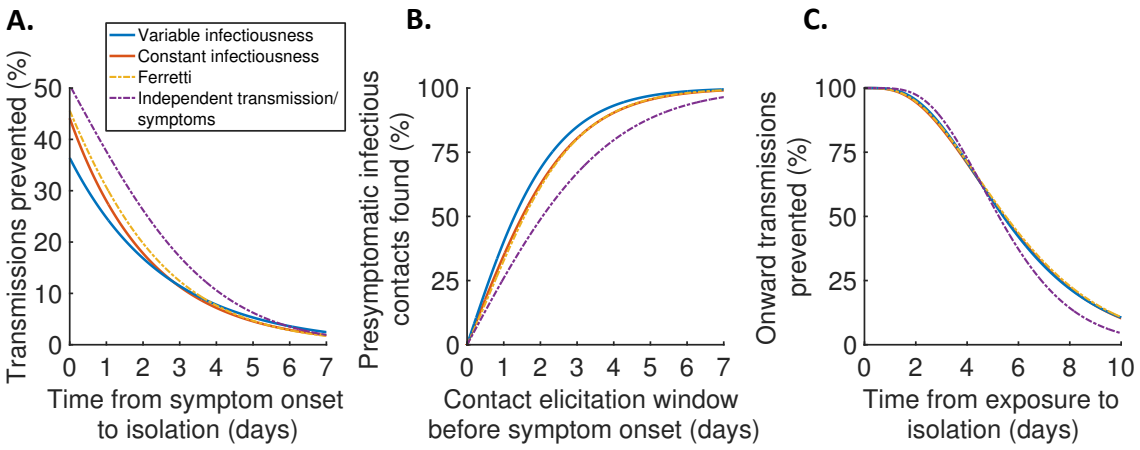


Figure R3.8: Implications for isolation and contact tracing A. Effect of the timing of isolation of symptomatic index cases: the proportion of transmissions prevented through isolation, for different time periods between symptom onset and isolation. B. Effect of the contact elicitation window: the proportion of infectious contacts occurring before the index host developed symptoms that are identified through contact tracing, for different times up to which contacts are traced before the symptom onset time of the index host. C. Effect of the timing of isolation of infected contacts: the proportion of onward transmissions by the contacts that are prevented by isolation of those contacts, for different time periods between exposure to the index host and isolation of the contacts. In all panels, lines represent predictions obtained using point estimate parameters for the variable infectiousness model (blue), constant infectiousness model (red), Ferretti model (orange dashed), and independent transmission and symptoms model (purple dashed). Here, isolation and contact tracing are both assumed to be perfect; equivalent panels in which the efficiency of isolation and/or contact tracing is less than 100% are shown in Figure R3.12.

Under the best-fitting variable infectiousness model, 25% (95% CrI 18–32%) of all transmissions that would be generated by a host are prevented if the host is isolated one day after developing symptoms (Figure R3.8A, blue). This compares to a higher estimate of 38% (95% CrI 32–44%) with the standard independent transmission and symptoms assumption (Figure R3.8A, purple dashed) and intermediate estimates for the constant infectiousness (Figure R3.8A, red) and Ferretti (Figure R3.8A, orange dashed) models. The limited impact of isolation of symptomatic hosts alone under

the variable infectiousness model, which is due to the high predicted proportion of presymptomatic transmissions (Figure R3.7A), highlights the importance of also conducting contact tracing.

The variable infectiousness model indicates that 69% (95% CrI 57–79%) of presymptomatic infectious contacts are identified if a contact elicitation window of up to two days before the index host developed symptoms is used (as in the UK [228] and USA [44]), compared to only 49% (95% CrI 44–53%) for the independent transmission and symptoms model (Figure R3.8B). If the contact elicitation window is extended to four days, then 93% (95% CrI 88–96%) of presymptomatic infectious contacts are identified under the variable infectiousness model.

For practical assessments of contact tracing and isolation effectiveness, it may be necessary to consider the combined effects of different delays at each stage of the contact tracing and isolation process. For example, if there is a delay of two days between an index case infecting a contact and the index case showing symptoms, and a further delay of two days between the index case showing symptoms and the contact being traced and isolated, then this corresponds to a total delay of four days between the contact being infected and isolated (assuming that the contact elicitation window is at least two days, so that the contact is traced). Under the variable infectiousness model, 71% of onward transmissions from the contact would then be expected to be prevented after this delay (Figure R3.8C). In contrast, for an infectious contact that occurred four days before the index host developed symptoms (so that the total delay between the contact being infected and isolated is six days, assuming that the contact elicitation window is at least four days so the contact is traced), only 42% of the contact's onward transmissions would be expected to be prevented (Figure R3.8C).

R3.3.6 Sensitivity analyses

R3.3.6.1 Effect of the incubation period distribution

When we fitted the different models to data, we assumed a specified incubation period distribution [143]. To explore the effect of the exact incubation period distribution on our findings, we repeated the analyses shown in Figures R3.5, R3.7 and R3.8 but using an alternative incubation period distribution [146] (Figures R3.9–R3.11), which had a slightly shorter mean but a higher standard deviation than the incubation period distribution we assumed in our main analyses [143] (the two incubation distributions are given in Section R3.2.3).

The exact incubation period distribution assumed did affect some of our estimates. For example, when we used the alternative incubation period estimate from [146],

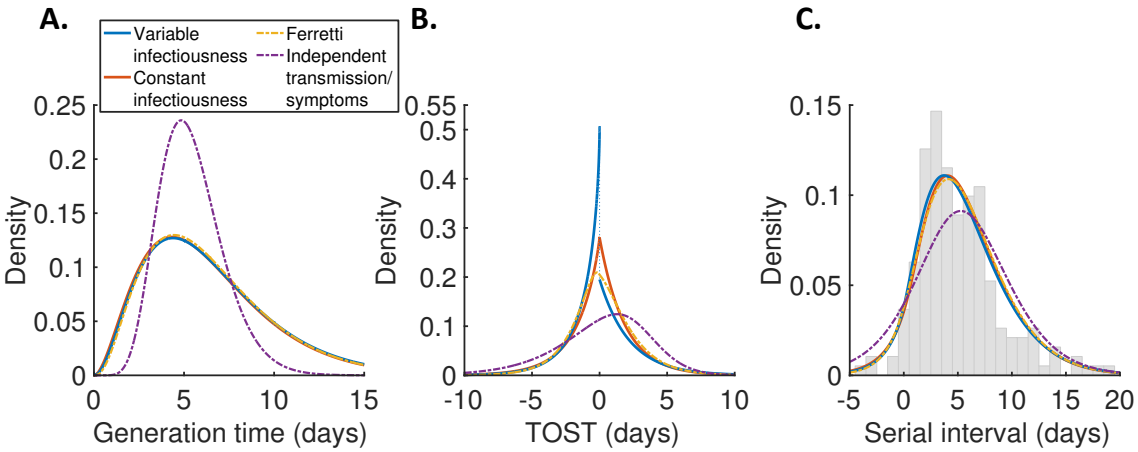


Figure R3.9: Distributions of epidemiological time intervals assuming an alternative incubation period distribution. Equivalent panels to Figure R3.5 but assuming a different incubation period distribution [146].

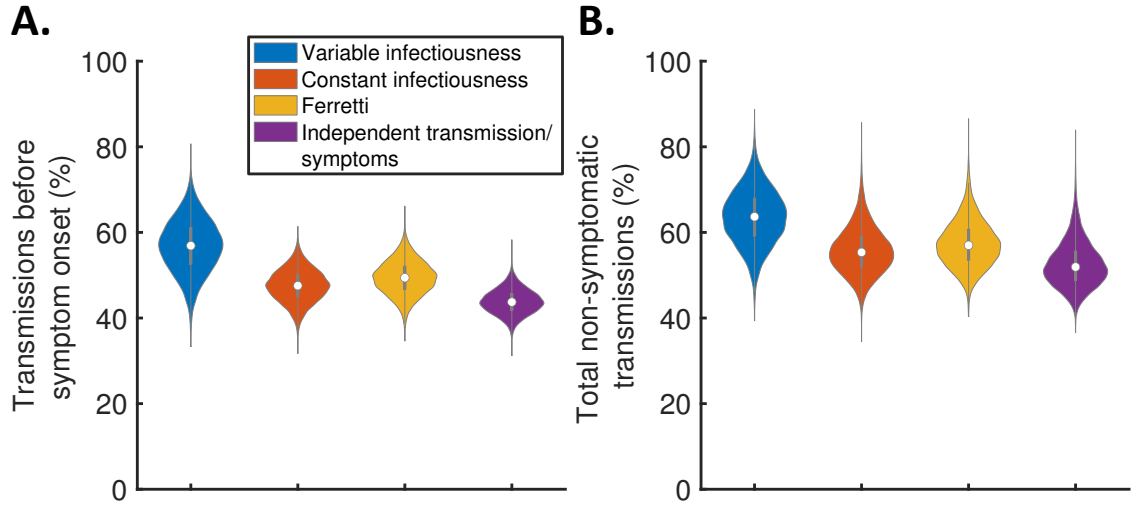


Figure R3.10: The contribution of non-symptomatic infectious individuals to transmission assuming an alternative incubation period distribution. Equivalent panels to Figure R3.7 but assuming a different incubation period distribution [146].

each model we considered suggested a lower proportion of transmissions occurring before symptom onset for infectors who develop symptoms (Figure R3.10; e.g. the median estimate for the variable infectiousness model was 57%, 95% CrI 44–69%) than that estimated when we used the incubation period distribution from [143] (Figure R3.10; the median estimate for the variable infectiousness model was 64%, 95% CrI 52–74%), albeit with overlapping credible intervals for each model.

However, our main qualitative findings were robust to the choice of incubation period distribution. For both incubation period distributions that we considered [143, 146], out of the four models that we used, the variable infectiousness model provided the best fit to the transmission pair data and the commonly used independent

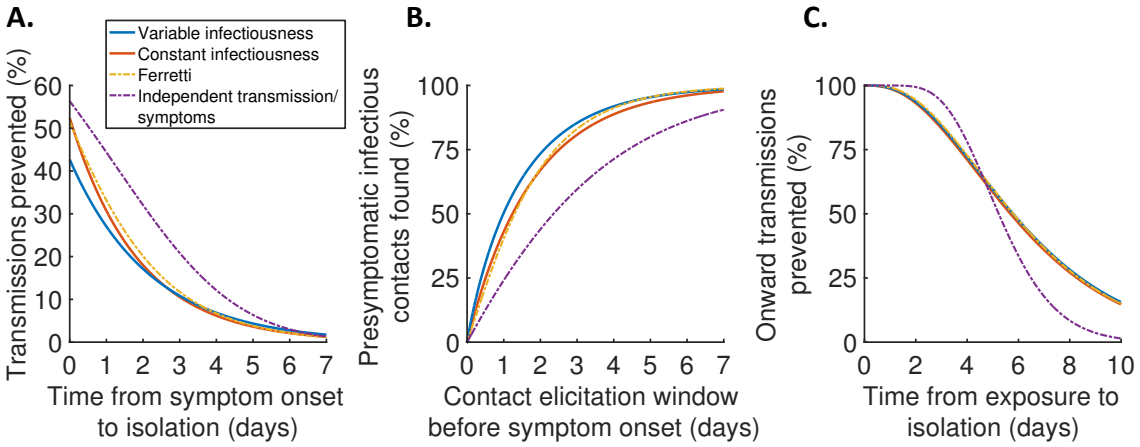


Figure R3.11: Implications for isolation and contact tracing assuming an alternative incubation period distribution. Equivalent panels to Figure R3.8 but assuming a different incubation period distribution [146].

transmission and symptoms model provided the worst fit – the ΔAIC values under the alternative incubation period distribution considered in this section [146] are 0 (variable infectiousness model), 5.1 (constant infectiousness model), 10.2 (Ferretti model), and 64.5 (independent transmission and symptoms model).

For each incubation period distribution, compared to the independent transmission and symptoms model, the variable infectiousness model also predicted:

- A more dispersed generation time distribution (Figure R3.5A and Figure R3.9A).
- A TOST distribution that is more concentrated around the time of symptom onset, with the variable infectiousness model predicting a decrease in infectiousness following symptom onset (Figure R3.5B and Figure R3.9B).
- A higher proportion of presymptomatic transmissions (Figure R3.7A and Figure R3.10A).
- A lower effectiveness of the isolation of symptomatic individuals (assuming isolation occurs within five days of symptom onset; Figure R3.8A and Figure R3.11A).
- If a two-day contact elicitation window is used, a higher proportion of contacts infected by an index host before the index host developed symptoms can be identified through contact tracing (Figure R3.8B and Figure R3.11B).

R3.3.6.2 Effect of the efficiency of isolation and contact tracing

In our main analysis of isolation and contact tracing (Figure R3.12), we assumed both isolation and contact identification to be perfect. The results in Figure R3.12

therefore represent an upper bound for the effectiveness of isolation (for a specified time between infection or symptom onset and isolation) and contact tracing (for a specified contact elicitation window). We considered imperfect isolation and contact tracing in Figure R3.12.

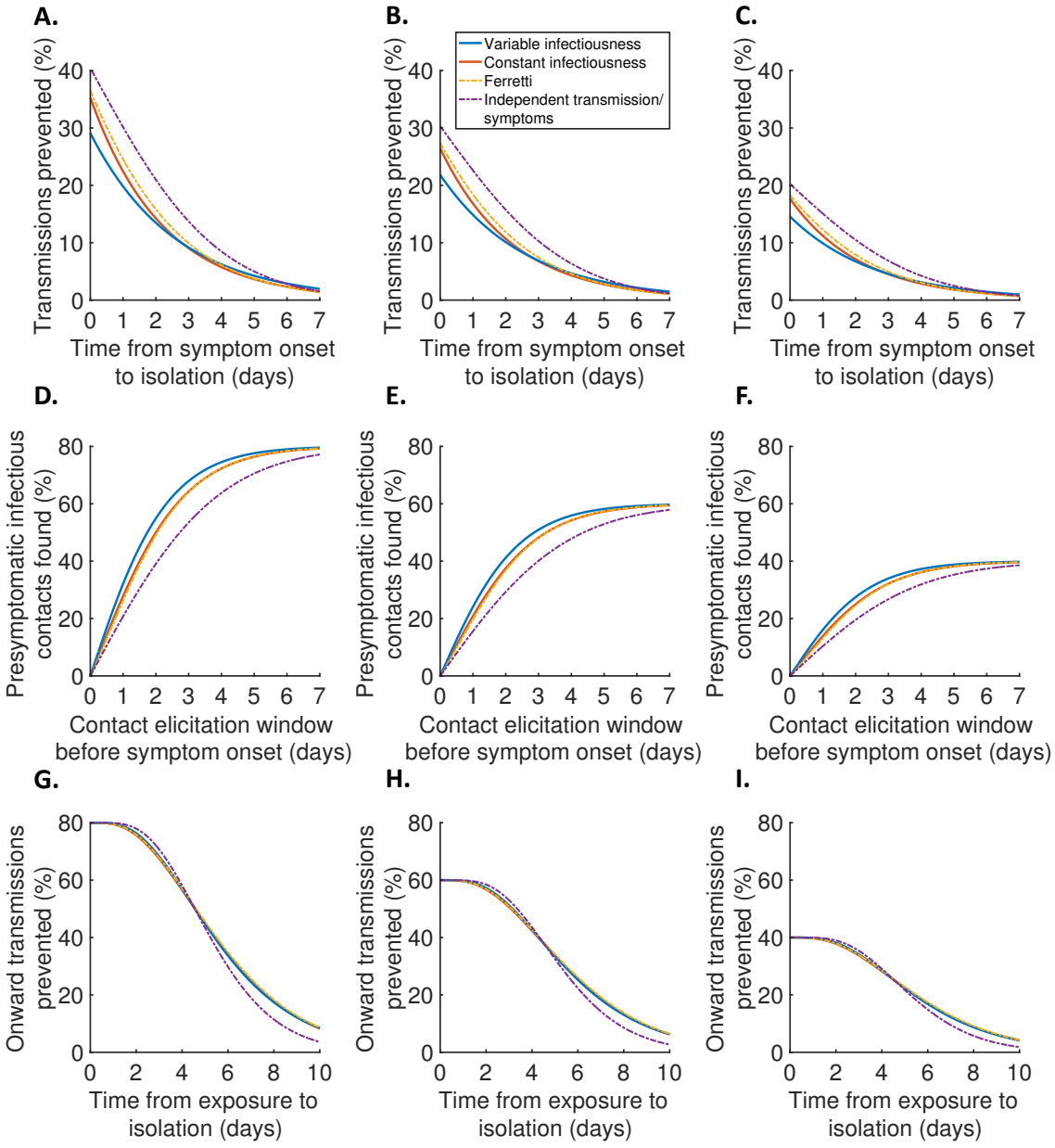


Figure R3.12: Effect of the efficiency of isolation and contact tracing. A–C. Equivalent panels to Figure R3.8A for different values of the isolation efficiency, ε_1. A. $\varepsilon_1 = 0.8$. B. $\varepsilon_1 = 0.6$. C. $\varepsilon_1 = 0.4$. D–F. Equivalent panels to Figure R3.8B for different values of the contact identification efficiency, ε_2. D. $\varepsilon_2 = 0.8$. E. $\varepsilon_2 = 0.6$. F. $\varepsilon_2 = 0.4$. G–I. Equivalent panels to Figure R3.8C for different values of the contact isolation efficiency, ε_3. G. $\varepsilon_3 = 0.8$. H. $\varepsilon_3 = 0.6$. I. $\varepsilon_3 = 0.4$. Note that perfect isolation and contact tracing (as considered in Figure R3.8) corresponds to $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = 1$; see Section R3.2.11 for the exact definition of ε_1, ε_2 and ε_3.

As would be expected, lower values of the efficiency of isolation lead to a smaller proportion of transmissions being prevented by isolation when an infected host is isolated at a given time since symptom onset (Figure R3.12A–C) or infection (Figure R3.12G–H). Similarly, a lower efficiency of contact identification leads to a lower proportion of presymptomatic infectious contacts being identified for a given contact elicitation window (Figure R3.12D–F). Nonetheless, for any assumed isolation efficiency, the best-fitting variable infectiousness model always indicated that isolation of symptomatic individuals (at a specified time within five days of symptom onset) will prevent fewer transmissions than suggested by the commonly used independent transmission and symptoms model. Similarly, the variable infectiousness model always predicted that a two-day contact elicitation window is more effective for identifying contacts infected by an index case before the index case developed symptoms, than suggested by the independent transmission and symptoms model.

R3.4 Discussion

In this research chapter, we have considered a range of approaches for estimating epidemiological time intervals using data from SARS-CoV-2 infector-infectee transmission pairs. Our novel mechanistic approach (the variable infectiousness and constant infectiousness models), based on compartmental modelling, provides an improved fit to data compared to a model predicated on the assumption that the infectiousness of an infected host is independent of exactly when they develop symptoms (the independent transmission and symptoms model). Despite neglecting potential relationships between viral shedding and symptoms, as well as behavioural changes in response to symptoms [160], the assumption of independent transmission and symptoms underlies most previous studies in which the SARS-CoV-2 generation time distribution has been estimated [74, 85, 93, 137]. Our variable infectiousness and constant infectiousness models relax this assumption by providing a mechanistic underpinning to the relationship between infectiousness and symptoms.

The infectiousness profile relative to the time of symptom onset (i.e. the TOST distribution) obtained using our best-fitting variable infectiousness model is more concentrated around the time of onset, compared to that estimated under the standard assumption of independent transmission and symptoms (Figure R3.5B). Under the variable infectiousness model, infectiousness is high immediately before symptom onset, with a decrease in infectiousness occurring immediately after symptom onset that is attributable to behavioural responses to symptoms (e.g. individuals deciding to stay at home as soon as they develop symptoms; we discuss this model artefact

further below). The variable infectiousness model also predicts a higher proportion of presymptomatic transmissions among infectors who develop symptoms (64%, 95% CrI 52–74%) than the independent transmission and symptoms model (49%, 95% CrI 43–56%).

As a result of the findings summarised in the previous paragraph, the variable infectiousness model predicts that isolation of symptomatic hosts alone is less effective at preventing transmissions than suggested by the independent transmission and symptom models. On the other hand, if contact tracing is carried out, then a two-day contact elicitation window (as used in the UK [228] and USA [44] earlier in the COVID-19 pandemic) is predicted to be more effective under the variable infectiousness model than under the independent transmission and symptoms model – specifically, the variable infectiousness model predicts that a higher proportion of contacts infected by an index case before the index case developed symptoms can be identified using a two-day window, since the infectiousness profile for that model is more concentrated around the time of symptom onset. Although choosing a contact elicitation window that is longer than two days would enable more infected contacts to be identified, it would also require more contacts to be traced, many of whom are likely to be uninfected. This effect is enhanced by the fact that index cases are expected to have been less infectious at longer time periods prior to symptom onset (Figure R3.5B), so that (assuming contacts occur at a constant rate prior to symptom onset) contacts traced from longer times before symptom onset are less likely to have been infected.

While the assumption of independent transmission and symptoms is common, some previous studies in which the generation time [71, 84] and/or TOST distributions [14, 84, 118] were estimated have considered an alternative assumption – that infectiousness depends only on the time since symptom onset, independent of the time of infection. If the serial interval is always positive, which is not the case for COVID-19 [78], this is equivalent to assuming that the serial interval and generation time distributions are identical [27, 62, 145]. In one article [84], a non-mechanistic model (the Ferretti model) was developed in which a host's infectiousness could depend on both the time since infection and the time since symptom onset. In that study [84], the Ferretti model was found to provide a better fit to data compared to models in which infectiousness depends either only on the time since infection (i.e. the independent transmission and symptoms model) or only on the time since symptom onset. However, as we have demonstrated, our mechanistic approach provides an improved fit to data compared to the Ferretti model. In addition, since the "$E/P/I$" stages of infection in our mechanistic approach correspond naturally to

compartments in an "SEPIR" population-scale compartmental model, our method is useful for parameterising such epidemic models – specifically, the estimates of stage durations and the relative infectiousness of presymptomatic and symptomatic infectious hosts for the variable infectiousness model in Table A3.2 could be used directly in an SEPIR compartmental model (we note that while we did not impose that the shape parameters of stage duration distributions should take integer values, we expect that gamma distributions with integer shape parameters would provide close approximations to our estimates, enabling the method of stages to be used to solve such a compartmental model numerically).

An important assumption underlying the "$E/P/I$" structure of the best-fitting variable infectiousness model (Figure R3.2B, solid line) is that infectiousness may change when individuals develop symptoms. The relative infectiousness of presymptomatic and symptomatic infectious individuals was then estimated from the data. Here, we attributed the inferred reduction in transmission following symptom onset found in Figure R3.5B (blue line) to behavioural factors. However, in practice behavioural changes may not occur immediately after symptoms appear, particularly if initial symptoms are mild or non-specific. A delay between symptom onset and a change in infectiousness could, in principle, be incorporated into our mechanistic framework by adding an additional stage of infection. This would generate a continuous TOST profile. However, we did not take this approach here since such increased model complexity would require additional parameters to be estimated, likely requiring further data.

Whereas in this chapter we inferred the infectiousness profile of SARS-CoV-2 infections using transmission pair data, an alternative approach would be to use viral load data instead (employing an approach similar to our method for estimating the influenza infectiousness profile in Chapter R2). Several studies have suggested that, for SARS-CoV-2, infectiousness does correlate with viral load [45, 125, 128, 161]. However, factors other than viral load may also be important in determining transmission risk, such as behavioural responses to symptoms (which are explicitly captured in our mechanistic model). An interesting topic for future work would be to inform estimates of the SARS-CoV-2 infectiousness profile using both viral load data and transmission pair data simultaneously (see also the discussion in Chapter R2 about including symptom scores in within-host models).

One limitation of this study is that our estimates were obtained using data collected early in the COVID-19 pandemic (December 2019–March 2020), predominantly from Asia. Since local case numbers were then increasing in locations where some (although not all) of the data were collected [84], shorter serial intervals may have

been over-represented in the dataset [27]. On the other hand, studies from China have indicated a shortening of the generation time [209] and serial interval [6] within the same time period, which was attributed to the effect of non-pharmaceutical interventions (for example, early isolation of infected hosts is likely to reduce the probability of transmission in the later stages of infection). Differences in isolation policies are also likely to affect predictions of the contribution of presymptomatic transmission [39, 209].

In our analyses of isolation and contact tracing (similarly to other studies [14, 15, 85]) we did not explicitly account for public health interventions already in place when the transmission pair data were collected. Therefore (for example), if out-of-home isolation policies in place at the time of data collection (e.g. isolation in hotels or other dedicated facilities) reduced the relative frequency of symptomatic transmissions in the data, then we may have under-estimated the overall effectiveness of isolating symptomatic infected hosts. Nonetheless, these analyses still give a useful indication of the reduction of transmission possible through further or more stringent interventions beyond the policies in place at the time of data collection, and in addition to voluntary reduction of contacts or self-isolation following symptom onset (such behavioural responses to symptoms are explicitly captured in our variable infectiousness model). We also emphasise our general, qualitative overall conclusions: for example, the best-fitting variable infectiousness model predicts fewer transmissions can be prevented through (more stringent) isolation of symptomatic hosts compared to the commonly used independent transmission and symptoms model.

As a result of changes in epidemic dynamics and public health policy, as well as the emergence of novel SARS-CoV-2 variants [114], epidemiological quantities such as the generation time distribution may have changed since the early months of the pandemic. Similarly, the generation time may vary between different locations. Our main aim in this research chapter was to compare estimates of important epidemiological quantities under different modelling assumptions, and therefore transmission data from early in the pandemic were sufficient for our analyses. However, in Chapter R4 we provide updated generation time estimates using more recent data from a UK household study.

In summary, using a novel mechanistic approach in combination with data from SARS-CoV-2 infector-infectee pairs to infer key epidemiological quantities indicates that a higher proportion of transmissions occur prior to symptoms than predicted by existing methods. A significant proportion of these transmissions arise immediately before symptom onset. This shows that, while the impact of isolation of symptomatic hosts alone may be limited, combining this with contact tracing and isolation of

presymptomatic infected contacts is valuable even if the contact elicitation window is short.

Chapter R4

Inference of the SARS-CoV-2 generation time using UK household data

We have published the research underlying this chapter in the journal *eLife* [115].

R4.1 Introduction

Inferring the generation time distribution of SARS-CoV-2 is important in order to predict the effects of non-pharmaceutical interventions, such as contact tracing and quarantine (see the analyses in Chapter R3, as well as other studies – e.g. [15, 85]). In addition, the generation time distribution is widely used in epidemiological models for estimating the time-dependent reproduction number from case notification data [1, 88, 101, 224] and is crucial for understanding the relationship between the reproduction number and the epidemic growth rate [88, 187, 190, 235].

The SARS-CoV-2 generation time distribution has previously been estimated using data from known infector-infectee transmission pairs [84, 85] or entire clusters of cases [93, 122, 209], including in Chapter R3 of this thesis. These studies all involved data [50, 85, 93, 118, 248, 250] collected between December 2019 and April 2020, almost entirely from countries in East and Southeast Asia (with the exception of four transmission pairs from Germany and four from Italy in [85]). Evidence from January and February 2020 in China suggested a temporal reduction in the mean generation time due to non-pharmaceutical interventions [209]. Specifically, effective isolation of infected individuals is likely to have reduced the proportion of transmissions occurring when potential infectors were in the later stages of infection, thereby shortening the generation time [209]. Similarly, two other studies found a decrease in the serial interval [6] and an increase in the proportion of presymptomatic

transmissions [31] in China over the same time period, which can be attributed to symptomatic hosts being isolated increasingly quickly over time.

Despite estimation of the SARS-CoV-2 generation time in Asia early in the pandemic, relatively little is known about the generation time distribution outside Asia, and whether or not any changes have occurred in the generation time since the early months of the pandemic. In particular, at the time of conducting the analyses described in this chapter, we were aware of only one study in which the generation time was estimated using data from the UK [46]. In that study [46], data describing symptom onset dates for 50 infector-infectee pairs, collected by Public Health England (PHE, now the UK Health Security Agency (UKHSA)) between January and March 2020 as part of the "First Few Hundred" case protocol [24, 199], were used to infer the generation time distribution. However, since these transmission pairs mostly consisted of international travellers and their household contacts, the authors concluded that their estimates of the generation time may have been biased downwards due to enhanced surveillance and isolation of these cases [46].

In this research chapter, we use data from a household study [168], conducted in the UK between March and November 2020, to obtain updated estimates of the SARS-CoV-2 generation time distribution. We consider two different underlying models of infectiousness (which were introduced in Chapter R3). In the first model (the "independent transmission and symptoms model"), a parsimonious assumption is made that the generation time and the incubation period of the infector are independent (i.e. there is no link between the times at which infectors transmit the virus and the times at which they develop symptoms), as has often been employed in studies in which the SARS-CoV-2 generation time has been estimated [46, 84, 85, 93, 116, 145] (Table R4.1). In the second model (the "mechanistic model" – note that this corresponds to the "variable infectiousness model" in Chapter R3), we use the mechanistic approach developed in Chapter R3 in which potential infectors progress through different stages of infection so that infectiousness is explicitly linked to symptoms.

Here, we adapt the approach of Chapter R3 in order to estimate the generation time using data from households rather than transmission pairs, accounting for susceptible depletion and uncertainty in the exact chains of transmissions within households. By fitting separately to data from three different time intervals within the study period, we explore whether or not there was a detectable temporal change in the generation time distribution. Our study represents the first analysis of the SARS-CoV-2 generation time outside Asia conducted using data from after the earliest stages of the pandemic.

Study	Location	Time period	Mean generation time (days)	Standard deviation of generation time distribution (days)
Chapter R3 of this thesis [116]	Various	December 2019–March 2020	5.6 (5.1–6.1)	2.3 (1.8–2.9)
Ferretti *et al.* [85]	Various	December 2019–February 2020	5.0	1.9
Ferretti *et al.* [84]	Various	December 2019–March 2020	5.5	1.8
Ganyani *et al.* [93]	Singapore	January–February 2020	5.2 (3.8–6.8)	1.7 (0.9–3.9)
Ganyani *et al.* [93]	China	January–February 2020	4.0 (3.0–4.9)	1.5 (0.7–3.0)

Table R4.1: Previous SARS-CoV-2 generation time estimates. Estimates of the mean and standard deviation of the generation time distribution, obtained under the assumption of independent transmission and symptoms. 95% credible intervals are shown in brackets where available.

R4.2 Methods

R4.2.1 Data

Data were obtained from a household study [168] conducted in 172 UK households (with 603 household members in total) by PHE between March and November 2020 (these data are provided alongside the publication from which this research chapter is adapted [115]). In each household, an index case was recruited following a positive polymerase chain reaction (PCR) test. The following data were then recorded for each household member:

- The timing and outcome of (up to) two subsequent PCR tests.
- The outcome of an antibody test (carried out for 541 individuals – 90% of the study cohort).
- Whether or not the household member developed symptoms.
- The date of symptom onset (only for symptomatic individuals with a positive PCR or antibody test).

In the study, all household members who tested positive in either a PCR or antibody test were assumed to have been infected. Conversely, all individuals who tested negative for antibodies (and where the two PCR tests were either negative or were not carried out) were assumed to have remained uninfected, irrespective of symptom status. For 34 individuals (6% of the study cohort), no antibody test was carried out and any PCR tests carried out were negative. Since it was considered unknown whether or not these individuals were infected, they were excluded from our main analyses (but were counted in the household size), although we considered the sensitivity of our results to this assumption (Figure A4.8).

In two households, at least one household member developed symptoms 55–56 days prior to the symptom onset date of the index case, with no other household members developing symptoms (or returning a positive PCR or antibody test) between these dates. In contrast, the maximum gap between successive symptom onset dates in the remaining households was 25 days (Figure A4.1). Data from these two households were excluded from our analyses, on the basis that the virus was most likely introduced multiple times into these households. Three other households were also excluded from our analyses because, other than the index cases in each household, all other household members were of unknown infection status (i.e. they were among the individuals for whom no antibody test was carried out and any PCR tests were negative).

Overall, aside from the five excluded households, the 167 remaining households comprised 587 individuals, of whom 330 became infected and developed symptoms, 27 became infected but remained asymptomatic, 200 remained uninfected, and the remaining 30 were of unknown infection status. The number of households and individuals recruited into the study by month and by household size is shown in Figure R4.1.

R4.2.2 General infectiousness modelling framework

We denote the expected force of infection exerted by an infected host onto each susceptible member of their household, at time τ since infection, by $\beta(\tau)$, where we assumed

$$\beta(\tau) = (\beta_0/n)\, f_{gen}(\tau), \tag{R4.1}$$

for a host who develops symptoms, and

$$\beta(\tau) = \alpha_A\, (\beta_0/n)\, f_{gen}(\tau), \tag{R4.2}$$

for a host who remains asymptomatic throughout infection. Here:

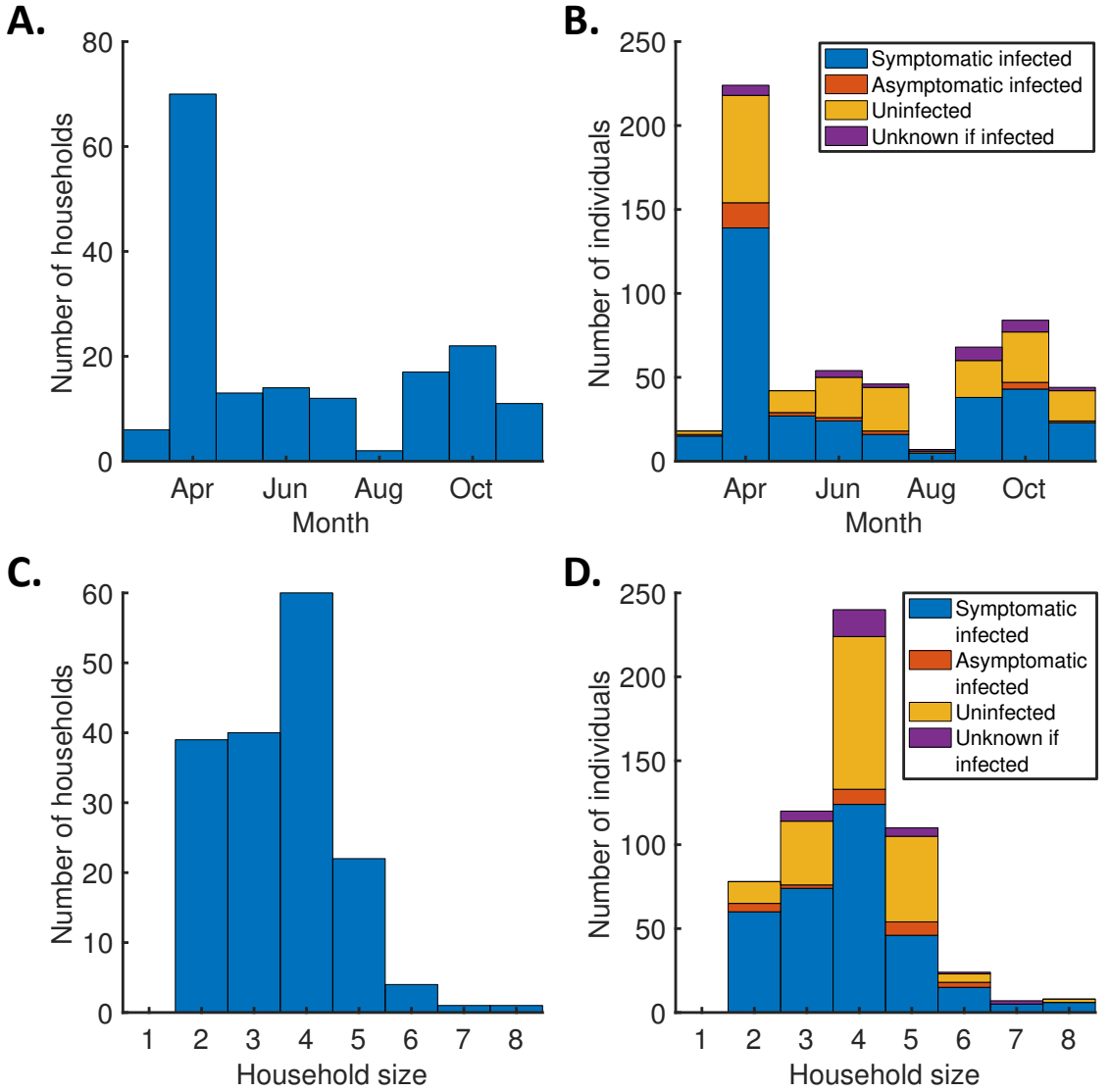


Figure R4.1: Number of study households and household members by recruitment month and household size. A. Bars indicating the number of households recruited each month into the study from which we obtained the household transmission data used in our analyses, from March to November 2020 (excluding the five households that were excluded from our analyses). B. Stacked bars indicating the total number of individuals within the households recruited each month, distinguishing between those who were infected and developed symptoms (blue), those who were infected but never developed symptoms (red), those who remained uninfected (yellow), and those for whom it was unknown whether or not they were infected (purple). C–D. Equivalent panels to A–B, but instead showing the number of households (C) and individuals (D) recruited by household size.

- β_0 is the overall infectiousness parameter, describing the expected number of household transmissions generated by a single infected host (who develops symptoms) in a large, otherwise entirely susceptible, household.

- n is the household size. The scaling of $\beta(\tau)$ with $1/n$ corresponds to frequency-dependent transmission, as assumed in [40, 42], although we carried out sensitivity analyses (Figures R4.13 and A4.6) in which we considered alternative possibilities where household transmission is density-dependent (without the scaling factor $1/n$), scales with $1/n^{0.5}$ [80], or scales with $1/(n-1)$.

- $f_{gen}(\tau)$ is the generation time distribution.

- α_A is the relative infectiousness of infected hosts who remain asymptomatic throughout infection. We assumed a value of 0.35 [30] in most of our analyses, although we considered different values of α_A in a sensitivity analysis (Figure A4.7).

Note that since we explicitly model the dependence of overall transmissibility on household size in this chapter, here β_0/n corresponds to β_0 in Chapter R3.

Except where otherwise stated, we focussed on the generation time distribution assuming a constant supply of susceptibles are available to infect throughout infection, $f_{gen}(\tau)$, which corresponds to the normalised expected infectiousness profile (with respect to the time since infection). However, realised generation times within a household may be shorter than predicted by this distribution due to the depletion of susceptible household members before longer generation times can be attained [41, 88, 114, 191]. For example, if infected hosts are (on average) equally infectious at two times since infection, $\tau_1 < \tau_2$, then $f_{gen}(\tau_1) = f_{gen}(\tau_2)$. However, since either the host under consideration or another infected household member may transmit the virus within the household in the intervening time, the number of susceptible household members may decrease between the times τ_1 and τ_2. As a result transmission is, in fact, more likely to occur in a household at the earlier time, τ_1, when more susceptibles are available (when this effect arises due to multiple infected hosts "competing" to infect the available susceptible individuals, with only the first potential transmission being realised, it is sometimes called a "race to infect" susceptibles [132, 148]). We therefore also predicted the mean and standard deviation of realised generation times within the study households in Figure R4.5.

We considered two different models of infectiousness (the independent transmission and symptoms model and the mechanistic model), which were introduced in Chapter R3 and are outlined below. For the two models, expressions are given for the generation time distribution, f_{gen}, and the TOST distribution, f_{tost}. As shown in Chapter R3 (Section R3.2.1), the serial interval distribution, f_{ser}, for each model

is given by the convolution of the TOST distribution, f_{tost}, and incubation period distribution, f_{inc} (which was assumed to be known – see Section R4.2.6 below), i.e.

$$f_{ser}\left(x_{ser}\right) = \int_0^\infty f_{tost}\left(x_{ser} - \tau\right) f_{inc}\left(\tau\right) \mathrm{d}\tau. \tag{R4.3}$$

In addition, formulae are given for the mean and standard deviation of the generation time distribution, and the proportion of presymptomatic transmissions (out of all transmissions generated by infectors who develop symptoms), for each model. These expressions were used to calculate posterior distributions and point estimates (posterior medians) of those quantities.

For the mechanistic model, we also give an expression for the expected infectiousness, $\beta\left(\tau \mid \tau_{inc}\right)$, of a host (who develops symptoms) at each time since infection, τ, conditional on their incubation period, τ_{inc}. In addition, an expression is given for the cumulative conditional infectiousness,

$$B\left(\tau \mid \tau_{inc}\right) = \int_0^\tau \beta\left(\tilde{\tau} \mid \tau_{inc}\right) \mathrm{d}\tilde{\tau}, \tag{R4.4}$$

under that model. The functions $\beta\left(\tau \mid \tau_{inc}\right)$ and $B\left(\tau \mid \tau_{inc}\right)$ were required in the likelihood function that we used when we fitted the mechanistic model to the household transmission data (see Section R4.2.5 below).

R4.2.3 Independent transmission and symptoms model

In this model, the infectiousness of an infected host (who does not remain asymptomatic throughout infection) at a given time since infection, τ, is assumed to be independent of exactly when the host develops symptoms – i.e. the generation time and incubation period are independent. In our main analyses using this model, we assumed that the generation time distribution, $f_{gen}\left(\tau\right)$, is the probability density function of a lognormal distribution [82] (an alternative case of a gamma-distributed generation time is considered in Figure A4.4). The generation time distribution was assumed to be the same for entirely asymptomatic infected hosts as for those who develop symptoms.

The mean, m_{gen}, and standard deviation, s_{gen}, of the generation time distribution, in addition to the overall infectiousness parameter, β_0, were estimated when we fitted the model to the household transmission data. Note that the mean and standard deviation are given, respectively, by

$$m_{gen} = \exp\left(\mu_{gen} + 0.5\sigma_{gen}^2\right), \tag{R4.5}$$

$$s_{gen} = \exp\left(\mu_{gen} + 0.5\sigma_{gen}^2\right)\sqrt{\exp\left(\sigma_{gen}^2\right) - 1}, \tag{R4.6}$$

where μ_{gen} and σ_{gen} are the parameters of the lognormal distribution (representing the mean and standard deviation of the natural logarithm of the generation time, respectively).

For this model, the TOST distribution, f_{tost}, is given by the convolution

$$f_{tost}(x_{tost}) = \int_0^\infty f_{gen}(x_{tost} + \tau) f_{inc}(\tau) \, \mathrm{d}\tau, \tag{R4.7}$$

where f_{inc} is the incubation period distribution. The proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) is given by [85, 90]

$$\int_0^\infty f_{gen}(\tau)(1 - F_{inc}(\tau)) \, \mathrm{d}\tau, \tag{R4.8}$$

where F_{inc} is the cumulative distribution function of the incubation period. Derivations of these two expressions are given in Chapter R3 (Section R3.2.4).

R4.2.3.1 Prior distributions for estimated parameters

The vector of unknown model parameters, $\theta = (m_{gen}, s_{gen}, \beta_0)$, was estimated when we fitted the independent transmission and symptoms model to the household transmission data. Independent lognormal prior distributions were assumed for fitted model parameters. We assumed the same prior distributions for the mean and standard deviation of the generation time distribution as we used for this model in Chapter R3 (see Section R3.2.4.4). The exact prior distributions we used, including the prior distribution for β_0, are detailed in Table A4.2.

R4.2.4 Mechanistic model

In this model, each infected host (who develops symptoms) progresses through independent latent (E), presymptomatic infectious (P) and symptomatic infectious (I) stages of infection. The infectiousness of the host during the P and I stages is denoted by β_P and β_I, respectively, and we denote the ratio $\alpha_P = \beta_P/\beta_I$. We also denote the stage durations by $y_{E/P/I}$, their probability density functions by $f_{E/P/I}$, and their cumulative distribution functions by $F_{E/P/I}$,

We assumed the duration of each stage to be gamma-distributed:

$$y_E \sim \text{Gamma}(k_E, 1/(k_{inc}\gamma)), \tag{R4.9}$$

$$y_P \sim \text{Gamma}(k_P, 1/(k_{inc}\gamma)), \tag{R4.10}$$

$$y_I \sim \text{Gamma}(k_I, 1/(k_I\mu)), \tag{R4.11}$$

where we write $X \sim \text{Gamma}(a, b)$ for a gamma-distributed random variable with shape parameter a and scale parameter b. We assumed that $k_E + k_P = k_{inc}$, so that the incubation period, $\tau_{inc} = y_E + y_P$, is gamma-distributed, with

$$\tau_{inc} \sim \text{Gamma}\left(k_{inc}, 1/(k_{inc}\gamma)\right). \tag{R4.12}$$

We fixed the values of the parameters k_{inc} and $1/\gamma$ (which represent the shape parameter of the incubation period distribution and the mean incubation period, respectively) in order to obtain the incubation period distribution specified in Section R4.2.6 below (the exact values of k_{inc} and $1/\gamma$ are given in Table A4.1). For simplicity, we also assumed that $k_I = 1$, so the symptomatic infectious period is exponentially distributed. We estimated k_E/k_{inc} (the ratio between the shape parameters of the latent (E) and incubation (combined E and P) period distributions), $1/\mu$ (the mean symptomatic infectious (I) period) and α_P (the ratio between the transmission rates of hosts in the P and I stages) when we fitted the model to the household transmission data.

Hosts who remain asymptomatic throughout infection were assumed to follow the same $E/P/I$ stages, although in this case the distinction between the P and I stages has no epidemiological meaning. Stage durations, as well as the value of α_P, were assumed to be identical for entirely asymptomatic hosts and those who develop symptoms, so that the generation time distribution is the same for all hosts.

R4.2.4.1 Generation time and TOST distributions

The generation time of an individual transmission, τ_{gen}, can be written as

$$\tau_{gen} = y_E + y^*, \tag{R4.13}$$

where y_E is the length of the latent (E) stage, and y^* is the time from the start of the presymptomatic infectious (P) stage to the transmission occurring. The generation time distribution is therefore given by the convolution

$$f_{gen}(\tau_{gen}) = \int_0^{\tau_{gen}} f^*(\tau_{gen} - y_E) f_E(y_E)\, \mathrm{d}y_E, \tag{R4.14}$$

where, as shown in Chapter R3 (Section R3.2.6), if the effect of susceptible depletion during infection is neglected, then y^* has density

$$f^*(y^*) = C\left(\alpha_P(1 - F_P(y^*)) + \int_0^{y^*} (1 - F_I(y^* - y_P)) f_P(y_P)\, \mathrm{d}y_P\right), \tag{R4.15}$$

where

$$C = \frac{k_{inc}\gamma\mu}{\alpha_P k_P \mu + k_{inc}\gamma}. \tag{R4.16}$$

The TOST distribution is given by

$$f_{tost}\left(x_{tost}\right) = \begin{cases} \alpha_P C\left(1 - F_P\left(-x_{tost}\right)\right), & x_{tost} < 0, \\ C\left(1 - F_I\left(x_{tost}\right)\right), & x_{tost} \geq 0. \end{cases} \tag{R4.17}$$

This expression was derived in Chapter R3 (Section R3.2.6).

R4.2.4.2 Mean and standard deviation of generation time distribution

Using the density function given in equation (R4.15), it can be shown that the moments of the distribution of y^* (the time from the start of the P stage to a transmission occurring) are

$$\mathbb{E}[(y^*)^m] = \frac{C}{m+1}\left(\alpha_P \mathbb{E}\left[{y_P}^{m+1}\right] + \mathbb{E}\left[\left(y_P + y_I\right)^{m+1} - {y_P}^{m+1}\right]\right), \tag{R4.18}$$

for each positive integer m. In particular,

$$\mathbb{E}[y^*] = \frac{C}{2}\left(\alpha_P \mathbb{E}\left[{y_P}^2\right] + 2\mathbb{E}[y_P]\,\mathbb{E}[y_I] + \mathbb{E}\left[{y_I}^2\right]\right), \tag{R4.19}$$

and

$$\begin{aligned} \mathrm{Var}\left[y^*\right] = \frac{C}{3}\left(\alpha_P \mathbb{E}\left[{y_P}^3\right] + 3\mathbb{E}\left[{y_P}^2\right]\mathbb{E}[y_I] + 3\mathbb{E}[y_P]\,\mathbb{E}\left[{y_I}^2\right] + \mathbb{E}\left[{y_I}^3\right]\right) \\ - \left(\mathbb{E}[y^*]\right)^2. \end{aligned} \tag{R4.20}$$

Now, the moments of a gamma-distributed random variable, $X \sim \mathrm{Gamma}\left(a, b\right)$, are given by

$$\mathbb{E}[X^m] = \frac{\Gamma\left(a+m\right)}{\Gamma\left(a\right)} b^m = a\left(a+1\right)\ldots\left(a+\left(m-1\right)\right) b^m. \tag{R4.21}$$

Therefore, for gamma-distributed stage durations, explicit expressions can be obtained for the mean and standard deviation of the generation time distribution,

$$m_{gen} = \mathbb{E}[\tau_{gen}] = \mathbb{E}[y_E] + \mathbb{E}[y^*], \tag{R4.22}$$

$$s_{gen} = \sqrt{\mathrm{Var}\left[\tau_{gen}\right]} = \sqrt{\mathrm{Var}\left[y_E\right] + \mathrm{Var}\left[y^*\right]}, \tag{R4.23}$$

where the last equality holds because of the assumed independence of y_E and y^*.

R4.2.4.3 Proportion of presymptomatic transmissions

Among infectors who develop symptoms, the proportion of transmissions occurring prior to symptom onset (neglecting the effect of susceptible depletion during infection) is given by

$$q_P = \frac{\left(\dfrac{\beta_P k_P}{k_{inc}\gamma}\right)}{\left(\dfrac{\beta_P k_P}{k_{inc}\gamma} + \dfrac{\beta_I}{\mu}\right)} = \frac{\alpha_P k_P \mu}{\alpha_P k_P \mu + k_{inc}\gamma}. \tag{R4.24}$$

This expression was derived in Chapter R3 (Section R3.2.6).

R4.2.4.4 Conditional infectiousness and cumulative conditional infectiousness

The expected infectiousness of a host (who develops symptoms) at time τ since infection, conditional on incubation period τ_{inc}, is given by

$$\beta\left(\tau \mid \tau_{inc}\right) = \begin{cases} \alpha_P C\left(\beta_0/n\right)\left(1 - F_{Beta}\left(1 - \tau/\tau_{inc}; k_P, k_E\right)\right), & 0 < \tau < \tau_{inc}, \\ C\left(\beta_0/n\right)\left(1 - F_I\left(\tau - \tau_{inc}\right)\right), & \tau > \tau_{inc}, \end{cases} \tag{R4.25}$$

where β_0 is the overall infectiousness parameter, n is the household size and $F_{Beta}\left(x; a, b\right)$ is the cumulative distribution function of a beta-distributed random variable with shape parameters a and b. This expression was derived in Chapter R3 (Section R3.2.6).

The cumulative conditional infectiousness can therefore be calculated to be

$$B\left(\tau \mid \tau_{inc}\right) = \int_0^{\tau} \beta\left(\tilde{\tau} \mid \tau_{inc}\right) \mathrm{d}\tilde{\tau} = \left(\tau - \tau_{inc}\right)\beta\left(\tau \mid \tau_{inc}\right) + B_1\left(\tau \mid \tau_{inc}\right), \tag{R4.26}$$

where

$$\begin{aligned} &B_1\left(\tau \mid \tau_{inc}\right) \\ &\quad = \begin{cases} \dfrac{\alpha_P C \beta_0}{n}\left[\dfrac{k_p \tau_{inc}}{k_{inc}}\left(1 - F_{Beta}\left(1 - \tau/\tau_{inc}; k_P + 1, k_E\right)\right)\right], & 0 \leq \tau < \tau_{inc}, \\ \dfrac{C\beta_0}{n}\left[\dfrac{\alpha k_p \tau_{inc}}{k_{inc}} + \dfrac{1}{\mu} F_{Gamma}\left(\tau - \tau_{inc}; k_I + 1, \dfrac{1}{k_I \mu}\right)\right], & \tau \geq \tau_{inc}. \end{cases} \end{aligned} \tag{R4.27}$$

Here, $F_{Gamma}\left(x; a, b\right)$ is the cumulative distribution of a gamma-distributed random variable with shape parameter a and scale parameter b. The total force of infection exerted on each household member by the host under consideration (over the course of infection) is then

$$B\left(\infty \mid \tau_{inc}\right) = \frac{\beta_0}{n}\left(\frac{\alpha_P k_P \gamma \mu \tau_{inc} + k_{inc}\gamma}{\alpha_P k_P \mu + k_{inc}\gamma}\right). \tag{R4.28}$$

The mean of this expression over the incubation period distribution is β_0/n.

For an infected host who remains asymptomatic throughout infection, the infectiousness, $\beta\left(\tau \mid \tau_{inc}\right)$, conditional on the combined duration of the E and P stages, $\tau_{inc} = y_E + y_P$, is given by α_A times the corresponding expression for a host who develops symptoms. We note that in this case, τ_{inc} has no epidemiological interpretation, but this conditional infectiousness was useful when fitting the model to the household transmission data. Specifically, because $\beta\left(\tau \mid \tau_{inc}\right)$ is relatively straightforward to calculate compared to $\beta(\tau)$ (which is proportional to the generation time distribution – see equations (R4.14) and (R4.15)), we augmented the observed data with the

transition times of asymptomatic infected hosts into the I stage of infection so that only $\beta(\tau \mid \tau_{inc})$ was needed in the likelihood function (see Section R4.2.5 below) that was used to fit the model to the data.

R4.2.4.5 Prior distributions for estimated parameters

The vector of model parameters,

$$\theta = (k_E/k_{inc}, 1/\mu, \alpha_P, \beta_0), \tag{R4.29}$$

was estimated by fitting the mechanistic model to the household transmission data. We assumed independent prior distributions for each entry of θ, using the same prior distributions for k_E/k_{inc}, $1/\mu$ and α_P as we used for this model in Chapter R3 (see Section R3.2.6.6; note that this model is referred to as the "variable infectiousness" model in Chapter R3). The exact prior distributions we used, including the prior distribution for β_0, are detailed in Table A4.2.

R4.2.5 Likelihood

Here, we consider a household of size n, in which n_I household members become infected (of whom n_S develop symptoms and n_A remain asymptomatic throughout infection) and $n_U = n - n_I$ remain uninfected. We derive an expression for the likelihood of the parameters of either model of infectiousness, given the entire sequence of infection times of individuals in the household ($t_1 < \ldots < t_{n_I}$) as well as the precise symptom onset time ($t_{s,j}$) of each host (j) who develops symptoms (where infected household members are here labelled in order of their infection times). For the mechanistic model, the likelihood also depends on the times at which entirely asymptomatic infected hosts enter the I stage of infection (these times are also denoted by $t_{s,j}$, although for asymptomatic infected individuals these times have no epidemiological meaning). Since exact infection times and the order of infections were not available within study households, and it was unknown exactly when each symptomatic infected host developed symptoms within their recorded symptom onset date, we used data augmentation MCMC to fit the two models to the UK household transmission data using this likelihood function (see Section R4.2.7 below).

When deriving the likelihood, we made several simplifying assumptions:

- The virus is introduced once into the household (i.e. no subsequent infections from the community occur following the infection of the primary case).
- No co-primary cases (we relaxed this assumption in Appendix A4.3 and Figures R4.10 and R4.11).

- Potential bias towards more recent introduction of the virus into the household if community prevalence is increasing, or less recent if prevalence is decreasing [27, 85, 145], was neglected.

We denote the expected infectiousness of infected household member j, at time since infection τ, by $\beta_j(\tau)$. For the mechanistic model in which transmission and symptoms are not independent, this infectiousness is taken to be conditional on the duration of the incubation period, $t_{s,j} - t_j$, for a host who develops symptoms, and similarly is also conditional on $t_{s,j} - t_j$ for an entirely asymptomatic infected host (although this interval has no epidemiological meaning for such individuals). The total (instantaneous) force of infection exerted at time t on each susceptible household member is then

$$\lambda(t) = \sum_{j=1}^{n_I} \beta_j(t - t_j), \tag{R4.30}$$

where $\beta_j(t - t_j) = 0$ for $t \leq t_j$, and the cumulative force of infection is

$$\Lambda(t) = \int_{-\infty}^{t} \lambda(s)\,\mathrm{d}s. \tag{R4.31}$$

For $k = 2, \ldots, n_I$, conditional on the sequence of infection times up to time t_k, the probability that individual k becomes infected at time t_k is given by

$$\lambda(t_k)\exp(-\Lambda(t_k)), \tag{R4.32}$$

where $\exp(-\Lambda(t_k))$ represents the probability of individual k avoiding infection from household contacts that occurred before their actual time of infection, t_k [40, 82]. This factor of $\exp(-\Lambda(t_k))$ was not included in the likelihood when we estimated the generation time using data from infector-infectee transmission pairs in Chapter R3. The exclusion of such a term in Chapter R3 and other studies estimating the SARS-CoV-2 generation time (e.g. [84, 85, 93]) amounts to an implicit assumption that contacts between transmission pairs are sporadic and of short duration. Since that assumption is not true for household contacts, this factor is required here to correct for the frequency of household contacts in order to ensure widely applicable generation time estimates (note that this factor is equal to one in the limit of a very small overall household infectiousness parameter, β_0).

For $k = n_I + 1, \ldots, n$, conditional on the entire sequence of infection times, $t_1, \ldots, t_{n_I}$, the probability of individual k never being infected is $\exp(-\Lambda(\infty))$. For the independent transmission and symptoms model, we have

$$\exp(-\Lambda(\infty)) = \exp(-\beta_0(n_S + \alpha_A n_A)/n), \tag{R4.33}$$

whereas for the mechanistic model, $\exp(-\Lambda(\infty))$ instead depends on the incubation periods of those hosts who develop symptoms, as well as the corresponding time periods for entirely asymptomatic infected hosts (see equation (R4.28)).

The likelihood contribution from the household, $L(\theta)$, where θ is the vector of unknown model parameters, can therefore be written as

$$L(\theta) = \prod_{k=1}^{n} L_{k,1}(\theta) L_{k,2}(\theta). \tag{R4.34}$$

Here, $L_{k,1}(\theta)$ is the contribution to the likelihood from the transmission, or absence of transmission, to individual k, i.e.

$$L_{k,1}(\theta) = \begin{cases} 1, & \text{for } k = 1; \\ \lambda(t_k) \exp(-\Lambda(t_k)), & \text{for } k = 2, \ldots, n_I; \\ \exp(-\Lambda(\infty)), & \text{for } k = n_I + 1, \ldots, n. \end{cases} \tag{R4.35}$$

The second term, $L_{k,2}(\theta)$, is the contribution from the incubation period of individual k (where applicable), i.e. for the independent transmission and symptoms model,

$$L_{k,2}(\theta) = \begin{cases} f_{inc}(t_{s,k} - t_k), & \text{if individual } k \text{ becomes infected and develops symptoms;} \\ 1, & \text{otherwise;} \end{cases} \tag{R4.36}$$

where f_{inc} is the probability density function of the incubation period (this was assumed to be known; the exact incubation period distribution we used is given below). For the mechanistic model, we also have a contribution to the likelihood from the (in this case not epidemiologically meaningful) times $t_{s,k} - t_k$ for entirely asymptomatic infected hosts, so that

$$L_{k,2}(\theta) = \begin{cases} f_{inc}(t_{s,k} - t_k), & \text{for } k = 1, \ldots, n_I; \\ 1, & \text{for } k = n_I + 1, \ldots, n. \end{cases} \tag{R4.37}$$

R4.2.6 Incubation period

For the independent transmission and symptoms model, we assumed a lognormal incubation period distribution with mean 5.8 days and standard deviation 3.1 days [162]. For the mechanistic model, we assumed a gamma-distributed incubation period with the same mean and standard deviation; this was for mathematical convenience, since the incubation period could then be decomposed into the sum of independent gamma-distributed latent and presymptomatic infectious periods. Results for the independent transmission and symptoms model using a gamma-distributed incubation period are shown in Figure A4.5, and uncertainty in the exact parameters of the incubation period distribution is accounted for in Figure R4.12.

R4.2.7 Parameter estimation

The values of unknown model parameters were estimated using data augmentation MCMC. The observed data comprised information about whether or not individuals were ever infected and/or displayed symptoms, symptom onset dates, and for some individuals an upper bound on their infection time (corresponding to the date of a positive PCR test). These data were augmented with precise times of infection and symptom onset (where applicable) for each infected host. For the mechanistic model only, augmented data also included the times of passage of entirely asymptomatic infected hosts from the P stage of infection to the I stage. No prior assumptions were made about the order of transmissions within each household.

Here (for a given model), we denote the vector of estimated model parameters by θ, and the augmented data by

$$\boldsymbol{t} = \left(\boldsymbol{t}^{(1)}, \ldots, \boldsymbol{t}^{(M)}\right), \tag{R4.38}$$

where $\boldsymbol{t}^{(m)}$ represents the augmented data from household $m = 1, \ldots, M$, and M is the total number of households. We write the (overall) likelihood as

$$L\left(\theta; \boldsymbol{t}\right) = \prod_{m=1}^{M} L^{(m)}\left(\theta; \boldsymbol{t}^{(m)}\right), \tag{R4.39}$$

where the likelihood contribution, $L^{(m)}\left(\theta; \boldsymbol{t}^{(m)}\right)$, from each household, m, was computed as described in the previous section (i.e. all households in the study were assumed to be independent), and we denote the prior density of θ by $\pi\left(\theta\right)$.

The parameter fitting procedure for the independent transmission and symptoms model is described in Algorithm R4.1. The procedure for the mechanistic model is similar and is given in Appendix A4.1 (Algorithm A4.1). For each model, the chain was run for `steps_total` = 10,000,000 iterations, of which the first 2,000,000 iterations were discarded as burn-in. Posteriors were obtained by recording every 100 iterations of the chain.

Algorithm R4.1: The data augmentation MCMC algorithm used to fit the independent transmission and symptoms model to the household transmission data.

Initialise θ and $\boldsymbol{t}$;
for `step_number` $= 1, \ldots,$ `steps_total` **do**
  if `step_number` $\equiv 1 \pmod 4$ **then**
    Propose new values for each entry of the vector of model parameters, θ, using independent normal proposal distributions for each parameter (around the corresponding parameter values in the previous step of the chain);
    Accept the proposed parameters, θ_{prop}, with probability

$$\min\left(\frac{L\left(\theta_{prop};\boldsymbol{t}\right)\pi\left(\theta_{prop}\right)}{L\left(\theta_{old};\boldsymbol{t}\right)\pi\left(\theta_{old}\right)},1\right),$$

    where θ_{old} denotes the vector of parameter values from the previous step of the chain, and where the augmented data, $\boldsymbol{t}$, remain unchanged in this step;
  else
    if `step_number` $\equiv 2 \pmod 4$ **then**
      Propose new values for the precise symptom onset times of each symptomatic infected host, using independent uniform proposal distributions (within the day of symptom of onset for each host);
    else if `step_number` $\equiv 3 \pmod 4$ **then**
      Propose new values for the infection time of one randomly chosen symptomatic infected host in each household (in households where there was at least one), using independent normal proposal distributions (around the equivalent times in the previous step of the chain);
    else
      Propose new values for the infection time of one randomly chosen asymptomatic infected host in each household (in households where there was at least one), using independent normal proposal distributions (around the equivalent times in the previous step of the chain);
    end
    for $m = 1, \ldots M$ **do**
      Accept the proposed augmented data, $\boldsymbol{t}^{(m)}_{prop}$, for household m, with probability

$$\min\left(\frac{L^{(m)}\left(\theta;\boldsymbol{t}^{(m)}_{prop}\right)}{L^{(m)}\left(\theta;\boldsymbol{t}^{(m)}_{old}\right)},1\right),$$

      where $\boldsymbol{t}^{(m)}_{old}$ denotes the corresponding augmented data from the previous step of the chain, and where the model parameters, θ, remain unchanged in this step;
    end
  end
end

R4.3 Results

R4.3.1 Models and parameter estimation

We fitted two models of infectiousness (the independent transmission and symptoms model and the mechanistic model) to data collected from 172 UK households in a study conducted by PHE between March and November 2020 [115, 168]. Each household was recruited to the study following a confirmed SARS-CoV-2 infection, and all household members were then followed to investigate whether or not they became infected (this was determined through PCR and antibody testing). If a household member was infected and developed symptoms, then their symptom onset date was recorded.

In Chapter R3 we fitted the same two models of infectiousness to data from infector-infectee transmission pairs collected in the early months of the COVID-19 pandemic. Here, we adapted the approach developed in Chapter R3 in order to estimate the generation time using household transmission data. Specifically, we used data augmentation MCMC [40, 41, 82, 204], augmenting the observed data with both estimated times of infection and (estimated) precise times at which symptomatic infected hosts developed symptoms (whereas in Chapter R3, augmented data comprised only symptom onset times). This enabled us (in the likelihood function) to account for uncertainty about exactly who-infected-whom within a household by summing together likelihood contributions corresponding to infection by different possible infectors. In addition, we corrected for the regularity of household contacts (which leads to susceptible depletion within households) to derive more widely applicable estimates of the generation time. We did this by including a factor in the likelihood to account for each infected individual avoiding infection from household contacts that occurred prior to their actual time of infection (see Section R4.2.5).

Initially, we fitted the models to the data from the entire study period. Posterior distributions of all fitted model parameters are shown in Figures R4.2 and R4.3, and point estimates and 95% credible intervals are given in Table A4.2. Since only the likelihood with respect to augmented data was calculated in the MCMC procedure, direct comparisons of the goodness of fit between the models were not readily available. However, comparing model predictions of the distribution of the interval between successive symptom onset dates within households to the analogous distribution in the household data suggested that both models provided a similar fit to the data (Figure A4.1).

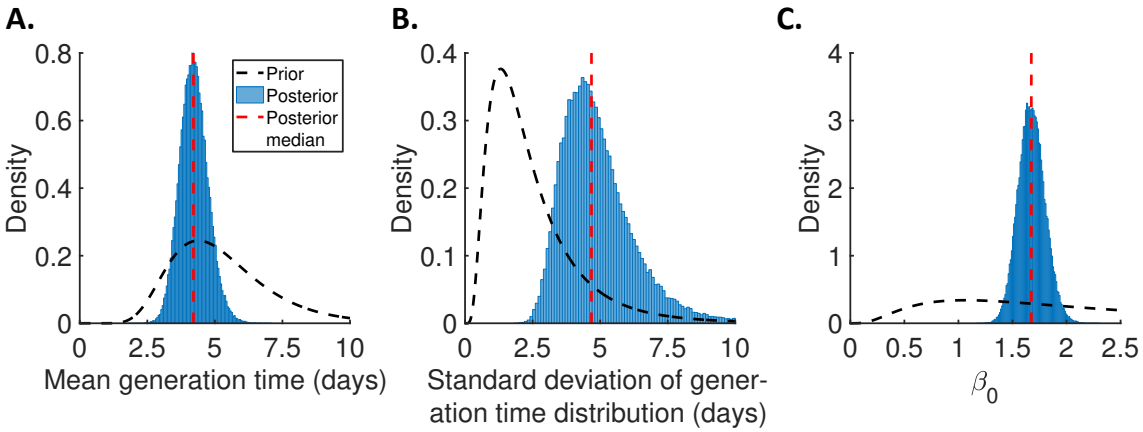


Figure R4.2: Posterior distributions of fitted parameters for the independent transmission and symptoms model. Prior distributions (black dashed lines), posterior distributions (blue bars) and posterior medians (vertical red dashed lines) of fitted parameters in the independent transmission and symptoms model. A. Mean generation time. B. Standard deviation of the generation time distribution. C. Overall infectiousness, β_0.

R4.3.2 Comparison of posterior estimates

For the two fitted models, we calculated posterior estimates of the mean (Figure R4.4A) and standard deviation (Figure R4.4B) of the generation time distribution, in addition to the proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms; Figure R4.4C) and the overall infectiousness parameter, β_0 (Figure R4.4D). Under the commonly used independent transmission and symptoms model, we obtained a point estimate (posterior median) of 4.2 days (95% CrI 3.3–5.3 days) for the mean generation time (Figure R4.4A, blue violin). This value is similar to a previous estimate obtained using data from China [93] (Table R4.1). It is slightly lower than our estimate using the independent transmission and symptoms model in Chapter R3, as well as estimates for Singapore obtained in [93] and for several countries (predominantly in Asia) in [84, 85], although there is some overlap between our credible intervals in Chapter R3 and here (Table R4.1). On the other hand, our estimated standard deviation (under the assumption of independent transmission and symptoms) of 4.7 days (95% CrI 3.0–8.3 days; Figure R4.4B, blue violin) is substantially higher than previous estimates, with no overlap between the credible interval here and our corresponding credible interval in Chapter R3 (Table R4.1). Using our mechanistic model, we obtained a higher estimate for the mean generation time of 6.0 days (95% CrI 5.2–7.0 days; Figure R4.4A, red violin), and a similar estimate for the standard deviation (4.8 days, 95% CrI 4.0–6.3 days; Figure R4.4B, red violin), compared to those predicted by the independent transmission and symptoms model.

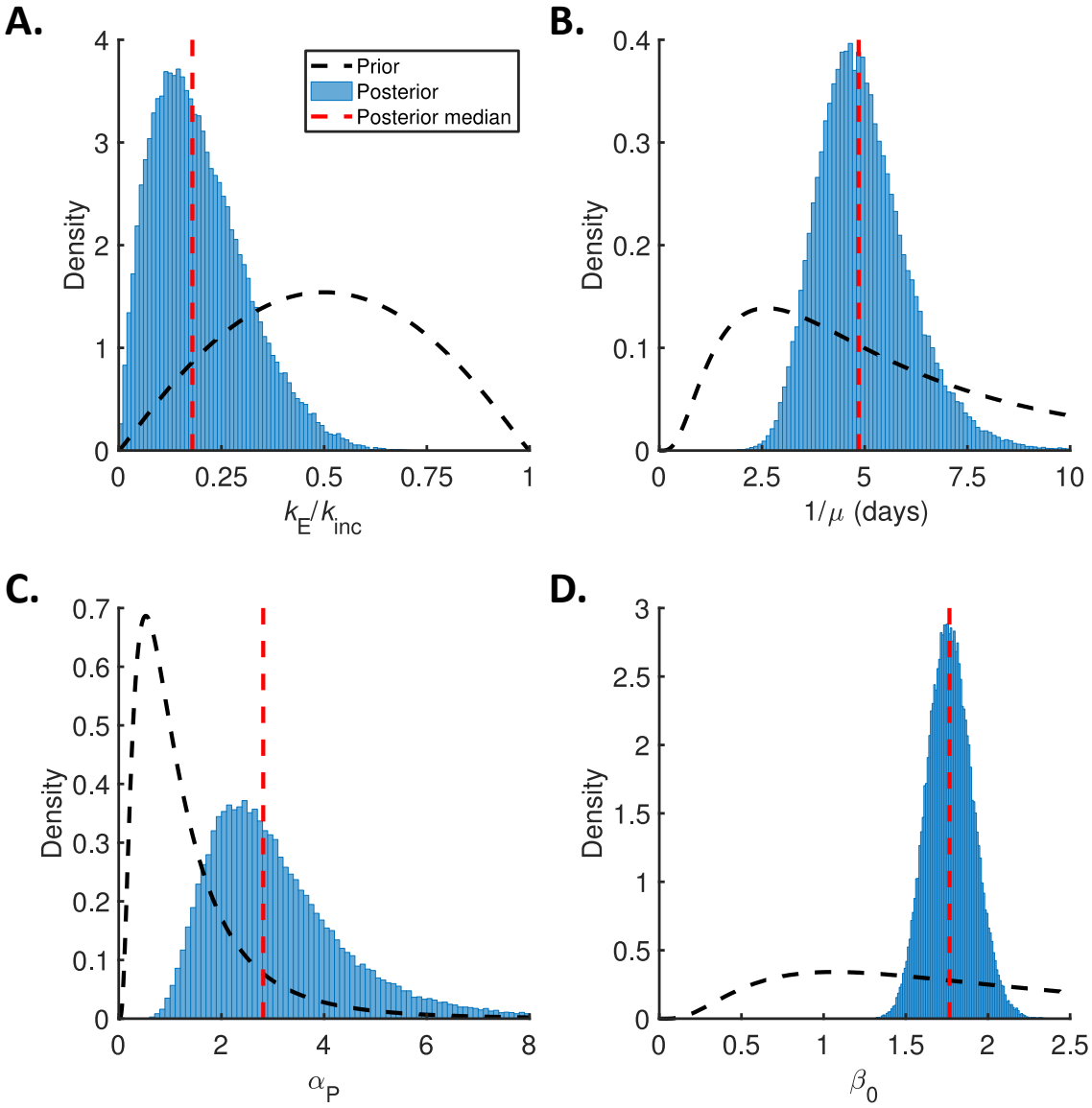


Figure R4.3: Posterior distributions of fitted parameters for the mechanistic model. Prior distributions (black dashed lines), posterior distributions (blue bars) and posterior medians (vertical red dashed lines) of fitted parameters in the mechanistic model. A. Ratio of mean durations of the latent (E) and incubation (combined E and P) periods, k_E/k_{inc}. B. Mean symptomatic infectious (I) period, $1/\mu$. C. Ratio of transmission rates in the presymptomatic infectious (P) and symptomatic infectious (I) stages of infection, α_P. D. Overall infectiousness, β_0.

The two models gave similar posterior distributions for the proportion of transmissions prior to symptom onset (Figure R4.4C). Specifically, point estimates for the proportion of transmissions prior to symptom onset were 72% (95% CrI 63–80%) for the independent transmission and symptoms model, and 73% (95% CrI 61–83%) for the mechanistic model. These estimates are higher than obtained in some previous studies in which the infectiousness profile of SARS-CoV-2 infected hosts at each time

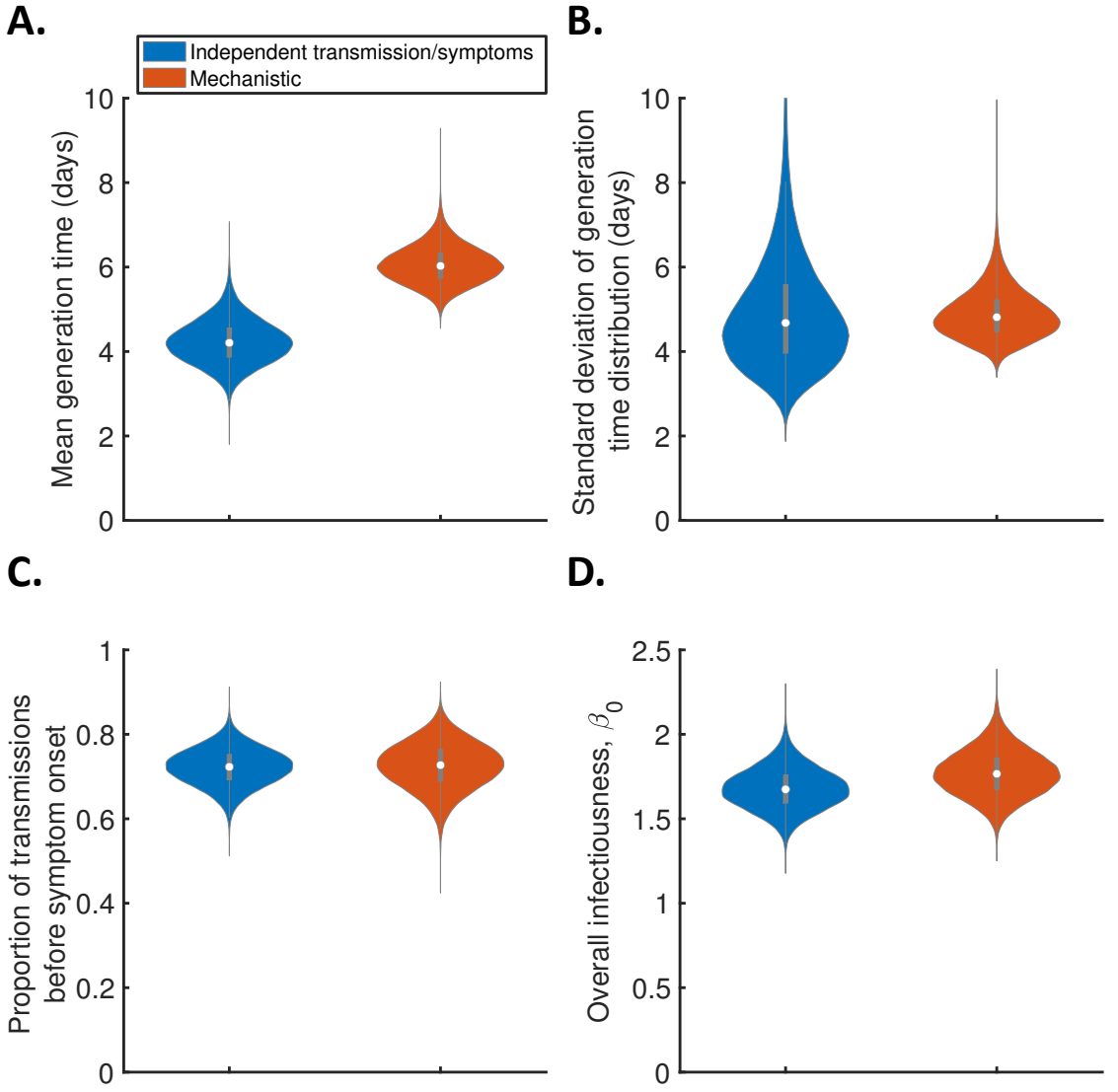


Figure R4.4: Comparison of posterior estimates between the two models. Violin plots indicating posterior distributions of the mean (A) and standard deviation (B) of the generation time distribution, proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) (C), and overall infectiousness parameter, β_0 (describing the expected number of household transmissions generated by a single infected host (who develops symptoms) in a large, otherwise entirely susceptible, household (D). Results are shown for both the independent transmission and symptoms model (blue), and the mechanistic model (red).

since infection and/or time since symptom onset has been estimated [14, 84, 85, 118]. On the other hand, our point estimates for the two models both lie within the 95% credible interval obtained for the mechanistic model in Chapter R3 (52–74%, point estimate 64%). Similar or higher estimates also exist in the wider literature [39, 93, 226].

R4.3.3 Estimation of realised generation times within households

In Figure R4.4 (and elsewhere, unless otherwise stated) we characterise the generation time distribution assuming that a constant supply of susceptible individuals are available to infect throughout the course of infection. This distribution corresponds to the normalised expected infectiousness profile of an infected host at each time since infection, and is widely applicable to transmission outside of, as well as within, households. However, realised household generation times are expected to be shorter than the estimates shown in Figure R4.4. This is due to the depletion of susceptible household members before longer generation times can be realised (especially in small households) [41, 88, 191]. As a result, we also predicted the mean and standard deviation of realised generation times within the study households (Figure R4.5A–B), accounting for the precise distribution of household sizes in the study.

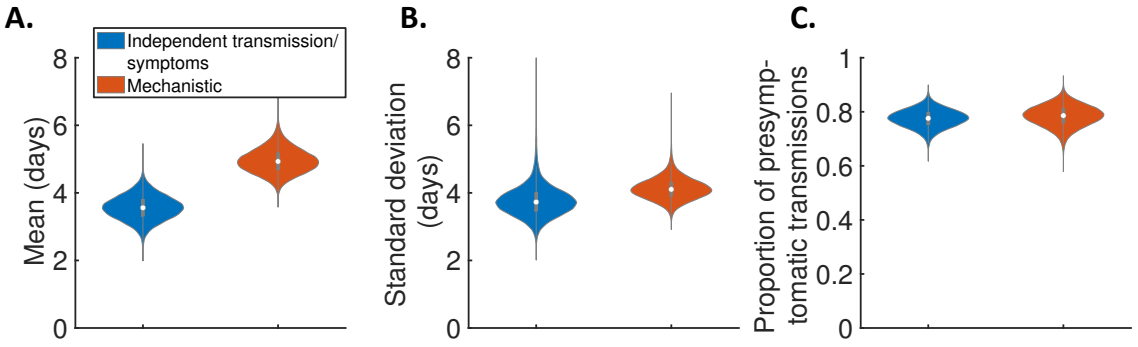


Figure R4.5: Estimation of realised generation times within study households. Violin plots indicating posterior distributions of the mean (A) and standard deviation (B) of realised generation times in the study households, and the realised proportion of transmissions occurring prior to symptom onset (among infectors who developed symptoms) (C), for the independent transmission and symptoms model (blue) and mechanistic model (red).

For both the independent transmission and symptoms model and the mechanistic model, the mean (point estimates (95% CrIs) 3.6 days (2.8–4.3 days) and 4.9 days (4.3–5.7 days) for the two models, respectively) and standard deviation (3.7 days (3.0–4.9 days) and 4.1 days (3.5–4.8 days)) of realised household generation times were lower than our main generation time estimates shown in Figure R4.4. Since household transmission typically occurs earlier in the infector's course of infection than indicated by the estimates shown in Figure R4.4, we predicted a higher proportion of presymptomatic transmissions within the study households (Figure R4.5C) compared to the estimates in Figure R4.4C.

R4.3.4 Distributions of epidemiological time intervals

For both models, we also used point estimates of fitted model parameters to infer the distributions of the generation time (Figure R4.6A), the TOST (Figure R4.6B) and the serial interval (Figure R4.6C). The TOST distribution (which, we recall, characterises the relative expected infectiousness of a host (who develops symptoms) at each time from symptom onset, as opposed to from infection [14, 84, 118, 145, 240]) obtained using the mechanistic model was more concentrated around the time of symptom onset compared to that predicted assuming independent transmission and symptoms (Figure R4.6B), as we found in Chapter R3 (Figure R3.5). In contrast, the estimated serial interval distributions were similar for the two models (Figure R4.6C).

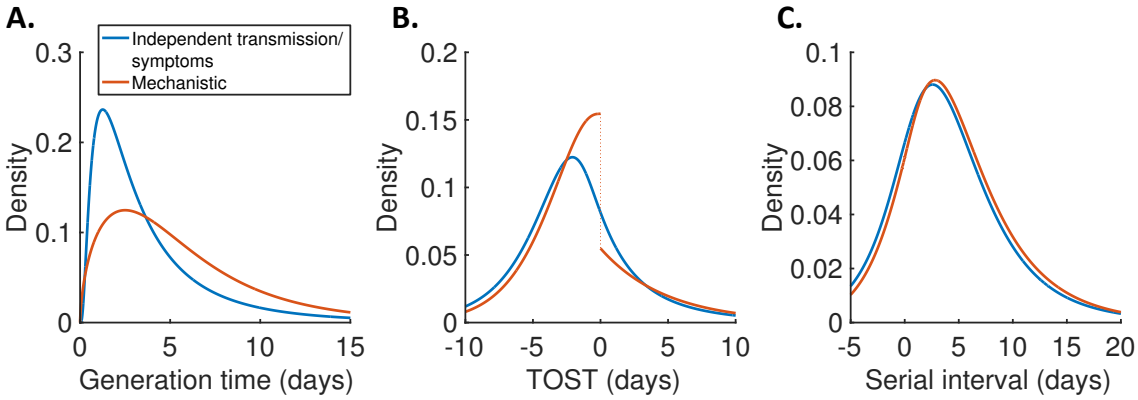


Figure R4.6: Generation time, TOST and serial interval distributions. Inferred generation time (A), TOST (B) and serial interval (C) distributions for the two models, obtained using point estimate (posterior median) parameters. The means and standard deviations of these distributions are given in Table A4.3. Similarly to Chapter R3 (Figure R3.5), the discontinuity in the red curve in Figure R4.6B occurs because different transmission rates were fitted for infectors in the presymptomatic infectious and symptomatic infectious periods. The reduction in transmission following symptom onset can be attributed to changes in behaviour in response to symptoms [160].

The means and standard deviations of the distributions shown in Figure R4.6 are given in Table A4.3. For the independent transmission and symptoms model, the generation time and serial interval distributions have the same mean, as is commonly assumed to be the case [145]. However, this is not true for the mechanistic model, in which hosts with longer presymptomatic infectious periods generate (on average) a higher number of transmissions. As a result, under the mechanistic model, a randomly chosen transmission is more likely to arise from an infector with a longer incubation period, thereby leading to a longer mean generation time than serial interval (since the difference between the generation time and the serial interval is equal to the difference between the incubation periods of the infector and infectee).

In Appendix A4.2 we consider the relationship between the generation time, TOST and serial interval in more detail, and derive an analytical expression for the exact difference between the mean generation time and serial interval for the mechanistic model.

R4.3.5 Temporal variation in the generation time distribution

To explore whether or not the generation time distribution changed during the study period, we then separately fitted both models to the data from households in which the index case was recruited in (i) March–April, (ii) May–August, or (iii) September–November 2020 (Figures R4.7 and R4.8). We chose these time periods to ensure the numbers of households recruited into the study during each interval were similar (Figure R4.1).

The results for the independent transmission and symptoms model shown in Figure R4.7A suggest a shorter mean generation time in September–November 2020 (2.9 days, 95% CrI 1.8–4.3 days) compared to earlier months (4.8 days, 95% CrI 3.6–6.3 days, for March–April and 5.1 days, 95% CrI 3.4–7.2 days, for May–August). Comparing the posterior estimates for May–August and September–November (the red and orange violins in Figure R4.7A, respectively) indicated a 97% posterior probability of a shorter mean generation time in the later of these two time periods. A similar temporal reduction in the mean generation time was found when we instead fitted the mechanistic model to the data from the three time intervals (Figure R4.8). Estimates of the mean generation time using the mechanistic model were 6.6 days (95% CrI 5.6–8.1 days) for March–April, 7.2 days (95% CrI 5.7–9.6 days) for May–August, and 5.1 days (95% CrI 4.3–6.4 days) for September–November, with a 98% posterior probability of a shorter mean generation time in September–November than May–August.

We also used point estimates of model parameters to compare the distributions of the generation time, TOST and serial interval between the time periods (Figure A4.2). Both models indicated that the transmission risk peaked earlier in infection for individuals infected in September–November compared to earlier months (Figure A4.2A,D).

Figure R4.7C shows posterior estimates for the proportion of transmissions occurring prior to symptom onset (among symptomatic infectors) across the three time periods for the independent transmission and symptoms model, indicating a very high proportion of presymptomatic transmissions in September–November (84%, 95% CrI 72–93%) compared to lower estimates for March–April (64%, 95% CrI

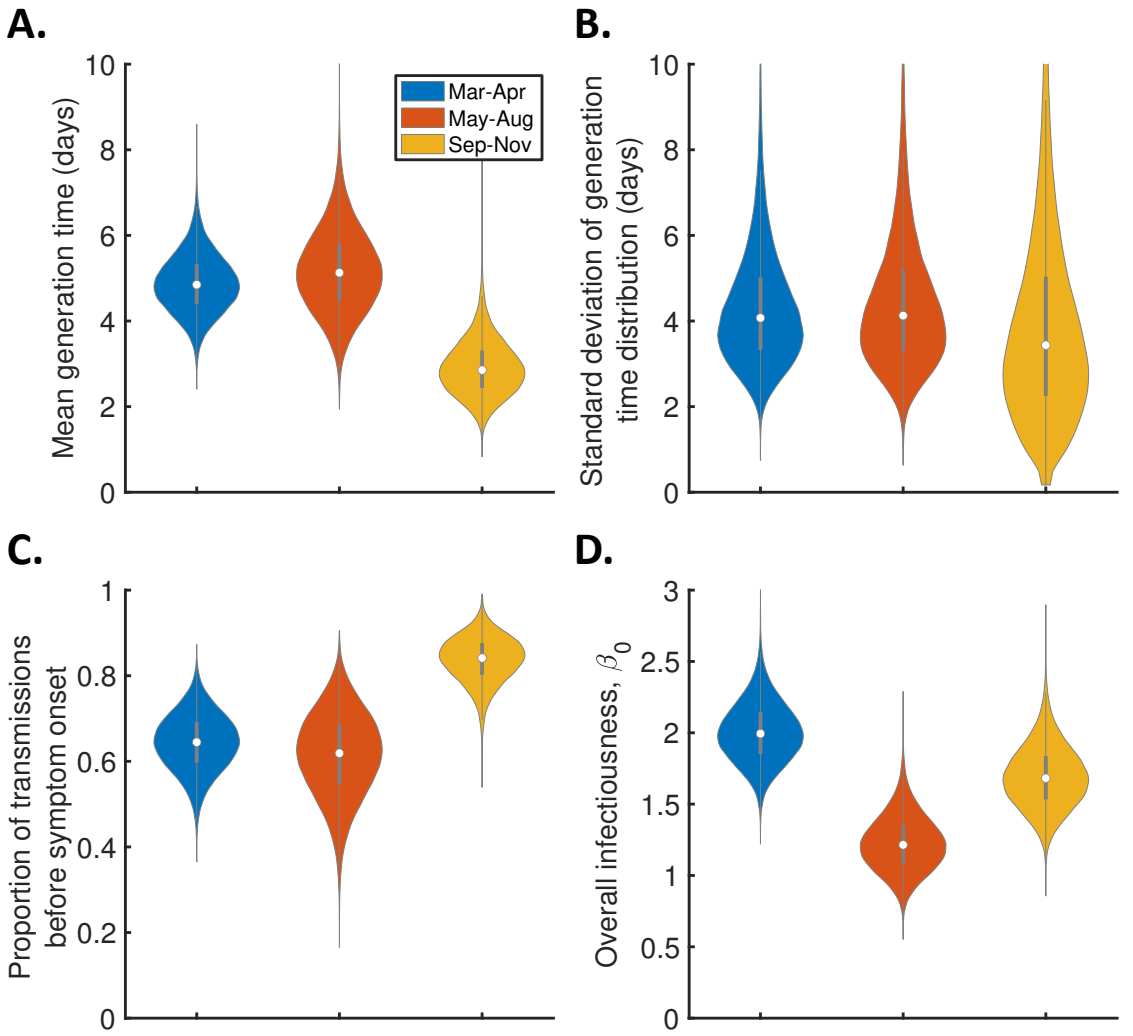


Figure R4.7: Temporal changes in the generation time for the independent transmission and symptoms model. Violin plots indicating posterior distributions of the mean (A) and standard deviation (B) of the generation time distribution, proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) (C), and overall infectiousness parameter, β_0 (D), for the independent transmission and symptoms model fitted to data from March–April (blue), May–August (red) or September–November 2020 (orange).

51–77%) and May–August (62%, 95% CrI 41–79%). Our results for the mechanistic model indicate a similar temporal increase in the proportion of presymptomatic transmissions during the study period (Figure R4.8C).

To explore the shorter estimated generation time for September–November further, we also fitted the independent transmission and symptoms model to the data from each of these months individually (Figure R4.9). The shorter estimated generation time compared to earlier in the pandemic was consistent across each of the three months (Figure R4.9A). We note that, while the Alpha (B.1.1.7) variant had begun to emerge in the UK by the end of the study period [200], genomic surveillance

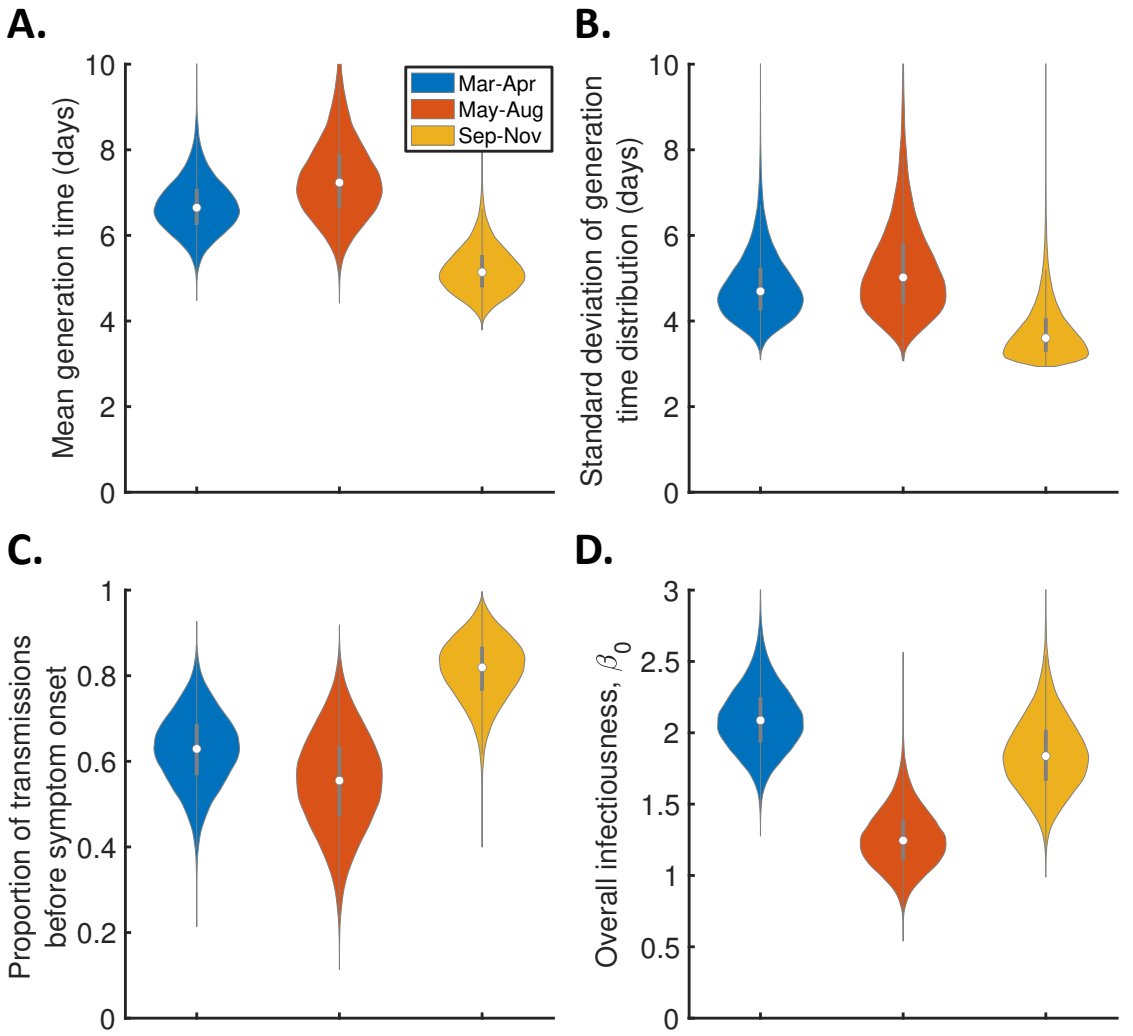


Figure R4.8: Temporal changes in the generation time for the mechanistic model. Violin plots indicating posterior distributions of the mean (A) and standard deviation (B) of the generation time distribution, proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) (C), and overall infectiousness parameter, β_0 (D), for the mechanistic model fitted to data from March–April (blue), May–August (red) or September–November 2020 (orange).

showed that this variant caused infections in only two study households. This variant was therefore unlikely to have been responsible for the temporal reduction in the generation time that we observed.

In Figure A4.3 we show the posterior distributions of the fitted parameters for the mechanistic model over the different time periods. The fitted parameters represent the mean duration of the symptomatic infectious period (Figure A4.3B), the relative infectiousness of presymptomatic infectious hosts compared to those with symptoms (Figure A4.3C), and the mean duration of the latent period (as a proportion of the mean incubation period; Figure A4.3A). However, there was substantial overlap

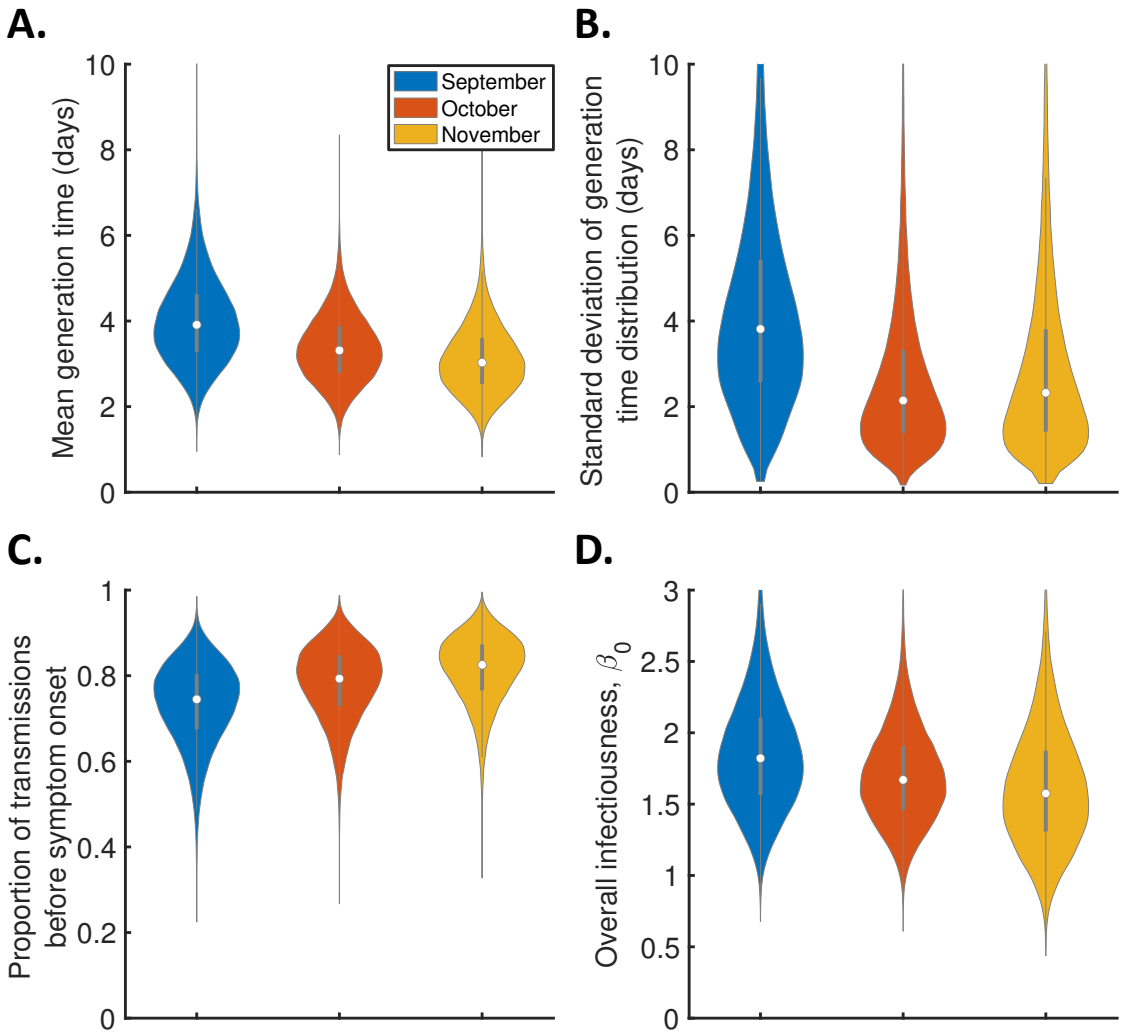


Figure R4.9: Monthly changes in the generation time from September–November 2020. Violin plots indicating posterior distributions of the mean (A) and standard deviation (B) of the generation time distribution, proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) (C), and overall infectiousness parameter, β_0 (D), for the independent transmission and symptoms model fitted to data from September (blue), October (red) or November 2020 (orange).

in the credible intervals for each of these three parameters between the three time periods. We were therefore unable to identify the precise parameter(s) responsible for the decrease in generation time and increase in the proportion of presymptomatic transmissions that we observed.

R4.3.6 Sensitivity analyses

R4.3.6.1 Effect of accounting for the possibility of co-primary cases

When we fitted the two models to the household transmission data, we assumed that each household transmission chain was initiated by a single primary case and all

other infected household members were infected from within the household. However, we also extended our framework to account for the possibility of co-primary cases (see Appendix A4.3 for details of this extension).

Accounting for co-primary cases led to slightly higher estimates of the mean generation time (Figure R4.10A) under each model compared to the corresponding estimates shown in Figure R4.4A, with point estimates of 4.8 days (95% CrI 3.6–6.3 days) for the independent transmission and symptoms model, and 6.9 days (95% CrI 5.7–8.6 days) for the mechanistic model. Estimates of the standard deviation of the generation time distribution were similar to those in Figure R4.4B (Figure R4.10B; point estimates were 4.5 days (95% CrI 2.9–7.9 days) for the independent transmission and symptoms model and 5.1 days (95% CrI 4.0–6.9 days) for the mechanistic model). As part of the fitting procedure, we estimated the probability that each household member was infected during the primary transmission event (Figure R4.10E), obtaining point estimates of 0.17 (95% CrI 0.02–0.33) under the independent transmission and symptoms model, and 0.28 (95% CrI 0.10–0.41) under the mechanistic model.

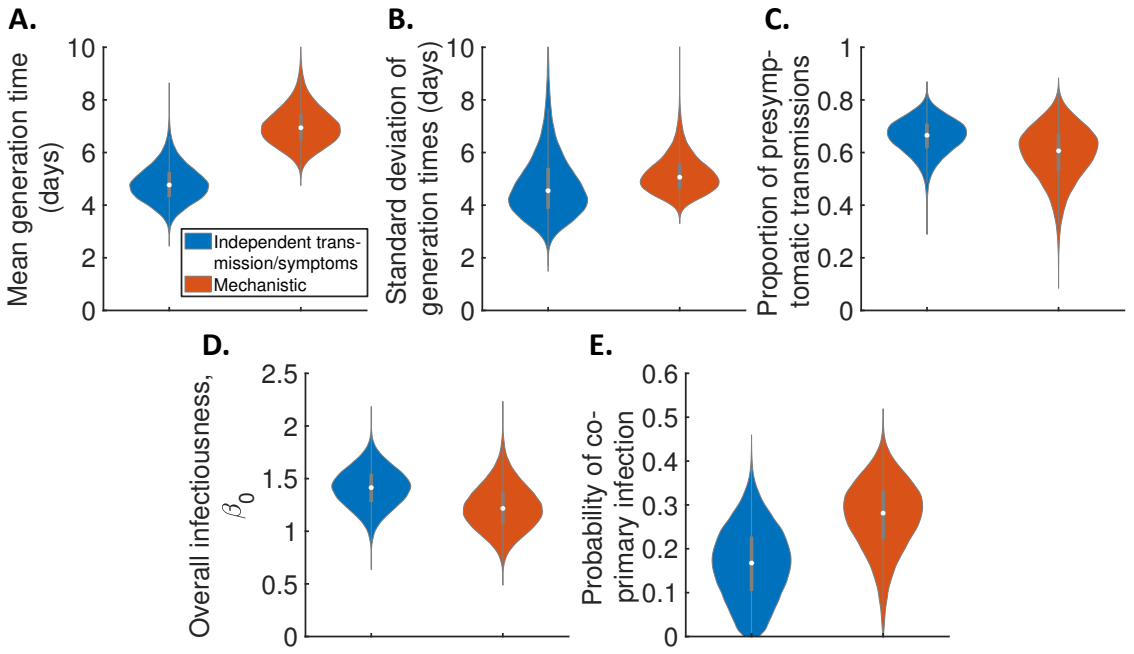


Figure R4.10: Effect of accounting for the possibility of co-primary cases on estimates of the generation time. Violin plots indicating posterior distributions of the mean (A) and standard deviation (B) of the generation time distribution, proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) (C), overall infectiousness parameter, β_0 (D), and probability of co-primary infection (E), for the independent transmission and symptoms model (blue) and mechanistic model (red), when the possibility of co-primary cases was included in our approach.

We also repeated the analyses in Figure R4.7 but accounting for the possibility of co-primary cases (Figure R4.11). Our main finding remained unchanged: the generation time was found to decrease during the study period (Figure R4.11A).

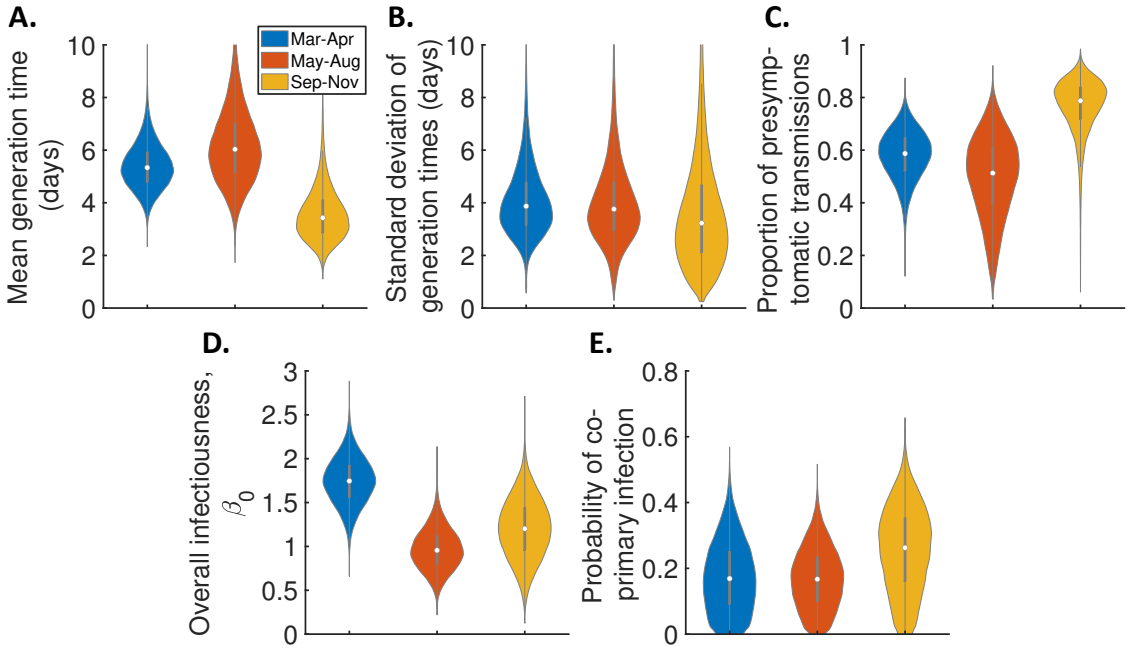


Figure R4.11: Temporal changes in the generation time for the independent transmission and symptoms model, accounting for the possibility of co-primary cases. Violin plots indicating posterior distributions of the mean (A) and standard deviation (B) of the generation time distribution, proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) (C), overall infectiousness parameter, β_0 (D), and probability of co-primary infection (E), for the independent transmission and symptoms model fitted to data from March–April (blue), May–August (red) or September–November 2020 (orange), when the possibility of co-primary cases was included in our approach.

R4.3.6.2 Effect of the functional forms of the generation time and incubation period distributions

In the independent transmission and symptoms model, we assumed that both the generation time and incubation period followed lognormal distributions. The mean and standard deviation of the generation time distribution were estimated by fitting the model to the household transmission data. In the fitting procedure, we assumed that the incubation period followed a lognormal distribution that was obtained in a previous meta-analysis [162]. In contrast, we assumed in our mechanistic approach that each infection could be decomposed into three gamma-distributed stages (latent, presymptomatic infectious and symptomatic infectious), so that the incubation period was also gamma-distributed (with the same mean and standard deviation as

the lognormal distribution obtained in [162]). An expression for the generation time distribution in the mechanistic model, which did not take a simple parametric form, is given in Section R4.2.4.1. However, we conducted supplementary analyses in which we instead assumed that either the generation time (Figure A4.4) or incubation period (Figure A4.5) in the independent transmission and symptoms model was gamma-distributed. In both cases, we obtained similar results to those shown for that model in Figure R4.4.

R4.3.6.3 Effect of accounting for uncertainty in the incubation period distribution

In Figure R4.12 we relaxed the assumption of a fixed incubation period distribution, instead using the confidence intervals obtained in [162] to account for uncertainty in the incubation period distribution (Figure R4.12A–B). Details of this analysis are given in Appendix A4.4. For both the independent transmission and symptoms model and the mechanistic model, accounting for uncertainty in the incubation period distribution did not substantially affect posterior estimates of either the mean (Figure R4.12C) or the standard deviation (Figure R4.12D) of the generation time distribution.

R4.3.6.4 Effect of the dependency of transmission on the household size

In our main analyses, we assumed that household transmission was frequency-dependent, so that the force of infection exerted by an infected household member on each susceptible household member scaled with $1/n$, where n is the household size [40, 42]. However, since some studies of influenza virus transmission in households have found transmission to lie somewhere in between frequency- and density-dependent [80, 82], we also considered alternative possibilities where infectiousness scaled with $n^{-\rho}$, for different values of ρ. In Figure R4.13A–C we compared estimates under our baseline value of $\rho = 1$ (frequency-dependent transmission) with those obtained assuming either $\rho = 0$ (density-dependent transmission) or the intermediate possibility of $\rho = 0.5$ considered in [80]. We also conducted an analysis in which the dependency, ρ, was estimated alongside other model parameters (Figure R4.13D).

We found that our estimates of the mean and standard deviation of the generation time distribution were robust to the assumed value of ρ (Figure R4.13A–B). However, when the value ρ was fitted (Figure R4.13D), we estimated a value of 1.0 (95% CrI 0.6–1.5). This supported our assumption of frequency-dependent transmission, although the credible interval was relatively wide. In addition, we considered the possibility that infectiousness instead scaled with $1/(n-1)$, so that the infector

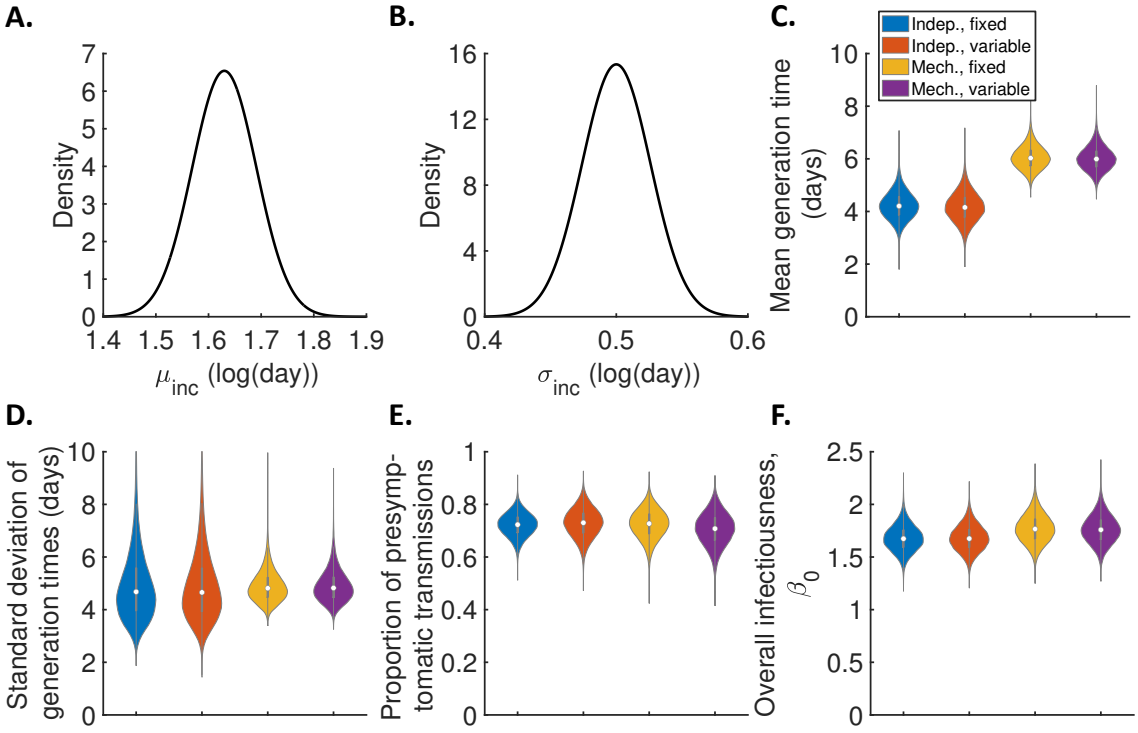


Figure R4.12: Effect of accounting for uncertainty in the incubation period distribution on estimates of the generation time. A–B. Prior distributions of the parameters, μ_{inc} (A) and σ_{inc} (B), of the lognormal incubation period distribution in the independent transmission and symptoms model (μ_{inc} and σ_{inc} represent the mean and standard deviation of the natural logarithm of the incubation period, respectively). In the mechanistic model, we assumed a gamma-distributed incubation period with the same mean and standard deviation as a lognormal distribution with parameters μ_{inc} and σ_{inc}. See Appendix A4.4 for details of how we accounted for uncertainty in the incubation period distribution using these prior distributions. C–F. Violin plots indicating posterior distributions of the mean (C) and standard deviation (D) of the generation time distribution, proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) (E), and overall infectiousness parameter, β_0 (F), for the independent transmission and symptoms model with either a fixed (as in our main analyses; blue) or variable (i.e. accounting for uncertainty in μ_{inc} and σ_{inc} as described above; red) incubation period distribution, and for the mechanistic model with a fixed (orange) or variable (purple) incubation period distribution.

under consideration is not included in this scaling, again obtaining similar estimates of the mean and standard deviation of the generation time distribution compared to those shown in Figure R4.4 (Figure A4.6).

R4.3.6.5 Effect of the relative infectiousness of entirely asymptomatic infected hosts

We also considered the sensitivity of our results to the assumed relative infectiousness of asymptomatic infected hosts (Figure A4.7). In most of our analyses, we assumed

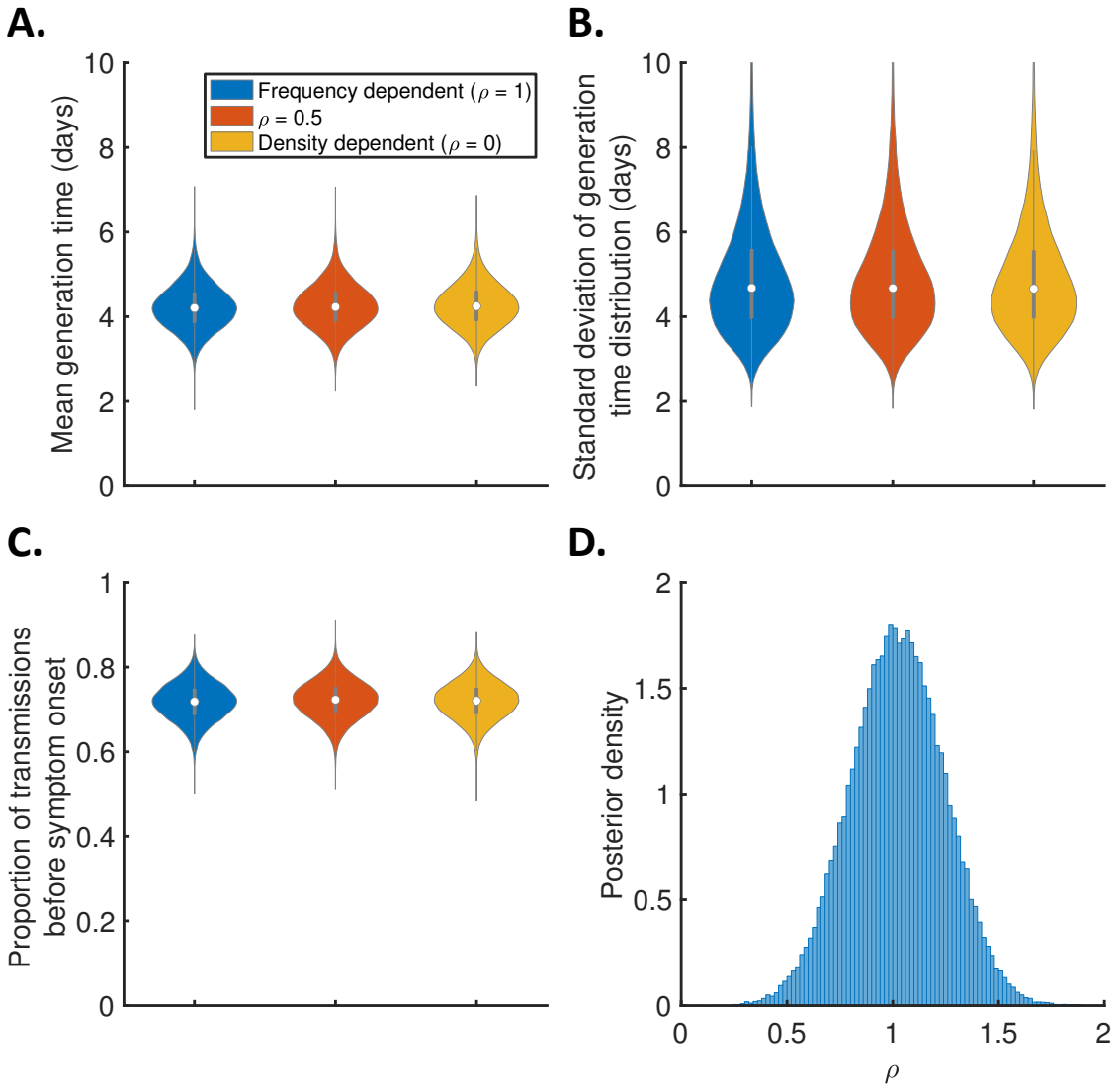


Figure R4.13: Effect of the dependency of transmission on the household size on estimates of the generation time. A–C. Violin plots indicating the posterior distributions of the mean (A) and standard deviation (B) of the generation time distribution, and proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) (C), for the independent transmission and symptoms model under different assumptions about the dependency of transmission on the household size. In these panels, infectiousness is assumed to scale with $n^{-\rho}$, where n is the household size, for $\rho = 1$ (frequency-dependent transmission; blue), $\rho = 0.5$ (red) and $\rho = 0$ (density-dependent transmission; orange). D. Posterior distribution of the dependency, ρ, when it was fitted to the data (alongside other model parameters), assuming a uniform prior distribution for ρ.

that the expected infectiousness of an infected host who remains asymptomatic throughout infection is a factor $\alpha_A = 0.35$ times that of a host who develops symptoms, at each time since infection [30]. However, similar estimates of the mean (Figure A4.7A) and standard deviation (Figure A4.7B) of the generation time

distribution were obtained when we instead assumed $\alpha_A = 0.1$ or $\alpha_A = 1.27$ (these values correspond to the lower and upper confidence bounds obtained in [30]). Lower values of α_A did lead to slightly higher estimates of the overall infectiousness of infectors who develop symptoms, β_0 (Figure A4.7D). However, this effect was minimal, most likely because very few cases in the household study were asymptomatic (27 out of 357).

R4.3.6.6 Effect of the exclusion of individuals of unknown infection status

Finally, we explored the robustness of our results to the exclusion of individuals of unknown infection status, considering the extreme possibilities where these individuals were instead assumed to have either all remained uninfected, or all become infected (Figure A4.8). Although the estimates of β_0 were affected by this assumption (Figure A4.8D), the estimated generation time distribution was robust to the assumed infection status of these individuals (Figure A4.8A–B).

R4.4 Discussion

In this research chapter, we estimated the generation time distribution of SARS-CoV-2 in the UK by fitting two different models to data describing the infection status and symptom onset dates of individuals in 172 households. The first model was predicated on an assumption that transmission and symptoms are independent. While this assumption has often been made in previous studies in which the SARS-CoV-2 generation time has been estimated [46, 74, 85, 93, 137], it is not an accurate reflection of the underlying epidemiology [16, 145]. Therefore, we also considered the mechanistic model (based on compartmental modelling) that we developed in Chapter R3, and which was shown in Chapter R3 to provide an improved fit to data from 191 SARS-CoV-2 infector-infectee pairs compared to previous models that have been used to estimate the generation time.

Here, infection times and the order of transmissions within households were unknown, whereas in Chapter R3 the direction of transmission was assumed to be known for each infector-infectee pair. We therefore extended the statistical inference methods underlying Chapter R3 to fit the two models to household data, adapting the likelihood function derived in Chapter R3 to account for both uncertainty in who-infected-whom and the regularity of household contacts. Similarly to previous studies of household influenza virus transmission [40, 41, 82, 204], we used data augmentation MCMC to fit the models to the household data using this adapted

likelihood function, augmenting the observed data with both (estimated) times of infection and precise symptom onset times within recorded onset dates (whereas only symptom onset times were imputed in Chapter R3).

Under the model assuming independent transmission and symptoms, we estimated a mean generation time of 4.2 days (95% CrI 3.3–5.3 days) and a standard deviation of 4.7 days (95% CrI 3.0–8.3 days). The estimate of the mean generation time was comparable to previous estimates obtained under this assumption using data from elsewhere in Chapter R3 and other studies [84, 85, 93] (Table R4.1). On the other hand, while our 95% credible interval for the standard deviation was wide, even the lower credible bound here exceeded the upper credible bound of our corresponding estimate in Chapter R3 (Table R4.1). One potential cause of this disparity in estimates of the standard deviation is the difference in isolation policies for symptomatic hosts between countries. In particular, the UK's policy of self-isolation may be expected to lead to a longer-tailed generation time distribution compared to countries with a policy of isolation outside the home, since under home isolation, some within-household transmission is likely to occur even following isolation. Isolation outside the home was commonplace in the East and Southeast Asian countries where the majority of the data underlying the estimates shown in Table R4.1 [84, 85, 93] were collected.

Using the mechanistic model, we predicted a higher mean generation time of 6.0 days (95% CrI 5.2–7.0 days) compared to the value estimated under the assumption of independent transmission and symptoms. On the other hand, the inferred serial intervals for the independent transmission and symptoms model and the mechanistic model were more similar (Figure R4.6C), with means of 4.2 days and 4.7 days, respectively. Temporal information in our household transmission data consisted mostly of symptom onset dates, with very few individuals testing positive before developing symptoms. Therefore, the variation in estimates of the generation time between the models can be attributed to differences in the assumed relationships between the generation time and serial interval under those models. Specifically, the generation time and serial interval have the same mean in the independent transmission and symptoms model (as is commonly assumed to be the case [145]), but in the mechanistic model the generation time is longer on average than the serial interval because of the realistic assumption that individuals with longer incubation periods generate more transmissions on average (see Section R4.3.4 and Appendix A4.2).

Our results do not indicate any clear difference in goodness of fit to the household data between the two models (Figure A4.1). A range of factors should therefore be considered when deciding which of our estimates of epidemiological parameters to

use in subsequent analyses. Although any model requires simplifying assumptions to be made, as discussed in Chapter R3 our mechanistic approach allows the standard assumption of independent transmission and symptoms to be relaxed by providing a mechanistic underpinning to the relationship between infectiousness and symptoms. Furthermore, as described above, this model was shown in Chapter R3 to provide a better fit to an earlier SARS-CoV-2 dataset than a model assuming independence between transmission and symptoms (in Chapter R3, the simpler setting of transmission pairs rather than households facilitated direct model comparison). On the other hand, the independent transmission and symptoms model has the advantage of producing an estimated generation time distribution with a simple parametric form. The choice of estimates to use may also depend on precisely what the estimates are being used for. For example, the generation time distribution inferred under the assumption of independent transmission and symptoms may be better suited for use in some models for estimating the time-dependent reproduction number, since those models often also involve the assumption that transmission and symptoms are independent [1]. In contrast, the parameter estimates from our mechanistic approach (Table A4.2) correspond naturally to parameters in compartmental epidemic models.

By fitting separately to data from three different intervals within the study period (March–November 2020), we investigated whether or not the generation time distribution in the UK changed as the pandemic progressed. Our results indicate a shorter mean generation time in September–November compared to earlier months (Figure R4.7A). One possible explanation for this is a higher proportion of time spent indoors in colder months leading to an increased transmission risk, particularly in the early stages of infection before symptoms develop (since symptomatic infected hosts were still expected to self-isolate). This explanation is consistent with our finding in Figure R4.7C of a higher proportion of transmissions occurring prior to symptom onset in September–November compared to March–April and May–August.

While behavioural changes may have been responsible for our finding of a temporal decrease in the generation time, an alternative explanation could be that evolutionary changes in the SARS-CoV-2 virus that occurred during the study period affected the generation time. For example, the B.1.177 lineage emerged in Spain in early summer 2020, and became the dominant SARS-CoV lineage in the UK around the beginning of October 2020 [234]. Subsequently, the Alpha (B.1.1.7) variant, which was first detected in September 2020, became dominant in the UK in December 2020 [200]. The Alpha variant has been shown to possess different characteristics compared to earlier variants, causing an increased epidemic growth rate in the UK that has been attributed to an increase in transmissibility of 43–90% [68]. While, in principle,

evolutionary changes could explain the variation in the generation time that we observed, sequencing data show that the Alpha variant was responsible for infections in only two households within our dataset. Consequently, the Alpha variant was not responsible for our main finding of a temporally decreasing generation time, and additional data are required to quantify the impact of the emergence of that variant (and subsequent variants, such as the Delta (B.1.617.2) and Omicron (B.1.1.529) variants) on the SARS-CoV-2 generation time (see Section D.1 where we discuss work we have now carried out since conducting the analyses described in this research chapter, in which we provide generation time estimates for the Alpha and Delta variants [114]).

In data collected from infector-infectee transmission pairs, shorter generation times are expected to be over-represented at times when case numbers are rising [27, 85, 145] and vice versa. While we used data from households (rather than transmission pairs) in our analyses, a similar effect may have contributed to our shorter estimated mean generation time for September–November 2020 (national case numbers were mostly increasing in September–October 2020) compared to earlier months of the study (in which case numbers were mostly decreasing) [138, 197]. However, we estimated the mean generation time to be similar in November (when case numbers were mostly decreasing [138, 197]) compared to September and October (Figure R4.9), suggesting that this effect of background epidemic dynamics alone did not drive the temporal changes in the generation time that we observed. We note, however, that sample sizes for individual months were small (Figure R4.1). Extending our household inference framework to explicitly account for background epidemic dynamics in generation time estimates (similar to methods that have been developed for transmission pair data [27, 85]) is an avenue for future work.

Our finding of a temporal decrease in the mean generation time during the study period highlights the importance of obtaining up-to-date generation time estimates specific to the location under study. Should variations in the generation time distribution occur and not be accounted for, estimates of the time-dependent reproduction number may be incorrect [189, 235]. Specifically, if the mean generation time is shorter than assumed, then the true value of the time-dependent reproduction number is likely to be closer to one than the inferred value [235] (and vice versa).

One advantage of our approach compared to previous studies in which the SARS-CoV-2 generation time has been estimated [84, 85, 93, 116] is that we were able to include the contribution of asymptomatic infected hosts to household transmission chains in our analyses. We showed that our estimated generation time distribution was robust to the assumed relative infectiousness of infected hosts who remained

asymptomatic, α_A (Figure A4.7). Similarly, while we assumed frequency-dependent household transmission in most of our analyses, we found that the exact relationship between transmission and the household size had little effect on our estimates of the mean and standard deviation of the generation time distribution (Figures R4.13 and A4.6). We also considered estimating the exponent governing the dependency of transmission on household size (Figure R4.13D). This supported our assumption of frequency-dependent transmission, and is consistent with the finding of an inverse relationship between household size and secondary attack rate in the household study underlying our analyses [168]. In previous studies of influenza transmission within households, evidence has been found both in favour of [40] and against [80] frequency-dependent transmission.

While our generation time estimates were robust to the assumed relative infectiousness of infected hosts who remained asymptomatic and whether transmission was assumed to be frequency- or density-dependent, extending our approach to account for the possibility that household transmission chains originated with multiple co-primary cases led to slightly higher estimates of the generation time (Figure R4.10) compared to our main estimates (Figure R4.4). Despite the overall higher estimated generation time, our main qualitative finding of a temporal decrease in the generation time still held when co-primary cases were incorporated (Figure R4.11).

Like any study, the research presented in this chapter has some limitations. We used household data in our analyses, whereas some characteristics of wider community transmission may differ from those of transmission within households. However, we corrected for the regularity of household contacts to estimate the (expected) infectiousness profile of an infected host at each time since infection, which provides a widely applicable generation time estimate (Figure R4.4). Specifically, the infectiousness profile gives the generation time distribution under the assumption that a constant supply of susceptible individuals are available to infect throughout the course of infection. This distribution can be conditioned to estimate realised generation times within specific population structures (in scenarios where susceptible depletion cannot be neglected), as we demonstrated by estimating the realised generation time distribution within the study households (Figure R4.5). The household generation time estimates shown in Figure R4.5 are shorter than our main generation time estimates (Figure R4.4), due to the regularity of household contacts and the depletion of susceptible individuals within households before longer generation times can be realised.

We also note that, while our dataset involved a larger sample size than used in most previous studies in which the SARS-CoV-2 generation time was estimated

[84, 85, 93, 116], the demographics of the study households may not have been completely representative of the wider population. Exploring heterogeneity in the generation time distribution between individuals and/or households with different characteristics is an important topic for future work. This could involve, but is not limited to, estimating the generation time distribution for individuals of different age, sex, ethnicity and socio-economic status. Nonetheless, as well as providing updated SARS-CoV-2 generation time estimates, our study demonstrates that changes in the generation time can be detected using data from household studies. Our finding that the generation time has become shorter highlights both the importance of continued monitoring of the generation time and the role of household studies in such monitoring efforts, particularly in light of the more recent emergence of novel SARS-CoV-2 variants.

In summary, in this chapter we have presented research in which we inferred the SARS-CoV-2 generation time distribution in the UK using household data and two different transmission models. A key output of this research is one of the first estimates of the SARS-CoV-2 generation time outside Asia. Another crucial feature of our analyses is that we used data from beyond the first few months of the pandemic. Since this research suggests that the generation time may be changing, continued data collection and analysis is of clear importance.

Discussion

Infectious disease epidemics and pandemics in human populations continue to have devastating consequences [59, 124, 172, 211, 220]. For example, the 2014–2016 Ebola epidemic in West Africa was responsible for an estimated 11,323 deaths [246]. The influenza and SARS-CoV-2 viruses, meanwhile, have been responsible for millions of deaths each in the 21$^{\text{th}}$ century so far [127, 185]. Mathematical models are widely used for projecting possible future epidemic dynamics and to plan control [220], and compartmental modelling remains a commonly used approach [131, 203]. Although other modelling approaches, such as individual-based [242] and/or network-based [129] models, may provide a more detailed description of epidemiological dynamics, compartmental models often require fewer parameter values to be estimated or assumed than in these alternative approaches. Compartmental models also have the advantages of being straightforward to solve numerically [63, 91, 123, 131] and easy to adapt in order to include different epidemiological details [63, 112, 131]. Because of their relative simplicity, compartmental models are particularly useful for providing projections and assessing the impacts of control measures during public health emergencies, when analyses are often required very quickly [173]. This thesis has explored how compartmental epidemic models can be adapted for two distinct but related purposes: first to account for variations in symptoms (Chapter R1) and infectiousness (Chapter R2) during infection in population-scale projections, and second to estimate how infectiousness varies during an infection using data from infector-infectee transmission pairs (Chapter R3) and households (Chapter R4).

In the remainder of this overall discussion, we first summarise the research presented in this thesis, before discussing important areas of ongoing and future research, and finally ending with some general conclusions.

D.1 Summary of findings

In the first part of this thesis (Chapters R1 and R2), we focussed on showing how details of individual host-scale dynamics can be included in population-scale compartmental epidemic models. First, in Chapter R1 we showed how the existing

"method of stages" [152, 237] – where, for example, the infected compartment in the SEIR compartmental epidemic model is divided into multiple sub-compartments to obtain a gamma-distributed infectious period – can be adapted to represent three distinct phases of Ebola infection, in which (following an incubation period) infected individuals first display non-specific symptoms, before entering a gastrointestinal phase and finally a deterioration phase [20] (we also considered an extended model that accounts for the possibility of individuals recovering before entering the deterioration phase). Unlike most models of Ebola interventions (e.g. [52, 140, 163, 165]), this allowed us to account for a higher probability of infected individuals with more severe symptoms being detected through surveillance efforts compared to individuals with less severe symptoms.

To explore the potential importance of including such epidemiological detail in models of Ebola dynamics and surveillance, we compared two different models: the "constant symptoms model", in which all infected hosts are equally likely to be detected, and the "variable symptoms model", in which the detection probability increases during the course of infection. We fitted these models to case data from the two largest Ebola epidemics in history [246] (where we considered the two datasets in separate analyses) – the 2014–2016 epidemic in West Africa, and the 2018–2020 epidemic in the DRC – finding for each dataset that the two models provided an indistinguishable fit to the data. Then, to assess the impact of intensifying control interventions, we increased the surveillance frequency in the two models (fitted to either of the two datasets) from a baseline level that was assumed when fitting the models to data. We modelled surveillance as an active process, where detected individuals are assumed to be isolated and unable to transmit the pathogen, which we contend to be more realistic than alternative approaches in which surveillance does not affect epidemiological dynamics [164, 225]. As a result, the projected number of cases was lower under intensified than baseline surveillance, since more frequent surveillance led to improved epidemic control. This remained true when we extended our model to account for the possibility of individuals recovering before either entering the deterioration phase of infection or being detected (i.e. such individuals are never detected, so in the extended model, intensified surveillance will lead to an increased proportion of cases being detected; in comparison, in our simplest model, all infected individuals are assumed to eventually be detected, either through surveillance or upon death).

Even though the models provided a very similar fit to the data, projections using the two models diverged when we identically altered the surveillance frequency in the two models (fitted to either of the two datasets) to assess the impact of

intensifying control interventions. Specifically, the constant symptoms model underestimated the number of cases under intensified surveillance compared to the more epidemiologically detailed variable symptoms model (i.e. the projections using the constant symptoms model were overly optimistic). This is likely because, while more frequent surveillance was always found to reduce transmission through improved case detection and isolation, detection occurs (on average) later in infection in the more realistic variable symptoms model, by which time fewer transmissions can be prevented through isolation. Such a difference in projections between the two models remained when we extended our models to account for further epidemiological detail, and when we considered the sensitivity of our findings to important assumptions. Therefore, our main conclusion in Chapter R1 was that accounting for varying symptoms during Ebola infection may be necessary in order to ensure accurate assessments of the impacts of control interventions.

More generally, the results in Chapter R1 demonstrate the potential importance of methods that can be used to inform population-scale projections with information about individual host-scale dynamics. While our focus in Chapter R1 was variations in symptoms during infection, another important way in which most compartmental population-scale models neglect details of host-scale dynamics is through an assumption that the infectiousness of each infected host does not change during the infectious period. Therefore, in Chapter R2 we turned our attention to multi-scale epidemic models, in which within-host data and/or models (describing the viral load of individual infected hosts at different times since infection) are used to estimate the expected infectiousness profile of infected hosts (relative to the time since infection), which is, in turn, used to inform a population-scale projection model. A review concluded that while numerous studies have considered such a multi-scale framework, relatively few of those studies included substantial use of data [51]. We found that previous multi-scale modelling approaches were mathematically and/or computationally complex, often involving IDEs [61, 73, 92, 108, 155, 158, 167, 174] or individual-based stochastic models [49, 82, 154, 181, 249], which may have contributed to the lack of multi-scale modelling studies using data.

We were therefore inspired in Chapter R2 to develop an alternative, easy-to-use, multi-scale epidemic modelling framework based on compartmental modelling. As we proved, our approach, which utilises a multi-stage compartmental population-scale model with a large number of infected compartments to approximate epidemic dynamics under a specified infectiousness profile, is mathematically equivalent in the large-compartment limit to an existing IDE approach (involving the K&M model [133]). However, our compartmental framework has the advantage that, unlike

IDEs, compartmental models are straightforward to solve numerically using standard numerical routines and software packages [63, 91, 123, 131].

To demonstrate our compartmental multi-scale framework in a concrete setting, we considered the case study of influenza A in humans. We showed how the viral load profiles of infected hosts could be estimated by fitting the TCL within-host model (which has previously been used to model influenza infection [17, 36, 207]) to daily viral load measurements (over the course of infection) from each of a cohort of hosts. We used simulated data in our analyses so we could test the performance of our method in a range of scenarios, e.g. when different amounts of host-scale data are available. Individual viral load profiles were combined to estimate the expected infectiousness profile, under the common assumption of a linear relationship between viral load and infectiousness [35, 49, 108, 111, 174] (although, in principle, other relationships could just as easily be used in our framework, as we showed by carrying out sensitivity analyses in which infectiousness was assumed to either scale with the logarithm of the viral load [49, 108] or saturate at high viral loads [9, 58]). Using this infectiousness profile, our compartmental approach was then used to estimate the population-scale dynamics of an influenza outbreak.

As expected, when sufficiently many compartments were used in our compartmental framework, population-scale projections were indistinguishable from the corresponding projections obtained using the K&M IDE model. The accuracy of projections can always be improved by using a larger number of compartments in our framework (at some cost to computational speed), since this is simply a choice for the user to make, although there is likely to be a threshold beyond which the benefit of using more compartments is negligible (particularly in light of the other sources of errors in population-scale projections described below).

As noted above, since we used simulated data, we were also able to analyse how the accuracy of projections obtained using multi-scale epidemic models depends on the availability and quality of data. We explored errors in population-scale projections arising due to both variability in host-scale dynamics between individuals [17, 35, 233] (which may lead to errors in estimates of the expected infectiousness profile when data are only available from a small cohort of hosts) and measurement errors in viral load data [174, 180], first considering these two factors in isolation before analysing their combined effect. In the analysis where we included both sources of error, we found that data from a cohort of at least 40 hosts were required in order to ensure a mean error of no more than 10% in population-scale projections. We also found that the population-scale projection error depends on the exact cohort of hosts from which data are collected, as well as the frequency of observations and other factors,

such as the extent of both the variability in host-scale dynamics between individuals and the measurement error in viral load observations. Nonetheless, our results in Chapter R2 demonstrate how host-scale data can be used to inform population-scale epidemic projections. In addition, this work also shows how our multi-scale modelling approach can be used to design cohort studies by analysing what host-scale data need to be collected to ensure sufficiently accurate population-scale projections based on such data.

While we focussed on the case study of influenza in Chapter R2, in principle our multi-scale framework can be used to model outbreaks of a wide range of pathogens, provided that the host-scale dynamics are well-characterised. Alternatively, our approach could be adapted to inform population-scale projections using an infectiousness profile that is estimated using data describing symptom onset times in infector-infectee transmission pairs [85, 93] or households [40, 82] (instead of using viral load data).

In the second part of this thesis (Chapters R3 and R4), we focussed on estimating the expected infectiousness profile of SARS-CoV-2 infected individuals using data from infector-infectee pairs (Chapter R3) and households (Chapter R4), using an approach based on compartmental modelling. The (normalised) infectiousness profile relative to the time since infection is equal to the generation time distribution (the distribution of times between infection events in infector-infectee pairs), at least in settings where local susceptible depletion has a negligible effect on the timing of realised transmission [90, 191] (this is not the case in households, as discussed below). We also estimated the infectiousness profile relative to the time since symptom onset, which is equal to the TOST distribution [14, 84, 118], in both Chapters R3 and R4 (although the generation time distribution was our main focus in Chapter R4). In addition to the possibility of using the estimated generation time distribution to inform population-scale projections (while not explored in this thesis, this is an interesting area of future work – see Section D.2 below), this distribution is a crucial input to models widely used to estimate the time-dependent reproduction number from case notification data [1, 88, 101, 224], while both the generation time and TOST distributions are used to assess the effectiveness of interventions such as isolation and contact tracing [15, 71, 85, 90, 240].

Usually, only the serial interval (the difference between the symptom onset times of infectors and infectees) is observed directly in infector-infectee transmission pair data [78, 84], and as we demonstrated, the serial interval distribution may differ from that of the generation time (in household data, even the serial interval of each transmission may not be observed directly if there is uncertainty in exactly

who-infected-whom within each household). Estimation of the generation time using such datasets often therefore involves an assumption (sometimes implicit) regarding the relationship between transmission risk and symptoms. For SARS-CoV-2, a common assumption is that the transmission risk is independent of exactly when an infected individual develops symptoms (or equivalently, that the generation time of an infector-infectee pair and the infector's incubation period are independent) [74, 85, 93, 137, 145], but this assumption is not epidemiologically justifiable [16, 145].

We developed an alternative inference approach which includes a mechanistic description of the relationship between transmission and symptoms. This method is based on compartmental modelling, with each infected individual assumed to progress through three stages of infection (with gamma-distributed durations) – latent (E), presymptomatic infectious (P) and symptomatic infectious (I). We note, however, that this approach does not involve direct simulation of a "SEPIR" compartmental model. Instead, we derived expressions for the likelihood of event times in infector-infectee pairs (Chapter R3) and households (Chapter R4) being observed under the modelling assumptions. These likelihood functions (as well as the corresponding likelihood expressions for other models of infectiousness that we considered) depend on event times which were not observed directly (e.g. exact symptom onset times within recorded symptom onset dates, as well as times of infection, although we integrated over possible infection times in Chapter R3). Therefore, we used data augmentation MCMC [31, 40, 41, 82, 204, 219] to fit the models considered in Chapters R3 and R4 to data, augmenting the observed data with unobserved event times so that these likelihood functions could be used. In Chapter R3 we also considered maximum likelihood estimates under an assumption that each infected host developed symptoms exactly in the middle of their onset date, which were useful since they enabled straightforward comparison of the relative goodness of fit of different models to the data using the AIC [4].

In Chapter R3 we fitted four different models to data from 191 SARS-CoV-2 infector-infectee pairs [50, 84, 85, 118, 248, 250]: (i) our mechanistic approach, estimating the relative transmission rates in the P and I stages of infection (the "variable infectiousness model"); (ii) our mechanistic approach, assuming equal transmission rates in the P and I stages (the "constant infectiousness model"); (iii) a previously used statistical approach [84] in which the presymptomatic portion of an individual's infectiousness profile is scaled (horizontally) depending on the duration of their incubation period (the "Ferretti model"); and (iv) the standard approach in which transmission risk is assumed independent of symptom status (the "independent transmission and symptoms model"). We found that our variable

infectiousness model provided the best fit to the data, and the commonly used independent transmission and symptoms model provided the worst fit (with the constant infectiousness and Ferretti models giving the second and third best fit, respectively).

Using the fitted models, we derived estimates of important epidemiological quantities and time intervals, and compared these between models. Compared to the independent transmission and symptoms model, the best-fitting variable infectiousness model suggested a slightly more dispersed generation time distribution (i.e. a higher transmission risk at large times since infection). However, the estimated proportion of presymptomatic transmissions (among infectors who develop symptoms) was highest when using the variable infectiousness model and lowest when using the independent transmission and symptoms model. Furthermore, the TOST distribution for the variable infectiousness model was highly concentrated around the time of symptom onset, with a decrease in the transmission risk immediately following symptom onset that can be attributed to behavioural responses to symptoms, indicating that many transmissions occur shortly before symptom onset.

We then analysed the impact of the results described in the previous paragraph on assessments of the effectiveness of isolation and contact tracing. Due to the high proportion of presymptomatic transmissions, the best-fitting variable infectiousness model indicates that relatively few transmissions can be prevented by isolating symptomatic hosts alone (even if isolation occurs shortly after symptom onset). On the other hand, since many transmissions occur shortly before symptom onset under the variable infectiousness model, that model indicates contact identification using a fixed contact elicitation window (e.g. tracing contacts of a symptomatic index case that occurred up to two days before the index case developed symptoms, as in the UK [228] and USA [44] earlier in the COVID-19 pandemic) may be more effective than suggested by the independent transmission and symptoms model.

While more infected contacts of each index case can always be identified using a longer contact elicitation window, this would increase strain on contact tracing programmes and (depending on the exact isolation policy) may lead to increased numbers of uninfected contacts being instructed to isolate unnecessarily, particularly since the index case is likely to have been less infectious at longer times before symptom onset. In addition, contacts who were infected at longer times before the index host developed symptoms may already have been isolated by the time contact tracing takes place (e.g. if they developed symptoms themselves), or otherwise a long delay between a contact's infection and isolation will reduce the impact of isolating such contacts. The exact effect of contact tracing on transmission will also

depend on the speed of contact identification and isolation (a detailed analysis was beyond the scope of this work). Nonetheless, our main conclusion in Chapter R3 was clear: although the isolation of symptomatic hosts alone may have a limited effect on transmission, combining this with contact tracing and isolation of presymptomatic infected contacts is beneficial, even if a short contact elicitation window is used (since infected individuals were found to generate a large number of transmissions just before symptom onset).

The data underlying our analyses in Chapter R3 were collected during the early months of the COVID-19 pandemic (December 2019–March 2020) [50, 85, 118, 248, 250]. Similarly, most other SARS-CoV-2 generation time estimates (at the time of conducting the analyses described in Chapter R3) also used data from those months, primarily collected in countries in East and Southeast Asia [84, 85, 93, 122, 209]. However, differences in non-pharmaceutical interventions and in individuals' behaviour, as well as evolutionary changes to the SARS-CoV-2 virus (see the discussion about estimating the generation time for novel variants in Section D.2 below), may be expected to lead to differences in the generation time between time periods and locations – for example, even over the course of January–February 2020, one study [209] found a temporal decrease in the generation time in China, attributing this decrease to the intensification of interventions such as case isolation.

The relative lack of generation time estimates from beyond the very early months of the pandemic motivated us to provide updated estimates of the generation time in Chapter R4 when more data became available. Specifically, in Chapter R4 we estimated the generation time using data collected from 172 UK households by PHE (now the UKHSA) between March and November 2020 [168]. We considered two of the models of infectiousness introduced in Chapter R3: the commonly used independent transmission and symptoms model, and the version of our mechanistic approach in which the relative transmission rates during the P and I stages of infection are estimated (i.e. the variable infectiousness model which provided the best fit to the data in Chapter R3; this model is referred to as the "mechanistic model" in Chapter R4).

Inference of the generation time using data from households, rather than known transmission pairs, is more challenging due to uncertainty in exactly who-infected-whom within households. In addition, realised generation times in households are likely to be shorter than the expected infectiousness profile would suggest [90, 191]. This is because household contacts occur frequently, often resulting in household transmissions occurring. Since the number of susceptible household members may decrease during the course of each infection (if either the infected

host under consideration, or another household member, transmits the virus within the household over that time), transmission is relatively less likely to occur later in infection than predicted by the expected infectiousness profile. Where this effect arises due to multiple infected individuals "competing" to infect the available susceptibles, it is sometimes called a "race to infect" [132, 148].

Therefore, similarly to some household influenza studies [40, 82], we adapted the likelihood functions used in Chapter R3 in order to estimate the SARS-CoV-2 generation time using household data. Specifically, in Chapter R4 we summed over likelihood contributions corresponding to infection by different possible infectors. We also corrected for the regularity of household contacts and susceptible depletion by including a factor corresponding to the probability of individuals avoiding being infected before their actual time of infection (or avoiding ever being infected, for uninfected individuals). The exclusion of such a factor in Chapter R3, and in other studies estimating the generation time using data from infector-infectee pairs [84, 85, 93], amounts to an implicit assumption that contacts between infector-infectee pairs are sporadic and of short duration (so that individuals are unlikely to have been infected before their actual time of infection), but this assumption is unlikely to be true in households. These likelihood adaptations enabled us to derive estimates of the (normalised) expected infectiousness profile (with respect to the time since infection) from the household data. The infectiousness profile is widely applicable to transmission both within and outside households, and gives the generation time distribution under the assumption of a constant supply of susceptibles being available to infect throughout the course of each infection [191]. However, we also estimated the realised generation time distribution in the study households, finding (as expected) that these estimates are shorter than our main generation time estimates described below, which are based on the infectiousness profile.

Under the standard assumption of independent transmission and symptoms, we obtained an overall mean generation time estimate (based on the infectiousness profile) that was comparable to estimates obtained under that assumption in Chapter R3 and other studies [84, 85, 93], but with a higher standard deviation than estimated in those other studies. One possible explanation for the higher standard deviation is that self-isolating individuals may still be at risk of transmitting the virus in their household following detection, leading to a higher relative transmission risk late in infection than would occur under a policy of out-of-home isolation (which was commonplace in the East and Southeast Asian countries where most of the data underlying previous SARS-CoV-2 generation time estimates were collected). The

mechanistic model indicated a higher mean generation time and similar standard deviation compared to the independent transmission and symptoms model.

Although direct comparisons of goodness of fit were not readily available in Chapter R4, the two models appeared to provide a similar fit to the data. Therefore, several factors should be taken into account when deciding which of our estimates to use in subsequent analyses, including the relative epidemiological plausibility of the two models (we contend that the mechanistic model should be preferred in this regard since this approach relaxes the unrealistic assumption of independent transmission and symptoms), the fact that the mechanistic model did provide an improved fit to data in Chapter R3, and exactly what the estimates are to be used for (e.g. the estimates assuming independent transmission and symptoms may be more suitable for models used to estimate the time-dependent reproduction number in which this same assumption is made [1]).

As well as obtaining overall generation time estimates using data from the duration of the household study, we also explored whether or not there was a detectable change in the generation time distribution during the study period. To do so, we fitted the models separately to data from March–April, May–August and September–November 2020. This suggested a shorter generation time in autumn 2020 compared to earlier months. One explanation for a temporal decrease in the generation time is that a higher level of indoor mixing in colder autumn months, particularly among infected individuals who had not yet developed symptoms, may have increased the relative probability of transmission earlier in infection. However, other explanations, such as evolutionary changes to the SARS-CoV-2 virus, are also possible (see the discussion on estimating the generation time for novel variants in Section D.2 below).

Regardless of its exact cause, one reason why our finding of a temporal decrease in the generation time is important is because, if a change in the generation time occurs but is not accounted for, then estimates of the time-dependent reproduction number are likely to be incorrect. For example, if the generation time is shorter than assumed, then in a growing epidemic the reproduction number is likely to be overestimated, which could potentially lead to unnecessarily stringent interventions being introduced. A shorter generation time may also mean that interventions such as isolation and contact tracing are less effective than predicted. The finding that the generation time became shorter is particularly significant because most previous SARS-CoV-2 generation time estimates have used data collected at the very beginning of the COVID-19 pandemic [84, 85, 93, 122, 209]. Our main conclusion in Chapter R4 was therefore that the SARS-CoV-2 generation time appears to be changing, highlighting the need for continued data collection and analysis.

D.2 Ongoing and future research

In each research chapter of this thesis we have outlined numerous areas in which future research may build upon our methods and findings. Here, we discuss several extensions that connect, or build upon, multiple research chapters, as well as describing one piece of research that we have already conducted that updates the results in Chapter R4.

Our conclusion in Chapter R1 that it may be important to include variations in symptoms during infection in models of Ebola interventions motivated us to develop a compartmental framework for using host-scale data to inform population-scale projections in Chapter R2. However, the focus in Chapter R2 was on variations in infectiousness rather than symptoms during infection. Therefore, it would be useful to confirm whether or not the conclusions of Chapter R1 apply more generally, for example by analysing how population-scale projections obtained using our multi-scale framework from Chapter R2 may differ from corresponding projections obtained using a simple model in which variations in infectiousness are neglected. While including additional epidemiological detail can improve the accuracy of model projections, this is not always the case [215]. Furthermore, simple models often offer greater tractability than more complex models, allowing more straightforward parameterisation and/or understanding. Therefore, we hope to explore in which scenarios it is necessary to use a model that accounts for detailed host-scale dynamics instead of a simpler model without that complexity.

We expect that accounting for variations in infectiousness during infection may be particularly important when modelling the population-scale impacts of interventions specifically targeting infected individuals (as opposed to interventions affecting all individuals within a population) such as case isolation or the therapeutic use of antiviral drugs. This is because such interventions often involve delays between individuals being infected and detected, and therefore, for example, estimates of the overall reduction in transmission from an infected host due to isolation (for a given delay) will depend on how infectiousness is assumed to change during infection, as we demonstrated in Chapter R3. In turn, the effectiveness of isolation is likely to affect population-scale projections – see Figure R1.7 for a simple example in which projections under intensified surveillance (leading to more infected hosts being isolated) differed under two assumptions regarding how infectiousness changes during an Ebola infection.

Because interventions such as isolation are often targeted towards symptomatic infected hosts (i.e. the delay between infection and detection often includes the incubation period), modelling such interventions in our multi-scale framework would likely require an understanding of how infectiousness and symptoms are linked.

Furthermore, in Chapter R1 we demonstrated the potential importance of accounting for variable symptoms (specifically the fact that infected hosts later in infection, with more severe symptoms, are more likely to be detected and therefore isolated, compared to hosts with milder initial symptoms) in models of Ebola interventions. Therefore, an interesting extension of both Chapters R1 and R2 would be to incorporate variations in symptoms (as well as infectiousness) during infection into our multi-scale framework. One possibility is to include the level of symptom expression in a within-host model (fitted to data including both viral load measurements and symptom scores; see, for example, the model in [35]), in order to simultaneously estimate how infectiousness and symptoms vary during an infection. Such an approach would also have the advantage of enabling the effect of behavioural responses to symptoms on transmission risk to be accounted for. For example, in Chapter R3 we estimated a substantial reduction in the SARS-CoV-2 transmission risk immediately following symptom onset, but in principle, a more complex approach could be used, linking the extent of behavioural responses at different stages of infection to the severity of symptoms.

For SARS-CoV-2, an alternative approach to include variations in infectiousness during infection in population-scale projections would be to, instead, use the infectiousness profiles that we derived from transmission pair data in Chapter R3 and from household data in Chapter R4. Specifically, the infectiousness profile estimates obtained using our mechanistic approach in Chapter R3 and/or Chapter R4 may be useful for modelling the population-scale effect of interventions such as isolation, since transmission is naturally linked to symptoms in that approach. The compartmental framework that we developed in Chapter R2 can be used to project population-scale dynamics using any specified infectiousness profile (with respect to the time since infection). However, our mechanistic approach in Chapters R3 and R4 has the advantage that our parameter estimates can be used directly in an "SEPIR" compartmental model with gamma-distributed stage durations (where the expected infectiousness profile underlying such a compartmental model will be given by our corresponding estimate). This may therefore provide a more straightforward approach to estimating population-scale dynamics.

While the approach described in the previous paragraph may be useful for modelling the population-scale impacts of non-pharmaceutical interventions, such as case isolation, there are some scenarios in which an approach that explicitly links host- and population-scale SARS-CoV-2 dynamics (e.g. our approach in Chapter R2) may be useful. For example, within-host models have been used to assess the impact of antiviral drugs on SARS-CoV-2 viral dynamics [66, 100, 103, 134, 177, 184, 236],

and (at the time of writing) effective antivirals have recently become available (at least with respect to prevention of severe disease) [159]. It would be interesting to use a multi-scale framework to explore the potential population-scale impact of widespread antiviral deployment. We note that our estimates of a high proportion of presymptomatic transmissions in Chapters R3 and R4 suggest a potentially limited effect on transmission of antiviral drugs administered even shortly after symptom onset, although it is possible that the relaxation of self-isolation rules [229] will affect the balance between presymptomatic and symptomatic transmission. Our estimates from Chapter R3 and/or Chapter R4 could be used alongside viral load data in a study of multi-scale SARS-CoV-2 dynamics, to infer a potential relationship between viral load and infectiousness, and to inform the relationship between transmission and symptoms, similarly to an approach used in [86].

Our main qualitative finding in Chapter R4 – that the SARS-CoV-2 generation time appears to be changing – highlights the need for generation time estimates to continue to be updated as more data become available. One particularly important question is whether or not the generation time is different for novel SARS-CoV-2 variants than for the original form of the virus that was responsible for infections in humans in early 2020. Specifically, the Alpha variant became the dominant variant in the UK in December 2020, subsequently succeeded, in turn, by the Delta and Omicron variants [200, 231, 234]. Each of these three variants has been associated with faster epidemic growth compared to the previously dominant variant [2, 34, 68, 126]. A higher epidemic growth rate could be attributed to either increased transmissibility or a shorter generation time, or both [189], and therefore estimating the generation time for these variants is important to understand ongoing transmission (as well as the effectiveness of interventions). Since the Alpha variant was responsible for infections in only two households in the household study underlying our results in Chapter R4, while the Delta and Omicron variants had not yet emerged when that household study concluded (November 2020), we left estimation of the generation time for those variants to future work.

After conducting the research described in Chapter R4, we were provided with data from a more recent household study by the UKHSA. Because data were available from February–August 2021, and genomic sequencing was conducted as part of the household study, we were able to estimate the generation time for both the Alpha and Delta SARS-CoV-2 variants (the Delta variant replaced the Alpha variant as the dominant variant in the UK in May 2021 [234]) in a study that is now published in *The Lancet Infectious Diseases* [114]. Using our mechanistic approach for estimating the generation time (and the household inference framework

we developed in Chapter R4), we first obtained generation time estimates for each variant based on the expected infectiousness profile (i.e. the generation time assuming a constant supply of susceptibles to be available throughout infection [191], which we called the "intrinsic generation time" in [114]). As in Chapter R4, we also estimated the realised generation time distribution within study households for each variant (denoted the "household generation time" in [114]; as discussed in Chapter R4 and Section D.1 above, the household generation time is shorter (on average) than the intrinsic generation time due to susceptible depletion in households [90, 191]). In addition, we compared household generation times on the basis of vaccination status (note that COVID-19 vaccines were not yet widely available during the period of the household study from which we obtained data in Chapter R4), age and date.

We estimated a 15% (95% CrI (−7)–31%) shorter mean intrinsic generation time for the Delta variant (4.7 days, 95% CrI 4.1–5.6 days) than the Alpha variant (5.5 days, 95% CrI 4.7–6.5 days), with a 92% posterior probability that this quantity is shorter for the Delta variant. In contrast, the mean household generation time was 28% (95% CrI 0–48%) shorter for the Delta variant (3.2 days, 95% CrI 2.5–4.2 days) than the Alpha variant (4.5 days, 95% CrI 3.7–5.4 days), with a 97.5% posterior probability of a shorter mean household generation time for the Delta variant. In other words, we found a bigger difference between the two variants in the household generation time than would be expected from the difference in the intrinsic generation time alone, which we attributed to higher transmissibility of the Delta variant leading to faster depletion of susceptibles in households. The variant responsible for infections was found to have a bigger impact on the household generation time than either vaccination status or age.

The finding that the Delta variant transmits more quickly in households than the Alpha variant is significant for estimates of the time-dependent reproduction number, suggesting that the difference in transmissibility between the two variants may be smaller than would be estimated assuming no difference in the generation time [189]. A shorter household generation time may also suggest that interventions such as self-isolation are less effective at preventing household transmission of the Delta variant. Estimation of the generation time distribution of the Omicron variant, which emerged after we conducted the analyses in [114], and (at the time of writing) is currently dominant in the UK, remains an important topic for future work.

D.3 Concluding remarks

In this thesis, we have shown how compartmental epidemic models can be used or adapted both to inform projections of population-scale dynamics with details of individual host-scale dynamics, and to estimate how infectiousness varies during an infection. Specifically, we first demonstrated how variations in symptoms (Chapter R1) and infectiousness (Chapter R2) during infection can be included in compartmental population-scale models, before showing how an approach based on compartmental modelling could be used to estimate the infectiousness profile of SARS-CoV-2 infections using data collected from infector-infectee pairs (Chapter R3) and households (Chapter R4). Additionally, we demonstrated how assessments of the possible impacts of control interventions may differ when using the methods we have developed, compared to widely used approaches. We found in Chapter R1 that relaxing the standard assumption of constant symptoms during Ebola infection led to more pessimistic projections of the effect of intensifying surveillance. In Chapter R3, our mechanistic approach for inference of the SARS-CoV-2 infectiousness profile indicated that isolation of symptomatic hosts may prevent fewer transmissions, whereas contact tracing using a specified contact elicitation window may enable more presymptomatic infectious contacts to be identified, than suggested by a commonly used approach. Overall, we have both developed widely applicable methods based on compartmental epidemic modelling (such as our framework for informing population-scale projections using host-scale data in Chapter R2), and applied these methods to three specific pathogens – Ebola, influenza and SARS-CoV-2 – to derive concrete epidemiological conclusions (such as our finding of a temporal decrease in the SARS-CoV-2 generation time in Chapter R4).

Infectious disease epidemics in human populations will remain a significant threat going forwards, and it is crucial to develop flexible epidemiological modelling frameworks that can be applied quickly during public health emergencies. Compartmental modelling, including the techniques we have described and developed here, is likely to remain important for understanding epidemic dynamics at different scales and for guiding control interventions.

Appendix A1

Appendix to Chapter R1

A1.1 Model extensions

A1.1.1 Model with variable infectiousness during infection

In most of our analyses, to isolate the effect of variable symptoms alone on epidemic projections, we assumed that the transmission rate was equal for all infectious hosts (at any given time in the outbreak). However, in reality the infectiousness of a host is likely to vary throughout an Ebola infection. We therefore conducted an additional analysis (Figures R1.7 and A1.3), in which infectiousness was assumed to vary during an infection. Here we describe the model used to perform that analysis, and its parameterisation.

Extending the SEI_3RC model so that individuals in compartment I_i have transmission rate $\beta^{(i)}(t)$ (for $i = 1$, 2 or 3) leads to the model

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -S\left(\beta^{(1)}(t)I_1 + \beta^{(2)}(t)I_2 + \beta^{(3)}(t)I_3\right), \tag{A1.1}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = S\left(\beta^{(1)}(t)I_1 + \beta^{(2)}(t)I_2 + \beta^{(3)}(t)I_3\right) - \gamma E, \tag{A1.2}$$

$$\frac{\mathrm{d}I_1}{\mathrm{d}t} = \gamma E - \mu I_1 - \delta_1 I_1, \tag{A1.3}$$

$$\frac{\mathrm{d}I_2}{\mathrm{d}t} = \mu I_1 - \ \mu I_2 - \delta_2 I_2, \tag{A1.4}$$

$$\frac{\mathrm{d}I_3}{\mathrm{d}t} = \mu I_2 - \ \mu I_3 - \delta_3 I_3, \tag{A1.5}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \mu I_3, \tag{A1.6}$$

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \delta_1 I_1 + \delta_2 I_2 + \delta_3 I_3. \tag{A1.7}$$

We considered three particular forms of this model:

i. The model used in our main analyses in which the transmission rate is the

same for individuals in each stage of infection, i.e.

$$\beta^{(1)}(t) = \beta^{(2)}(t) = \beta^{(3)}(t) = \beta(t). \tag{A1.8}$$

ii. An increasing infectiousness model in which

$$\beta^{(1)}(t) = 0.5\beta(t), \quad \beta^{(2)}(t) = \beta(t), \quad \beta^{(3)}(t) = 1.5\beta(t). \tag{A1.9}$$

iii. A decreasing infectiousness model in which

$$\beta^{(1)}(t) = 1.5\beta(t), \quad \beta^{(2)}(t) = \beta(t), \quad \beta^{(3)}(t) = 0.5\beta(t). \tag{A1.10}$$

In each case, we still assumed that the transmission rate of all infectious individuals is reduced at some time T during the epidemic (e.g. due to changes in behavioural responses or alterations to interventions), so that

$$\beta(t) = \begin{cases} \beta_0, & \text{for } t \leq T, \\ \beta_1, & \text{for } t > T. \end{cases} \tag{A1.11}$$

The values of the parameters β_0, β_1 and T, in addition to the epidemic start date T_0, were fitted to the epidemic data under consideration, either under the assumption of constant symptoms ($\delta_1 = \delta_2 = \delta_3 = \delta$) or variable symptoms ($\delta_1 < \delta_2 < \delta_3$). All other parameter values are given in Table A1.1.

A1.1.2 Model in which some hosts recover prior to the deterioration phase of an Ebola infection

Here, we describe an extension of the SEI_3RC model used in most of our analyses, in which some individuals recover instead of progressing to the deterioration phase. In reality, an individual progressing to the deterioration phase of an Ebola infection is very likely to die from the infection. In the absence of surveillance, we assumed in the extended model that a proportion, q, of infected hosts recover during the gastrointestinal phase of infection (represented by the I_2 compartment), while the remainder continue to the deterioration phase (the I_3 compartment) and subsequently die. This leads to the extended model,

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta(t)S\left(I_1 + I_2 + I_3\right), \tag{A1.12}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \beta(t)S\left(I_1 + I_2 + I_3\right) - \gamma E, \tag{A1.13}$$

$$\frac{\mathrm{d}I_1}{\mathrm{d}t} = \gamma E - \mu I_1 - \delta_1 I_1, \tag{A1.14}$$

$$\frac{\mathrm{d}I_2}{\mathrm{d}t} = \mu I_1 - \mu I_2 - \delta_2 I_2, \tag{A1.15}$$

$$\frac{\mathrm{d}I_3}{\mathrm{d}t} = (1-q)\,\mu I_2 - \mu I_3 - \delta_3 I_3, \tag{A1.16}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = q\mu I_2, \tag{A1.17}$$

$$\frac{\mathrm{d}D}{\mathrm{d}t} = \mu I_3, \tag{A1.18}$$

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \delta_1 I_1 + \delta_2 I_2 + \delta_3 I_3. \tag{A1.19}$$

In this new $\mathrm{SEI_3RDC}$ model, hosts that have recovered without being detected (R) are explicitly distinguished from those that are detected at death (D).

As in our main analyses, we estimated the parameters β_0, β_1, T and T_0 (defined analogously to the corresponding parameters in the baseline model) by fitting the model to the Beni and Liberia datasets, under the assumptions of constant symptoms and variable symptoms (in turn). When fitting the model, we assumed that the number of observed cases corresponds to the sum of C and D in the model. The values of other model parameters were as in Table A1.1.

A1.2 Fitting models to non-cumulative data

Here, we describe how we fitted the constant symptoms and variable symptoms models to non-cumulative incidence data in the analysis shown in Figure R1.11 (in all of our other analyses, the models were fitted to cumulative data as described in Section R1.2.4). In this analysis, we inferred the values of the parameters β_0, β_1, T and T_0 in each model by matching the number of new cases in each interval in the model (i.e. the difference in $C + R$ between the start and end of each approximately weekly reporting interval) to those observed in reality. Specifically, we calculated the parameter values which minimised the sum of squares residual,

$$SS_{res} = \sum_{i=(1-m)}^{n} \left[C\left(t_i\right) + R\left(t_i\right) - \left(C\left(t_{i-1}\right) + R\left(t_{i-1}\right)\right) - \left(y_i - y_{i-1}\right)\right]^2, \tag{A1.20}$$

Here $C(t_i) + R(t_i)$ is the cumulative number of detected or removed hosts in the model under consideration and y_i is the cumulative number of cases in the data at each reporting time, t_i. We accounted for cases predicted by the model to have occurred in weekly time intervals between times t_{-m} and t_0, where t_{-m} is the start of the weekly interval containing the model-predicted epidemic start date (T_0), and t_0 is the start of the weekly interval that ended with the first date on which cases were actually reported, t_1 (where t_1 was 3$^{\mathrm{rd}}$ August 2018 for the outbreak in the DRC and

23rd March 2014 for the epidemic in Liberia). All cases in the data recorded at time t_1 were assumed to have been detected during the preceding week (i.e. between times t_0 and t_1). Therefore, y_i was taken to be equal to zero for $i \leq 0$. The total number of dates on which cases were reported (from time t_1 onward) is denoted by n.

All parameter values other than the values of β_0, β_1, T and T_0 were assumed to be known exactly, and were fixed at the values given in Table A1.1.

A1.3 Supplementary Figures

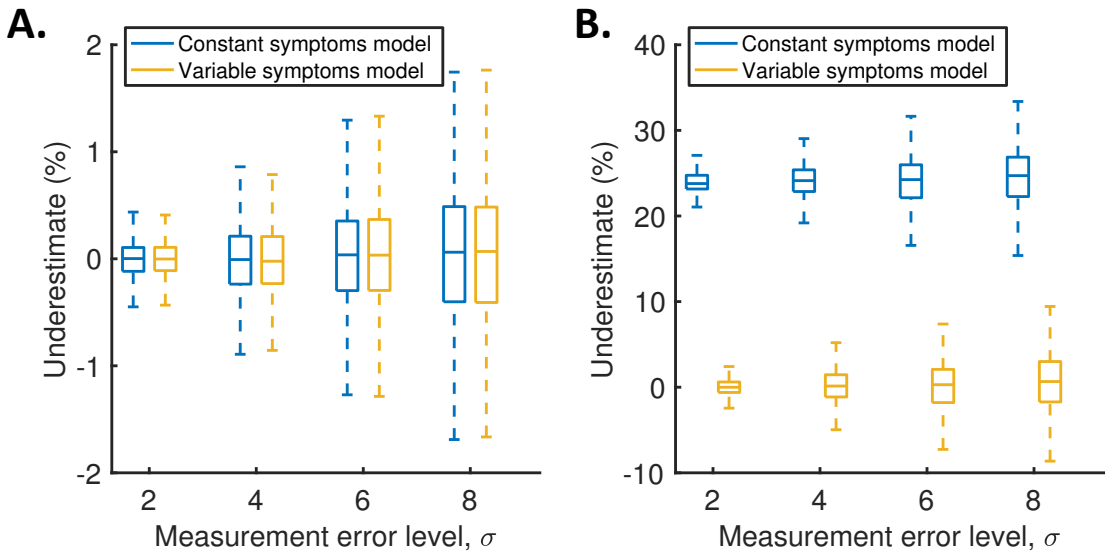


Figure A1.1: Effect of the extent of measurement error in the simulated data. A. Box-and-whisker plots indicating the distribution, over 1,000 simulated datasets for each of a range of values of the measurement error level, σ, of underestimates in the total cumulative number of observed cases projected when the constant symptoms model (blue) and variable symptoms model (yellow) were fitted to the simulated data (separately for each simulated dataset), compared to the number of cases projected by the variable symptoms model with the "true" parameter values used to generate the simulated data. The boxes indicate the lower quartile, median and upper quartile, and the maximum length of each whisker is 1.5 times the interquartile range. B. The corresponding errors once the level of surveillance is intensified.

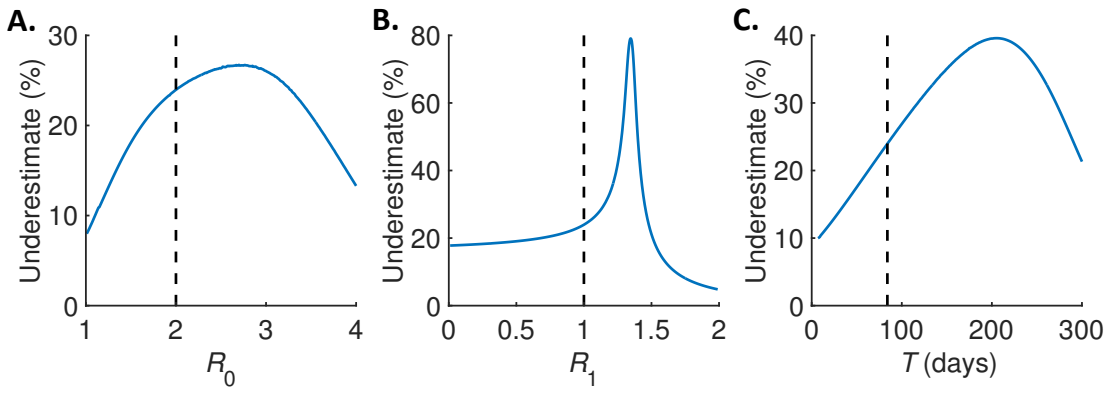


Figure A1.2: Robustness of our results to different values of parameters used to generate the synthetic data. Varying the parameters $R_0 = 3\beta_0 N/\mu$ (A), $R_1 = 3\beta_1 N/\mu$ (B) and T (C) one-by-one, for each set of "true" parameters we generated synthetic data using the variable symptoms model (for simplicity taking the measurement error to be zero in this analysis) and fitted the constant symptoms model to the data. We then calculated the percentage by which the fitted constant symptoms model underestimated the total cumulative number of cases under intensified surveillance (compared to the corresponding projection using the variable symptoms model). In each panel, the dashed vertical line indicates the parameter value used in our other analyses using simulated data.

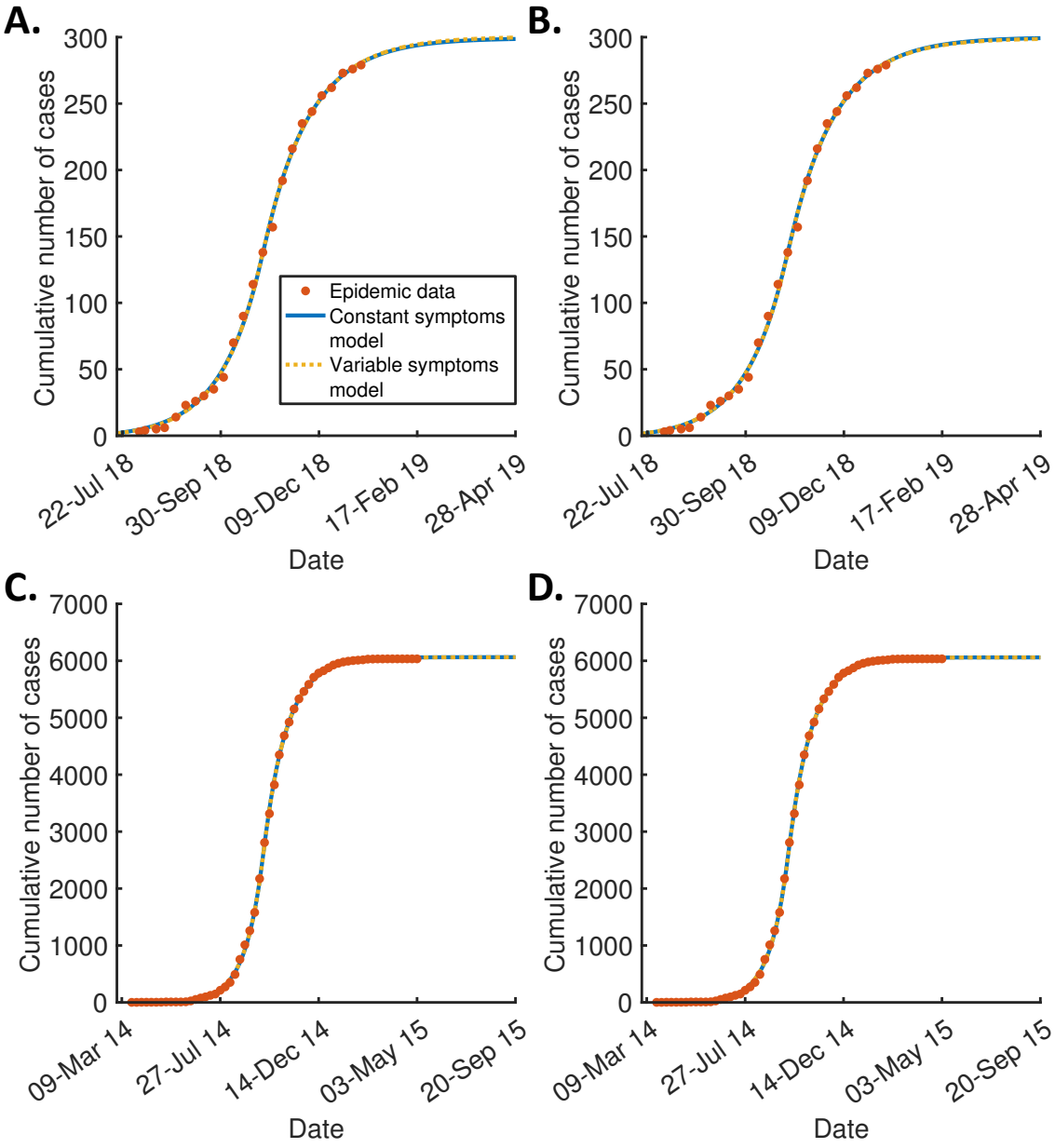


Figure A1.3: Model fits for the model with variable infectiousness. A–B. Model fits to the data from the 2018–2020 epidemic in the Beni and Kalunguta health zones in the DRC (red dots) for the constant symptoms (blue) and variable symptoms (yellow dotted) models, when infectiousness was assumed to either increase (A) or decrease (B) as symptoms progress. C–D. Equivalent panels showing model fits to the data from the 2014–2016 Ebola epidemic in Liberia in the two scenarios.

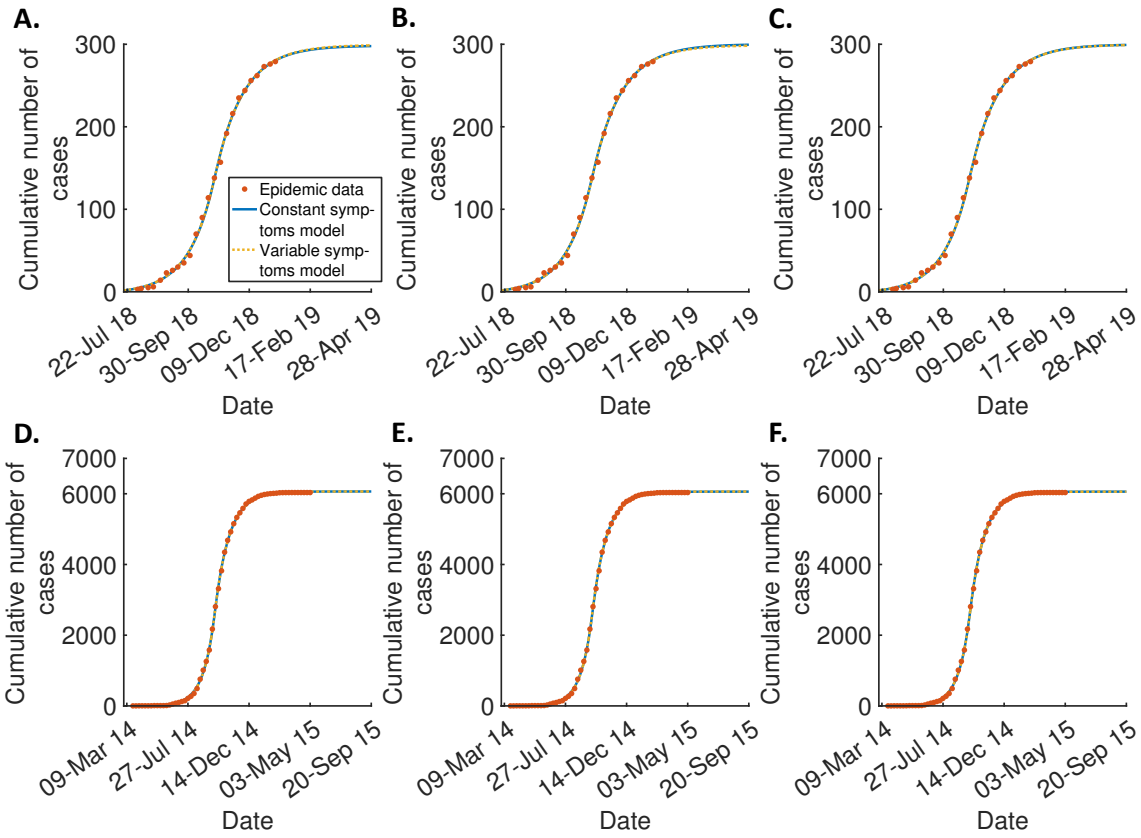


Figure A1.4: Model fits for different surveillance frequencies. A–C. Model fits to the data from the 2018–2020 epidemic in the Beni and Kalunguta health zones in the DRC (red dots) for the constant symptoms (blue) and variable symptoms (yellow dotted) models, when the surveillance frequency was assumed to be $\Delta = 7$ days (A), $\Delta = 14$ days (B) and $\Delta = 28$ days (C). D–F. Equivalent panels showing model fits to the data from the 2014–2016 Ebola epidemic in Liberia for each value of Δ.

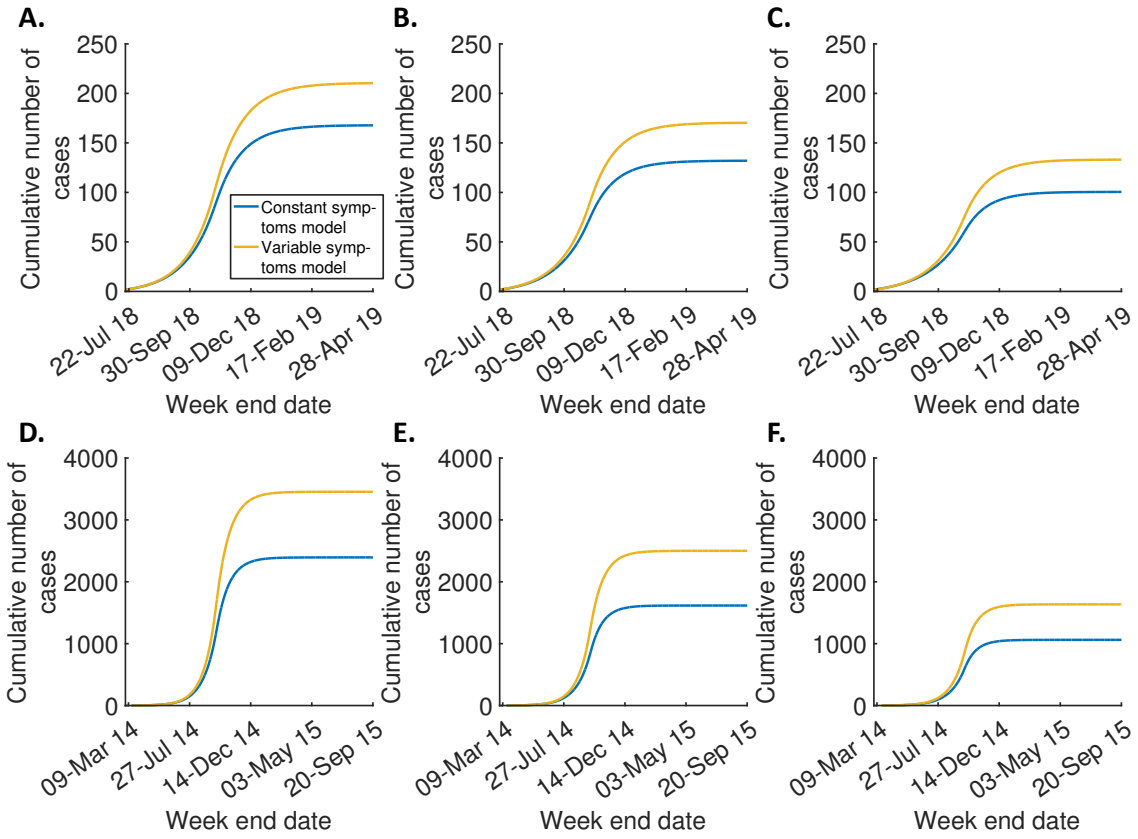


Figure A1.5: Effect of the mean detection probability. We varied the detection probability, here denoted p, in the constant symptoms model, considering three possibilities of low ($p = 0.367$), medium ($p = 0.5$) and high ($p = 0.663$) mean detection probability. In the variable symptoms model, the values of p_1 and p_3 were held fixed at 0.1 and 0.9, respectively, with p_2 varied so that the mean value of p_1, p_2 and p_3 was equal to p. This corresponded to values of $p_2 = 0.1$, 0.5 and 0.9 in the low, medium and high mean detection probability scenarios, respectively. A–C. The projected number of cases under intensified surveillance under the low (A), medium (B) and high (C) detection probability scenarios, using the constant symptoms model (blue) and variable symptoms model (yellow) fitted to data from the 2018–2020 epidemic in the DRC (model fits to data are not shown). D–F. Equivalent figures to A–C, using the data from the 2014–2016 Ebola epidemic in Liberia.

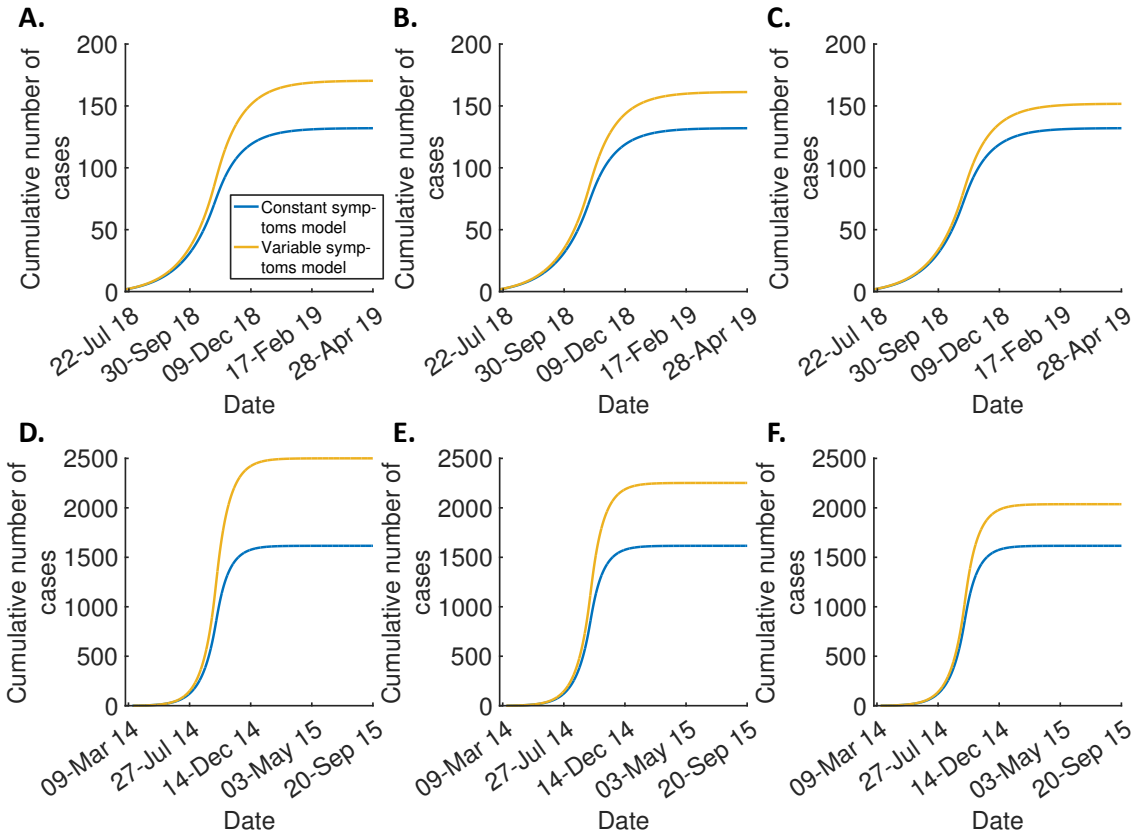


Figure A1.6: Effect of the extent of variation in symptom expression during infection in the variable symptoms model. We fixed the detection probability in the constant symptoms model to be $p_1 = p_2 = p_3 = 0.5$, and took the value $p_2 = 0.5$ in the variable symptoms model. In the variable symptoms model, we then took $p_1 = \alpha$ and $p_3 = (1 - \alpha)$, and considered different values of the parameter α. A–C. The projected number of cases under intensified surveillance for $\alpha = 0.1$ (A), $\alpha = 0.2$ (B) and $\alpha = 0.3$ (C), using the constant symptoms model (blue) and variable symptoms model (yellow) fitted to data from the 2018–2020 epidemic in the DRC (model fits to data are not shown). D–F. Equivalent figures to A–C, using the data from the 2014–2016 Ebola epidemic in Liberia.

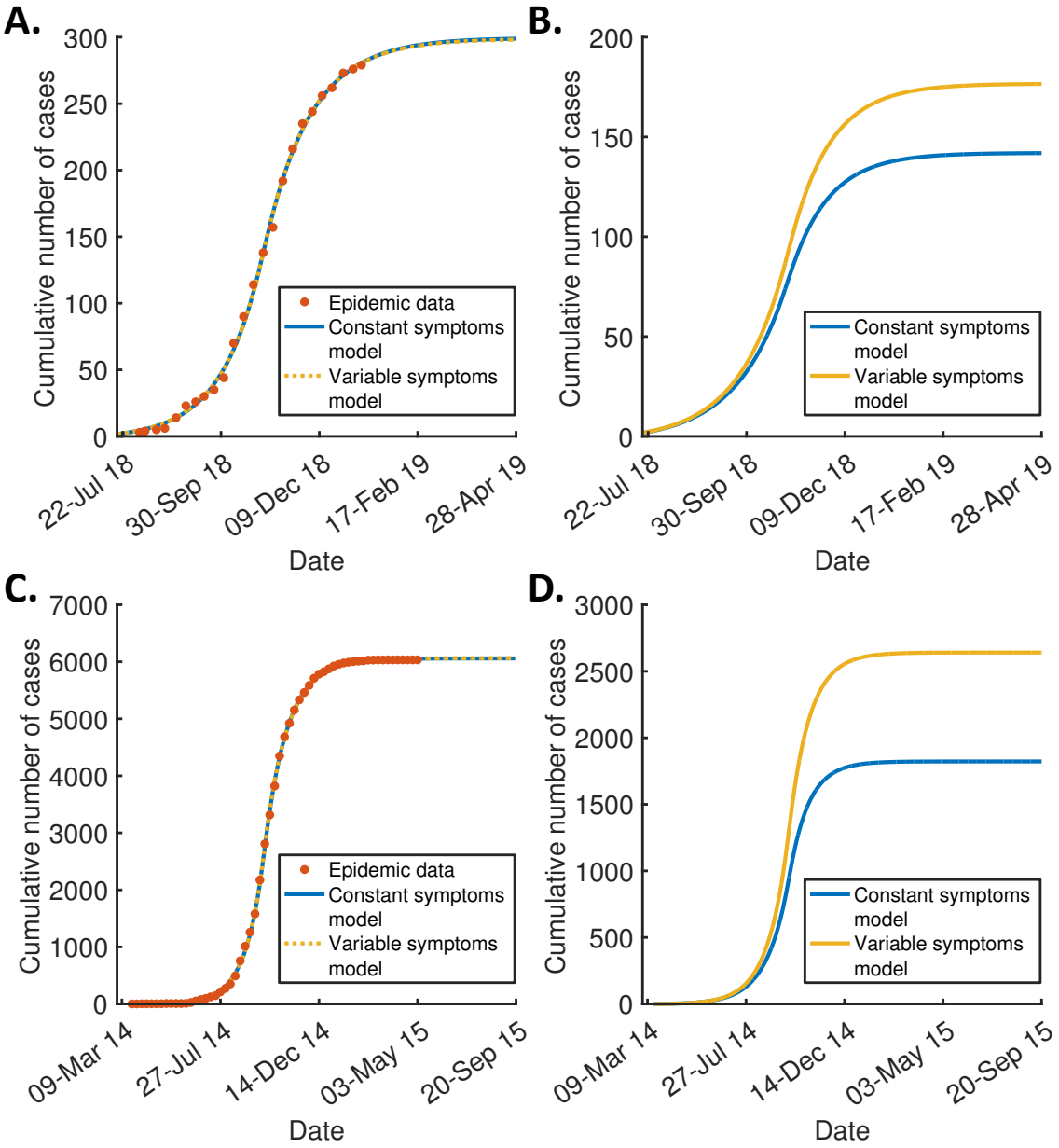


Figure A1.7: Alternative dependence of the detection rate on surveillance. A. Model fits to the data from the 2018–2020 epidemic in the DRC (red dots), using the constant symptoms model (blue) and variable symptoms model (yellow dotted), assuming an exponentially distributed interval between surveillance rounds (see Section R1.2.3). B. Projections of the effect of intensified surveillance, using the constant symptoms model (blue) and variable symptoms model (yellow). C–D. Equivalent figures to A–B, using the data from the 2014–2016 Ebola epidemic in Liberia.

A1.4 Supplementary Tables

Parameter	Interpretation	Value	Justification
N	Population size	230,000 (DRC)	Estimate for city of Beni from [247]
		4,500,000 (Liberia)	2015 estimate from [212]
		200,000 (simulation study)	Assumed
$1/\gamma$	Mean latent/incubation period	7 days	Taken from [163]
$3/\mu$	Mean infectious period	9.8 days	Taken from [163]
p_1	Detection probability in early infection	0.6 (constant symptoms model) 0.1 (variable symptoms model)	Assumed (other values considered in a sensitivity analysis)
p_2	Detection probability in mid infection	0.6 (constant symptoms model) 0.8 (variable symptoms model)	Assumed (other values considered in a sensitivity analysis)
p_3	Detection probability in mid infection	0.6 (constant symptoms model) 0.9 (variable symptoms model)	Assumed (other values considered in a sensitivity analysis)
Δ	Sampling frequency	21 days (baseline) 14 days (intensified surveillance)	Assumed (other values considered in a sensitivity analysis)

Table A1.1: Assumed parameter values. Default values of assumed (not fitted) model parameters used in our analyses. Except where explicitly indicated otherwise, model parameters were taken to be equal for the constant symptoms and variable symptoms models, and in our analyses using either simulated data or real data from the 2018–2020 DRC epidemic or the 2014–2016 epidemic in Liberia.

Quantity	**Value (constant symptoms model)**	**Value (variable symptoms model)**
Proportion of infected hosts detected and isolated, q_C	31% (baseline surveillance) 42% (intensified surveillance)	35% (baseline surveillance) 46% (intensified surveillance)
Proportion of transmissions prevented by detection and isolation, ε_C	22% (baseline surveillance) 30% (intensified surveillance)	18% (baseline surveillance) 25% (intensified surveillance)

Table A1.2: Effect of surveillance. Calculated values of the proportion of infected individuals who are ever detected and isolated during the course of infection, q_C (see equation (R1.12)), and proportion of transmissions prevented from each infected host by detection and isolation, ε_C (see equation (R1.14)), using the parameter values in Table A1.1.

Parameter	**True value (variable symptoms model)**	**Estimated value (constant symptoms model)**	**Estimated value (variable symptoms model)**
β_0	1.0×10^{-6} day^{-1}	1.1×10^{-6} day^{-1}	1.0×10^{-6} day^{-1}
β_1	5.1×10^{-7} day^{-1}	5.3×10^{-7} day^{-1}	5.1×10^{-7} day^{-1}
T	84 days	84 days	84 days
T_0	-14 days	-12 days	-13 days

Table A1.3: True and fitted parameter values in our analysis of simulated data. The parameter values used to generate the synthetic data using the variable symptoms model (denoted the "true" parameter values), and estimated parameters for both the constant symptoms model and the variable symptoms model when the two models were fitted to the data. Definitions of the parameters are given in Table A1.4. Times are relative to the beginning of the week in which the first case was reported.

Parameter	Interpretation	Estimated value (constant symptoms model)	Estimated value (variable symptoms model)
β_0	Transmission rate early in epidemic	8.4×10^{-7} day^{-1}	8.1×10^{-7} day^{-1}
β_1	Transmission rate late in epidemic	3.7×10^{-7} day^{-1}	3.6×10^{-7} day^{-1}
T	Transmission rate switch time	25$^{\text{th}}$ October 2018	24$^{\text{th}}$ October 2018
T_0	Start date of epidemic	25$^{\text{th}}$ June 2018	25$^{\text{th}}$ June 2018

Table A1.4: Fitted model parameters in our analysis of data from the 2018–2020 DRC epidemic. Estimated values of model parameters for both the constant symptoms model and the variable symptoms model when the two models were fitted to the data.

Parameter	Interpretation	Estimated value (constant symptoms model)	Estimated value (variable symptoms model)
β_0	Transmission rate early in epidemic	4.5×10^{-8} day^{-1}	4.3×10^{-8} day^{-1}
β_1	Transmission rate late in epidemic	1.9×10^{-8} day^{-1}	1.8×10^{-8} day^{-1}
T	Transmission rate switch time	21$^{\text{st}}$ September 2014	21$^{\text{st}}$ September 2014
T_0	Start date of epidemic	20$^{\text{th}}$ March 2014	18$^{\text{th}}$ March 2014

Table A1.5: Fitted model parameters in our analysis of data from the 2014–2016 epidemic in Liberia. Estimated values of model parameters for both the constant symptoms model and the variable symptoms model when the two models were fitted to the data.

Appendix A2

Appendix to Chapter R2

A2.1 Explicitly including variability in the infectiousness time-courses of different infected hosts in the population-scale model

In this section we consider an $\mathrm{SI}_{n,m}\mathrm{R}$ compartmental model which explicitly incorporates variability in the individual infectiousness profiles of different infected hosts. We show that, as was the case for the $\mathrm{SI}_n\mathrm{R}$ model, the K&M model is obtained in the limit of infinitely many stages of infection when the $\mathrm{SI}_{n,m}\mathrm{R}$ model is used to estimate population-scale dynamics.

A2.1.1 The $\mathrm{SI}_{n,m}\mathrm{R}$ model

Here we consider a multi-stage compartmental model in which there are m possible "types" of infection, which are acquired with probabilities p_j, $j = 1 \dots m$ (where these probabilities sum to 1). An infected individual with infection type j progresses through n infected compartments, $I_{i,j}$, $i = 1, \dots, n$. The model takes the form

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -S \sum_{k=1}^{n} \sum_{l=1}^{m} \beta_{k,l} I_{k,l}, \tag{A2.1}$$

$$\frac{\mathrm{d}I_{1,j}}{\mathrm{d}t} = p_j S \sum_{k=1}^{n} \sum_{l=1}^{m} \beta_{k,l} I_{k,l} - \mu_{1,j} I_{1,j}, \tag{A2.2}$$

$$\frac{\mathrm{d}I_{i,j}}{\mathrm{d}t} = \mu_{i-1,j} I_{i-1,j} - \mu_{i,j} I_{i,j}, \quad i = 2, \dots, n, \tag{A2.3}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \sum_{l=1}^{m} \mu_{n,l} I_{n,l}. \tag{A2.4}$$

Here, $\beta_{i,j}$ is the transmission rate parameter of individuals in compartment $I_{i,j}$, and the $1/\mu_{i,j}$ is the mean duration of the i^{th} infected stage for an individual with infection type j. The equation for R decouples and is ignored below.

In the limit of a continuum of possible infection types, denoted by the real variable y, we obtain the system of equations

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -S\sum_{k=1}^{n}\int_{-\infty}^{\infty}\beta_k(y)I_k(t,y)\mathrm{d}y, \tag{A2.5}$$

$$\frac{\partial I_1(t,y)}{\partial t} = f(y)S\sum_{k=1}^{n}\int_{-\infty}^{\infty}\beta_k(y)I_k(t,y)\mathrm{d}y - \mu_1(y)I_1(t,y), \tag{A2.6}$$

$$\frac{\partial I_i(t,y)}{\partial t} = \mu_{i-1}(y)I_{i-1}(t,y) - \mu_i(y)I_i(t,y), \quad i = 2,\ldots,n, \tag{A2.7}$$

where $f(y)$ is the probability density of acquiring a type y infection, $I_i(t,y)$ is the density of individuals with a type y infection and in the i^{th} infected stage, $\beta_i(y)$ is the transmission rate of such individuals, and $1/\mu_i(y)$ is the mean duration of the i^{th} infected stage for an individual with a type y infection. This easily generalises to a multi-parameter distribution of infection types.

A2.1.2 Approximating the population-scale dynamics within the $\mathrm{SI}_{n,m}\mathrm{R}$ framework when there are multiple infection types

Suppose now that there are m possible infection types, acquired with probabilities p_j, and that each host with a type j infection follows a known infectiousness profile, $\beta_j(\tau)$. In this case, similarly to the parameterisation of the $\mathrm{SI}_n\mathrm{R}$ model described in Section R2.2.3.1, we can approximate the population-scale dynamics within the $\mathrm{SI}_{n,m}\mathrm{R}$ framework by taking

$$\mu_{ij} = 1/\lambda(n), \tag{A2.8}$$

$$\beta_{ij} = \frac{1}{\lambda(n)}\int_{(i-1)\lambda(n)}^{i\lambda(n)}\beta_j(\tau)\mathrm{d}\tau, \qquad i = 1,\ldots,n-1, \tag{A2.9}$$

$$\beta_{nj} = \frac{1}{\lambda(n)}\int_{(n-1)\lambda(n)}^{\infty}\beta_j(\tau)\mathrm{d}\tau, \tag{A2.10}$$

where we will again take $\lambda(n) \to 0$ as the number of compartments, $n \to \infty$.

If we instead have a continuum of infection types, denoted by y, with corresponding infectiousness profiles $\beta(\tau,y)$, then we take

$$\mu_i(y) = 1/\lambda(n), \tag{A2.11}$$

$$\beta_i(y) = \frac{1}{\lambda(n)}\int_{(i-1)\lambda(n)}^{i\lambda(n)}\beta(\tau,y)\mathrm{d}\tau, \qquad i = 1,\ldots,n-1, \tag{A2.12}$$

$$\beta_n(y) = \frac{1}{\lambda(n)}\int_{(n-1)\lambda(n)}^{\infty}\beta(\tau,y)\mathrm{d}\tau, \tag{A2.13}$$

in equations (A2.5)–(A2.7).

A2.1.3 The limit of infinitely many compartments

Now, suppose there are m possible infectiousness profiles, and we parameterise the $\mathrm{SI}_{n,m}\mathrm{R}$ model as outlined above. Analogously to our definitions in Sections R2.2.3.1 and R2.3.2, we define $\beta_j(\tau;n)$ as the piecewise constant approximation to the j^{th} infectiousness profile, and $I_j(t,\tau;n)$ as the approximate density of individuals with a type j infection who have been infected for time τ, so that the $\mathrm{SI}_{n,m}\mathrm{R}$ model can be written in the form

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -S\sum_{l=1}^{m}\int_0^{n\lambda(n)}\beta_l(\tau;n)I_l(t,\tau;n)\mathrm{d}\tau, \tag{A2.14}$$

$$\frac{\partial I_j(t,\tau;n)}{\partial t} = \begin{cases} -\dfrac{1}{\lambda(n)}\left(p_j\dfrac{\mathrm{d}S}{\mathrm{d}t} + I_j(t,\tau;n)\right), & \text{for } 0\le\tau<\lambda(n), \\ \dfrac{1}{\lambda(n)}\left(I_j(t,\tau-\lambda(n);n) - I_j(t,\tau;n)\right), & \text{for } \lambda(n)\le\tau<n\lambda(n). \end{cases} \tag{A2.15}$$

In the limit $n\to\infty$, we obtain

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -S\sum_{l=1}^{m}\int_0^{\infty}\beta_l(\tau)I_l(t,\tau)\mathrm{d}\tau, \tag{A2.16}$$

$$\frac{\partial I_j(t,\tau)}{\partial t} + \frac{\partial I_j(t,\tau)}{\partial \tau} = 0, \quad \tau>0, \tag{A2.17}$$

$$I_j(t,0) = -p_j\frac{\mathrm{d}S}{\mathrm{d}t}. \tag{A2.18}$$

The PDE (A2.17), subject to boundary condition (A2.18) and general initial condition $I_j(0,\tau)=g_j(\tau)$, has solution

$$I_j(t,\tau) = \begin{cases} g_j(\tau-t), & \text{for } t<\tau, \\ -p_j\dfrac{\mathrm{d}S(t-\tau)}{\mathrm{d}t}, & \text{for } t>\tau. \end{cases} \tag{A2.19}$$

We then find that

$$\frac{\mathrm{d}S}{\mathrm{d}t} = S\sum_{j=1}^{m}\left[\int_0^t p_j\beta_j(\tau)\frac{\mathrm{d}S(t-\tau)}{\mathrm{d}t}\mathrm{d}\tau - \int_t^{\infty}\beta_j(\tau)g_j(\tau-t)\mathrm{d}\tau\right]. \tag{A2.20}$$

If we now define the expected infectiousness profile to be

$$\beta(\tau) \coloneqq \sum_{j=1}^{m}p_j\beta_j(\tau), \tag{A2.21}$$

then we have that

$$\frac{\mathrm{d}S}{\mathrm{d}t} = S\left[\int_0^t \beta(\tau)\frac{\mathrm{d}S(t-\tau)}{\mathrm{d}t}\mathrm{d}\tau - \sum_{j=1}^{m}\int_t^{\infty}\beta_j(\tau)g_j(\tau-t)\mathrm{d}\tau\right]. \tag{A2.22}$$

Assuming that the term generated by the initial conditions has little effect once the outbreak has taken off (certainly if there exists some $T > 0$ such that for each j, $\beta_j(\tau) = 0$ for $\tau > T$, then the second term in equation (A2.22) vanishes for time $t > T$), then this is simply the standard K&M model.

Therefore, the population-scale dynamics may be calculated by assuming that all infected individuals follow infectiousness profile $\beta(\tau)$. In the limit of a continuum of possible infection types, this analysis still holds, and equation (A2.21) becomes

$$\beta(\tau) = \int_{-\infty}^{\infty} f(y)\beta(\tau, y)\mathrm{d}y. \tag{A2.23}$$

This may be generalised further to multi-parameter distributions of infection types. In our case study where host-scale influenza dynamics are described by the TCL model, the variation of the within-host model parameters δ and $V(0)$ leads to a two-parameter distribution of possible infectiousness profiles.

A2.2 Alternative relationships between pathogen load and infectiousness

In our main analyses, we assumed that the infectiousness, $\beta^{(i)}(\tau)$, of an influenza-infected host (where i represents the particular host under consideration) is proportional to their viral load, $V^{(i)}(\tau)$, at any time since infection, τ. In Figure R2.10, we considered two alternative possibilities in which we instead assumed

$$\beta^{(i)}(\tau) = k \times F\left(V^{(i)}(\tau)\right), \tag{A2.24}$$

where $F(V)$ is a known function of viral load, V, and k is a constant. In particular, we considered the following two possibilities:

i. A scenario where infectiousness scales with the logarithm of the viral load, in which we took

$$F(V) = \max\{\log_{10}\left(V/V^*\right), 0\}, \tag{A2.25}$$

where we assumed that $V^* = 10^3$ TCID_{50}/ml.

ii. A scenario where infectiousness saturates at high viral loads, in which we took

$$F(V) = \frac{V}{K + V}, \tag{A2.26}$$

where we assumed that $K = 10^5$ TCID_{50}/ml.

Assuming the basic reproduction number, R_0, to be known, the expected infectiousness profile was therefore given in each case by

$$\beta(\tau) = \frac{R_0}{N\int_0^\infty \overline{F\left(V^{(i)}(\tau)\right)}\mathrm{d}\tau}\overline{F\left(V^{(i)}(\tau)\right)}, \tag{A2.27}$$

where the bar denotes the average (at a given time since infection, τ), calculated over a large number of realisations of the within-host model. To compute the "true" expected infectiousness profile, $\beta(\tau)$, we calculated the average of $F\left(V^{(i)}(\tau)\right)$ over 10,000 within-host model realisations (Figure R2.10A,D). In cases where data were available from d hosts, each sampled once daily, we fitted the TCL within-host model to the data for each host in order to estimate $\beta(\tau)$ using equation (A2.27).

A2.3 Supplementary Figures

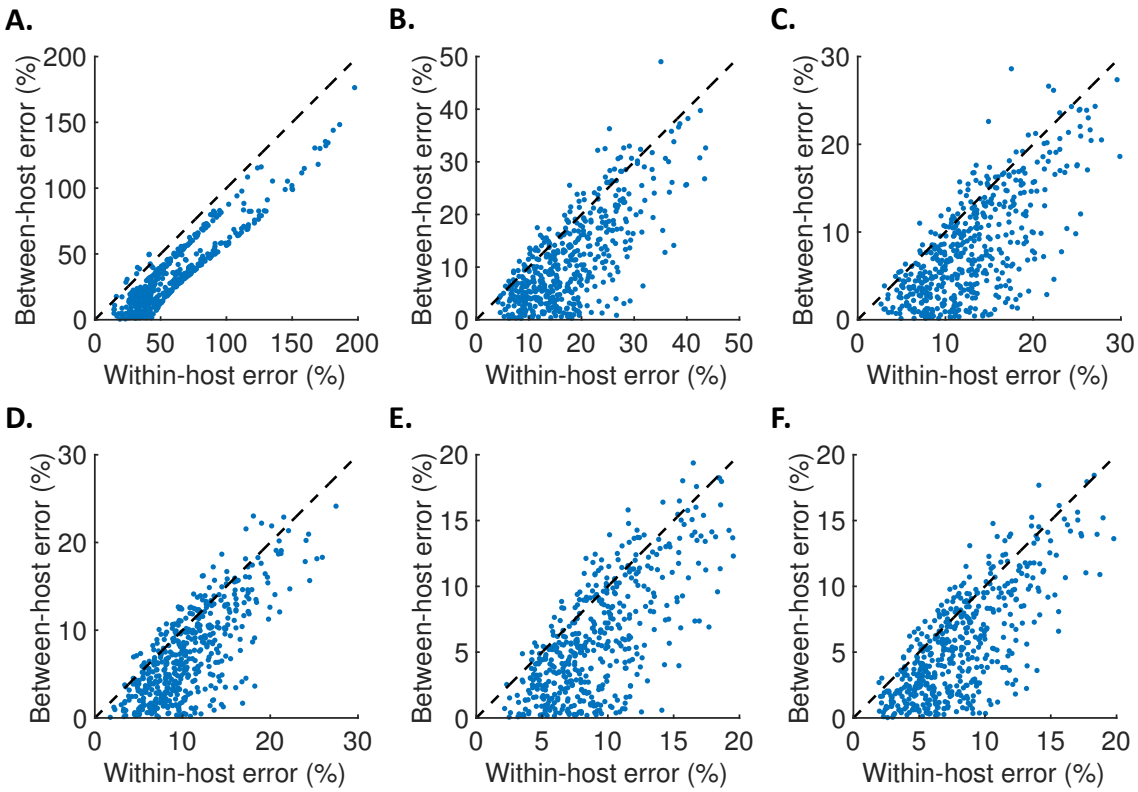


Figure A2.1: Comparison of errors at different scales. Scatter plots comparing the errors at host- and population-scale from Figure R2.5D, for cohort sizes of $d = 1$ (A), $d = 10$ (B), $d = 20$ (C), $d = 30$ (D), $d = 40$ (E) and $d = 50$ (F). Note that for each cohort size, errors are plotted here for only one in every ten repeats used to generate Figure R2.5D (500 repeats out of 5,000). The dashed line in each panel indicates where the errors at the two scales are equal.

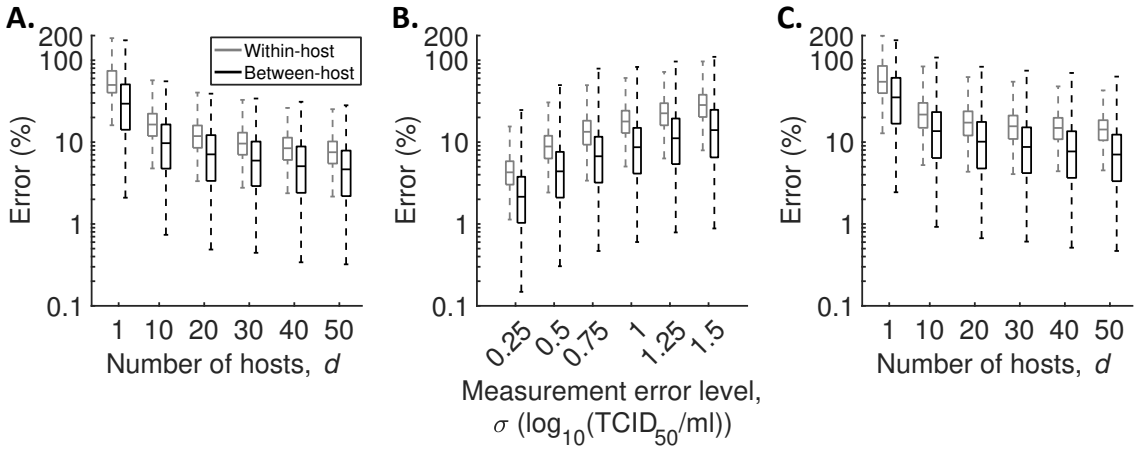


Figure A2.2: Results using the IDE method to transition from host-scale to population-scale dynamics. A–C are equivalent, respectively, to Figure R2.5D, Figure R2.6D and Figure R2.7A, but using the IDE method instead of our compartmental approach to calculate population-scale dynamics.

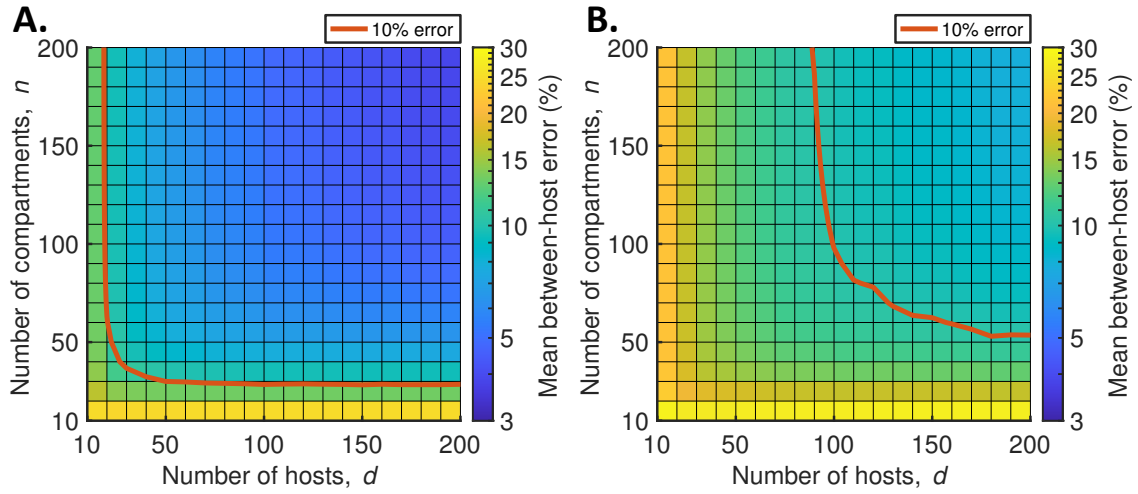


Figure A2.3: Effect of the extent of measurement error in population-scale data. A. The expected error in the population-scale dynamics when $\sigma = 0.5$ $\log_{10}(\text{TCID}_{50}/\text{ml})$, against the number of compartments, n, and the number of hosts, d. The red line indicates where the error is 10%. B. Equivalent figure to A, for $\sigma = 1.5$ $\log_{10}(\text{TCID}_{50}/\text{ml})$.

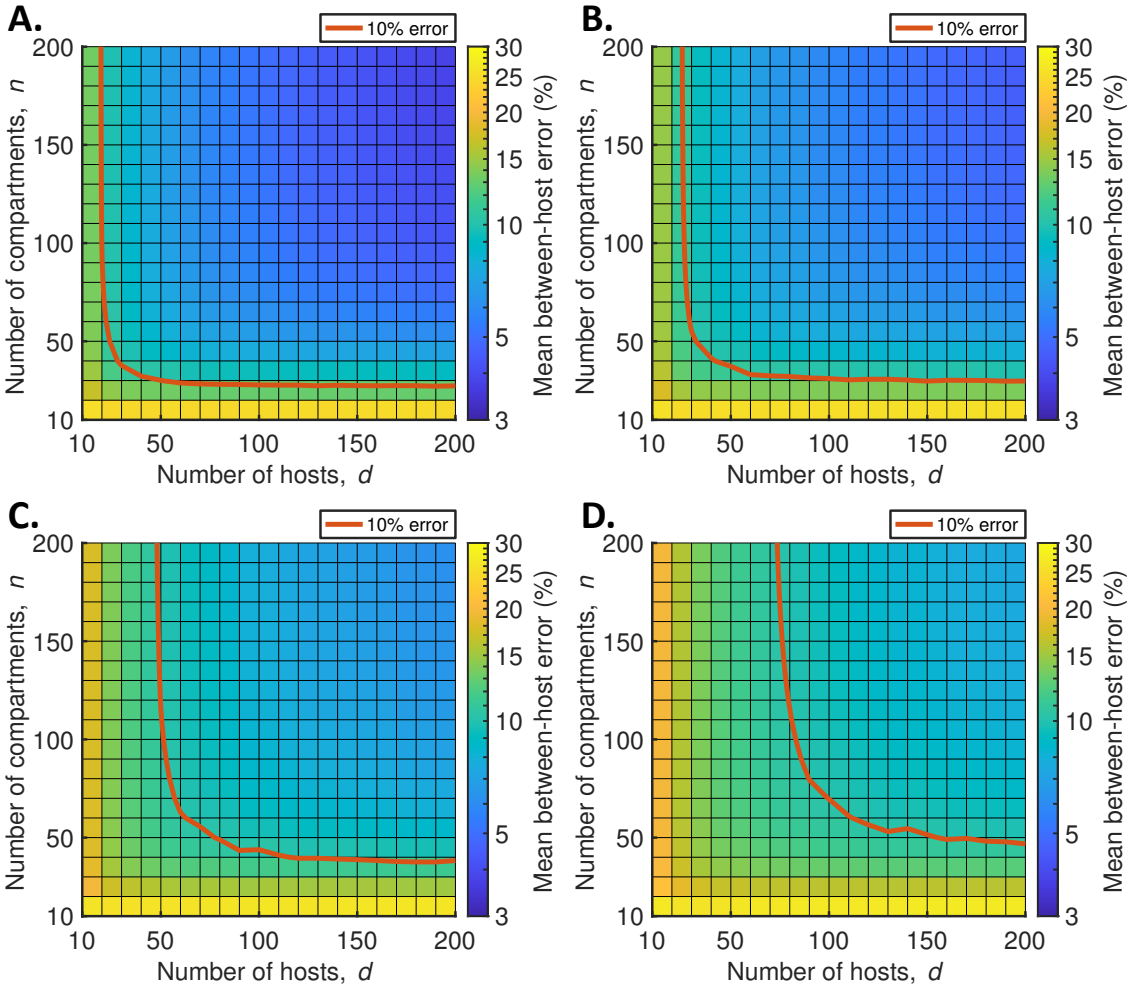


Figure A2.4: Effect of the extent of variability in host-scale parameter values. A. The expected error in the population-scale dynamics when the extent of parameter variability was assumed to be $\alpha = 0.8$ times that considered in our main analyses (see the text in Section R2.3.5.7 for the exact definition of α), against the number of compartments, n, and the number of hosts, d. The red line indicates where the error is 10%. B–D. Equivalent figures to A, for $\alpha = 0.9$ (B), $\alpha = 1.1$ (C) and $\alpha = 1.2$ (D).

Appendix A3

Appendix to Chapter R3

A3.1 Discretisation of the serial interval distribution

Here, we summarise the approach for discretising a specified continuous-time serial interval distribution, f_{ser}, proposed in [62], that we used to obtain the discretised serial interval distributions shown in Figure R3.6.

If x is the exact (continuous) serial interval for a given transmission pair, and d is the discrete serial interval (i.e. the number of days between the start of the date on which the infector developed symptoms and the start of the day on which the infectee developed symptoms), then x and d are related by

$$r_1 + x = d + r_2, \tag{A3.1}$$

where r_1 is the interval between the start of the date on which the infector developed symptoms and their exact time of symptom onset, and r_2 is the corresponding interval for the infectee. Note that r_1 and r_2 must both lie between zero and one day, giving the inequality $d - 1 \leq x \leq d + 1$ (where here, and in the following equations in this section, all time units are in days).

An expression for the probability mass function, $p(d)$, of the discretised serial interval distribution, can then be derived under the assumption that the interval r_1 is uniformly distributed between zero and one day (independently of the continuous serial interval, x). Specifically, supposing that the continuous serial interval is x, then a discrete serial interval of length d (where $d - 1 \leq x \leq d + 1$, so $x - 1 \leq d \leq x + 1$ and therefore, assuming x is not exactly an integer, two d values of $\lfloor x \rfloor$ and $\lceil x \rceil$ are possible) requires $0 \leq r_1 \leq 1$ and $0 \leq r_2 = r_1 + x - d \leq 1$, i.e.

$$\max\{0, d - x\} \leq r_1 \leq \min\{1, 1 + d - x\}. \tag{A3.2}$$

Since r_1 is assumed to be uniformly distributed, a discrete serial of interval of length d then occurs (for given x) with probability

$$p\left(d \mid x\right) = \min\left\{1, 1+d-x\right\} - \max\left\{0, d-x\right\} = 1 - |d-x|, \tag{A3.3}$$

for $d = \lfloor x \rfloor$ or $\lceil x \rceil$. Finally, conditioning on the continuous serial interval gives

$$\begin{aligned} p(d) &= \int_{d-1}^{d+1} p\left(d \mid x\right) f_{ser}(x)\mathrm{d}x & \text{(A3.4)} \\ &= \int_{d-1}^{d+1} \left(1 - |d-x|\right) f_{ser}(x)\mathrm{d}x & \text{(A3.5)} \\ &= (1+d)F_{ser}(d+1) - 2dF_{ser}(d) + (d-1)F(d-1) \\ &\quad + \int_{d-1}^{d} x f_{ser}(x)\mathrm{d}x - \int_{d}^{d+1} x f_{ser}(x)\mathrm{d}x, & \text{(A3.6)} \end{aligned}$$

where F_{ser} is the cumulative distribution function of the continuous serial interval distribution. This formula was evaluated numerically to obtain the discretised serial interval distributions shown in Figure R3.6.

A3.2 Supplementary Figures

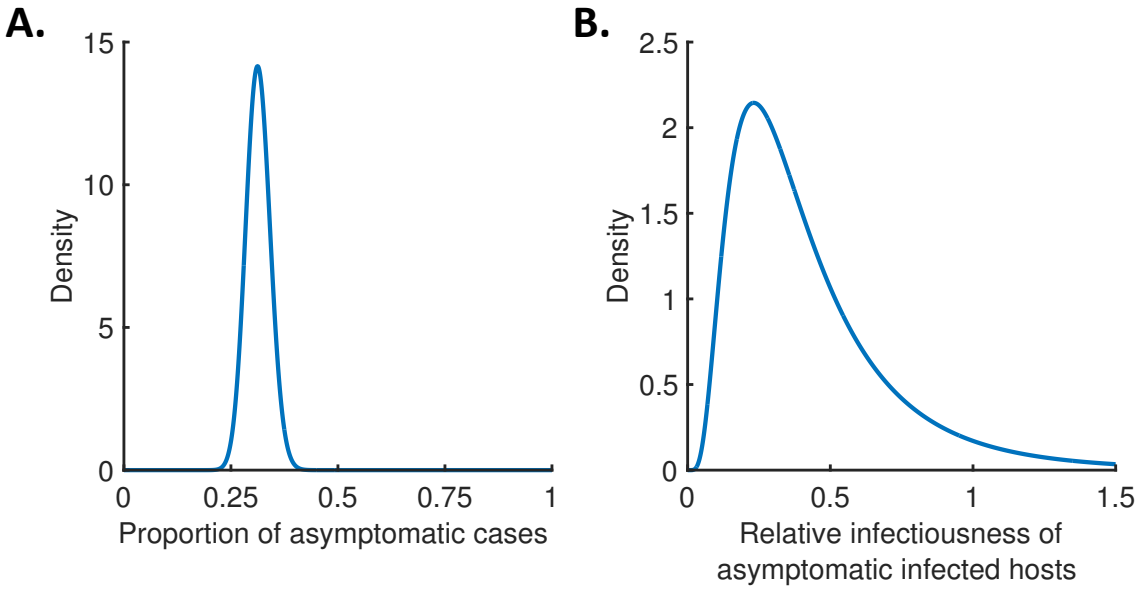


Figure A3.1: The contribution of asymptomatic cases to transmission. A. Assumed distribution for the proportion of asymptomatic cases, p_A. B. Assumed distribution for the relative infectiousness of asymptomatic infected hosts, α_A. These distributions were chosen for consistency with the confidence intervals obtained in [30] (see Section R3.2.10).

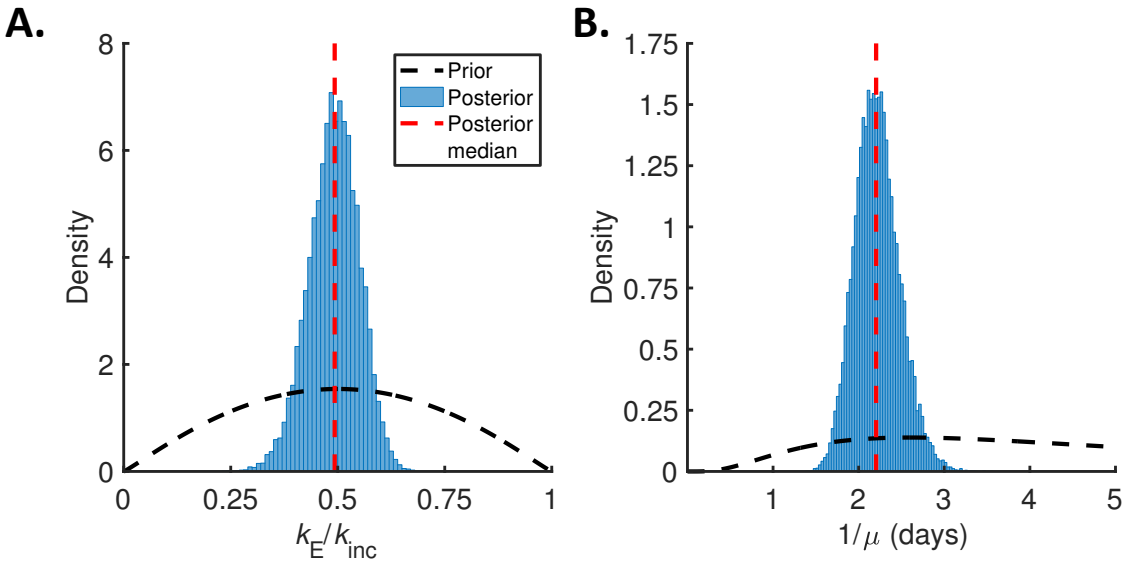


Figure A3.2: Posterior distributions of fitted parameters for the constant infectiousness model. Prior distributions (black dashed lines) of estimated parameters in the constant infectiousness model, and posterior distributions (blue bars) and posterior medians (vertical red dashed lines) obtained when we used data augmentation MCMC to fit this model to the transmission pair data. A. Ratio of mean durations of the latent (E) and incubation (combined E and P) periods, k_E/k_{inc}. B. Mean symptomatic infectious (I) period, $1/\mu$.

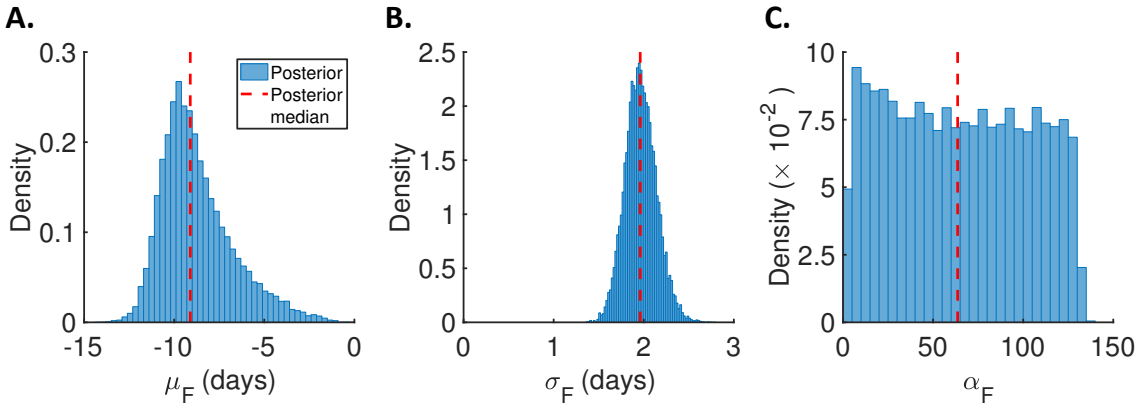


Figure A3.3: Posterior distributions of fitted parameters for the Ferretti model. Posterior distributions (blue bars) and posterior medians (vertical red dashed lines) of estimated parameters obtained when we used data augmentation MCMC to fit the Ferretti model to the transmission pair data. A. μ_F. B. σ_F. C. α_F. Uniform prior distributions were assumed for each model parameter, imposing only that the values of σ_F and α_F were positive (whereas μ_F could take either positive or negative values). Note that the parameters μ_F and α_F do not have the same epidemiological interpretations as the parameters μ and α_P in our mechanistic approach.

A3.3 Supplementary Tables

Model(s)	Parameter	Value	Justification
All	Shape parameter of (gamma) incubation period distribution, k_{inc}	5.807	Taken from [143] (alternative incubation period distribution considered in a sensitivity analysis)
	Scale parameter of (gamma) incubation period distribution	0.948	Taken from [143] (alternative incubation period distribution considered in a sensitivity analysis)
	Mean incubation period, $m_{inc} = 1/\gamma$	5.5 days	Calculated using stated shape and scale parameters (alternative incubation period distribution considered in a sensitivity analysis)
	Standard deviation of incubation period distribution	2.3 days	Calculated using stated shape and scale parameters (alternative incubation period distribution considered in a sensitivity analysis)
Variable infectiousness and constant infectiousness	Shape parameter of (gamma) symptomatic infectious (I) period distribution, k_I	1	Assumed
Constant infectiousness	Ratio of transmission rates in the P and I stages, α_P	1	Assumed

Table A3.1: Assumed parameter values. Assumed (not fitted) parameter values used in the models that we considered.

Model	Parameter	Prior distribution	Posterior median (95% CrI)
Variable infectiousness	Ratio of mean durations of the latent (E) and incubation (combined E and P) periods, k_E/k_{inc}	Beta(2.1,2.1) [prior median 0.5, 95% CrI 0.1–0.9]	0.57 (0.42–0.70)
	Mean symptomatic infectious (I) period, $1/\mu$	Lognormal(1.6,0.8) [prior median 5.0 days, 95% CrI 1.0–23.8 days]	2.6 days (1.9–3.6 days)
	Ratio of transmission rates in the P and I stages, α_P	Lognormal(0,0.8) [prior median 1.0, 95% CrI 0.2–4.8]	1.9 (0.9–4.3)
Constant infectiousness	Ratio of mean durations of the latent (E) and incubation (combined E and P) periods, k_E/k_{inc}	Beta(2.1,2.1) [prior median 0.5, 95% CrI 0.1–0.9]	0.49 (0.37–0.60)
	Mean symptomatic infectious (I) period, $1/\mu$	Lognormal(1.6,0.8) [prior median 5.0 days, 95% CrI 1.0–23.8 days]	2.2 days (1.7–2.8 days)
Ferretti	μ_F	Uniform on $(-\infty, \infty)$	-9.1 days ((-11.7)–(-3.8) days)
	σ_F	Uniform on $(0, \infty)$	2.0 days (1.6–2.3 days)
	α_F	Uniform on $(0, \infty)$	64 (5–127)
Independent transmission and symptoms	Mean generation time, m_{gen}	Lognormal(1.6,0.35) [prior median 5.0 days, 95% CrI 2.5–9.8 days]	5.6 days (5.1–6.1 days)
	Standard deviation of generation time distribution, s_{gen}	Lognormal(0.7,0.65) [prior median 2.0 days, 95% CrI 0.6–7.2 days]	2.3 days (1.8–2.9 days)

Table A3.2: Fitted parameter values using data augmentation MCMC. The prior distributions used, and posterior medians and 95% credible intervals obtained in the data augmentation MCMC procedure (which accounted for uncertainty in precise symptom onset times within recorded onset dates), are given for fitted parameters in each model we considered. Note that the parameters μ_F and α_F in the Ferretti model do not have the same epidemiological interpretations as μ and α_P in our mechanistic approach.

Model	**ΔAIC value**	**Parameter**	**Maximum likelihood estimate**
Variable infectiousness	0	Ratio of mean durations of the latent (E) and incubation (combined E and P) periods, k_E/k_{inc}	0.61
		Mean symptomatic infectious (I) period, $1/\mu$	2.7 days
		Ratio of transmission rates in the P and I stages, α_P	2.4
Constant infectiousness	1.3	Ratio of mean durations of the latent (E) and incubation (combined E and P) periods, k_E/k_{inc}	0.50
		Mean symptomatic infectious (I) period, $1/\mu$	2.2 days
Ferretti	5.1	μ_F	-3.9 days
		σ_F	1.8 days
		α_F	5.3
Independent transmission and symptoms	38.9	Mean generation time, m_{gen}	5.6 days
		Standard deviation of generation time distribution, s_{gen}	2.4 days

Table A3.3: Maximum likelihood parameter estimates and ΔAIC values. Maximum likelihood parameter estimates, obtained under the assumption that each individual developed symptoms in the middle of their symptom onset date, are given for each fitted parameter in the four models that we considered. Note that the parameters μ_F and α_F in the Ferretti model do not have the same epidemiological interpretations as the parameters μ and α_P in our mechanistic approach. ΔAIC values, relative to the value for the variable infectiousness model (which provided the best fit to the data) are also given to indicate the relative goodness of fit of each model.

Model	Distribution	Mean	Standard deviation
Variable infectiousness	Generation time	6.0 days	3.2 days
	TOST	−0.1 days	2.8 days
	Serial interval	5.4 days	3.6 days
Constant infectiousness	Generation time	6.0 days	3.1 days
	TOST	−0.1 days	2.7 days
	Serial interval	5.4 days	3.6 days
Ferretti	Generation time	6.0 days	3.1 days
	TOST	0.0 days	2.8 days
	Serial interval	5.5 days	3.6 days
Independent transmission and symptoms	Generation time	5.6 days	2.3 days
	TOST	0.0 days	3.3 days
	Serial interval	5.6 days	4.0 days

Table A3.4: Mean and standard deviation of distributions of epidemiological time periods. The means and standard deviations of the generation time, TOST and serial interval distributions shown in Figure R3.5. Other than the generation time distribution for the independent transmission and symptoms model (which is lognormal with the specified mean and standard deviation), none of the remaining distributions take a simple parametric form.

Appendix A4

Appendix to Chapter R4

A4.1 Details of parameter estimation algorithm for the mechanistic model

The data augmentation MCMC procedure we used to fit the mechanistic model to the household transmission data is given in Algorithm A4.1 (see Section R4.2.7).

A4.2 Relationship between the generation time, TOST and serial interval

Here, we consider a randomly chosen infector-infectee pair (in which both the infector and the infectee develop symptoms) within a large, well-mixed population, of which only a small proportion is infected. In that setting, the observed generation time distribution is equal to the normalised infectiousness profile, which will not be true within a household (cf. Figures R4.4 and R4.5). We define:

$$\tau_{inc,1} = (\text{incubation period of the infector}), \tag{A4.1}$$

$$\tau_{inc,2} = (\text{incubation period of the infectee}), \tag{A4.2}$$

$$\tau_{gen} = (\text{generation time}), \tag{A4.3}$$

$$x_{tost} = (\text{TOST}), \tag{A4.4}$$

$$x_{ser} = (\text{serial interval}), \tag{A4.5}$$

where we use τ for time intervals relative to the time of infection and x for those relative to the time of symptom onset. We denote the probability density functions of these time periods by $f_{inc,1}$, $f_{inc,2}$, f_{gen}, f_{tost} and f_{ser}, respectively. Note that

$$x_{tost} = \tau_{gen} - \tau_{inc,1}, \tag{A4.6}$$

Algorithm A4.1: The data augmentation MCMC algorithm used to fit the mechanistic model to the household transmission data.

Initialise θ and $\boldsymbol{t}$;
for `step_number` $= 1, \ldots,$ `steps_total` **do**
  if `step_number` $\equiv 1 \pmod 4$ **then**
    Propose new values for each entry of the vector of model parameters, θ, using independent normal proposal distributions for each parameter (around the corresponding parameter values in the previous step of the chain);
    Accept the proposed parameters, θ_{prop}, with probability

$$\min\left(\frac{L\left(\theta_{prop};\boldsymbol{t}\right)\pi\left(\theta_{prop}\right)}{L\left(\theta_{old};\boldsymbol{t}\right)\pi\left(\theta_{old}\right)},1\right),$$

    where θ_{old} denotes the vector of parameter values from the previous step of the chain, and where the augmented data, $\boldsymbol{t}$, remain unchanged in this step;
  else
    if `step_number` $\equiv 2 \pmod 4$ **then**
      Propose new values for the precise symptom onset times of each symptomatic infected host, using independent uniform proposal distributions (within the day of symptom of onset for each host);
    else if `step_number` $\equiv 3 \pmod 4$ **then**
      Propose new values for the infection time of one randomly chosen infected host in each household (either symptomatic or asymptomatic), using independent normal proposal distributions (around the equivalent times in the previous step of the chain);
    else
      Propose new values for both the infection time, t, and the time of the start of the I stage, t_s, holding $t_s - t$ constant, for one randomly chosen asymptomatic infected host in each household (in households where there was at least one), using independent normal proposal distributions (around the equivalent times in the previous step of the chain);
    end
    for $m = 1, \ldots M$ **do**
      Accept the proposed augmented data, $\boldsymbol{t}^{(m)}_{prop}$, for household m, with probability

$$\min\left(\frac{L^{(m)}\left(\theta;\boldsymbol{t}^{(m)}_{prop}\right)}{L^{(m)}\left(\theta;\boldsymbol{t}^{(m)}_{old}\right)},1\right),$$

      where $\boldsymbol{t}^{(m)}_{old}$ denotes the corresponding augmented data from the previous step of the chain, and where the model parameters, θ, remain unchanged in this step;
    end
  end
end

and

$$x_{ser} = x_{tost} + \tau_{inc,2}, \tag{A4.7}$$

so that

$$x_{ser} = \tau_{gen} + \tau_{inc,2} - \tau_{inc,1}. \tag{A4.8}$$

In the independent transmission and symptoms model, τ_{gen} and $\tau_{inc,1}$ are assumed to be independent, and the incubation periods of the infector and infectee are assumed to be drawn independently from the population incubation period distribution, $f_{inc} = f_{inc,1} = f_{inc,2}$. Therefore, the TOST distribution is given by the convolution

$$f_{tost}\left(x_{tost}\right) = \int_0^\infty f_{gen}\left(x_{tost} + \tau\right) f_{inc}\left(\tau\right) \mathrm{d}\tau. \tag{A4.9}$$

Assuming that x_{tost} and $\tau_{inc,2}$ are independent, the serial interval distribution can be calculated from the TOST distribution as

$$f_{ser}\left(x_{ser}\right) = \int_0^\infty f_{tost}\left(x_{ser} - \tau\right) f_{inc}\left(\tau\right) \mathrm{d}\tau. \tag{A4.10}$$

Note that

$$\mathbb{E}[x_{ser}] = \mathbb{E}[\tau_{gen}] + \mathbb{E}[\tau_{inc,2}] - \mathbb{E}[\tau_{inc,1}] = \mathbb{E}[\tau_{gen}], \tag{A4.11}$$

i.e. the generation time and serial interval distributions have the same mean.

For the mechanistic model, we still have $f_{inc,2} = f_{inc}$, and the serial interval distribution can be calculated from the TOST distribution using equation (A4.10). On the other hand, τ_{gen} and $\tau_{inc,1}$ are not independent, so equation (A4.9) connecting the TOST and generation time distributions for the independent transmission and symptoms model does not hold for the mechanistic model. As shown in Chapter R3 (Section R3.2.6), the TOST distribution for the mechanistic model is, instead, given by

$$f_{tost}\left(x_{tost}\right) = \begin{cases} \alpha_P C\left(1 - F_P\left(-x_{tost}\right)\right), & x_{tost} < 0, \\ C\left(1 - F_I\left(x_{tost}\right)\right), & x_{tost} \geq 0. \end{cases} \tag{A4.12}$$

Further, under the mechanistic model, the expected number of presymptomatic transmissions generated by an infected host is dependent on their incubation period. As a result, the infector's incubation period does not follow the same distribution as that of the infectee. In particular, by Bayes' theorem, we have

$$f_{inc,1}\left(\tau_{inc,1}\right) = p\left(\tau_{inc,1} \mid 1 \to 2\right) = \frac{p\left(1 \to 2 \mid \tau_{inc,1}\right) f_{inc}\left(\tau_{inc,1}\right)}{p\left(1 \to 2\right)}, \tag{A4.13}$$

where we write $1 \to 2$ to denote the occurrence of the transmission from the infector to the infectee. Because we are here considering a large population, the probability of

the transmission occurring is proportional to the overall infectiousness of the infector (integrated over the course of infection), $B(\infty)$, so we have

$$f_{inc,1}(\tau_{inc,1}) = \frac{B(\infty \mid \tau_{inc,1}) f_{inc}(\tau_{inc,1})}{B(\infty)} \tag{A4.14}$$

$$= \left(\frac{\alpha_P k_P \gamma \mu \tau_{inc,1} + k_{inc}\gamma}{\alpha_P k_P \mu + k_{inc}\gamma} \right) f_{inc}(\tau_{inc,1}). \tag{A4.15}$$

The expected incubation period of the infector is then

$$\mathbb{E}[\tau_{inc,1}] = \frac{1}{\gamma} + \frac{\alpha_P k_P \mu}{k_{inc}\gamma(\alpha_P k_P \mu + k_{inc}\gamma)} = \mathbb{E}[\tau_{inc,2}] + \frac{q_P}{k_{inc}\gamma}, \tag{A4.16}$$

where q_P is the proportion of transmissions occurring prior to symptom onset.

As a result of the above, the expected values of the generation time and serial interval in the mechanistic model are not equal. Instead, we have

$$\mathbb{E}[x_{ser}] = \mathbb{E}[\tau_{gen}] - \frac{q_P}{k_{inc}\gamma}. \tag{A4.17}$$

Under the values of k_{inc} and γ that we assumed (Table A4.1), this gives a mean generation time that is approximately $(1.6 \times q_P)$ days longer than the mean serial interval.

A4.3 Extension of framework to account for co-primary cases

In most of our analyses, we assumed that each household transmission chain was initiated by a single primary case, so that all other infected household members were infected from within the household. However, we also relaxed this assumption by extending our framework to account for the possibility of co-primary cases (Figures R4.10 and R4.11). Rather than assuming that all co-primary cases were infected at exactly the same time, we instead assumed that each household member could be infected at any time during a primary infection event that was taken to last one day (the choice of one day was arbitrary but, in principle, any duration could be used). This enabled us to easily incorporate the possibility of co-primary cases into our data augmentation MCMC approach by adapting the likelihood function as described below.

As in Section R4.2.5, we here consider a household (of size n) in which n_I household members become infected (of whom n_S develop symptoms and n_A remain asymptomatic throughout infection) and n_U remain uninfected. We derive an amended likelihood function which accounts for co-primary cases, again conditional

on the following data (which were imputed during the parameter fitting procedure): the entire sequence of infection times of individuals in the household ($t_1 < \ldots < t_{n_I}$); the precise symptom onset time ($t_{s,j}$) of each host (j) who develops symptoms (where infected household members are here labelled in order of their infection times); for the mechanistic model only, the times at which entirely asymptomatic infected hosts enter the I stage of infection (also denoted by $t_{s,j}$).

Under either the independent transmission and symptoms model or the mechanistic model, we now denote the total force of infection exerted on each susceptible member of the household by other household members at time t by $\lambda_h(t)$, and the cumulative force of infection by $\Lambda_h(t)$ (i.e. these correspond to the quantities denoted by $\lambda(t)$ and $\Lambda(t)$, respectively, in Section R4.2.5). Assuming each (susceptible) household member is also subject to a constant force of infection, β_p, during a primary transmission event taking place between times $t_{p,start}$ and $t_{p,end}$, the total force of infection exerted on each susceptible household member at time t is

$$\lambda(t) = \lambda_p(t) + \lambda_h(t), \tag{A4.18}$$

where

$$\lambda_p(t) = \begin{cases} \beta_p, & t_{p,start} \leq t \leq t_{p,end}; \\ 0, & \text{otherwise.} \end{cases} \tag{A4.19}$$

The cumulative force of infection is

$$\Lambda(t) = \Lambda_p(t) + \Lambda_h(t), \tag{A4.20}$$

where

$$\Lambda_p(t) = \int_{-\infty}^{t} \lambda_p(s)\,\mathrm{d}s = \frac{\beta_p}{2}\left(t_{p,end} - t_{p,start} + |t - t_{p,start}| - |t_{p,end} - t|\right). \tag{A4.21}$$

We took $t_{p,start}$ and $t_{p,end}$ to be the start and end of the day of the first household member becoming infected, respectively.

The likelihood contribution from the household, $L(\theta)$, where θ is the vector of unknown model parameters, is then given by

$$L(\theta) = \frac{1}{1 - \exp\left(-n\beta_p\left(t_{p,end} - t_{p,start}\right)\right)} \prod_{k=1}^{n} L_{k,1}(\theta)\, L_{k,2}(\theta). \tag{A4.22}$$

Here,

$$L_{k,1}(\theta) = \begin{cases} \lambda(t_k)\exp\left(-\Lambda(t_k)\right), & \text{for } k = 1, \ldots, n_I; \\ \exp\left(-\Lambda(\infty)\right), & \text{for } k = n_I + 1, \ldots, n; \end{cases} \tag{A4.23}$$

and for the independent transmission and symptoms model,

$$L_{k,2}(\theta) = \begin{cases} f_{inc}(t_{s,k} - t_k), & \text{if individual } k \text{ becomes infected and develops symptoms;} \\ 1, & \text{otherwise;} \end{cases} \tag{A4.24}$$

where f_{inc} is the probability density function of the incubation period, while for the mechanistic model,

$$L_{k,2}(\theta) = \begin{cases} f_{inc}(t_{s,k} - t_k), & \text{for } k = 1, \ldots, n_I; \\ 1, & \text{for } k = n_I + 1, \ldots, n. \end{cases} \tag{A4.25}$$

The factor

$$\frac{1}{1 - \exp(-n\beta_p(t_{p,end} - t_{p,start}))} \tag{A4.26}$$

is included to condition on at least one household member becoming infected during the primary transmission event.

Using this likelihood function, we fitted both the independent transmission and symptoms model and the mechanistic model to the household data using the same data augmentation MCMC algorithms as in our main analyses (see Section R4.2.7 and Algorithms R4.1 and A4.1). Alongside other model parameters, we estimated the probability of infection during the primary transmission event,

$$1 - \exp(-\beta_p(t_{p,end} - t_{p,start})) = 1 - \exp(-\beta_p \times (1 \text{ day})), \tag{A4.27}$$

during the MCMC procedure. A uniform prior distribution was assumed for the probability of primary infection.

A4.4 Accounting for uncertainty in the incubation period distribution

While in our main analyses we assumed a fixed incubation period distribution that was obtained in a meta-analysis [162] (see Section R4.2.6), we also conducted a sensitivity analysis in which we accounted for uncertainty in the incubation period distribution (Figure R4.12). In this sensitivity analysis, we updated the parameters, μ_{inc} and σ_{inc}, of a lognormal incubation period distribution (these parameters represent the mean and standard deviation of the natural logarithm of the incubation period, respectively) alongside unknown model parameters during the MCMC procedure. Independent normal prior distributions (truncated at zero) consistent with the 95% confidence intervals obtained in [162] were assumed for μ_{inc} (Figure R4.12A; prior mean 1.63 log(day), standard deviation 0.061 log(day), 95% CrI 1.51–1.75) and σ_{inc}

(Figure R4.12B; prior mean 0.5 log(day), standard deviation 0.026 log(day), 95% CrI 0.45–0.55). This incubation period was then used directly when evaluating the likelihood in the independent transmission and symptoms model. In the mechanistic model, we instead assumed a gamma-distributed incubation period with the same mean and standard deviation as a lognormal distribution with parameters μ_{inc} and σ_{inc}.

A4.5 Supplementary Figures

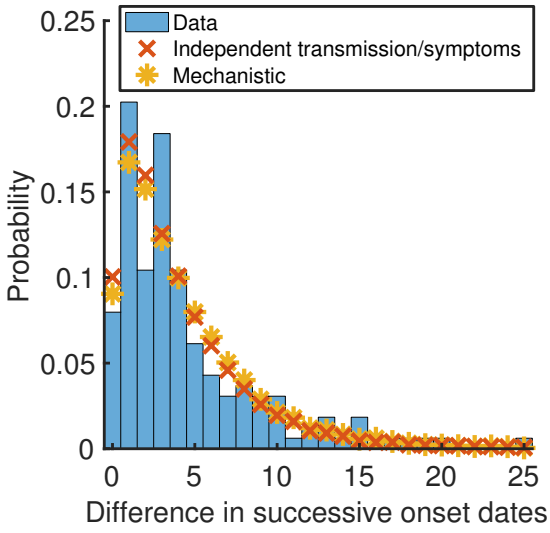


Figure A4.1: Observed and model-predicted distributions of intervals between successive symptom onset dates within households. Using posterior median parameter estimates (Table A4.2), we predicted the distribution of the difference in successive symptom onset dates (this differs from the serial interval because an infected individual may not have been infected by the previous household member to develop symptoms) under the fitted independent transmission and symptoms model (red crosses) and mechanistic model (orange stars). These distributions were compared to the UK household data (blue bars). The distributions for the fitted models were obtained by generating synthetic data from 16,700 households using the same distribution of household sizes as in the UK data.

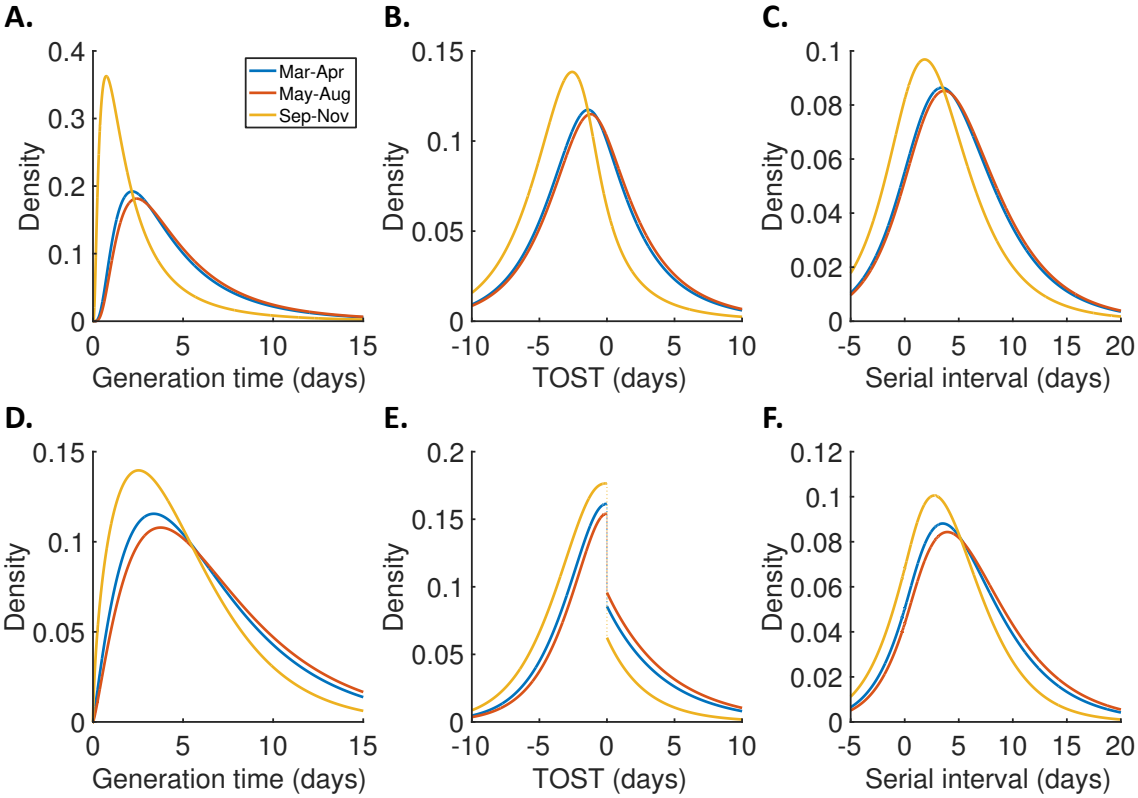


Figure A4.2: Temporal changes in the generation time, TOST and serial interval distributions. A–C. Inferred generation time (A), TOST (B) and serial interval (C) distributions, obtained using point estimate (posterior median) parameters for the independent transmission and symptoms model fitted to data from March–April (blue), May–August (red) or September–November 2020 (orange). D–F. Equivalent panels to A–C for the mechanistic model.

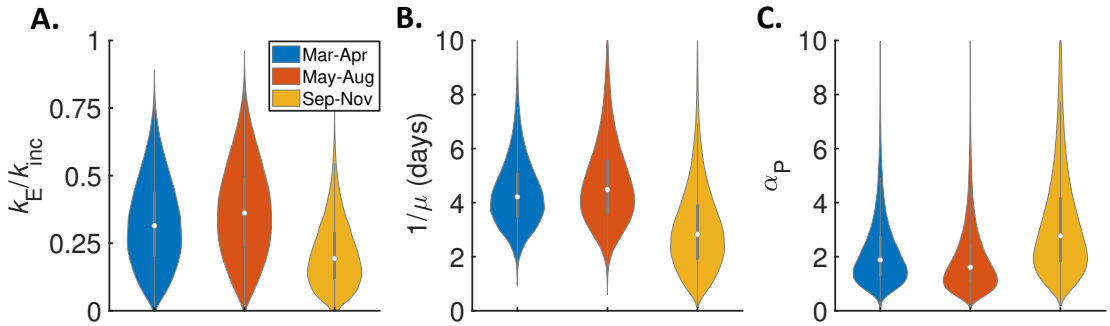


Figure A4.3: Temporal changes in fitted parameters for the mechanistic model. Violin plots indicating posterior distributions of estimated model parameters for the mechanistic model fitted to data from March–April (blue), May–August (red) or September–November 2020 (orange). A. Ratio of mean durations of the latent (E) and incubation (combined E and P) periods, k_E/k_{inc}. B. Mean symptomatic infectious (I) period, $1/\mu$. C. Ratio of transmission rates in the presymptomatic infectious (P) and symptomatic infectious (I) stages of infection, α_P. Corresponding estimates of the overall infectiousness, β_0 (the other fitted model parameter) are shown in Figure R4.8D.

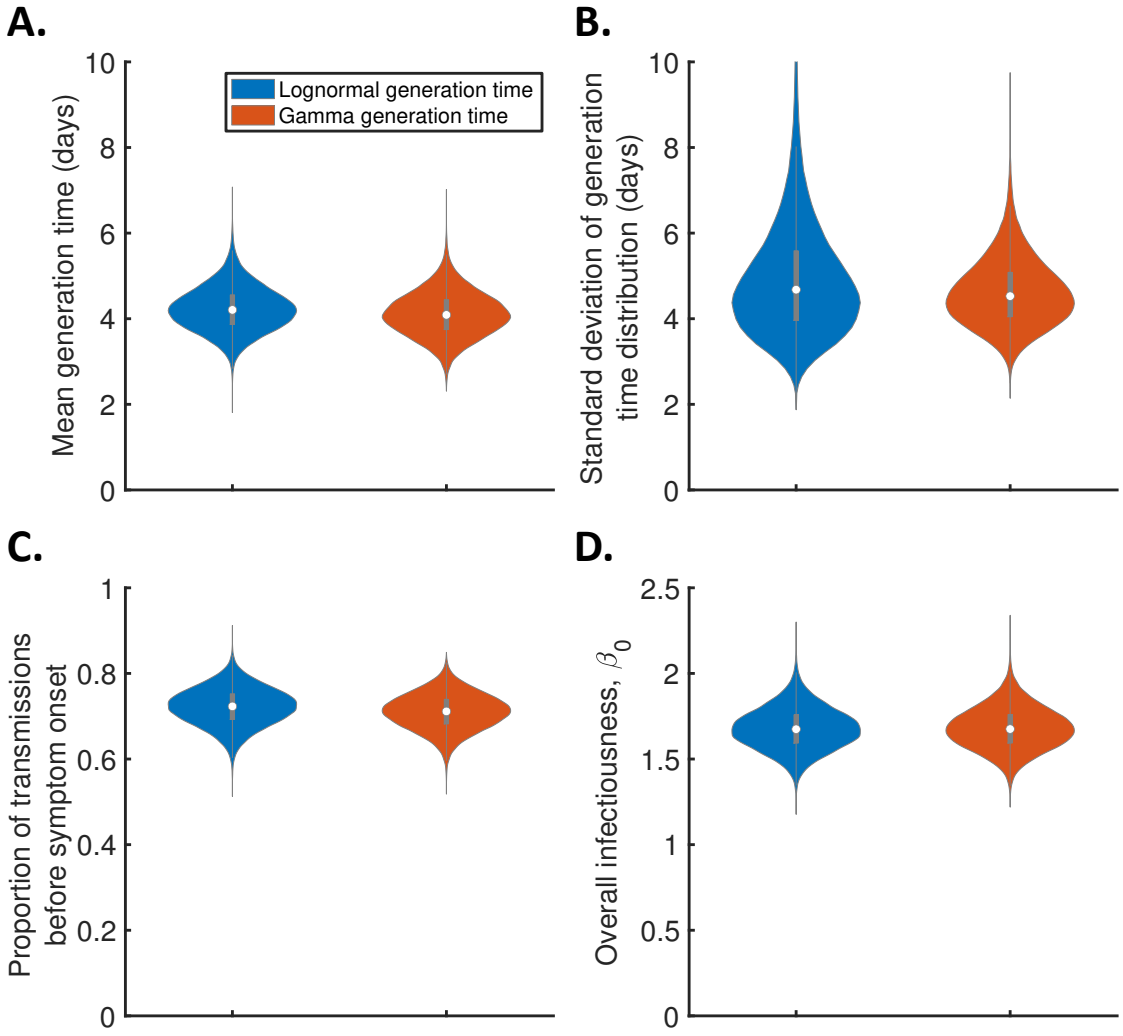


Figure A4.4: Effect of the functional form of the generation time distribution for the independent transmission and symptoms model on estimates of the generation time. Violin plots indicating posterior distributions of the mean (A) and standard deviation (B) of the generation time distribution, proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) (C), and overall infectiousness parameter, β_0 (D), for the independent transmission and symptoms model, when the generation time was assumed to follow either a lognormal (blue, as in our main analyses) or a gamma (red) distribution, and the incubation period distribution followed a lognormal distribution (as in our main analyses).

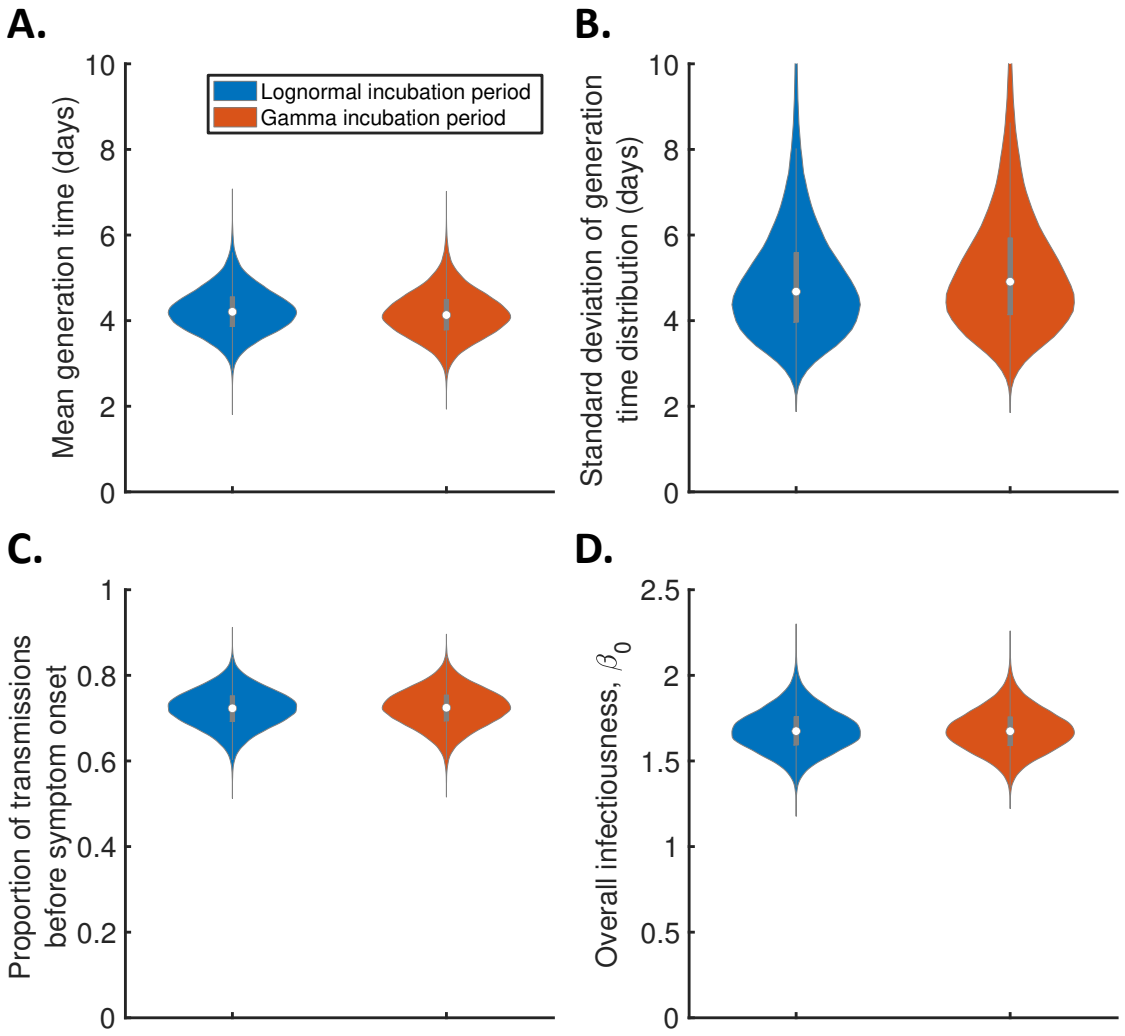


Figure A4.5: Effect of the functional form of the incubation period distribution for the independent transmission and symptoms model on estimates of the generation time. Violin plots indicating posterior distributions of the mean (A) and standard deviation (B) of the generation time distribution, proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) (C), and overall infectiousness parameter, β_0 (D), for the independent transmission and symptoms model, when the incubation period was assumed to follow either a lognormal distribution (blue, as in our main analyses) or a gamma distribution with the same mean and standard deviation (red), and the generation time followed a lognormal distribution (as in our main analyses).

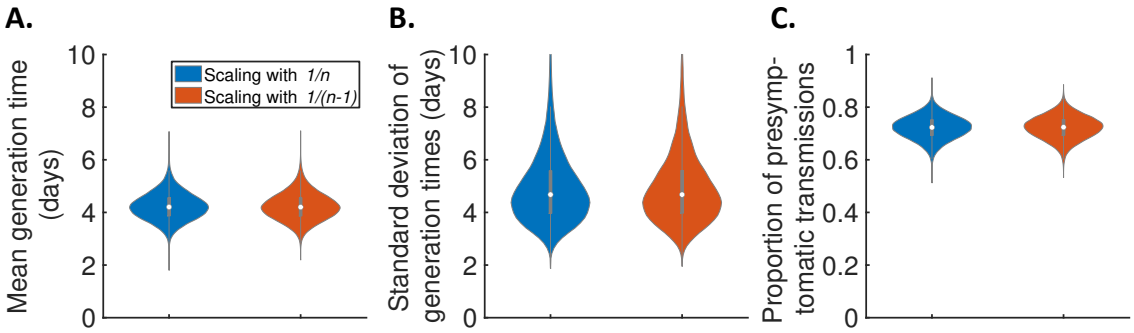


Figure A4.6: Further analysis of the effect of the dependency of transmission on the household size on estimates of the generation time. Violin plots indicating the posterior distributions of the mean (A) and standard deviation (B) of the generation time distribution, and proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) (C), for the independent transmission and symptoms model under different assumptions about the dependency of transmission on the household size. In these panels, infectiousness is assumed to scale with either $1/n$ (blue) or $1/(n-1)$ (red).

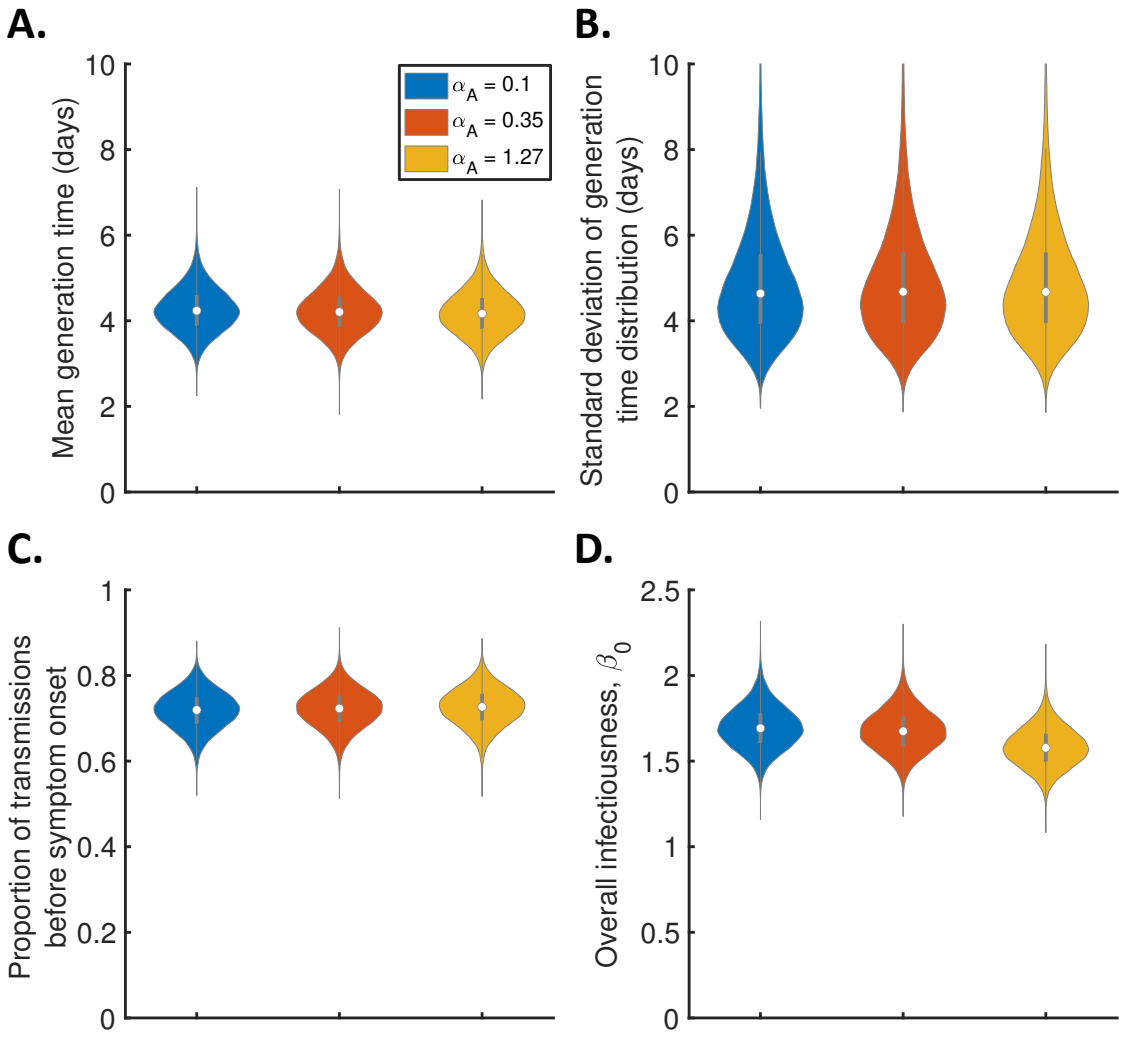


Figure A4.7: Effect of the relative infectiousness of entirely asymptomatic infected hosts on estimates of the generation time. Violin plots indicating posterior distributions of the mean (A) and standard deviation (B) of the generation time distribution, proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) (C), and overall infectiousness parameter, β_0 (D), for the independent transmission and symptoms model, when the relative infectiousness of asymptomatic hosts (compared to a host who develops symptoms, at the same time since infection) was assumed to be $\alpha_A = 0.1$ (blue), $\alpha_A = 0.35$ (red) and $\alpha_A = 1.27$ (orange).

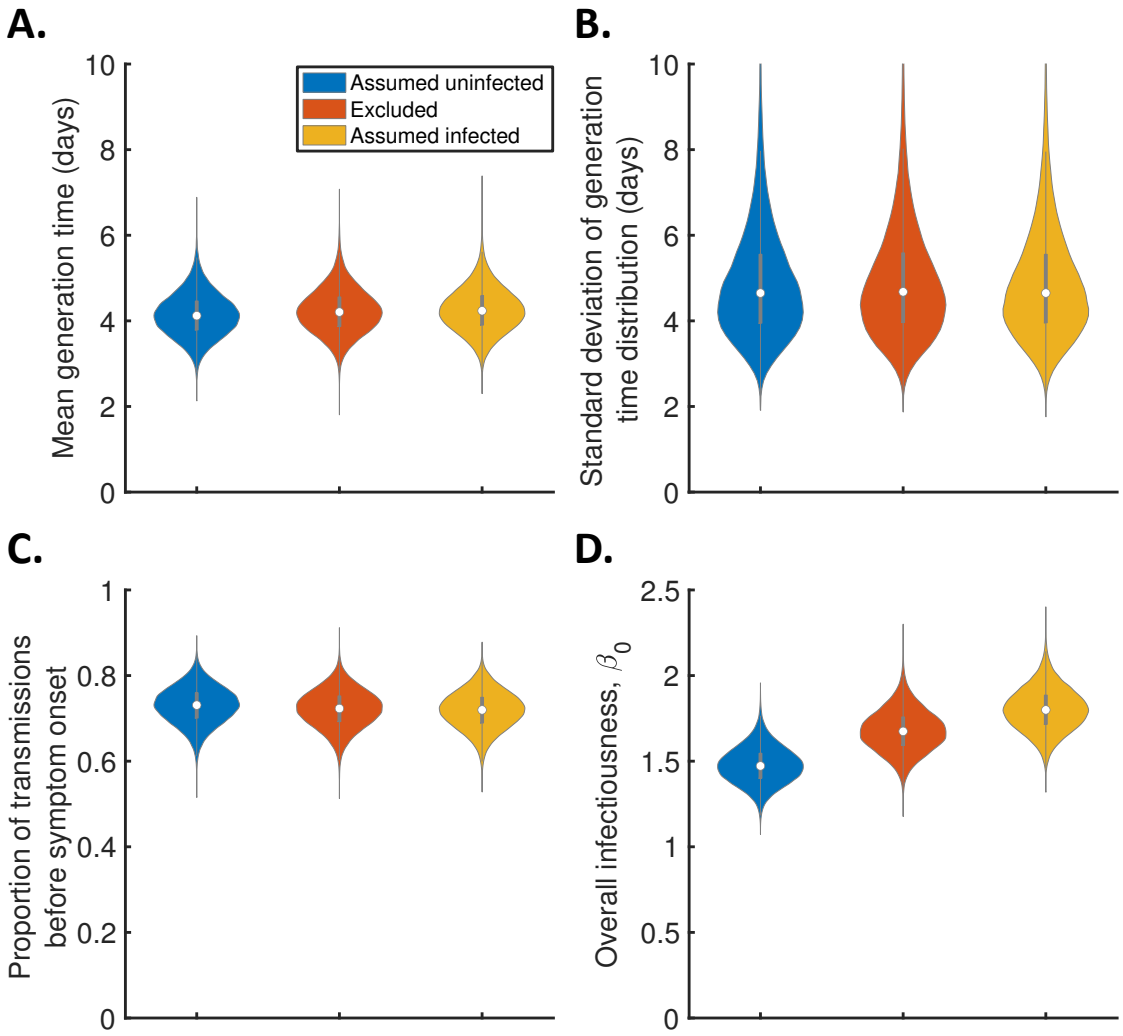


Figure A4.8: Effect of the exclusion of individuals of unknown infection status on estimates of the generation time. Violin plots indicating the posterior distributions of the mean (A) and standard deviation (B) of the generation time distribution, proportion of transmissions occurring prior to symptom onset (among infectors who develop symptoms) (C), and overall infectiousness parameter, β_0 (D), for the independent transmission and symptoms model, when individuals of unknown infection status were assumed all uninfected (blue), excluded (red), or assumed all infected (orange).

A4.6 Supplementary Tables

Model	Parameter	Value	Justification
Both	Relative infectiousness of entirely asymptomatic infected hosts, α_A	0.35	Taken from [30] (other values considered in a sensitivity analysis)
	Mean incubation period, $1/\gamma$	5.8 days	Consistent with [162] (uncertainty in incubation period distribution considered in a sensitivity analysis)
	Standard deviation of incubation period distribution	3.1 days	Consistent with [162] (uncertainty in incubation period distribution considered in a sensitivity analysis)
Independent transmission and symptoms	Mean of natural logarithm of incubation period (parameter of lognormal incubation period distribution)	1.63 log(day)	Taken from [162] (uncertainty in this value considered in a sensitivity analysis)
	Standard deviation of natural logarithm of incubation period (parameter of lognormal incubation period distribution)	0.50 log(day)	Taken from [162] (uncertainty in this value considered in a sensitivity analysis)
Mechanistic	Shape parameter of (gamma) incubation period distribution, k_{inc}	3.5	Consistent with stated mean and standard deviation from [162]
	Shape parameter of (gamma) symptomatic infectious (I) period distribution, k_I	1	Assumed

Table A4.1: Assumed parameter values. Assumed (not fitted) parameter values used in the two models that we considered.

Model	Parameter	Prior distribution	Posterior median (95% CrI)
Independent transmission and symptoms	Mean generation time, m_{gen}	Lognormal(1.6,0.35) [prior median 5.0 days, 95% CrI 2.5–9.8 days]	4.2 days (3.3–5.3 days)
	Standard deviation of generation time distribution, s_{gen}	Lognormal(0.7,0.65) [prior median 2.0 days, 95% CrI 0.6–7.2 days]	4.7 days (3.0–8.3 days)
	Overall infectiousness parameter, β_0	Lognormal(0.7,0.8) [prior median 2.0, 95% CrI 0.4–9.7]	1.7 (1.4–1.9)
Mechanistic	Ratio of mean durations of the latent (E) and incubation (combined E and P) periods, k_E/k_{inc}	Beta(2.1,2.1) [prior median 0.5, 95% CrI 0.1–0.9]	0.18 (0.03–0.46)
	Mean symptomatic infectious (I) period, $1/\mu$	Lognormal(1.6,0.8) [prior median 5.0 days, 95% CrI 1.0–23.8 days]	4.9 days (3.2–7.5 days)
	Ratio of transmission rates in the P and I stages, α_P	Lognormal(0,0.8) [prior median 1.0, 95% CrI 0.2–4.8]	2.8 (1.2–6.9)
	Overall infectiousness parameter, β_0	Lognormal(0.7,0.8) [prior median 2.0, 95% CrI 0.4–9.7]	1.8 (1.5–2.1)

Table A4.2: Fitted parameter values. Descriptions of fitted parameters in the two models, the prior distributions used, and the posterior medians and 95% credible intervals obtained.

Model	Distribution	Mean	Standard deviation
Independent transmission and symptoms	Generation time	4.2 days	4.7 days
	TOST	−1.6 days	5.6 days
	Serial interval	4.2 days	6.4 days
Mechanistic	Generation time	5.9 days	4.8 days
	TOST	−1.0 days	4.9 days
	Serial interval	4.7 days	5.8 days

Table A4.3: Mean and standard deviation of distributions of epidemiological time periods. The means and standard deviations of the generation time, TOST and serial interval distributions shown in Figure R4.6. Other than the generation time distribution for the independent transmission and symptoms model (which is lognormal with the specified mean and standard deviation), none of the remaining distributions take a simple parametric form.

Bibliography

[1] S. Abbott, J. Hellewell, R. N. Thompson, K. Sherratt, H. P. Gibbs, N. I. Bosse, J. D. Munday, S. Meakin, E. L. Doughty, J. Y. Chun et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Research*, 5:112, 2020.

[2] S. Abbott, K. Sherratt, M. Gerstung and S. Funk. Estimation of the test to test distribution as a proxy for generation interval distribution for the Omicron variant in England. *medRxiv*:2022.01.08.22268920, 2022.

[3] F. B. Agusto, M. I. Teboh-Ewungkem and A. B. Gumel. Mathematical assessment of the effect of traditional beliefs and customs on the transmission dynamics of the 2014 Ebola outbreaks. *BMC Medicine*, 13(1):96, 2015.

[4] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[5] A. Aleta, D. Martín-Corral, A. Pastore y Piontti, M. Ajelli, M. Litvinova, M. Chinazzi, N. E. Dean, M. E. Halloran, I. M. Longini, S. Merler et al. Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. *Nature Human Behaviour*, 4(9):964–971, 2020.

[6] S. T. Ali, L. Wang, E. H. Lau, X. K. Xu, Z. Du, Y. Wu, G. M. Leung and B. J. Cowling. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. *Science*, 369(6507):1106–1109, 2020.

[7] S. Alizon and M. Van Baalen. Acute or chronic? Within-host models with immune dynamics, infection outcome, and parasite evolution. *American Naturalist*, 172(6):E244–E256, 2008.

[8] L. J. Allen and G. E. Lahodny. Extinction thresholds in deterministic and stochastic epidemic models. *Journal of Biological Dynamics*, 6(2):590–611, 2012.

[9] A. E. S. Almocera, V. K. Nguyen and E. A. Hernandez-Vargas. Multiscale model within-host and between-host for viral infectious diseases. *Journal of Mathematical Biology*, 77(4):1035–1057, 2018.

[10] C. L. Althaus. Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. *PLoS Currents*, 6, 2014.

[11] D. Anderson and R. Watson. On the spread of a disease with gamma distributed latent and infectious periods. *Biometrika*, 67(1):191–198, 1980.

[12] R. Anderson, C. Donnelly, D. Hollingsworth, M. Keeling, C. Vegvari, R. Baggaley et al. Reproduction number (R0) and growth rate (r) of the COVID-19 epidemic in the UK: methods of estimation, data sources, causes of heterogeneity, and use as a guide in policy formulation. Technical report, Royal Society, 2021.

[13] R. Anderson and R. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, 1992.

[14] P. Ashcroft, J. S. Huisman, S. Lehtinen, J. A. Bouman, C. L. Althaus, R. R. Regoes and S. Bonhoeffer. COVID-19 infectivity profile correction. *Swiss Medical Weekly*, 150(32):w20336, 2020.

[15] P. Ashcroft, S. Lehtinen, D. C. Angst, N. Low and S. Bonhoeffer. Quantifying the impact of quarantine duration on COVID-19 transmission. *eLife*, 10:e63704, 2021.

[16] S. Bacallado, Q. Zhao and N. Ju. Letter to the editor: Generation interval for COVID-19 based on symptom onset data. *Eurosurveillance*, 25(29):2001381, 2020.

[17] P. Baccam, C. Beauchemin, C. A. Macken, F. G. Hayden and A. S. Perelson. Kinetics of influenza A virus infection in humans. *Journal of Virology*, 80(15):7590–7599, 2006.

[18] M. V. Barbarossa, A. Dénes, G. Kiss, Y. Nakata, G. Röst and Z. Vizi. Transmission dynamics and final epidemic size of Ebola virus disease outbreaks with varying interventions. *PLoS ONE*, 10(7):1–21, 2015.

[19] C. A. Beauchemin and A. Handel. A review of mathematical models of influenza A infections within a host or cell culture: lessons learned and challenges ahead. *BMC Public Health*, 11(1):S7, 2011.

[20] N. J. Beeching, M. Fenech and C. F. Houlihan. Ebola virus disease. *BMJ*, 349:g7348, 2014.

[21] K. Best, J. Guedj, V. Madelain, X. de Lamballerie, S.-Y. Lim, C. E. Osuna, J. B. Whitney and A. S. Perelson. Zika plasma viral dynamics in nonhuman primates provides insights into early infection and antiviral strategies. *Proceedings of the National Academy of Sciences*, 114(33):8847–8852, 2017.

[22] Blavatnik School of Government (University of Oxford). Oxford COVID-19 Government Response Tracker. URL: `https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker`.

[23] S. P. Blythe and R. M. Anderson. Distributed incubation and infectious periods in models of the transmission dynamics of the Human Immunodeficiency Virus (HIV). *Mathematical Medicine and Biology*, 5(1):1–19, 1988.

[24] N. L. Boddington, A. Charlett, S. Elgohari, C. Byers, L. Coughlan, T. G. Vilaplana, R. Whillock, M. Sinnathamby, N. Panagiotopoulos, L. Letley et al. Epidemiological and clinical characteristics of early COVID-19 cases, United Kingdom of Great Britain and Northern Ireland. *Bulletin of the World Health Organization*, 99(3):178–189, 2021.

[25] B. Boldin and O. Diekmann. Superinfections can induce evolutionarily stable coexistence of pathogens. *Journal of Mathematical Biology*, 56(5):635–672, 2008.

[26] F. Brauer. The Kermack-McKendrick epidemic model revisited. *Mathematical Biosciences*, 198(2):119–131, 2005.

[27] T. Britton and G. S. Tomba. Estimation in emerging epidemics: Biases and remedies. *Journal of the Royal Society Interface*, 16(150):20180670, 2019.

[28] E. Brooks-Pollock, L. Danon, T. Jombart and L. Pellis. Modelling that shaped the early COVID-19 pandemic response in the UK. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1829):20210001, 2021.

[29] E. Brooks-Pollock, G. O. Roberts and M. J. Keeling. A dynamic model of bovine tuberculosis spread and control in Great Britain. *Nature*, 511(7508):228–231, 2014.

[30] D. Buitrago-Garcia, D. Egli-Gany, M. J. Counotte, S. Hossmann, H. Imeri, A. M. Ipekci, G. Salanti and N. Low. Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis. *PLoS medicine*, 17(9):e1003346, 2020.

[31] M. Bushman, C. Worby, H. H. Chang, M. U. Kraemer and W. P. Hanage. Transmission of SARS-CoV-2 before and after symptom onset: impact of nonpharmaceutical interventions in China. *European Journal of Epidemiology*, 36(4):429–439, 2021.

[32] E. H. Bussell, C. E. Dangerfield, C. A. Gilligan and N. J. Cunniffe. Applying optimal control theory to complex epidemiological models to inform real-world disease management. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1776):20180284, 2019.

[33] A. Camacho, A. Kucharski, Y. Aki-Sawyerr, M. A. White, S. Flasche, M. Baguelin, T. Pollington, J. R. Carney, R. Glover, E. Smout et al. Temporal changes in Ebola transmission in Sierra Leone and implications for control requirements: a real-time modelling study. *PLoS Currents*, 7, 2015.

[34] F. Campbell, B. Archer, H. Laurenson-Schafer, Y. Jinnai, F. Konings, N. Batra, B. Pavlin, K. Vandemaele, M. D. Van Kerkhove, T. Jombart et al. Increased transmissibility and global spread of SARSCoV- 2 variants of concern as at June 2021. *Eurosurveillance*, 26(24):2100509, 2021.

[35] L. Canini and F. Carrat. Population modeling of influenza A/H1N1 virus kinetics and symptom dynamics. *Journal of Virology*, 85(6):2764–2770, 2011.

[36] L. Canini and A. S. Perelson. Viral kinetic modeling: state of the art. *Journal of Pharmacokinetics and Pharmacodynamics*, 41(5):431–443, 2014.

[37] F. Carrat, E. Vergu, N. M. Ferguson, M. Lemaitre, S. Cauchemez, S. Leach and A. J. Valleron. Time lines of infection and disease in human influenza: a review of volunteer challenge studies. *American Journal of Epidemiology*, 167(7):775–785, 2008.

[38] R. Casagrandi, L. Bolzoni, S. A. Levin and V. Andreasen. The SIRC model and influenza A. *Mathematical Biosciences*, 200(2):152–169, 2006.

[39] M. Casey-Bryars, J. Griffin, C. McAloon, A. Byrne, J. Madden, D. Mc Evoy, Á. Collins, K. Hunt, A. Barber, F. Butler et al. Presymptomatic transmission of SARS-CoV-2 infection: a secondary analysis using published data. *BMJ Open*, 11(6):e041240, 2021.

[40] S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron and P. Y. Boëlle. A Bayesian MCMC approach to study transmission of influenza: Application to household longitudinal data. *Statistics in Medicine*, 23(22):3469–3487, 2004.

[41] S. Cauchemez, C. A. Donnelly, C. Reed, A. C. Ghani, C. Fraser, C. K. Kent, L. Finelli and N. M. Ferguson. Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. *New England Journal of Medicine*, 361(27):2619–2627, 2009.

[42] S. Cauchemez, N. M. Ferguson, A. Fox, L. Q. Mai, L. T. Thanh, P. Q. Thai, D. D. Thoang, T. N. Duong, L. N. Minh Hoa, N. Tran Hien et al. Determinants of influenza transmission in South East Asia: insights from a household cohort study in Vietnam. *PLoS Pathogens*, 10(8):e1004310, 2014.

[43] G. Celeux, F. Forbes, C. P. Robert and D. M. Titterington. Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–674, 2006.

[44] Centers for Disease Control and Prevention (CDC). Investigating a COVID-19 case, 2020. URL: `https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/contact-tracing-plan/investigating-covid-19-case.html`.

[45] M. Cevik, K. Kuppalli, J. Kindrachuk and M. Peiris. Virology, transmission, and pathogenesis of SARS-CoV-2. *The BMJ*, 371:m3862, 2020.

[46] R. Challen, E. Brooks-Pollock, K. Tsaneva-Atanasova and L. Danon. Meta-analysis of the severe acute respiratory syndrome coronavirus 2 serial intervals and the impact of parameter uncertainty on the coronavirus disease 2019 reproduction number. *Statistical Methods in Medical Research*:9622802211065159, 2021.

[47] D. Champredon, J. Dushoff and D. J. Earn. Equivalence of the Erlang-distributed SEIR epidemic model and the renewal equation. *SIAM Journal on Applied Mathematics*, 78(6):3258–3278, 2018.

[48] D. B. Chang and C. S. Young. Simple scaling laws for influenza A rise time, duration, and severity. *Journal of Theoretical Biology*, 246(4):621–635, 2007.

[49] D. L. Chao, M. E. Halloran, V. J. Obenchain and I. M. Longini. FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS Computational Biology*, 6(1):e1000656, 2010.

[50] H. Y. Cheng, S. W. Jian, D. P. Liu, T. C. Ng, W. T. Huang and H. H. Lin. Contact tracing assessment of COVID-19 transmission dynamics in Taiwan and risk at different exposure periods before and after symptom onset. *JAMA Internal Medicine*, 180(9):1156–1163, 2020.

[51] L. M. Childs, F. El Moustaid, Z. Gajewski, S. Kadelka, R. Nikin-Beers, J. W. Smith Jr, M. Walker and L. R. Johnson. Linked within-host and between-host models and data for infectious diseases: a systematic review. *PeerJ*, 7:e7057, 2019.

[52] D. Chowell, C. Castillo-Chavez, S. Krishna, X. Qiu and K. S. Anderson. Modelling the effect of early detection of Ebola. *The Lancet Infectious Diseases*, 15(2):148–149, 2015.

[53] G. Chowell, C. E. Ammon, N. W. Hengartner and J. M. Hyman. Estimation of the reproductive number of the Spanish flu epidemic in Geneva, Switzerland. *Vaccine*, 24(44-46):6747–6750, 2006.

[54] G. Chowell, N. W. Hengartner, C. Castillo-Chavez, P. W. Fenimore and J. M. Hyman. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *Journal of Theoretical Biology*, 229(1):119–126, 2004.

[55] G. Chowell and H. Nishiura. Transmission dynamics and control of Ebola virus disease (EVD): a review. *BMC Medicine*, 12(1):196, 2014.

[56] G. Chowell, H. Nishiura and L. M. Bettencourt. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *Journal of the Royal Society Interface*, 4(12):154–166, 2007.

[57] G. Chowell, L. Simonsen, C. Viboud and Y. Kuang. Is West Africa approaching a catastrophic phase or is the 2014 Ebola epidemic slowing down? Different models yield different answers for Liberia. *PLoS Currents*, 6, 2014.

[58] R. C. Christofferson, C. N. Mores and H. J. Wearing. Characterizing the likelihood of dengue emergence and detection in naïve populations. *Parasites and Vectors*, 7(1):282, 2014.

[59] B. J. Coburn, B. G. Wagner and S. Blower. Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1). *BMC Medicine*, 7:30, 2009.

[60] V. Colizza, A. Barrat, M. Barthelemy, A. J. Valleron and A. Vespignani. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS Medicine*, 4(1):e13, 2007.

[61] D. Coombs, M. A. Gilchrist and C. L. Ball. Evaluating the importance of within- and between-host selection pressures on the evolution of chronic pathogens. *Theoretical Population Biology*, 72(4):576–591, 2007.

[62] A. Cori, N. M. Ferguson, C. Fraser and S. Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9):1505–1512, 2013.

[63] N. J. Cunniffe, R. O. J. H. Stutt, F. van den Bosch and C. A. Gilligan. Time-dependent infectivity and flexible latent and infectious periods in compartmental models of plant disease. *Phytopathology*, 102(4):365–380, 2012.

[64] N. J. Cunniffe, R. C. Cobb, R. K. Meentemeyer, D. M. Rizzo and C. A. Gilligan. Modeling when, where, and how to manage a forest epidemic, motivated by sudden oak death in California. *Proceedings of the National Academy of Sciences of the United States of America*, 113(20):5640–5645, 2016.

[65] N. J. Cunniffe, F. F. Laranjeira, F. M. Neri, R. E. DeSimone and C. A. Gilligan. Cost-effective control of plant disease when epidemiological knowledge is incomplete: modelling Bahia bark scaling of citrus. *PLoS Computational Biology*, 10(8):e1003753, 2014.

[66] P. Czuppon, F. Débarre, A. Gonçalves, O. Tenaillon, A. S. Perelson, J. Guedj and F. Blanquart. Success of prophylactic antiviral therapy for SARS-CoV-2: Predicted critical efficacies and impact of different drug-specific mechanisms of action. *PLoS Computational Biology*, 17(3):e1008752, 2021.

[67] H. Dahari, R. M. Ribeiro and A. S. Perelson. Triphasic decline of hepatitis C virus RNA during antiviral therapy. *Hepatology*, 46(1):16–21, 2007.

[68] N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. B. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, 372(6538):eabg3055, 2021.

[69] N. G. Davies, P. Klepac, Y. Liu, K. Prem, M. Jit, C. A. Pearson, B. J. Quilty, A. J. Kucharski, H. Gibbs, S. Clifford et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nature Medicine*, 26(8):1205–1211, 2020.

[70] N. G. Davies, A. J. Kucharski, R. M. Eggo, A. Gimma, W. J. Edmunds, T. Jombart, K. O'Reilly, A. Endo, J. Hellewell, E. S. Nightingale et al. Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. *The Lancet Public Health*, 5(7):e375–e385, 2020.

[71] E. L. Davis, T. C. Lucas, A. Borlase, T. M. Pollington, S. Abbott, D. Ayabina, T. Crellen, J. Hellewell, L. Pi, R. Lowe et al. Contact tracing is an imperfect tool for controlling COVID-19 transmission and relies on population adherence. *Nature Communications*, 12:5412, 2021.

[72] A. Davison. *Statistical Models for Data Analysis*. Cambridge University Press, New York, 2003.

[73] T. Day, S. Alizon and N. Mideo. Bridging scales in the evolution of infectious disease life histories: theory. *Evolution*, 65(12):3448–3461, 2011.

[74] Y. Deng, C. You, Y. Liu, J. Qin and X. H. Zhou. Estimation of incubation period and generation time based on observed length-biased epidemic cohort with censoring for COVID-19 outbreak in China. *Biometrics*, 77(3):929–941, 2020.

[75] J. L. Devore. *Probability and Statistics for Engineering and the Sciences*. Brooks/Cole, Cengage Learning, Boston, 2011.

[76] O. Diekmann, M. Gyllenberg and J. A. Metz. Finite Dimensional State Representation of Linear and Nonlinear Delay Systems. *Journal of Dynamics and Differential Equations*, 30(4):1439–1467, 2018.

[77] O. Diekmann and J. A. P. Heesterbeek. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. John Wiley & Sons, Chichester, 2000.

[78] Z. Du, X. Xu, Y. Wu, L. Wang, B. J. Cowling and L. A. Meyers. Serial interval of COVID-19 among publicly reported confirmed cases. *Emerging Infectious Diseases*, 26(6):1341–1343, 2020.

[79] L. Dyson, E. M. Hill, S. Moore, J. Curran-Sebastian, M. J. Tildesley, K. A. Lythgoe, T. House, L. Pellis and M. J. Keeling. Possible future waves of SARS-CoV-2 infection generated by variants of concern with a range of characteristics. *Nature Communications*, 12:5730, 2021.

[80] A. Endo, M. Uchida, A. J. Kucharski and S. Funk. Fine-scale family structure shapes influenza transmission risk in households: Insights from primary schools in Matsumoto city, 2014/15. *PLoS Computational Biology*, 15(12):e1007589, 2019.

[81] N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg et al. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Technical report, Imperial College COVID-19 Response Team, 2020. DOI: 10.25561/77482. URL: https://doi.org/10.25561/77482..

[82] N. M. Ferguson, D. A. T. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn and D. S. Burke. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437(7056):209–214, 2005.

[83] N. M. Ferguson, C. A. Donnelly and R. M. Anderson. The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science*, 292(5519):1155–1160, 2001.

[84] L. Ferretti, A. Ledda, C. Wymant, L. Zhao, V. Ledda, L. Abeler-Dörner, M. Kendall, A. Nurtay, H.-Y. Cheng, T.-C. Ng et al. The timing of COVID-19 transmission. *medRxiv*:2020.09.04.20188516, 2020.

[85] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall and C. Fraser. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, 368(6491):eabb6936, 2020.

[86] L. Ferretti, C. Wymant, A. Nurtay, L. Zhao, R. Hinch, D. Bonsall, M. Kendall, J. Masel, J. Bell, S. Hopkins et al. Modelling the effectiveness and social costs of daily lateral flow antigen tests versus quarantine in preventing onward transmission of COVID-19 from traced contacts. *medRxiv*:2021.08.06.21261725, 2021.

[87] D. Fisman and A. Tuite. Projected impact of vaccination timing and dose availability on the course of the 2014 West African Ebola Epidemic. *PLoS Currents*, 6, 2014.

[88] C. Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS ONE*, 2(8):e758, 2007.

[89] C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons et al. Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science*, 324(5934):1557–1561, 2009.

[90] C. Fraser, S. Riley, R. M. Anderson and N. M. Ferguson. Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6146–6151, 2004.

[91] S. Frost. epirecipes. URL: `http://epirecip.es/epicookbook`.

[92] A. Gandolfi, A. Pugliese and C. Sinisgalli. Epidemic dynamics and host immune response: a nested approach. *Journal of Mathematical Biology*, 70(3):399–435, 2015.

[93] T. Ganyani, C. Kremer, D. Chen, A. Torneri, C. Faes, J. Wallinga and N. Hens. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance*, 25(17):2000257, 2020.

[94] M. Gatto, E. Bertuzzo, L. Mari, S. Miccoli, L. Carraro, R. Casagrandi and A. Rinaldo. Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences of the United States of America*, 117(19):10484–10491, 2020.

[95] M. A. Gilchrist and D. Coombs. Evolution of virulence: interdependence, constraints, and selection using nested models. *Theoretical Population Biology*, 69(2):145–153, 2006.

[96] W. R. Gilks, S. Richardson and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 1996.

[97] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[98] J. R. Gog, E. M. Hill, L. Danon and R. N. Thompson. Vaccine escape in a heterogeneous population: insights for SARS-CoV-2 from a simple model. *Royal Society Open Science*, 8:210530, 2021.

[99] J. R. Gog, L. Pellis, J. L. Wood, A. R. McLean, N. Arinaminpathy and J. O. Lloyd-Smith. Seven challenges in modeling pathogen dynamics within-host and across scales. *Epidemics*, 10:45–48, 2015.

[100] A. Gonçalves, J. Bertrand, R. Ke, E. Comets, X. de Lamballerie, D. Malvy, A. Pizzorno, O. Terrier, M. Rosa Calatrava, F. Mentré et al. Timing of antiviral treatment initiation is critical to reduce SARS-CoV-2 viral load. *CPT: Pharmacometrics and Systems Pharmacology*, 9(9):509–514, 2020.

[101] K. M. Gostic, L. McGough, E. B. Baskerville, S. Abbott, K. Joshi, C. Tedijanto, R. Kahn, R. Niehus, J. A. Hay, P. M. De Salazar et al. Practical considerations for measuring the effective reproductive number, Rt. *PLoS Computational Biology*, 16(12):e1008409, 2020.

[102] L. O. Gostin, M. M. Kavanagh and E. Cameron. Ebola and war in the Democratic Republic of Congo: avoiding failure and thinking ahead. *JAMA - Journal of the American Medical Association*, 321(3):243–244, 2019.

[103] A. Goyal, E. F. Cardozo-Ojeda and J. T. Schiffer. Potency and timing of antiviral therapy as determinants of duration of SARS-CoV-2 shedding and intensity of inflammatory response. *Science Advances*, 6(47):eabc7112, 2020.

[104] Y. H. Grad, J. C. Miller and M. Lipsitch. Cholera modeling: challenges to quantitative analysis and predicting the impact of interventions. *Epidemiology*, 23(4):523–530, 2012.

[105] J. Griffin, M. Casey, Á. Collins, K. Hunt, D. McEvoy, A. Byrne, C. McAloon, A. Barber, E. A. Lane and S. More. Rapid review of available evidence on the serial interval and generation time of COVID-19. *BMJ Open*, 10(11):e040263, 2020.

[106] T. Hale, N. Angrist, R. Goldszmidt, B. Kira, A. Petherick, T. Phillips, S. Webster, E. Cameron-Blake, L. Hallas, S. Majumdar et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, 5(4):529–538, 2021.

[107] B. Hancioglu, D. Swigon and G. Clermont. A dynamical model of human immune response to influenza A virus infection. *Journal of Theoretical Biology*, 246(1):70–86, 2007.

[108] A. Handel, J. Brown, D. Stallknecht and P. Rohani. A multi-scale analysis of influenza A virus fitness trade-offs due to temperature-dependent virus persistence. *PLoS Computational Biology*, 9(3):e1002989, 2013.

[109] A. Handel, I. M. Longini and R. Antia. Neuraminidase inhibitor resistance in influenza: assessing the danger of its generation and spread. *PLoS Computational Biology*, 3(12):2456–2464, 2007.

[110] A. Handel, I. M. Longini and R. Antia. Intervention strategies for an influenza pandemic taking into account secondary bacterial infections. *Epidemics*, 1(3):185–195, 2009.

[111] A. Handel and P. Rohani. Crossing the scale from within-host infection dynamics to between-host transmission fitness: a discussion of current assumptions and knowledge. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1675):20140302, 2015.

[112] W. S. Hart, L. F. Hochfilzer, N. J. Cunniffe, H. Lee, H. Nishiura and R. N. Thompson. Accurate forecasts of the effectiveness of interventions against Ebola may require models that account for variations in symptoms during infection. *Epidemics*, 29:100371, 2019.

[113] W. S. Hart, P. K. Maini, C. A. Yates and R. N. Thompson. A theoretical framework for transitioning from patient-level to population-scale epidemiological dynamics: influenza A as a case study. *Journal of the Royal Society, Interface*, 17(166):20200230, 2020.

[114] W. S. Hart, E. Miller, N. J. Andrews, P. Waight, P. K. Maini, S. Funk and R. N. Thompson. Generation time of the alpha and delta SARS-CoV-2 variants: an epidemiological analysis. *The Lancet Infectious Diseases*, 2022.

[115] W. S. Hart, S. Abbott, A. Endo, J. Hellewell, E. Miller, N. Andrews, P. K. Maini, S. Funk and R. N. Thompson. Inference of the SARS-CoV-2 generation time using UK household data. *eLife*, 11:e70767, 2022.

[116] W. S. Hart, P. K. Maini and R. N. Thompson. High infectiousness immediately before COVID-19 symptom onset highlights the importance of continued contact tracing. *eLife*, 10:e65534, 2021.

[117] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[118] X. He, E. H. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. C. Lau, J. Y. Wong, Y. Guan, X. Tan et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, 26(5):672–675, 2020.

[119] A. M. Henao-Restrepo, A. Camacho, I. M. Longini, C. H. Watson, W. J. Edmunds, M. Egger, M. W. Carroll, N. E. Dean, I. Diatta, M. Doumbia et al. Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!) *The Lancet*, 389(10068):505–518, 2017.

[120] E. A. Hernandez-Vargas, E. Wilk, L. Canini, F. R. Toapanta, S. C. Binder, A. Uvarovskii, T. M. Ross, C. A. Guzman, A. S. Perelson and M. Meyer-Hermann. Effects of aging on influenza virus infection dynamics. *Journal of Virology*, 88(8):4123–4131, 2014.

[121] H. W. Hethcote, J. W. Van Ark and I. M. Longini. A simulation model of AIDS in San Francisco: I. Model formulation and parameter estimation. *Mathematical Biosciences*, 106(2):203–222, 1991.

[122] S. Hu, W. Wang, Y. Wang, M. Litvinova, K. Luo, L. Ren, Q. Sun, X. X. Chen, G. Zeng, J. Li et al. Infectivity, susceptibility, and risk factors associated with SARS-CoV-2 transmission under intensive contact tracing in Hunan, China. *Nature Communications*, 12:1533, 2021.

[123] S. M. Jenness, S. M. Goodreau and M. Morris. Epimodel: An R package for mathematical modeling of infectious disease over networks. *Journal of Statistical Software*, 84:1–47, 2018.

[124] K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman and P. Daszak. Global trends in emerging infectious diseases. *Nature*, 451(7181):990–993, 2008.

[125] T. C. Jones, G. Biele, B. Mühlemann, T. Veith, J. Schneider, J. Beheim-Schwarzbach, T. Bleicker, J. Tesch, M. L. Schmidt, L. E. Sander et al. Estimating infectiousness throughout SARS-CoV-2 infection course. *Science*, 373(6551):eabi5273, 2021.

[126] S. S. Karim and Q. A. Karim. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. *The Lancet*, 398(10317):2126–2128, 2021.

[127] A. Karlinsky and D. Kobak. Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset. *eLife*, 10:e69336, 2021.

[128] R. Ke, C. Zitzmann, D. D. Ho, R. M. Ribeiro and A. S. Perelson. In vivo kinetics of SARS-CoV-2 infection and its relationship with a person's infectiousness. *Proceedings of the National Academy of Sciences of the United States of America*, 118(49):e2111477118, 2021.

[129] M. J. Keeling and K. T. Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307, 2005.

[130] M. J. Keeling, E. M. Hill, E. E. Gorsich, B. Penman, G. Guyver-Fletcher, A. Holmes, T. Leng, H. McKimm, M. Tamborrino, L. Dyson et al. Predictions of COVID-19 dynamics in the UK: short-term forecasting and analysis of potential exit strategies. *PLoS Computational Biology*, 17(1):e1008619, 2021.

[131] M. J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton, 2008.

[132] E. Kenah, M. Lipsitch and J. M. Robins. Generation interval contraction and epidemic data analysis. *Mathematical Biosciences*, 213(1):71–79, 2008.

[133] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721, 1927.

[134] K. S. Kim, K. Ejima, S. Iwanami, Y. Fujita, H. Ohashi, Y. Koizumi, Y. Asai, S. Nakaoka, K. Watashi, K. Aihara et al. A quantitative model used to compare within-host SARS-CoV-2, MERS-CoV, and SARS-CoV dynamics provides insights into the pathogenesis and treatment of SARS-CoV-2. *PLOS Biology*, 19(3):e3001128, 2021.

[135] A. A. King, M. D. De Cellés, F. M. Magpantay and P. Rohani. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proceedings of the Royal Society B: Biological Sciences*, 282(1806):20150347, 2015.

[136] M. Kiskowski and G. Chowell. Modeling household and community transmission of Ebola virus disease: epidemic growth, spatial dynamics and insights for epidemic control. *Virulence*, 7(2):163–173, 2016.

[137] J. Knight and S. Mishra. Estimating effective reproduction number using generation time versus serial interval, with application to COVID-19 in the Greater Toronto Area, Canada. *Infectious Disease Modelling*, 5:889–896, 2020.

[138] E. S. Knock, L. K. Whittles, J. A. Lees, P. N. Perez-Guzman, R. Verity, R. G. FitzJohn, K. A. Gaythorpe, N. Imai, W. Hinsley, L. C. Okell et al. Key epidemiological drivers and impact of interventions in the 2020 SARS-CoV-2 epidemic in England. *Science Translational Medicine*, 13(602):2021.01.11.21249564, 2021.

[139] A. M. Kramer, J. T. Pulliam, L. W. Alexander, A. W. Park, P. Rohani and J. M. Drake. Spatial spread of the West Africa Ebola epidemic. *Royal Society Open Science*, 3(8):160294, 2016.

[140] A. J. Kucharski, A. Camacho, S. Flasche, R. E. Glover, W. J. Edmunds, S. Funk and B. H. Singer. Measuring the impact of Ebola control measures in Sierra Leone. *Proceedings of the National Academy of Sciences of the United States of America*, 112(46):14366–14371, 2015.

[141] A. J. Kucharski, P. Klepac, A. J. Conlan, S. M. Kissler, M. L. Tang, H. Fry, J. R. Gog, W. J. Edmunds, J. C. Emery, G. Medley et al. Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(10):1151–1160, 2020.

[142] T. Kypraios. *Efficient Bayesian Inference for Partially Observed Stochastic Epidemics and A New Class of Semi-Parametric Time Series Models*. PhD thesis, Lancaster University, 2007. URL: `http://eprints.lancs.ac.uk/26392/`.

[143] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich and J. Lessler. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9):577–582, 2020.

[144] M. Legros and S. Bonhoeffer. A combined within-host and between-hosts modelling framework for the evolution of resistance to antimalarial drugs. *Journal of the Royal Society Interface*, 13(117):20160148, 2016.

[145] S. Lehtinen, P. Ashcroft and S. Bonhoeffer. On the relationship between serial interval, infectiousness profile and generation time. *Journal of the Royal Society Interface*, 18(174):20200756, 2021.

[146] N. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A. Akhmetzhanov, S.-m. Jung, B. Yuan, R. Kinoshita and H. Nishiura. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *Journal of Clinical Medicine*, 9(2):538, 2020.

[147] N. M. Linton, A. R. Akhmetzhanov and H. Nishiura. Correlation between times to SARS-CoV-2 symptom onset and secondary transmission undermines epidemic control efforts. *medRxiv*:2021.08.29.21262512, 2021.

[148] Q. H. Liu, M. Ajelli, A. Aleta, S. Merler, Y. Moreno and A. Vespignani. Measurability of the epidemic reproduction number in data-driven contact networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115(50):12680–12685, 2018.

[149] Y. Liu, Centre for Mathematical Modelling of Infectious Diseases nCoV Working Group, S. Funk and S. Flasche. The contribution of pre-symptomatic infection to the transmission dynamics of COVID-2019. *Wellcome Open Research*, 5:58, 2020.

[150] A. L. Lloyd. The dependence of viral parameter estimates on the assumed viral life cycle: limitations of studies of viral load data. *Proceedings of the Royal Society B: Biological Sciences*, 268(1469):847–854, 2001.

[151] A. L. Lloyd. Sensitivity of model-based epidemiological parameter estimation to model assumptions. In G. Chowell, J. M. Hyman, L. M. A. Bettencourt and C. Castillo-Chavez, editors, *Mathematical and Statistical Estimation Approaches in Epidemiology*, pages 123–141. Springer Netherlands, Dordrecht, 2009.

[152] A. L. Lloyd. Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. *Proceedings of the Royal Society B: Biological Sciences*, 268(1470):985–993, 2001.

[153] I. M. Longini, J. A. Røttingen, M. P. Kieny, W. J. Edmunds and A. M. Henao-Restrepo. Questionable efficacy of the rVSV-ZEBOV Ebola vaccine – Authors' reply. *The Lancet*, 391(10125):1021–1022, 2018.

[154] S. Lukens, J. Depasse, R. Rosenfeld, E. Ghedin, E. Mochan, S. T. Brown, J. Grefenstette, D. S. Burke, D. Swigon and G. Clermont. A large-scale immuno-epidemiological simulation of influenza A epidemics. *BMC Public Health*, 14(1):1019, 2014.

[155] K. A. Lythgoe, L. Pellis and C. Fraser. Is HIV short-sighted? Insights from a multistrain nested model. *Evolution*, 67(10):2769–2782, 2013.

[156] J. Ma, J. Dushoff, B. M. Bolker and D. J. Earn. Estimating initial epidemic growth Rates. *Bulletin of Mathematical Biology*, 76(1):245–260, 2014.

[157] J. Ma and D. J. Earn. Generality of the final size formula for an epidemic of a newly invading infectious disease. *Bulletin of Mathematical Biology*, 68(3):679–702, 2006.

[158] P. Magal, C. C. McCluskey and G. F. Webb. Lyapunov functional and global asymptotic stability for an infection-age model. *Applicable Analysis*, 89(7):1109–1140, 2010.

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

[160] P. Manfredi and A. D'Onofrio, editors. *Modeling the Interplay Between Human Behavior and the Spread of Infectious Diseases*. Springer, New York, 2013.

[161] A. Marc, M. Kerioui, F. Blanquart, J. Bertrand, O. Mitjà, M. Corbacho-Monné, M. Marks and J. Guedj. Quantifying the relationship between Sars-CoV-2 viral load and infectiousness. *eLife*, 10:e69302, 2021.

[162] C. McAloon, Á. Collins, K. Hunt, A. Barber, A. W. Byrne, F. Butler, M. Casey, J. Griffin, E. Lane, D. McEvoy et al. Incubation period of COVID-19: A rapid systematic review and meta-analysis of observational research. *BMJ Open*, 10(8):e039652, 2020.

[163] S. R. Meakin, M. J. Tildesley, E. Davis and M. J. Keeling. A metapopulation model for the 2018 Ebola virus disease outbreak in Equateur province in the Democratic Republic of the Congo. *bioRxiv*:465062, 2018.

[164] M. I. Meltzer, C. Y. Atkins, S. Santibanez, B. Knust, B. W. Petersen, E. D. Ervin, S. T. Nichol, I. K. Damon, M. L. Washington and Centers for Disease Control and Prevention (CDC). Estimating the future number of cases in the Ebola epidemic–Liberia and Sierra Leone, 2014-2015. *MMWR supplements*, 63(3):1–14, 2014.

[165] S. Merler, M. Ajelli, L. Fumanelli, M. F. Gomes, A. Pastore y Piontti, L. Rossi, D. L. Chao, I. M. Longini Jr, M. E. Halloran and A. Vespignani. Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. *The Lancet Infectious Diseases*, 15(2):204–211, 2015.

[166] W. G. Metzger and S. Vivas-Martínez. Questionable efficacy of the rVSV-ZEBOV Ebola vaccine. *The Lancet*, 391(10125):1021, 2018.

[167] N. Mideo, S. Alizon and T. Day. Linking within- and between-host dynamics in the evolutionary epidemiology of infectious diseases. *Trends in Ecology and Evolution*, 23(9):511–517, 2008.

[168] E. Miller, P. A. Waight, N. J. Andrews, K. McOwat, K. E. Brown, H. Katja, S. Ijaz, L. Letley, D. Haskins, M. Sinnathamby et al. Transmission of SARS-CoV-2 in the household setting: a prospective cohort study in children and adults in England. *Journal of Infection*, 83(4):483–489, 2021.

[169] C. E. Mills, J. M. Robins and M. Lipsitch. Transmissibility of 1918 pandemic influenza. *Nature*, 432(7019):904–906, 2004.

[170] L. Mitchell and J. V. Ross. A data-driven model for influenza transmission incorporating media effects. *Royal Society Open Science*, 3(10):160481, 2016.

[171] S. Moore, E. M. Hill, M. J. Tildesley, L. Dyson and M. J. Keeling. Vaccination and non-pharmaceutical interventions for COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 21(6):793–802, 2021.

[172] D. M. Morens, G. K. Folkers and A. S. Fauci. The challenge of emerging and re-emerging infectious diseases. *Nature*, 430(6996):242–249, 2004.

[173] O. Morgan. How decision makers can use quantitative approaches to guide outbreak responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374:20180365, 2019.

[174] L. N. Murillo, M. S. Murillo and A. S. Perelson. Towards multiscale modeling of influenza infection. *Journal of Theoretical Biology*, 332:267–290, 2013.

[175] B. R. Murphy, M. B. Rennels, R. G. Douglas, R. F. Betts, R. B. Couch, T. R. Cate, R. M. Chanock, A. P. Kendal, H. F. Maassab, S. Suwanagool et al. Evaluation of influenza A/Hong Kong/123/77 (H1N1) ts-1A2 and cold-adapted recombinant viruses in seronegative adult volunteers. *Infection and Immunity*, 29(2):348–355, 1980.

[176] E. Namukose, C. Bowah, I. Cole, G. Dahn, P. Nyanzee, R. Saye, M. Duworko, P. Nsubuga, M. Mawanda, N. Mahmoud et al. Active case finding for improved Ebola Virus Disease case detection in Nimba County, Liberia, 2014/2015: lessons learned. *Advances in Public Health*, 2018:6753519, 2018.

[177] N. Néant, G. Lingas, Q. Le Hingrat, J. Ghosn, I. Engelmann, Q. Lepiller, A. Gaymard, V. Ferré, C. Hartard, J.-C. Plantier et al. Modeling SARS-CoV-2 viral kinetics and association with mortality in hospitalized patients from the French COVID cohort. *Proceedings of the National Academy of Sciences*, 118(8):e2017962118, 2021.

[178] A. U. Neumann, N. P. Lam, H. Dahari, D. R. Gretch, T. E. Wiley, T. J. Layden and A. S. Perelson. Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-α therapy. *Science*, 282(5386):103–107, 1998.

[179] V. K. Nguyen and E. A. Hernandez-Vargas. Windows of opportunity for Ebola virus infection treatment and vaccination. *Scientific Reports*, 7(1):8975, 2017.

[180] V. K. Nguyen, F. Klawonn, R. Mikolajczyk and E. A. Hernandez-Vargas. Analysis of practical identifiability of a viral infection model. *PLoS ONE*, 11(12):e0167568, 2016.

[181] V. K. Nguyen, R. Mikolajczyk and E. A. Hernandez-Vargas. High-resolution epidemic simulation using within-host infection and contact data. *BMC Public Health*, 18(1):1–11, 2018.

[182] H. Nishiura and M. Eichner. Infectiousness of smallpox relative to disease age: estimates based on transmission network and incubation period. *Epidemiology and Infection*, 135(7):1145–1150, 2007.

[183] H. Nishiura, N. M. Linton and A. R. Akhmetzhanov. Serial interval of novel coronavirus (COVID-19) infections. *International Journal of Infectious Diseases*, 93:284–286, 2020.

[184] H. Ohashi, K. Watashi, W. Saso, K. Shionoya, S. Iwanami, T. Hirokawa, T. Shirai, S. Kanaya, Y. Ito, K. S. Kim et al. Potential anti-COVID-19 agents, cepharanthine and nelfinavir, and their usage for combination treatment. *iScience*, 24(4):102367, 2021.

[185] J. Paget, P. Spreeuwenberg, V. Charu, R. J. Taylor, A. D. Iuliano, J. Bresee, L. Simonsen and C. Viboud. Global mortality associated with seasonal influenza epidemics: new burden estimates and predictors from the GLaMOR Project. *Journal of Global Health*, 9(2):020421, 2019.

[186] A. Pandey, K. E. Atkins, J. Medlock, N. Wenzel, J. P. Townsend, J. E. Childs, T. G. Nyenswah, M. L. Ndeffo-Mbah and A. P. Galvani. Strategies for containing Ebola in West Africa. *Science*, 346(6212):991–995, 2014.

[187] K. V. Parag, R. N. Thompson and C. A. Donnelly. Are epidemic growth rates more informative than reproduction numbers? Technical report, 2021. URL: https://rss.org.uk/RSS/media/File-library/Publications/Special%20topic%20meeting/Final_Manuscript_Parag_Thompson_Donnelly.pdf.

[188] K. V. Parag and C. A. Donnelly. Using information theory to optimise epidemic models for real-time prediction and estimation. *PLOS Computational Biology*, 16(7):e1007990, 2020.

[189] S. W. Park, B. M. Bolker, S. Funk, C. Jessica, E. Metcalf, J. S. Weitz, B. T. Grenfell and J. Dushoff. Roles of generation-interval distributions in shaping relative epidemic strength, speed, and control of new SARS-CoV-2 variants. *medRxiv*:2021.05.03.21256545, 2021.

[190] S. W. Park, B. M. Bolker, D. Champredon, D. J. Earn, M. Li, J. S. Weitz, B. T. Grenfell and J. Dushoff. Reconciling early-outbreak estimates of the basic reproductive number and its uncertainty: framework and applications to the novel coronavirus (SARS-CoV-2) outbreak. *Journal of the Royal Society Interface*, 17(168):20200144, 2020.

[191] S. W. Park, D. Champredon and J. Dushoff. Inferring generation-interval distributions from contact-tracing data. *Journal of the Royal Society Interface*, 17(167):20190719, 2020.

[192] K. A. Pawelek, G. T. Huynh, M. Quinlivan, A. Cullinane, L. Rong and A. S. Perelson. Modeling within-host dynamics of influenza virus infection including immune responses. *PLoS Computational Biology*, 8(6):e1002588, 2012.

[193] A. S. Perelson. Modelling viral and immune system dynamics. *Nature Reviews Immunology*, 2(1):28–36, 2002.

[194] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard and D. D. Ho. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271(5255):1582–1586, 1996.

[195] Peter Salama on Twitter, 2019. URL: `https://web.archive.org/web/20190921181952/https://twitter.com/PeteSalama/status/1085265184162238465`.

[196] J. D. Peterson and R. Adhikari. Efficient and flexible methods for simulating models of time since infection. *Physical Review E*, 104(2):024410, 2021.

[197] K. B. Pouwels, T. House, E. Pritchard, J. V. Robotham, P. J. Birrell, A. Gelman, K. D. Vihta, N. Bowers, I. Boreham, H. Thomas et al. Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. *The Lancet Public Health*, 6(1):e30–e38, 2021.

[198] C. A. Prete, L. Buss, A. Dighe, V. B. Porto, D. da Silva Candido, F. Ghilardi, O. G. Pybus, W. K. de Oliveira, J. H. Croda, E. C. Sabino et al. Serial interval distribution of SARS-CoV-2 infection in Brazil. *Journal of Travel Medicine*, 28(2):taaa115, 2021.

[199] Public Health England. "The First Few Hundred (FF100)" enhanced case and contact protocol v12. Technical report, 2020. URL: `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/360190/2012_13_FF100_Protocol_H7N9_ver_12.pdf`.

[200] Public Health England. SARS-CoV-2 variants of concern and variants under investigation in England (Technical briefing 7). Technical report, 2021. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/972247/Variants_of_Concern_VOC_Technical_Briefing_7_England.pdf.

[201] R. C. Reiner, R. M. Barber, J. K. Collins, P. Zheng, C. Adolph, J. Albright, C. M. Antony, A. Y. Aravkin, S. D. Bachmeier, B. Bang-Jensen et al. Modeling COVID-19 scenarios for the United States. *Nature Medicine*, 27(1):94–105, 2021.

[202] C. M. Rivers, E. T. Lofgren, M. Marathe, S. Eubank and B. L. Lewis. Modeling the impact of interventions on an epidemic of Ebola in Sierra Leone and Liberia. *PLoS Currents*, 6, 2014.

[203] M. Roberts, V. Andreasen, A. Lloyd and L. Pellis. Nine challenges for deterministic epidemic models. *Epidemics*, 10:49–53, 2015.

[204] M. M. Saito, N. Hirotsu, H. Hamada, M. Takei, K. Honda, T. Baba, T. Hasegawa and Y. Kitanishi. Reconstructing the household transmission of influenza in the suburbs of Tokyo based on clinical cases. *Theoretical Biology and Medical Modelling*, 18:7, 2021.

[205] R. Sender, Y. M. Bar-On, S. W. Park, E. Noor, J. Dushoff and R. Milo. The unmitigated profile of COVID-19 infectiousness. *medRxiv*:2021.11.17.21266051, 2021.

[206] M. Shen, Y. Xiao and L. Rong. Modeling the effect of comprehensive interventions on Ebola virus transmission. *Scientific Reports*, 5:15818, 2015.

[207] A. M. Smith and A. S. Perelson. Influenza A virus infection kinetics: quantitative data and models. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(4):429–445, 2011.

[208] T. Strutz. *Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond*. Vieweg+Teubner, Wiesbaden, 2011.

[209] K. Sun, W. Wang, L. Gao, Y. Wang, K. Luo, L. Ren, Z. Zhan, X. Chen, S. Zhao, Y. Huang et al. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. *Science*, 371(6526):eabe2424, 2021.

[210] Å. Svensson. A note on generation times in epidemic models. *Mathematical Biosciences*, 208(1):300–311, 2007.

[211] L. H. Taylor, S. M. Latham and M. E. Woolhouse. Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 356(1411):983–989, 2001.

[212] The World Bank. Liberia — Data. URL: https://data.worldbank.org/country/liberia.

[213] R. N. Thompson, C. A. Gilligan and N. J. Cunniffe. Will an outbreak exceed available resources for control? Estimating the risk from invading pathogens using practical definitions of a severe epidemic. *Journal of the Royal Society Interface*, 17:20200690, 2020.

[214] R. N. Thompson, J. E. Stockwin, R. D. van Gaalen, J. A. Polonsky, Z. N. Kamvar, P. A. Demarsh, E. Dahlqwist, S. Li, E. Miguel, T. Jombart et al. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics*, 29:100356, 2019.

[215] R. N. Thompson and W. S. Hart. Effect of confusing symptoms and infectiousness on forecasting and control of Ebola outbreaks. *Clinical Infectious Diseases*, 67(9):1472–1474, 2018.

[216] R. N. Thompson, K. Jalava and U. Obolski. Sustained transmission of Ebola in new locations: more likely than previously thought. *Lancet Infectious Diseases*, 19(10):1058–1059, 2019.

[217] R. N. Thompson, C. Wymant, R. A. Spriggs, J. Raghwani, C. Fraser and K. A. Lythgoe. Link between the numbers of particles and variants founding new HIV-1 infections depends on the timing of transmission. *Virus Evolution*, 5(1):vey038, 2019.

[218] R. N. Thompson. Epidemiological models are important tools for guiding COVID-19 interventions. *BMC Medicine*, 18:152, 2020.

[219] R. N. Thompson. Novel coronavirus outbreak in Wuhan, China, 2020: intense surveillance is vital for preventing sustained transmission in new locations. *Journal of Clinical Medicine*, 9(2):498, 2020.

[220] R. N. Thompson and E. Brooks-Pollock. Detection, forecasting and control of infectious disease epidemics: modelling outbreaks in humans, animals and plants. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1775):20190038, 2019.

[221] R. N. Thompson, C. A. Gilligan and N. J. Cunniffe. Detecting presymptomatic infection is necessary to forecast major epidemics in the earliest stages of infectious disease outbreaks. *PLoS Computational Biology*, 12(4):e1004836, 2016.

[222] R. N. Thompson, C. A. Gilligan and N. J. Cunniffe. Control fast or control smart: when should invading pathogens be controlled? *PLoS Computational Biology*, 14(2):e1006014, 2018.

[223] R. N. Thompson, E. M. Hill and J. R. Gog. SARS-CoV-2 incidence and vaccine escape. *The Lancet Infectious Diseases*, 21(7):913–914, 2021.

[224] R. N. Thompson, T. D. Hollingsworth, V. Isham, D. Arribas-Bel, B. Ashby, T. Britton, P. Challenor, L. H. K. Chappell, H. Clapham, N. J. Cunniffe et al. Key questions for modelling COVID-19 exit strategies. *Proceedings of the Royal Society B: Biological Sciences*, 287(1932):20201405, 2020.

[225] R. N. Thompson, O. W. Morgan and K. Jalava. Rigorous surveillance is necessary for high confidence in end-of-outbreak declarations for Ebola and other infectious diseases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1776):20180431, 2019.

[226] L. C. Tindale, J. E. Stockdale, M. Coombe, E. S. Garlock, W. Y. V. Lau, M. Saraswat, L. Zhang, D. Chen, J. Wallinga and C. Colijn. Evidence for transmission of COVID-19 prior to symptom onset. *eLife*, 9:e57149, 2020.

[227] S. Towers, O. Patterson-Lomba and C. Castillo-Chavez. Temporal variations in the effective reproduction number of the 2014 West Africa Ebola outbreak. *PLoS Currents*, 6, 2014.

[228] UK Government. NHS test and trace: how it works, 2020. URL: `https://www.gov.uk/guidance/nhs-test-and-trace-how-it-works`.

[229] UK Government. COVID-19: people with COVID-19 and their contacts, 2022. URL: `https://www.gov.uk/government/publications/covid-19-people-with-covid-19-and-their-contacts/covid-19-people-with-covid-19-and-their-contacts`.

[230] UK Government. Deaths in the UK — Coronavirus in the UK. URL: `https://coronavirus.data.gov.uk/details/deaths`.

[231] UK Health Security Agency. SARS-CoV-2 variants of concern and variants under investigation in England (Technical briefing 36). Technical report, 2022. URL: `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1056487/Technical-Briefing-36-22.02.22.pdf`.

[232] C. Vegvari, S. Abbott, F. Ball, E. Brooks-Pollock, R. Challen, B. S. Collyer, C. Dangerfield, J. R. Gog, K. M. Gostic, J. M. Heffernan et al. Commentary on the use of the reproduction number R during the COVID-19 pandemic. *Statistical Methods in Medical Research*:9622802211037079, 2021.

[233] C. Vegvari, C. Hadjichrysanthou, E. Cauët, E. Lawrence, A. Cori, F. De Wolf and R. M. Anderson. How can viral dynamics models inform endpoint measures in clinical trials of therapies for acute viral infections? *PLoS ONE*, 11(7):e0158237, 2016.

[234] H. S. Vöhringer, T. Sanderson, M. Sinnott, N. De Maio, T. Nguyen, R. Goater, F. Schwach, I. Harrison, J. Hellewell, C. V. Ariani et al. Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature*, 600:506–511, 2021.

[235] J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604, 2007.

[236] S. Wang, Y. Pan, Q. Wang, H. Miao, A. N. Brown and L. Rong. Modeling the viral dynamics of SARS-CoV-2 infection. *Mathematical Biosciences*, 328:108438, 2020.

[237] H. J. Wearing, P. Rohani and M. J. Keeling. Appropriate models for the management of infectious diseases. *PLoS Medicine*, 2(7):0621–0627, 2005.

[238] G. Webb, C. Browne, X. Huo, O. Seydi, M. Seydi and P. Magal. A model of the 2014 Ebola epidemic in West Africa with contact tracing. *PLoS Currents*, 7, 2015.

[239] J. S. Weitz and J. Dushoff. Modeling post-death transmission of Ebola: Challenges for inference and opportunities for control. *Scientific Reports*, 5:8751, 2015.

[240] C. R. Wells, J. P. Townsend, A. Pandey, S. M. Moghadas, G. Krieger, B. Singer, R. H. McDonald, M. C. Fitzpatrick and A. P. Galvani. Optimal COVID-19 quarantine and testing strategies. *Nature Communications*, 12(1):356, 2021.

[241] WHO Ebola Response Team. Ebola virus disease in West Africa - the first 9 months of the epidemic and forward projections. *The New England Journal of Medicine*, 371(16):1481–95, 2014.

[242] L. Willem, F. Verelst, J. Bilcke, N. Hens and P. Beutels. Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015). *BMC Infectious Diseases*, 17:612, 2017.

[243] World Health Organization. Ebola data and statistics, 2016. URL: `https://apps.who.int/gho/data/view.ebola-sitrep.ebo`.

[244] World Health Organization. Disease outbreak news (DONs) on the Ebola outbreak in Democratic Republic of the Congo, 2019. URL: `https://www.who.int/ebola/situation-reports/drc-20`.

[245] World Health Organization. Ebola Virus Disease - Democratic Republic of the Congo (External Situation Report 98). Technical report, 2020. URL: `https://apps.who.int/iris/bitstream/handle/10665/332654/SITREP_EVD_DRC_20200623-eng.pdf`.

[246] World Health Organization. Ebola virus disease, 2021. URL: `https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease`.

[247] Worldometer. DR Congo Population (2021). URL: `https://www.worldometers.info/world-population/democratic-republic-of-the-congo-population/`.

[248] W. Xia, J. Liao, C. Li, Y. Y. Li, X. Qian, X. Sun, H. Xu, G. Mahai, X. Zhao, L. Shi et al. Transmission of corona virus disease 2019 during the incubation period may lead to a quarantine loophole. *medRxiv*:2020.03.06.20031955, 2020.

[249] D. Yamin, S. Gertler, M. L. Ndeffo-Mbah, L. A. Skrip, M. Fallah, T. G. Nyenswah, F. L. Altice and A. P. Galvani. Effect of Ebola progression on transmission and control in Liberia. *Annals of Internal Medicine*, 162(1):11–17, 2015.

[250] J. Zhang, M. Litvinova, W. Wang, Y. Wang, X. Deng, X. X. Chen, M. Li, W. Zheng, L. Yi, X. X. Chen et al. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. *The Lancet Infectious Diseases*, 20(7):793–802, 2020.

[251] W. Zhang. Estimating the presymptomatic transmission of COVID19 using incubation period and serial interval data. *medRxiv*:2020.04.02.20051318, 2020.

[252] C. Zitzmann and L. Kaderali. Mathematical analysis of viral replication dynamics and antiviral treatment strategies: from basic models to age-based multi-scale modeling. *Frontiers in Microbiology*, 9:1546, 2018.