METHODS FOR MODELING THE SPREAD OF INFECTIOUS DISEASE

METHODS FOR MODELING THE SPREAD OF INFECTIOUS DISEASE

By MICHAEL LI, Hon.B.Sc.,M.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2019) Hamilton, Ontario (Biology)
TITLE: Methods for modeling the spread of infectious disease
AUTHOR: Michael Li, Hon.B.Sc. (University of Toronto), M.Sc. (McMaster University)
SUPERVISOR: Professor Benjamin M. Bolker & Professor Jonathan Dushoff
NUMBER OF PAGES: ix, Main Text 85, Appendix 106.

Lay Abstract

Mathematical and statistical models are widely used in studying infectious disease. Over the last couple of decades, modeling techniques have advanced tremendously due to improvements in computational power, data availability and data accessibility; this enables researchers to use various modeling approaches that capture more realistic aspects of infectious disease epidemics. My work focuses on exploring and improving methods for modeling the spread of infectious disease; in particular, in simulations of hypothetical emerging disease outbreaks and in real-life epidemic outbreaks of canine rabies. I used a high-quality data set from an ongoing rabies study in Africa to show that variation among dogs biases transmission calculations, and that the parameters underlying spread in canine rabies are more complicated and less well understood than previously thought. I also developed a method to improve modeling trait relationships while incorporating phylogenetic relationships.

Abstract

Mathematical and statistical models are widely used in studying infectious disease. They provide important insights – including mechanisms of the spread of infectious disease, forecast epidemic size and duration, and effects of intervention strategies – which are useful in studying and combating infectious disease. Over the last couple of decades, modeling techniques have advanced tremendously due to improvements in computational power, data availability, and data accessibility; this enables researchers to use various modeling approaches to capture more realistic aspects of infectious disease epidemics. Despite having flexible modeling techniques, these approaches use different modeling assumptions to incorporate information and propagate uncertainty, often arriving at inconsistent conclusions. My work focuses on exploring and improving methods for modeling the spread of infectious disease; in particular, exploring the state of the art techniques for disease modeling in real epidemic outbreaks and simulation settings.

Motivated by a synthetic forecasting challenge inspired by the 2014 West African Ebola outbreak, I compared simple Markov chain Monte Carlo approaches to simulated epidemics (Chapter 2). Using high-resolution data from an ongoing rabies contact-tracing study, I apply robust techniques to reassess global historical risk estimates of canine rabies (Chapter 3), and show that disease trait correlations bias generation time estimates, with implications for conclusions about control (Chapter 4). In Chapter 5, I developed a method to improve modeling trait relationships while incorporating phylogenetic relationships by reformulating phylogenetic mixed models to improve flexibility and speed.

Acknowledgements

First and foremost, I would like to give my deepest gratitude to my supervisors Ben Bolker and Jonathan Dushoff, for their guidance throughout the years of my graduate studies. Thank you for guiding me above and beyond my academic work, but most importantly how to be real scientist (and human) with an unselfish passion for science, how to accumulate experience and developing my own unique philosophy to view the work, and the never-ending mentality to help others and support others to advance science to create a better world for everyone. I thank Ben, for taking a huge risk and gamble six years ago for taking me as a student during my Masters in statistics, and since then, supported me every day to the end of my Ph.D. We both took a huge gamble when you converted me into the "dark side" – from theoretical statistics to applied science in the field of biology – with an unimaginable bright future. I never would have imagined my life would turn out this way, and I sincerely thank you for offering me this opportunity to take a huge step in a different direction out of my comfort zone. I thank Jonathan, for showing and teaching me how to be "the human" – someone who faces challenges head-on regardless if the odds are against them, someone who is passionate about science and work towards improving the world, someone who is always ready to help others, and someone loves to have fun. You the Human JD.

I would also like to thank my supervisory committee member Mark Loeb for giving helpful comments at every committee meeting and recommending me books to catch up the basics in epidemiology.

Next, I would like to thank my collaborators Katie Hampson and David Earn for inviting me to collaborate on phenomenal projects. I am very grateful to Katie for the opportunity to be part of the rabies research team and for the endless support, resources, feedback, and networking with other rabies researchers around the world.

I thank David for the opportunity to collaborate in the RDC project; it is every data-scientist dream to work with a phenomenal database like the one you work so hard to get. Moreover, David's presentation skills are extraordinary sharp, clear and very enjoyable to watch; a skill I try really hard to accumulate over the years watching his presentations.

In addition, I would like to thank past and present members of MacTheoBio Lab: Guillaume Blanchet, David Champredon, Steve Cygu, Edgar González, Morgan Kain, Lindsay Keegan, Daniel Park, and Steve Walker. I also thank Morgan Kain and Jo Werba for always being there giving feedback and answering my silly questions every day. And of course, a special thanks to Chyun Shi, for keeping a friendly environment and organizing events in the lab.

A sincere thank you to my family and dearest friends outside of academia for providing me with endless support. I thank my parents for all the sacrifices they've made over the years for taking care of the family and supporting me to pursue my career. Over the years as I pursue deeply in my research, people come but mostly go, and for that, I thank these individuals who remain and believe I will succeed in my academic career: Andy Liang, Vincent Liang, Kenneth Peng, James Sau, and Jikwon Wang. In addition, a special thanks to Mariam Al-Musa, who believes I can be a professor one day; and Yanling Jin, who I promised to work hard to use my skills to help humans for the rest of my life.

Last but not least, to my late grandfather, who unfortunately didn't stay in this world long enough to see his grandson become a doctor.

Contents

List of Figures and Tables

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Abbreviations and Symbols

GI Generation Interval

SI Serial interval

r Initial growth rate

$\mathcal{R}_e$ Effective reproductive number

$\mathcal{R}_0$ Basic reproductive number

GLM Generalized Linear Model

GLMM Generalized Linear Mixed Effects Model

GLS Generalized Least Squares

HMC Hamiltonian Monte Carlo

LMIC Low- and Middle- Income Country

MCMC Markov Chain Monte Carlo

RABV Rabies virus

SIR Susceptible-Infected-Removed compartmental model

Declaration of Academic Achievement

This sandwich thesis contains an introduction (Chapter 1), one published paper (Chapter 2), two drafts of manuscripts in preparation for publication (Chapter 3 and Chapter 4), one draft ready for submission (Chapter 5) and a conclusion (Chapter 6). Preambles to Chapters 2-5 describe the authors' contributions. Chapters 1 and 6 were written by me.

Chapter 1: Introduction

Thesis overview

In this thesis, I show a series of works exploring and developing methods for modeling the spread of infectious disease and phylogenetic mixed models. In Chapter 2, I present a simulation study that compares different modern Bayesian Markov Chain Monte Carlo (MCMC) modeling approaches to the early stages of epidemic outbreaks where data is limited. In Chapter 3 and 4, I focus on canine rabies, a feared disease which causes an estimated 50,000 human deaths each year. Using historic outbreak case-incidence time series data from around the world and high-resolution contact-tracing data, I used newly developed modeling and estimation techniques to reassess epidemiological parameters and the risk of these historical outbreaks (Chapter 3). In Chapter 4, I focus on exploring the details and issues in estimating generation intervals. I carefully define these intervals, and their component parts, and construct equations that clarify the source of differences between generation intervals and the related serial intervals. Then I use rabies as a case study to explore the differences between constructed generation intervals and observed generation and serial intervals. Lastly, the work I present in chapter 5 is a general statistical framework in modeling traits while incorporating phylogenetic correlations. This work can be integrated into disease trait models with multiple species.

In addition to the results presented, I also focus on the importance of methodological validation and reproducibility. All methods in this thesis are validated with simulations, and all code is available, with reproducible examples, in public GitHub repositories.

Mathematical and statistical models

Mathematical and statistical models are widely used in studying infectious disease. They provide important insights – including evaluating mechanisms of spread, forecasting epidemics, and predicting effects of intervention strategies.

The use of mathematical models to study infectious diseases dates back at least to the 18th century when Daniel Bernoulli used mathematical analysis to encourage universal inoculation against smallpox (Bernoulli and Blower, 2004; Dietz and Heesterbeek, 2002). In the early 1900s, Ronald Ross developed influential modeling ideas while studying malaria (he was also the first to show that mosquitoes transmitted malaria) (Cox, 2010; Smith et al., 2012). In particular, Ross noted that a disease can be eliminated from a population as long as each case causes on average less than one new case, and used this idea to argue that malaria could be effectively controlled by reducing mosquito density.

Over the last few decades, increases in power and availability of computers have led to new developments in epidemiological modeling. Specifically, modern methods allow the application of statistical approaches to mathematical models which account for dynamics of disease spread. This progress is facilitated by improvements in the quality and availability of epidemiological data, particularly disease-incidence data, but also demographic, geographical, environmental, and even biological sequence data.

Reproductive numbers

Ross' insight about epidemic spread has become the foundation of a critical concept in mathematical epidemiology. The basic reproductive number $\mathcal{R}_0$, is defined as the expected number of new cases per cases in a fully susceptible population (Macdonald, 1952). $\mathcal{R}_0$ is often used to guide disease control: disease-control programs based on the idea that if $\mathcal{R}_0$ can be estimated, and transmission reduced by a factor of $\mathcal{R}_0$, the disease can be eliminated. For example, for a disease with effective vaccines, if less

than $1/\mathcal{R}_0$ of a population remains unvaccinated, the population is expected to be protected by 'herd immunity' (shared protection). Thus, the statistical estimation of $\mathcal{R}_0$ is often of interest.

One common way to calculate $\mathcal{R}_0$ is using two other quantities that are often easier to measure empirically: the rate of spread and the generation interval (Wallinga and Lipsitch, 2006). The rate of spread describes the speed the disease is spreading at the population level and is inferred primarily from case-incidence reports (Park et al., 2019). The generation interval is the time in between one individual getting infected and infecting another individual. The distribution of the generation interval describes how fast the disease spreads at the individual level and is typically inferred from contact tracing or estimated if data are limited.

Another common way to estimate $\mathcal{R}_0$ is based on the idea that diseases reach equilibrium when there is one case per case. If per-individual transmission (i.e., the number of cases per case) is controlled by the proportion of susceptibles in a population, then $\mathcal{R}_0$ can be estimated from the estimated proportion susceptible at equilibrium.

Rabies

Rabies has been feared throughout human history. It is highly virulent, with virtually every clinical case ending in death, and poses a high mortality burden, with an estimated 50,000 annual deaths in humans (Knobel et al., 2005). Rabies viruses (RABVs) co-circulate among a wide range of mammalian hosts (Bourhy et al., 2008); however, most rabies viruses are found in canine-associated clades and particularly in domestic dogs (*C. domesticus*), are responsible for more than 99% of human rabies deaths (Knobel et al., 2005). While rabies has been eliminated from domestic dog populations in Western Europe and North America, it remains a huge problem in many low- and middle-income countries (LMICs), primarily in Asia and Africa

(Cleaveland and Hampson, 2017).

Canine rabies can be effectively controlled by vaccinating domestic dog populations; high-income countries have used this approach to eliminate human deaths from dog-mediated rabies. A few LMICs in Africa and Asia have implemented mass dog vaccination at scale; however, they are not as effective and more challenging than expected. United Against Rabies recently launched a campaign to eliminate rabies by 2030 by partnering with World Health Organization (WHO), the Food and Agriculture Organization of the United Nations (FAO), the World Organisation for Animal Health (OIE) and the Global Alliance for Rabies Control (GARC); this is an exciting time to study rabies.

Chapter summaries

Chapter 2. Inspired by recent Ebola forecast challenge (Viboud et al., 2018), I explored the performance and limitations of different Bayesian Markov Chain Monte Carlo (MCMC) modeling approaches to estimating disease parameters and forecasting epidemics. I developed relatively simple MCMC approaches that were able to incorporate stochasticity in both transmission and observation, and applied them to data from simulated epidemics. The simulation design in this study mimics the early stages of emerging disease outbreak where data are limited. I learned two things in this study: first, modeling different processes with dispersion is a naive but effective way to add uncertainty in the model; and, approximating discrete latent state process with continuous processes can aid efficiency without losing robustness of fit.

Chapter 3. In an earlier, influential paper, estimates of $\mathcal{R}_0$ based on historical outbreaks of rabies have surprisingly low, typically between 1 and 2 (Hampson et al., 2009). I further investigated and re-assessed why these rabies $\mathcal{R}_0$ estimates have generally been low. I used a logistic model to estimate the initial growth rate r

in a Bayesian framework using Hamiltonian Monte Carlo and empirical generation intervals from contact tracing data. I developed a hybrid approach where I was able to incorporate uncertainties in both r and generation interval when estimating $\mathcal{R}_0$. $\mathcal{R}_0$ estimates using the hybrid approach are larger with wider confidence intervals than previously estimated. The results suggest more efforts are needed to control rabies and rabies is not as well known as previously thought.

Chapter 4. Following up on Chapter 3, I explored the difference between constructed estimated generation intervals (Hampson et al., 2009), realized generation intervals and serial intervals from contact tracing data. I used a simple generalized linear mixed model approach to model dogs' biting behaviour and time distributions and compared incubation periods simulated from the predictive model and random, independent resampling from the empirical data. The results showed that incubation periods are positively correlated with biting behaviour in rabies. I learned that neglecting correlations between time distributions and biting behaviour can bias generation interval estimates.

Chapter 5. Inspired by the inflexible and computationally demanding approaches to fit phylogenetic mixed models; I explored the performance and limitations of existing R packages. I developed an alternative approach in parametrizing the phylogenetic tree covariance matrix (similarity between species) with a species-branching matrix (branches shared between species) and implemented in R package `lme4` and `glmmTMB`. I compared the new approach with existing R packages that fit phylogenetic mixed models with simulated phylogenetic trees. The new approach is much more flexible in fitting phylogenetic random effects, magnitudes faster, and able to fit large volumes of data. This improvement offers researchers a flexible way to fit multi-species trait models.

References

Bernoulli, D. and S. Blower 2004. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. *Reviews in medical virology 14*(5), 275–288.

Bourhy, H., J.-M. Reynes, E. J. Dunham, L. Dacheux, F. Larrous, V. T. Q. Huong, G. Xu, J. Yan, M. E. G. Miranda, and E. C. Holmes 2008. The origin and phylogeography of dog rabies virus. *The Journal of general virology 89*(Pt 11), 2673.

Cleaveland, S. and K. Hampson 2017. Rabies elimination research: juxtaposing optimism, pragmatism and realism. *Proceedings of the Royal Society B: Biological Sciences 284*(1869), 20171880.

Cox, F. E. 2010. History of the discovery of the malaria parasites and their vectors. *Parasites & vectors 3*(1), 5.

Dietz, K. and J. Heesterbeek 2002. Daniel bernoulli's epidemiological model revisited. *Mathematical biosciences 180*(1-2), 1–21.

Hampson, K., J. Dushoff, S. Cleaveland, D. T. Haydon, M. Kaare, C. Packer, and A. Dobson 2009. Transmission dynamics and prospects for the elimination of canine rabies. *PLoS biology 7*(3), e1000053.

Knobel, D. L., S. Cleaveland, P. G. Coleman, E. M. Fèvre, M. I. Meltzer, M. E. G. Miranda, A. Shaw, J. Zinsstag, and F.-X. Meslin 2005. Re-evaluating the burden of rabies in africa and asia. *Bulletin of the World health Organization 83*, 360–368.

Macdonald, G. 1952. The analysis of equilibrium in malaria. *Tropical diseases bulletin 49*(9), 813.

Park, S. W., D. Champredon, J. S. Weitz, and J. Dushoff 2019. A practical generation-interval-based approach to inferring the strength of epidemics from their speed. *Epidemics 27*, 12–18.

Smith, D. L., K. E. Battle, S. I. Hay, C. M. Barker, T. W. Scott, and F. E. McKenzie 2012. Ross, macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. *PLoS pathogens 8*(4), e1002588.

Viboud, C., K. Sun, R. Gaffey, M. Ajelli, L. Fumanelli, S. Merler, Q. Zhang, G. Chowell, L. Simonsen, A. Vespignani, et al. 2018. The rapidd ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics 22*, 13–21.

Wallinga, J. and M. Lipsitch 2006. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences 274*(1609), 599–604.

Chapter 2: Fitting mechanistic epidemic models to data: A comparison of simple Markov Chain Monte Carlo approaches

In this chapter, I developed a simulation study comparing different Bayesian Markov Chain Monte Carlo modeling approaches fitting to simulated epidemics data. The simulation design mimics the data-stream at early stages of emerging disease outbreak where data are limited. The simulation model is a discrete-time SIR model that incorporates stochasticity and additional sources of variation in the form of overdispersion in both transmission and observation processes. I compared model approaches of varying complexity as well as different MCMC platforms. This study illustrates the importance of propagating uncertainty in at least at one level of and ways to aid efficiency without losing robustness of fit.

Author Contributions

ML performed all simulations and statistical analyses with helpful feedback from BMB and JD; ML wrote the first draft of the manuscript, and all authors revised the manuscript.

Acknowledgements

I thank the Ebola challenge organizers for organizing the Ebola model challenge that inspired this project.

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2018, Vol. 27(7) 1956–1967

DOI: 10.1177/0962280217747054
journals.sagepub.com/home/smm
SAGE

Fitting mechanistic epidemic models to data: A comparison of simple Markov chain Monte Carlo approaches

Michael Li,[1] Jonathan Dushoff[1,2,3] and Benjamin M Bolker[1,2,3]

Abstract

Simple mechanistic epidemic models are widely used for forecasting and parameter estimation of infectious diseases based on noisy case reporting data. Despite the widespread application of models to emerging infectious diseases, we know little about the comparative performance of standard computational-statistical frameworks in these contexts. Here we build a simple stochastic, discrete-time, discrete-state epidemic model with both process and observation error and use it to characterize the effectiveness of different flavours of Bayesian Markov chain Monte Carlo (MCMC) techniques. We use fits to simulated data, where parameters (and future behaviour) are known, to explore the limitations of different platforms and quantify parameter estimation accuracy, forecasting accuracy, and computational efficiency across combinations of modeling decisions (e.g. discrete vs. continuous latent states, levels of stochasticity) and computational platforms (JAGS, NIMBLE, Stan).

Keywords

Markov chain Monte Carlo, Hamiltonian Monte Carlo, discrete-time susceptible-infectious-removed model, dispersion, moment-matching

1 Introduction

Simple homogeneous population models have been widely used to study emerging infectious disease outbreaks. Although such models can provide important insights – including estimated epidemic sizes and predicted effects of intervention strategies, as well as short-term forecasts – they neglect important spatial, individual-level and other heterogeneities. Decades of work have created frameworks that enable researchers to construct models that capture many of these more realistic aspects of infectious disease epidemics. But many challenges remain. In particular, estimating parameters (and associated uncertainties) is always challenging, especially for models incorporating multiple forms of heterogeneity, and especially during the early stages of an epidemic when data are limited. Using complex models that are insufficiently supported by data can lead to imprecise and unstable parameter estimates[1] – in such cases, researchers often revert to simpler models for practical purposes.

In the past few decades, researchers have begun to adopt Bayesian approaches to disease modeling problems. Bayesian Markov Chain Monte Carlo (MCMC) is a powerful, widely used sampling-based estimation approach. Despite the widespread use of MCMC in epidemic modeling,[2,3] however, there have been relatively few systematic studies of the comparative performance of statistical frameworks for disease modeling.[4]

In this paper, we apply relatively simple MCMC approaches to data from simulated epidemics that incorporate stochasticity in both transmission and observation, as well as variable generation-interval distributions (not assumed to be known when fitting). We compare model approaches of varying complexity, including an estimation model that matches the simulation model. For each model, we quantify parameter estimation accuracy and forecasting accuracy; this sheds light on which phenomena are most important to include in models to be used for estimation and forecasting.

[1]Department of Biology, McMaster University, Hamilton, Ontario, Canada
[2]Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
[3]Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada

Corresponding author:
Michael Li, Department of Biology, McMaster University, Hamilton, Ontario, Canada.
Email: lim88@mcmaster.ca

We also compare three different MCMC platforms: JAGS,[5] NIMBLE[6] and Stan.[7] In principle, for any given model, any valid method of MCMC sampling should eventually converge on the same (correct) posterior distribution. However, even with the relatively simple models considered here, a theoretically valid software package can experience problems in practice: we wanted to investigate this phenomenon. Furthermore, even when different platforms converge to essentially the same result, they may show large differences in computational efficiency: we therefore also quantify efficiency for the models we study.

2 Methods

We generated test data using a simple framework that combines a *transmission process* based on a simple discrete-time model with an *observation process* to account for incomplete reporting. Both processes are assumed to be stochastic. We then fit the observed cases from these simulations using Bayesian methods that model the underlying true number of infections as a latent (i.e. unobserved) variable. Our Bayesian fitting models explore an approach that matches the assumptions of the simulation model, as well as various simplifications: in particular, we explore simpler methods of accounting for variation in both the transmission process and the observation process, and the use of continuous rather than discrete latent variables. For simplicity, we have here assumed that data are reported on the same discrete time scale on which the disease process is simulated (but not that the reporting period is the same as the generation time of the disease; see below). This assumption requires that the generation time be at least as long as the reporting period. It would be relatively straightforward to relax this assumption, for example by assuming that the epidemic dynamics occur on a finer time scale than the reporting interval, or by simulating in continuous time but fitting with a discrete-time model; we do not explore these questions here.

2.1 Simulation model

The transmission process of our dual-process framework is based on the Reed-Frost chain binomial model, which can also be described as a discrete-time, stochastic compartmental SIR model.[8] To account for the possibility that some fraction of the population may be beyond the scope of the epidemic – geographically or socially isolated, genetically resistant, vaccinated or immune due to previous exposure – we assume that only a proportion P_{eff} of the total census population is actually susceptible to infection. We further assume that, in every time step, only a proportion (randomly chosen with mean P_{rep}) of new infections are actually observed. We model both transmission and observation using a beta-binomial (rather than binomial) distribution to account for additional sources of variation (i.e. overdispersion) in both processes. The equations are

$$N_{\text{eff}} = P_{\text{eff}} N \tag{1}$$

$$S_1 = N_{\text{eff}} - I_1 \tag{2}$$

$$\Phi_t = \sum_{i=1}^{\ell} k(i) I_{t-\ell+i} \tag{3}$$

$$I_{t+1} \sim \text{BetaBin}\,(1 - e^{-\Phi_t}, S_t, \delta_P) \tag{4}$$

$$S_{t+1} = S_t - I_{t+1} \tag{5}$$

$$\text{Obs}_t \sim \text{BetaBin}\,(P_{\text{rep}}, I_t, \delta_{\text{obs}}) \tag{6}$$

where Φ_t is the force of infection at time t; N_{eff} is the effective population size; and ℓ is the number of lags.

The most common parameterization of the beta-binomial comprises three parameters: the binomial size parameter N plus two additional shape parameters (α and β) that describe the Beta distribution of the per-trial probability. Uses of the beta-binomial in statistical modeling instead typically transform the shape parameters into a pair of parameters that describe the per-trial probability and a dispersion parameter[9]; larger values of the dispersion parameter δ correspond to less variability. We use a slight modification of this parameterization (see Figure 1).

We extend the Reed-Frost model by allowing the infectious period to last longer than one step, and the infectivity to vary based on how long an individual has been infected; we do this by parameterizing a transmission kernel that describes the force of infection coming from individuals who were infected ℓ time steps ago. For convenience,

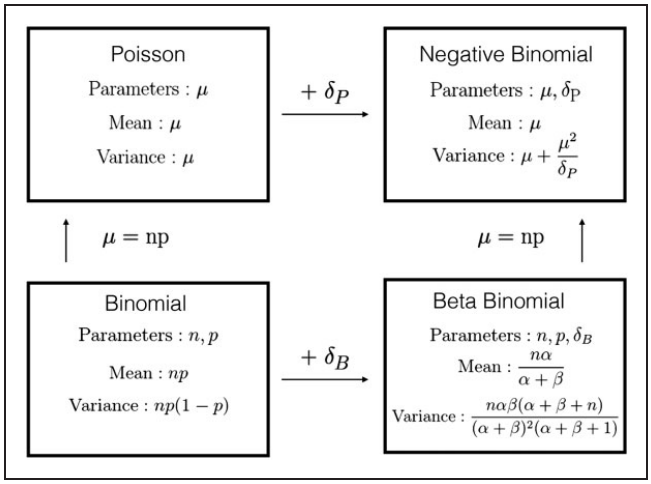

Figure 1. Discrete distribution relationships. For beta-binomial distribution (bottom right panel), we used an alternative parameterization α and β, where $\alpha = \frac{\delta_B}{1-p}$ and $\beta = \frac{\delta_B}{p}$. Moving from the top to bottom row adds a size parameter (replacing μ with np). Moving from left to right adds a dispersion parameter δ_P and δ_B for Poisson and Binomial distribution, respectively.

we assumed a fixed maximum window length ($\ell = 5$). We then based our transmission kernel on a negative binomial distribution, truncated to fit this window:

$$\tilde{k}(i) = i^{(G_S-1)} \times \exp\left(\frac{-i}{G_P \times \ell}\right), \quad i = 1, \ldots, \ell \tag{7}$$

$$k(i) = \frac{\mathcal{R}_0}{N_{\text{eff}}} \times \frac{\tilde{k}(i)}{\sum_{i=1}^{\ell} \tilde{k}(i)}, \quad i = 1, \ldots, \ell \tag{8}$$

Here, $\mathcal{R}_0$ represents the basic reproductive number and G_S and G_P are shape and position parameters, respectively.

2.2 Fitting model

2.2.1 Transmission and observational process errors

The transmission (equation(4)) and observation (equation (6)) processes in the simulation model are both defined as beta-binomial (BB) processes. In fitting, we used the BB to match the simulation model, but also tried several simpler alternatives: binomial (B), Poisson (P), and negative-binomial (NB) processes. Process B does not allow for overdispersion, while NB does not incorporate the size of the pool from which a value is chosen; that is, it is theoretically possible for a NB sample of the number of infections to be larger than the current susceptible population (although this is extremely unlikely when the *per capita* infection probability is small). Process P neglects both of these phenomena. Figure 1 illustrates the relationship of the four discrete distributions.

2.2.2 Multiple scale decorrelation

The proportion of the population assumed to be effectively susceptible (P_{eff}) and the reporting proportion (P_{rep}) has very similar effects on observed incidence. We therefore reparameterized the model so that it uses a single parameter P_{effrep} for their product, and a second to govern how the product is apportioned between the two quantities:

$$\hat{P}_{\text{eff}} = P_{\text{effrep}}^{1-\rho} \tag{9}$$

$$\hat{P}_{\text{rep}} = P_{\text{effrep}}^{\rho} \tag{10}$$

We expected a priori that this parameterization would improve statistical convergence, since it makes it possible to sample different values of the poorly constrained value of ρ without changing P_{effrep}. It is straightforward to

back-calculate P_{eff} and P_{rep} once the model is fitted. For similar reasons, we experimented with measuring infected individuals on a ''reporting'' scale in our continuous-variable models (see below).

2.2.3 *Continuous latent variables*

Another simplification we considered was treating the unobserved number of underlying cases as a continuous variable. To do this, we matched the first two moments of the discrete distribution to a Gamma distribution (Figure 2).

Equations (4) and (6) can be rewritten as

$$\hat{I}_{t+1} \sim \text{Gamma}\left(a, \frac{r}{P_{\text{rep}}}\right) \tag{11}$$

$$\text{Obs}_t \sim \text{NB}(\widehat{I}_t, \delta_{\text{obs}}) \tag{12}$$

One advantage of this continuous approximation approach is that it allows us to scale our latent variable to help with model convergence, so that infected individuals are measured on the reporting scale. Another advantage is that it allows us to use Hamiltonian Monte Carlo (HMC), which cannot easily use discrete latent variables.

2.3 Bayesian Markov Chain Monte Carlo

In Bayesian MCMC, model parameters are sampled from the posterior distribution by a reversible Markov chain whose stationary distribution is the target posterior distribution. Classical MCMC techniques include the Metropolis-Hasting algorithm,[10] Gibbs sampling,[11] and slice sampling.[12] Recently, convenient implementations of a powerful MCMC technique called Hamiltonian Monte Carlo (HMC: also called hybrid MC)[13] have become available. HMC uses the concept of Hamiltonian dynamics to create a proposal distribution for the M-H algorithm, together with the leap-frog algorithm and the No U-Turn sampler.[14] HMC requires more computational effort per sample step compared to other MCMC techniques, but because subsequent steps are less strongly correlated it also produces more effective samples per sample step.[7,14]

2.3.1 *Platforms*

Many software platforms implement the automatic construction of MCMC samplers for user-defined models. One of the most widely used platforms is JAGS (Just Another Gibbs Sampler); despite its name, it implements a variety of MCMC techniques to fit models. NIMBLE (Numerical Inference for Statistical Models for Bayesian and Likelihood Estimation) is a more recent platform that allows users to flexibly model and customize different algorithms and sampling techniques for MCMC. Neither JAGS nor NIMBLE has yet implemented HMC. One of the relatively few platforms that currently implements HMC is Stan, which provides full Bayesian inference for continuous-variable models based on the No-U-Turn sampler, an adaptive form of HMC.

Continuous Approximation (Hybridization) Gamma(shape $= a$, rate $= r$)	
Poisson $a = \mu$ $r = 1$	Negative Binomial $a = \mu r$ $r = \frac{\delta_P}{\delta_P + \mu}$
Binomial $a = npr$ $r = \frac{1}{1-p}$	Beta Binomial $a = npr$ $r = \frac{\delta_B + p(1-p)}{(1-p)(\delta_B + np(1-p))}$

Figure 2. Continuous approximation of discrete distributions via moment matching. Distributions in Figure 1 were matched to a Gamma distribution with equivalent first and second moments.

2.3.2 Simulation and evaluations

We evaluated our estimates of (1) total cases predicted over the forecast window (disaggregated forecasts are analyzed in the supplementary material) and (2) key model parameters, including the estimated mean generation interval $\left(\text{MGI : defined as } \frac{\sum_{i=1}^{\ell} i\hat{k}(i)}{\sum_{i=1}^{\ell} \hat{k}(i)}\right)$. We used bias, root mean square error (RMSE), and coverage to assess model fit. Bias and RMSE are based on proportional errors, defined as the log ratio of our estimate (taken as the median of the posterior sample) to the known true value from our simulations. Errors were compared on the log scale in order to allow comparison of the accuracy of estimation of different parameters that may be on very different scales. The median is a scale-invariant, robust summary statistic for the location parameter of a Bayesian posterior distribution.[15] Thus in order to compare different parameters in a consistent, unitless fashion, the errors were calculated as $\epsilon_i = \log(\text{med}(\widehat{\theta_i})/\theta_i)$. We then calculated bias (median(ϵ)) and RMSE ($\sqrt{\text{mean}(\epsilon_i^2)}$).

Coverage refers to the frequency with which the computed confidence intervals include the true values of parameters or simulated quantities such as the forecast number of cases. We used 90% quantile-based intervals to evaluate coverage (i.e. a range from the 0.05 to the 0.95 quantile of the sampled posterior distributions).[16]

Evaluating the coverage of Bayesian model estimates based on simulated parameters runs the risk of confounding two questions: How well does the modeling implementation work? and How appropriate are the prior distributions for the particular question? In particular, when tested parameters are from regions with high prior density, coverage is biased upwards (i.e. it will be higher than the nominal value when the method is working properly) – particularly problematic is that this bias may make fits look good when in fact they are under-covering. This scenario can easily occur if we follow the standard frequentist simulation scheme of simulating all epidemic realizations with the same fixed set of parameters, then choose Bayesian priors that are centered on or near the fixed parameters. One potential solution is to use only uninformative priors (so that the simulation parameter values do not have high prior density); this was both impractical, because completely uninformative priors led to numerical instability in our fitting procedures, and unrealistic, because it is likely that researchers would use informative priors in a real epidemic-fitting exercise.

As an alternative way to resolve this situation, we implemented an established Bayesian validation protocol[16] where we: draw parameters from our assumed prior distribution; generate data using the drawn parameters; and fit the Bayesian model with the same prior distributions. This scheme matches the assumptions of our model, and is therefore a fair way to evaluate how well the implementation works. We sampled 100 sets of the parameters from the same prior distribution that was used in the fitting process; for each parameter set, we simulated one realization of 15 time steps (10 for fitting and 5 to compare to forecasts). All model variants were used to fit each realization (Tables 1 and 2 in the online appendix give more detail about parameters and priors). We combined two convergence criteria to assess convergence for the main parameters ($\mathcal{R}_0$, P_{eff}, P_{rep}): we required a value of the Gelman and Rubin statistic $\hat{R} < 1.1$ and an effective sample size (ESS) greater than 400 for each replication. For each replication we sample four chains starting with 4000 iterations; we repeatedly double the number of iterations (with a upper threshold of one million iterations) until the convergence criteria are met. Forecasts were made by simulating incidence five time steps forward using parameters sampled from the fitted posterior distributions.

3 Results

The full model (which matches the simulation model) provides generally good forecasts and parameter estimates as assessed by bias (Figure 3) or RMSE (Figure 4), except for estimates of P_{eff} using JAGS.

In general, models with any kind of dispersion in the transmission process, or with negative binomial dispersion in the observation process, did well. The exception is that models that combined negative binomial transmission dispersal with beta binomial observation dispersal produced biased forecasts and estimates of $P_{\text{rep.}}$

There are no clear differences in the quality of model fit due to multi-scale decorrelation, latent continuous transmission process or platform.

Figure 5 shows the statistical coverage of our estimates. Similar to the results for bias and RMSE (Figures 3 and 4), we find generally good coverage (i.e. close to the nominal value of 0.9) for models with dispersion in the transmission process, except that the negative-binomial transmission process model undercovers across the board (coverage $\approx$0.8 for all observation process models and platforms) for forecasts and $P_{\text{rep.}}$ For models without dispersion in transmission, models with dispersion in the observation process have low coverage ($\approx$0.8) for most parameters, while the beta-binomial process model has low coverage ($\approx$0.4) for P_{rep} and models without any dispersion have uniformly low coverage.

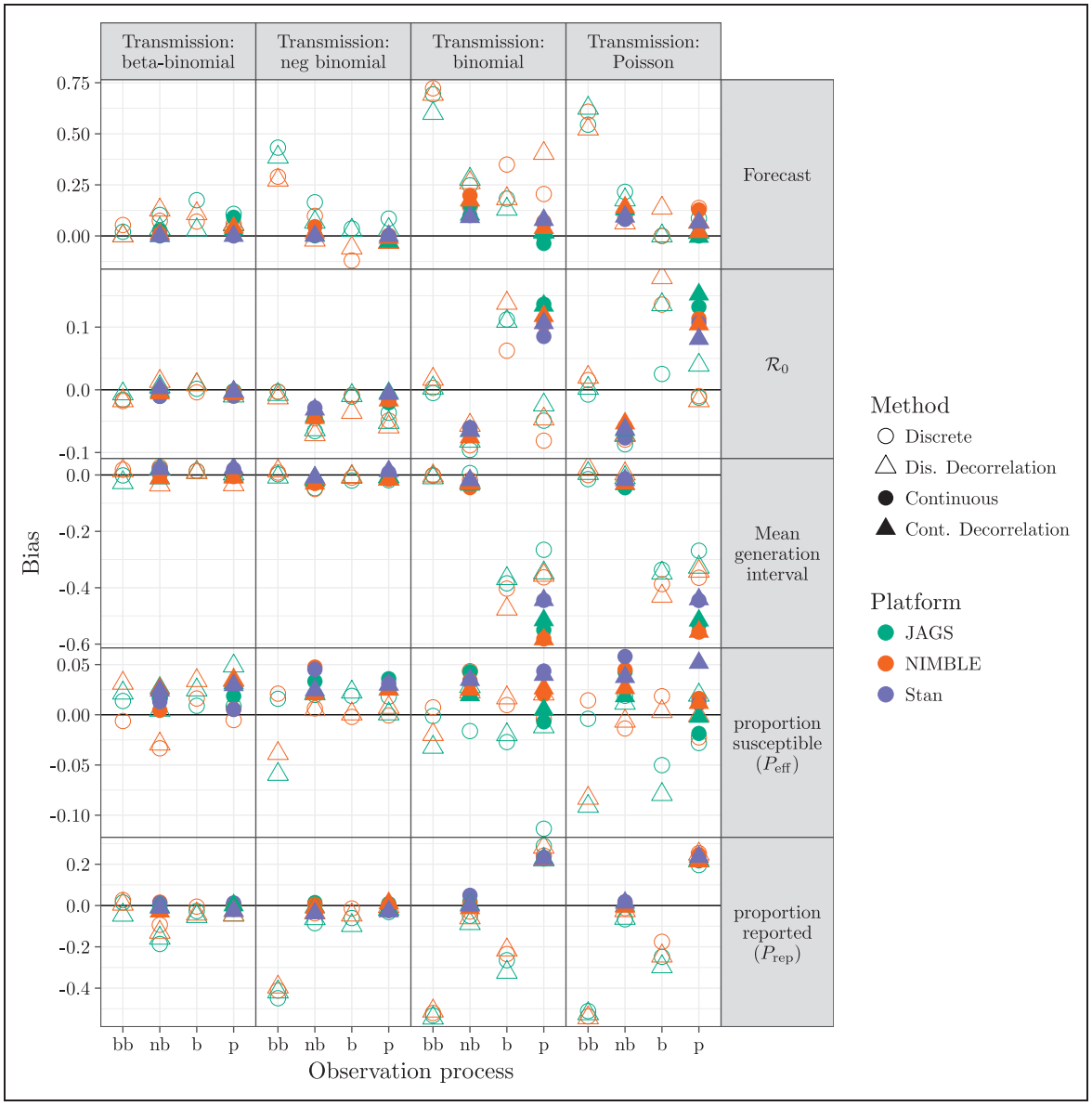

Figure 3. Comparison of bias (based on proportional errors) for forecasts and parameters using models described in section 2.2 across different platforms described in section 2.3.1. Models with overdispersion in the transmission process (BB and NB, leftmost and second-left columns of panels) and models with overdispersion in the observation process (BB and NB, leftmost and second-left x-axis ticks within each panel) have generally low bias. Continuous latent-state models (solid points) are only implemented for negative binomial and Poisson observational processes.

There are substantial efficiency differences between transmission-process approaches (continuous vs. discrete), as measured by time per effective sample size, shown in Figure 6. For a given platform, models using continuous latent variables are generally more efficient than discrete latent processes. Comparing models with continuous latent variables between platforms (Figure 5, second and fourth column of every panel), Stan (using HMC) is sightly more efficient for majority of the parameters, followed by NIMBLE and JAGS. Furthermore, continuous latent-variable models (especially using HMC in STAN) use fewer iterations (when meeting all convergence criteria described in section 2.3.2) than discrete latent-variable models.

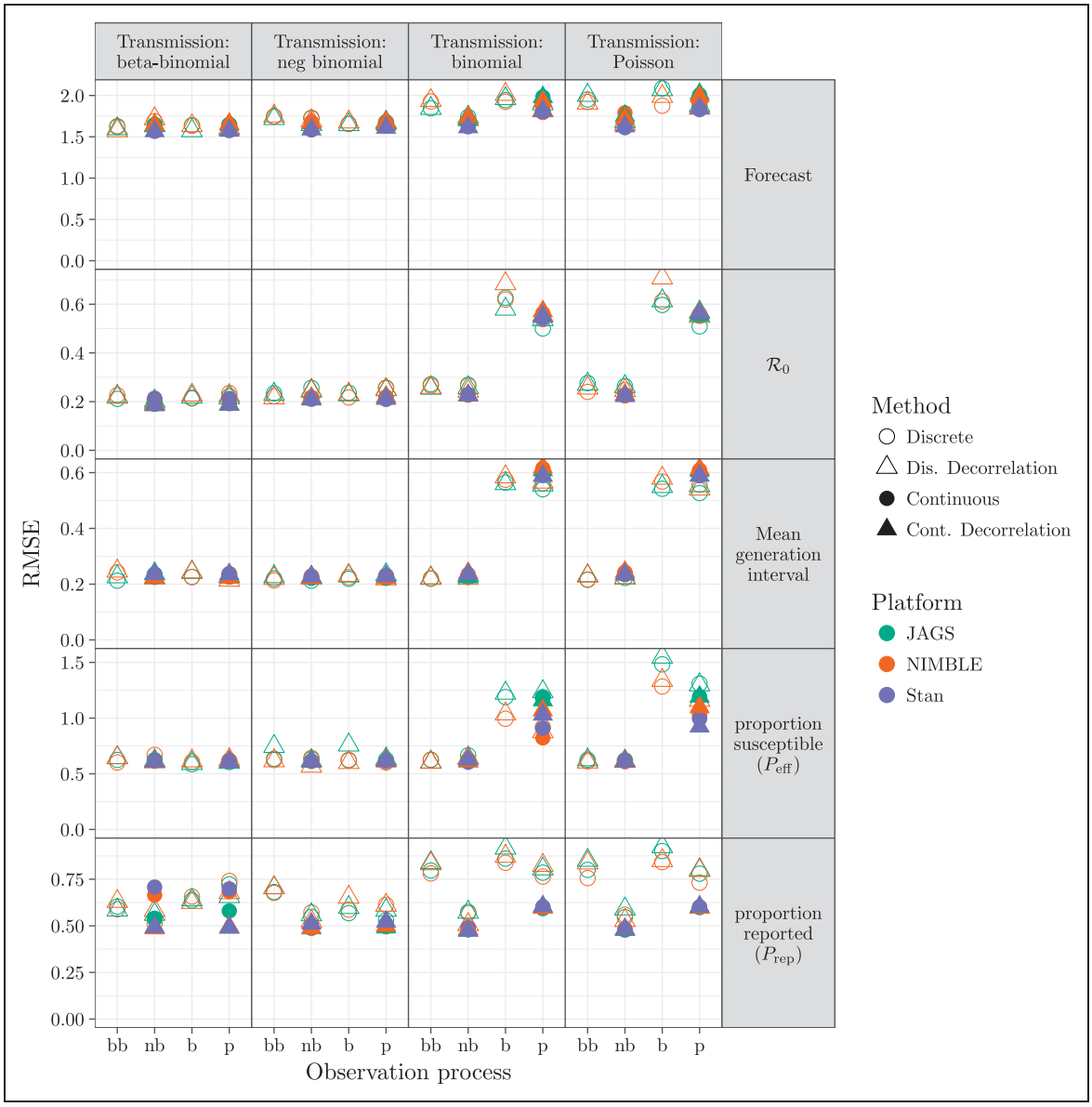

Figure 4. Comparison of RMSE (based on proportional errors) for all fitting model variants. The layout matches that of Figure 3. Patterns across models and platforms are similar to those seen in Figure 3. Short-term forecasts have generally high error, even when bias is low, reflecting inherent uncertainty in the system. The highly correlated parameters P_{eff} and P_{rep} also show high error but not high bias.

4 Discussion

We have fitted models varying in complexity to simulated epidemic data with multiple sources of heterogeneity, using several different platforms. Using models that include some form of overdispersion is necessary for robust fits, but models that include overdispersion only in the transmission process can work as well as or better than the full model. Including overdispersion only in the observation process (if implemented as a negative binomial distribution) also provides relatively robust fits to these data. Simplifying the models by using continuous rather than discrete latent variables increased efficiency with little effect on the quality of the fits.

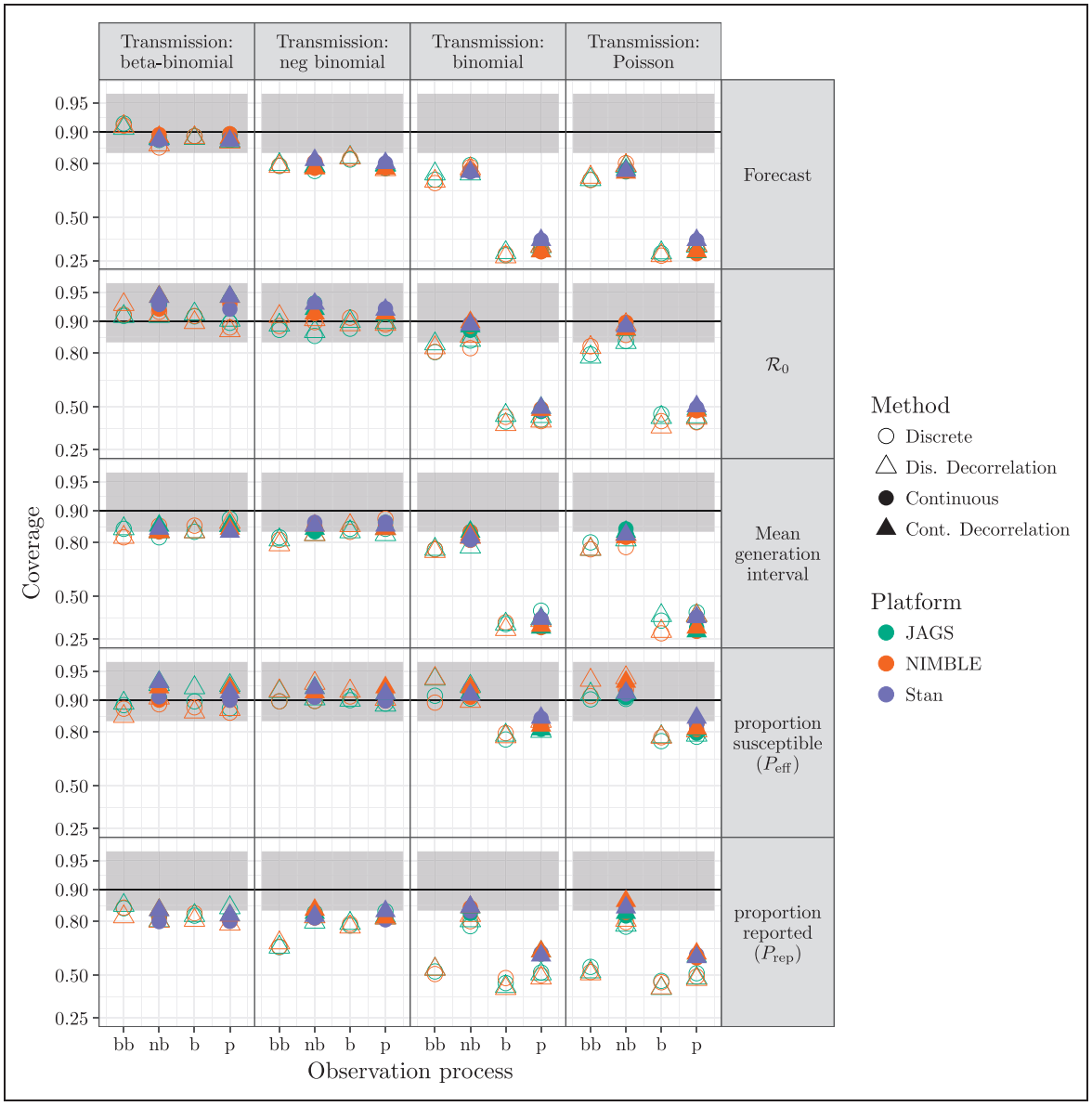

Figure 5. Comparison of coverage probability for forecast and parameters. Models with overdispersion in the transmission process (BB and NB, leftmost and second-left columns of panels) and models with overdispersion in the observation process (BB and NB, leftmost and second-left x-axis ticks within each panel) have coverage near the nominal value of 0.9 for all parameters and model variants. The black line shows the nominal coverage, and the grey ribbon the 95% binomial confidence interval based on 100 simulated fits. Vertical axis is plotted on a logit scale.

4.1 Ceilings

The effects of using distributions with ceilings (i.e. binomial and beta-binomial distributions) instead of their less realistic counterparts without ceilings (Poisson and negative binomial) were relatively small. In our framework, ceilings only apply in models with discrete latent variables; the primary effect of such ceilings is to reduce variance as probabilities (of infection or of sampling) become large. (Reporting-process models without ceilings also allow for false positives or over-reporting, which may be important in some contexts.)

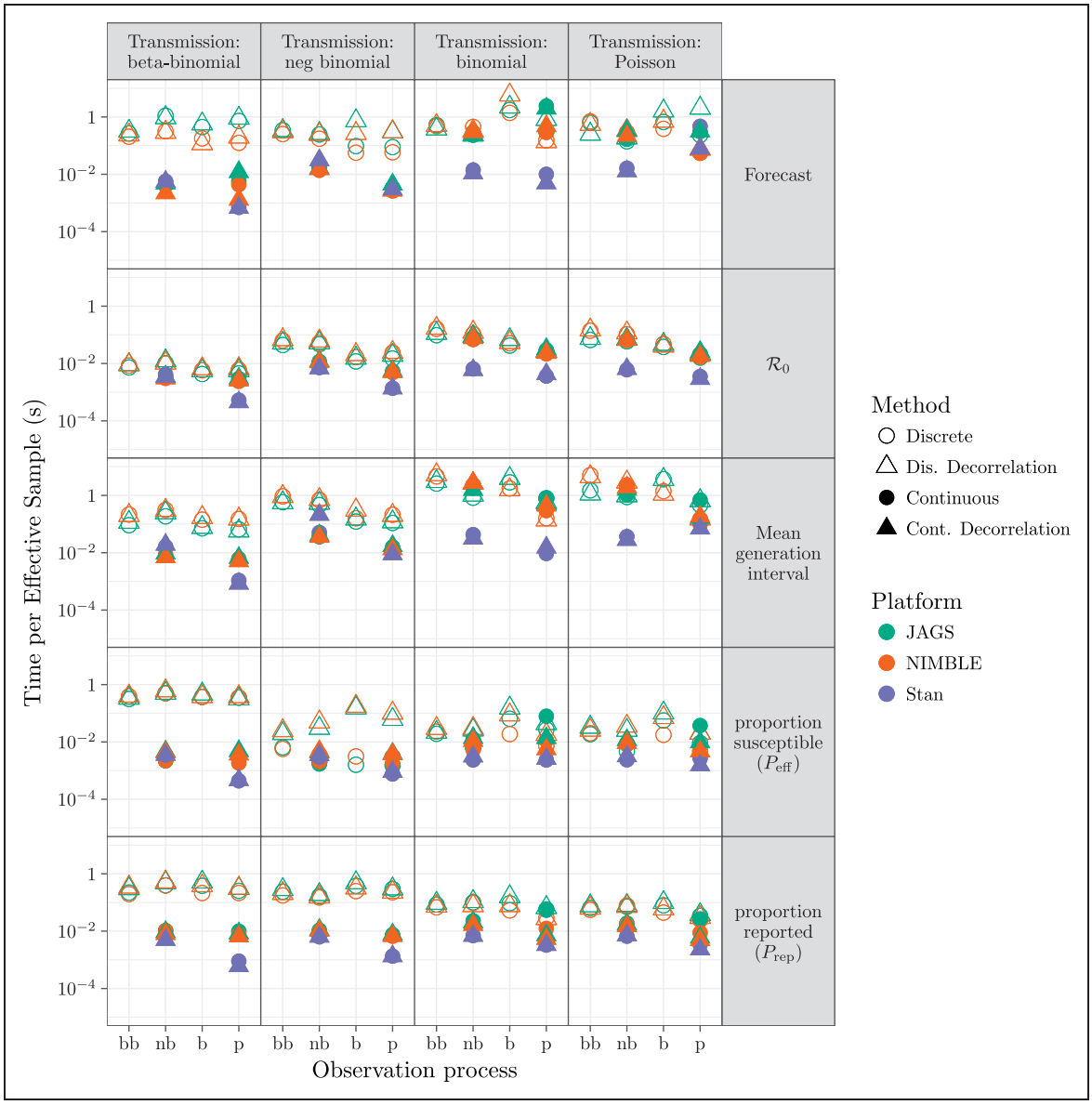

Figure 6. Comparison of efficiency for all fitting model variants: layout of models and platforms as in Figure 3.

4.2 Overdispersion

Accounting for overdispersion had more impact on our fits than the presence or absence of ceilings. In particular, models with no overdispersion in either process lacked flexibility and tended to be over-confident (that is, they showed low coverage). However, models that account for overdispersion in only one process (either transmission or observation) tended to be reliable for estimating parameters such as $\mathcal{R}_0$, mean generation interval, and short-term forecasts, particularly when overdispersion was implemented through the negative binomial (a less constrained distribution than the beta binomial). However, parameters that are closely tied to the details of a particular model structure (such as the overdispersion parameters for the observation and transmission processes) must change when the overdispersion model changes, in order to compensate for missing sources of variability.

Several authors[17,18] suggest that accounting for process as well as observation error in estimates of $\mathcal{R}_0$ and in forecasts is necessary in order to avoid over-confident estimates. Our exploration does not include any cases where process error is completely absent – even our "dispersion-free" processes incorporate sampling error in the process. However, we find that neglecting overdispersion can still lead to over-confident and unreliable estimates.

4.3 Reporting

In classic infectious disease models, reducing reporting rate and reducing the total effective population size have similar effects: reducing the observed size of the epidemic. While we want to make as few assumptions as possible about unobservable aspects of the epidemic, underreporting is of huge practical importance. Additionally, modeling observation error explicitly is required for reliable estimates of uncertainty.[17] If reporting error is modeled with a ceiling, then underreporting is a necessary component of reporting error (i.e. reporting will be biased downward in the presence of other sources of noise). Allowing overdispersion decouples the variance from the mean of the reporting process (i.e. the extra overdispersion parameter means that the variance is not determined by the mean).

Because reporting rate and effective population size play similar roles in epidemic dynamics, incorporating them both in a model may make their parameter estimates strongly correlated and hence difficult to identify: we may be very uncertain whether low observed epidemic incidence is driven by a small effective population size or a low reporting rate. We have addressed convergence problems arising from this issue by reparameterizing the model (Section 2.2.2). From a conceptual point of view, joint unidentifiability is not necessarily a serious problem, as long as the quantities we are most interested in (such as $\mathcal{R}_0$) are identifiable. In practice, however, weak identifiability can cause hard-to-detect convergence problems; known-parameter simulations like those implemented here are useful for validation in such cases.

4.4 Extensions and alternative approaches

Our analysis covers classical MC (i.e. conditional updating of parameters via conjugate, slice, and Metropolis-Hastings samplers) and HMC approaches. Even within this scope there is additional room for analysis, both in terms of exploring important heterogeneities that we have neglected here (such as spatial, age and social structure), and in improving sampling techniques (e.g. by adjusting the choice of samplers in JAGS or NIMBLE or by redundant parameterization[19]).

More broadly, a plethora of other model-fitting tools is available to researchers, from deterministic optimization tools based on the Laplace approximation[20,21] to sequential techniques such as iterated filtering and particle MC.[22–25] These techniques can in principle be combined flexibly with the methods we explore here, e.g. using HMC to sample top-level parameters while implementing a sequential MC technique for the latent states. It will be interesting to see how the single-technique methods here compete with hybrid approaches, and how flexible toolboxes such as NIMBLE will fare against more focused platforms like Stan.

4.5 Prior distributions

This paper focuses on evaluating Bayesian methods for fitting and forecasting epidemics. For the purposes of evaluation, we use parameter distributions for simulation that exactly match our Bayesian priors. We are assuming that researchers have a reasonable method of choosing appropriate Bayesian priors; in real applications this will be an important challenge.

5 Conclusion

We have presented a comparison of simple MCMC approaches to fit epidemic data. We learned two things about fitting epidemic data. First, modeling different processes with dispersion (BB and NB) is a naive but effective way to add uncertainty in the model; models that neglect such uncertainty are likely to be over-confident and less accurate at forecasting. Second, approximating discrete latent state process with continuous processes can aid efficiency without losing robustness of fit. This allows more efficient fitting in the classic framework (e.g. JAGS and NIMBLE), and also allows us to use the more advanced HMC technique (which we implemented via Stan).

Acknowledgements

We would like to thank the Ebola challenge organizers for organizing the Ebola model challenge that sparked our interest in this project, and Fred Adler and Michael Betancourt for thoughtful comments.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by NSERC Discovery Grant and CIHR Ebola Grant.

Supplementary material

In the main text, we present the bias, RMSE, coverage and efficiency plots for aggregated forecast, $\mathcal{R}_0$, MGI, $P_{\text{eff,}}$ and $P_{\text{rep.}}$ Here, we present plots showing the other parameters (shape G_S and position G_P of the transmission kernel and process and observation overdispersion parameters δ_P and δ_{obs}) and disaggregated forecasts (five forecast steps) that are excluded in the main text. We also add some representative plots of the simulated cases and forecast.

References

1. Ludwig D and Walters CJ. Are age-structured models appropriate for catch-effort data? *Can J Fisheries Aquatic Sci* 1985; **42**: 1066–1072. http://www.nrcresearchpress.com/doi/abs/10.1139/f85-132
2. O'Neill PD. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Math Biosci* 2002; **180**: 103–114.
3. Morton A and Finkenstädt BF. Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. *J Royal Stat Soc: Series C (Appl Stat)* 2005; **54**: 575–594.
4. O'Neill PD, Balding DJ, Becker NG, et al. Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *J Royal Stat Soc: Ser C (Appl Stat)* 2000; **49**: 517–542.
5. Plummer M, et al. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd international workshop on distributed statistical computing*, Vienna, Austria, 20–22 March 2003. vol. 124. p.125.
6. de Valpine P, Turek D, Paciorek CJ, et al. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *J Comput Graphic Stat* 2016; (just-accepted): 1–28.
7. Carpenter B, Gelman A, Hoffman M, et al. Stan: A probabilistic programming language. *J Stat Softw* 2016; 76: 1–32.
8. Ludwig D. Mathematical models for the spread of epidemics. *Comput Biol Med* 1973; **3**: 137–139.
9. Morris WF. Disentangling effects of induced plant defenses and food quantity on herbivores by fitting nonlinear models. *Am Nat* 1997; **150**: 299–327.
10. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; **57**: 97–109.
11. Geman S and Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transact Pattern Analys Mach Intelligence* 1984; **6**: 721–741.
12. Neal RM. Slice sampling. *Ann Stat* 2003; **31**: 705–741.
13. Duane S, Kennedy AD, Pendleton BJ, et al. Hybrid Monte Carlo. *Phys Lett B* 1987; **195**: 216–222.
14. Hoffman MD and Gelman A. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J Mach Learn Res* 2014; **15**: 1593–1623.
15. Minsker S, Srivastava S, Lin L, et al. Scalable and robust Bayesian inference via the median posterior. In: *International conference on machine learning*, Beijing, 21–26 June 2014, pp.1656–1664.
16. Cook SR, Gelman A and Rubin DB. Validation of software for Bayesian models using posterior quantiles. *J Computat Graph Stat* 2006; **15**: 675–692.
17. King AA, de Cells MD, Magpantay FMG, et al. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proc R Soc B* 2015; **282**: 20150347. http://rspb.royalsocietypublishing.org/content/282/1806/20150347
18. Taylor BP, Dushoff J and Weitz JS. Stochasticity and the limits to confidence when estimating of Ebola and other emerging infectious diseases. *J Theoret Biol* 2016; **408**: 145–154. http://www.sciencedirect.com/science/article/pii/S0022519316302466
19. Gelman A, Carlin JB, Stern HS, et al. *Bayesian data analysis*. vol. 2. Boca Raton, FL: Chapman & Hall, 2014.

20. Illian JB, Sørbye SH and Rue H. A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *Ann Appl Stat* 2012; **6**: 1499–1530.
21. Kristensen K, Nielsen A, Berg CW, et al. TMB: Automatic differentiation and Laplace approximation. *J Stat Software* 2016; 70. http://www.jstatsoft.org/v70/i05/
22. He D, Ionides EL and King AA. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *J Royal Soc Interface* 2009; **7**: 271–283.
23. Del Moral P, Doucet A and Jasra A. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat Comput* 2012; **22**: 1009–1020.
24. Yang W, Karspeck A and Shaman J. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS Comput Biol* 2014; **10**: e1003583.
25. Ionides EL, Bretó C and King A. Inference for nonlinear dynamical systems. *Proc Natl Acad Sci* 2006; **103**: 18438–18443.

Chapter 3: Reassessing global historical $\mathcal{R}_0$ estimates of canine rabies

In my first of two chapters on canine rabies, I present the results of $\mathcal{R}_0$ estimates of historical rabies outbreaks around the world. In this chapter, I revisiting different $\mathcal{R}_0$ estimation approaches for canine rabies. Using the same time series data, I used a more reliable approach to model the growth rate r. Using a hybrid approach that propagates uncertainties from both r estimates from a Bayesian framework and generation intervals from empirical data result in larger $\mathcal{R}_0$ estimates with wider confidence intervals.

The text I present here is a draft of a manuscript planned for submission for publication

Author Contributions

ML performed statistical analyses with helpful feedback from BMB, JD, and KH; ML wrote the first draft of the manuscript, and all authors revised the manuscript.

Acknowledgements

I thank Katie Hampson for sharing the data; Ben Bolker, Jonathan Dushoff, and Katie Hampson for helping me refine and present this work.

Reassessing global historical $\mathcal{R}_0$ estimates of canine rabies

Michael Li[1*], Katie Hampson[2], Benjamin M. Bolker[1,3,4], Jonathan Dushoff[1,3,4]

* Corresponding author: lim88@mcmaster.ca
1 Department of Biology, McMaster University, Hamilton, Ontario, Canada
2 Institute of BAH&CM, University of Glasgow, Glasgow, UK
3 Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
4 Institute for Infectious Diseases Research, McMaster University, Hamilton, Ontario, Canada

Keywords: incubation period, infectious period, generation interval

Abstract

Rabies spread by domestic dogs continues to cause tens of thousands of human deaths every year in low- and middle-income countries. Despite this heavy mortality burden, rabies is often neglected, perhaps because it has been eliminated from high-income countries through mass dog vaccination. Estimates of the intrinsic reproductive number ($\mathcal{R}_0$) of canine rabies from a wide range of times and locations are low (values <2), with narrow confidence intervals. The persistence of rabies in environments that vary enormously in ecological conditions is thus surprising. We combined incidence data from many historical outbreaks of canine rabies from around the world and used high-quality data from contact tracing from Tanzania (2002-present) to investigate initial growth rates (r), generation intervals (GIs) and reproductive numbers ($\mathcal{R}_0$). We used hybrid techniques to propagate uncertainties for $\mathcal{R}_0$, which accounts for uncertainties for both r and GI. Our $\mathcal{R}_0$ estimates and larger with wider confidence intervals compared to previous estimates. Our results suggest that in general $\mathcal{R}_0$ for rabies may be more uncertain and less well constrained by data than previously thought. This

hybrid approach of estimating $\mathcal{R}_0$ is applicable to other diseases systems that uses r and GIs to estimate $\mathcal{R}_0$.

Introduction

Canine rabies, primarily spread by domestic dogs, is a vaccine-preventable disease that continues to cause tens of thousands of human deaths every year in low- and middle-income countries (LMICs) (Minghui et al., 2018; Taylor et al., 2017). Canine rabies has been effectively eliminated from high-income countries by mass dog vaccination (Rupprecht et al., 2008). Despite the effectiveness of vaccinating dogs, rabies continues to cause considerable mortality and large economic losses in LMICs due to the limited implementation of rabies control strategies (Hampson et al., 2015). Over the past two decades, there has been an increase in efforts to control rabies – including dog vaccination campaigns and improvements in surveillance (Gibson et al., 2018; Kwoba et al., 2019; Mazeri et al., 2018; Mtema et al., 2016; Wallace et al., 2015). More recently, the World Health Organization (WHO) and partners (OIE, FAO, GARC) joined forces to support LMICs to eliminate human deaths from dog-mediated rabies by 2030 (Abela-Ridder et al., 2016; Minghui et al., 2018). In some LMICs mass dog vaccination campaigns have begun and are being scaled up (Castillo-Neyra et al., 2019; Evans et al., 2019). An understanding of rabies epidemiology — in particular, the basic reproductive number ($\mathcal{R}_0$), a quantitative measure of disease spread that is often used to guide vaccination strategies, could inform rabies control efforts.

$\mathcal{R}_0$ is defined as the expected number of secondary cases generated from each primary case in a fully susceptible population (Macdonald, 1952). Estimates of $\mathcal{R}_0$ using various methods (i.e., direct estimates from infection histories, epidemic tree reconstruction, and epidemic curve methods) based on historical outbreaks of rabies have generally been low, typically between 1 and 2 (Hampson et al., 2009; Kitala et al.,

2002; Kurosawa et al., 2017). In contrast to diseases with large $\mathcal{R}_0$ (e.g. measles with $\mathcal{R}_0$ >10 (Guerra et al., 2017)), $\mathcal{R}_0$ estimates for rabies imply that control through vaccination should be relatively easy (compared to e.g., rinderpest with $\mathcal{R}_0 \approx 4$ (Mariner et al., 2005)). Even in the absence of vaccination, one might expect rabies to fade out from behavioural control measures combined with stochastic fluctuations.

Our focus is to explore why rabies nonetheless continues to persist, often robustly, in many countries around the world. This persistence suggests that rabies potential for spread, and therefore the difficulty of control, may have been underestimated. In this chapter, I will use inferences from epidemic curves and a large number of observed generation intervals from a high-resolution contact tracing data to estimate rabies $\mathcal{R}_0$ around the world. Compared to other $\mathcal{R}_0$ estimation approaches, this approach is relatively robust to under-reporting; it also allows us to apply the generation interval data to historical rabies incidence data around the world. Re-assessing $\mathcal{R}_0$ estimates can improve the estimation of $\mathcal{R}_0$ and understanding of disease control more generally.

Materials and Methods

$\mathcal{R}_0$ is often estimated from two other epidemiological quantities: the initial growth rate of an epidemic (r) and the generation interval (GI) distribution, where a GI is defined as the time between successive infections along a transmission chain. r is often estimated by fitting a growth rate to time series data from the early stages of epidemics. GI is an individual level quantity that measures the time between an individual getting infected to infecting another individual. The generation interval distribution is the natural way to link r and $\mathcal{R}_0$ (Champredon and Dushoff, 2015; Wallinga and Lipsitch, 2006). During an outbreak in a fully susceptible population, $\mathcal{R}_0$ can be calculated from r and the GI distribution by the Euler-Lotka equation

(Wallinga and Lipsitch, 2006)

$$\mathcal{R}_0 = \frac{1}{\sum_{t=1}^{\infty} G(t)e^{-rt}}, \tag{1}$$

where t is time, and $G(t)$ is the generation interval distribution. This formula is convenient to calculate point estimates of $\mathcal{R}_0$; however, propagating uncertainty from estimates of r and the GI distribution can be hard.

Initial growth rate

Disease incidence typically increases approximately exponentially during the early stages of an epidemic. The initial growth rate r is often estimated by fitting exponential curves from near the beginning to near the peak of an epidemic. However, recent studies have shown that exponential models and their estimated growth rates are biased, overconfident, and sensitive to the choice of fitting windows (Ma et al., 2014). Alternative models such as the logistic model and generalized Richards model provide robust estimates of r in simulations (Chowell, 2017; Ma et al., 2014).

Here we instead assume that *cumulative* incidence follows a logistic function, but fit directly to incidence in each epidemic (to avoid statistical independence problems) (Ma et al., 2014). We select our fitting window consistently as follows: the starting point is the first detected case such that all time points (monthly incidence reports) after the starting point have at least one case; the final point in the first point after the observed global peak (the month with the highest incidence) of period incidence.

Observed Generation intervals

Transmission events are generally hard to observe for most diseases. In an earlier, influential paper, my colleagues and others constructed estimated generation intervals by summing two quantities: a latent period (the time from infection to infectiousness),

and a wait time (time from infectiousness to transmission). Since clinical signs and infectiousness appear at nearly the same time in rabies, the incubation period (the time from infection to clinical signs), is routinely used as a proxy for the latent period. In their analysis, latent (really, incubation) periods and infectious periods were randomly and independently resampled from empirically observed distributions (Hampson et al., 2009), and then wait times sampled uniformly from the selected infection periods.

My research has uncovered a previously overlooked problem with this approach for constructing *GI*s: random, independent resampling of incubation and infectious periods values does not account for the possibility of multiple transmissions of the same individual and the correlation between time distributions and biting behaviour. Figure 1 illustrates the generation intervals of a single transmission event from a rabid animal (comprising a single incubation period plus a waiting time) and multiple transmission events from a rabid animal (comprising a single incubation period and three waiting times). For diseases like rabies, where transmissions links (and generation intervals) are observable, multiple transmissions and possible correlation structures are all accounted for within the observation processes.

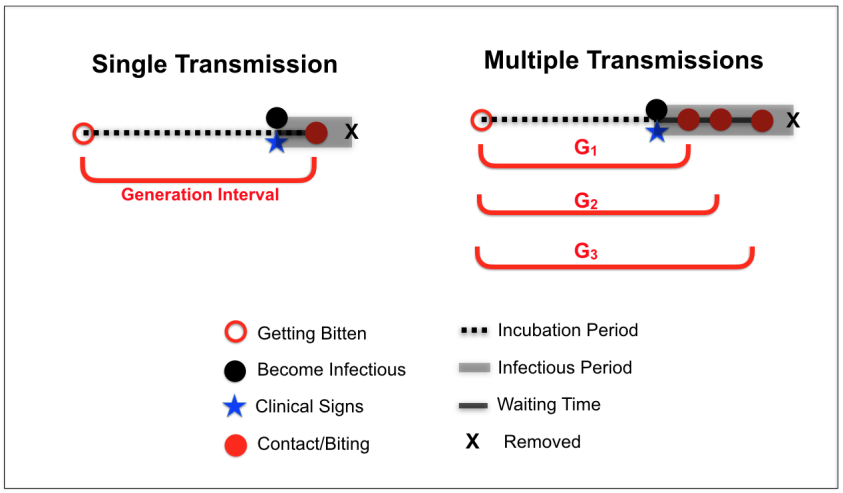

Figure 1: **Decomposing generation intervals.** Generation intervals start when a focal animal acquires infection (solid red circle); and end after virus replication (dashed line) when an animal shows clinical signs (blue star), becomes infectious (solid black circle) and infects another animal – in rabies the onset of clinical signs and of becoming infectious are closely synchronized. Once the infectious period (grey block) starts, there is a wait time (solid black line) until a susceptible host (solid red circle) is bitten. The focal host dies (black X) at the end of the infectious period. The generation interval is the interval between getting infected and infecting a new case (red interval between open and solid circles). (right) If a single biter transmits multiple times, the wait times are generally different, but the incubation period will be the same.

Correcting for vaccination

In a population where some animals are not susceptible, calculations based on estimates of r and the GI distribution (1) estimate the *realized* average number of cases per case, also known as the effective reproductive number $\mathcal{R}_e$. In the case of rabies, vaccination is the only known cause of immunity (case fatality in dogs is believed to be 100%). For a given population with ν vaccination proportion, $\mathcal{R}_e$ is:

$$\mathcal{R}_e = \mathcal{R}_0(1-\nu). \tag{2}$$

We therefore adjust our estimates of $\mathcal{R}_0$ in dog populations that have been vaccinated:

$$\mathcal{R}_0 = \frac{\mathcal{R}_e}{(1-\nu)}. \tag{3}$$

Data and material

We used data from January 2002 – May 2019, from an ongoing contact tracing project in Tanzania (Hampson et al., 2008, 2009). Since 2002, $\approx$ 12,000 transmission events ($\approx$ 10,000 animals and $\approx$ 1,300 humans), and $\approx$ 3,300 suspected rabid animals including $\approx$ 370 confirmed cases were observed.

Transmission events were documented through retrospective interviews with witnesses, applying diagnostic epidemiological and clinical criteria from the six-step method (Tepsumethanon et al., 2005). Each animal was given a unique identifier. The date of the bite and clinical signs were recorded if applicable and available. We restricted our analysis in this paper to domestic dog transmissions (i.e., dog to dog), and obtained 1179 directly observed generation intervals.

Fitting and Propagating Parameter Uncertainties

To propagate uncertainties for both r and GI, we used a hybrid approach. We first fitted logistic models, with negative binomial observation error, to incidence data using a Hamiltonian Monte Carlo (Duane et al., 1987) implemented in STAN (Carpenter et al., 2017). We obtained a posterior distribution for r, from a relatively broad prior (Normal(0.5/month,3)). We then calculate a sample of 500 $\hat{\mathcal{R}}_0$ using equation (3); for each value of $\hat{\mathcal{R}}_0$, we first draw a value of $\hat{r}$ from its posterior distribution and 1000 GIs (sampling with replacement from the empirical contact tracing data). Sampling from the empirical generation interval distribution accounts for the possibility of correlations between time distributions (incubation period and waiting times) and biting behaviour (dogs with multiple transmissions) that was previously overlooked.

Generation intervals are independent of transmission (secondary cases). This hybrid approach incorporates both sources of uncertainties from r (Bayesian posterior distribution) and GIs (frequentist empirical bootstrap) when calculating $\mathcal{R}_0$ estimates. Finally, we take the 2.5, 50, 97.5% quantiles $\mathcal{R}_0$ estimates for each rabies outbreak.

Results

Observed Generation intervals

Figure 2 shows the empirical distributions of the observed incubation periods, rabid dog biting frequency, and generation intervals from contact tracing data. The mean observed incubation period is 21.3 days (n = 1,134 dogs) and the weighted mean incubation period is 31.3 days (n = 248 biting dogs). The mean observed generation interval is 32.1 days (n = 250 primary infections resulting in 1179 secondary cases) is greater than the mean generation interval constructed from independently summing incubation periods and wait times 24.9 days. (Hampson et al., 2009). The weighted incubation period distribution resembles much closer to the generation interval distribution than the incubation period of all dogs.

We estimated r from historical outbreak data (see supplement Figure A1), and combined them with empirical GIs from our detailed Tanzanian data to produce $\mathcal{R}_e$ estimates. Our estimates of $\mathcal{R}_e$ ranged from 1.08 to 2.66, with upper confidence intervals greater than 2 for most locations. The hybrid approach provides larger values of $\mathcal{R}_e$ and wider confidence intervals than previous $\mathcal{R}_0$ estimates after propagating uncertainty from both r and generation interval distributions (see Figure 3).

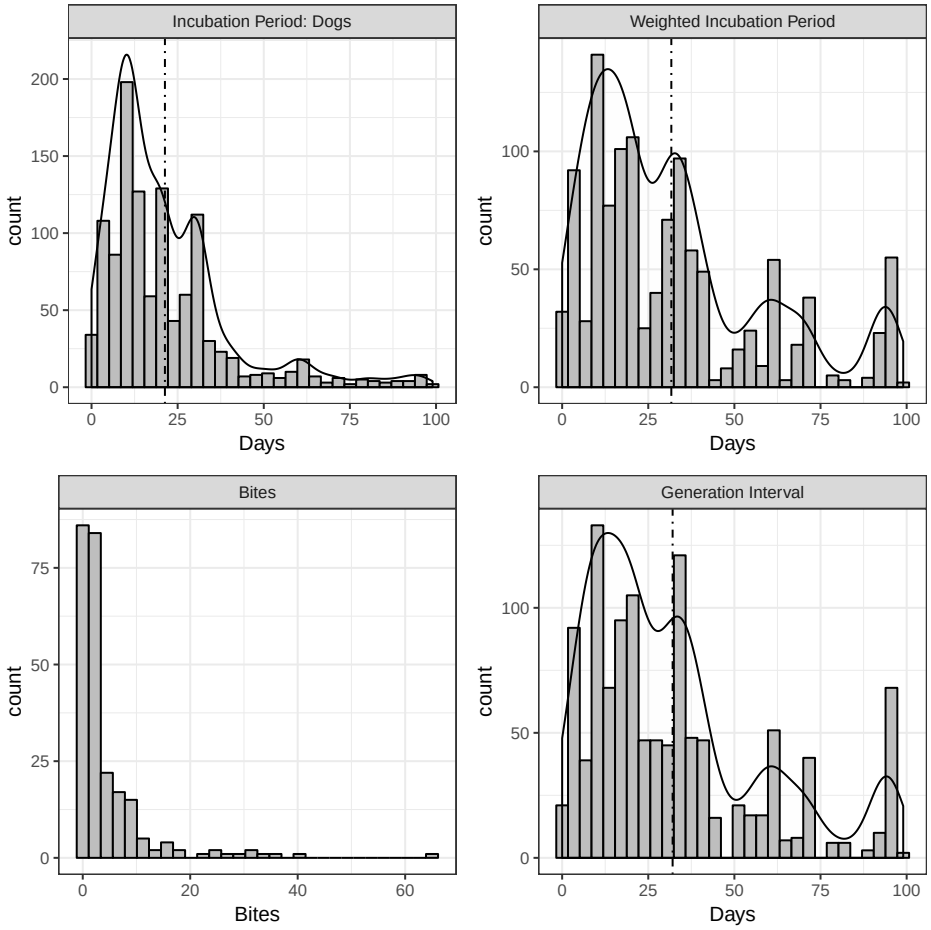

Figure 2: **Empirical distributions from contact tracing data.** Top left) the distribution of observed incubation periods. Top right) the distribution of infectious periods weighted by each dog's biting frequency (biting frequency shown bottom left). The weighted distribution corresponds to the contribution of incubation periods to generation intervals, which are shown bottom right. Dash-dotted lines show the means of each time-interval distribution.

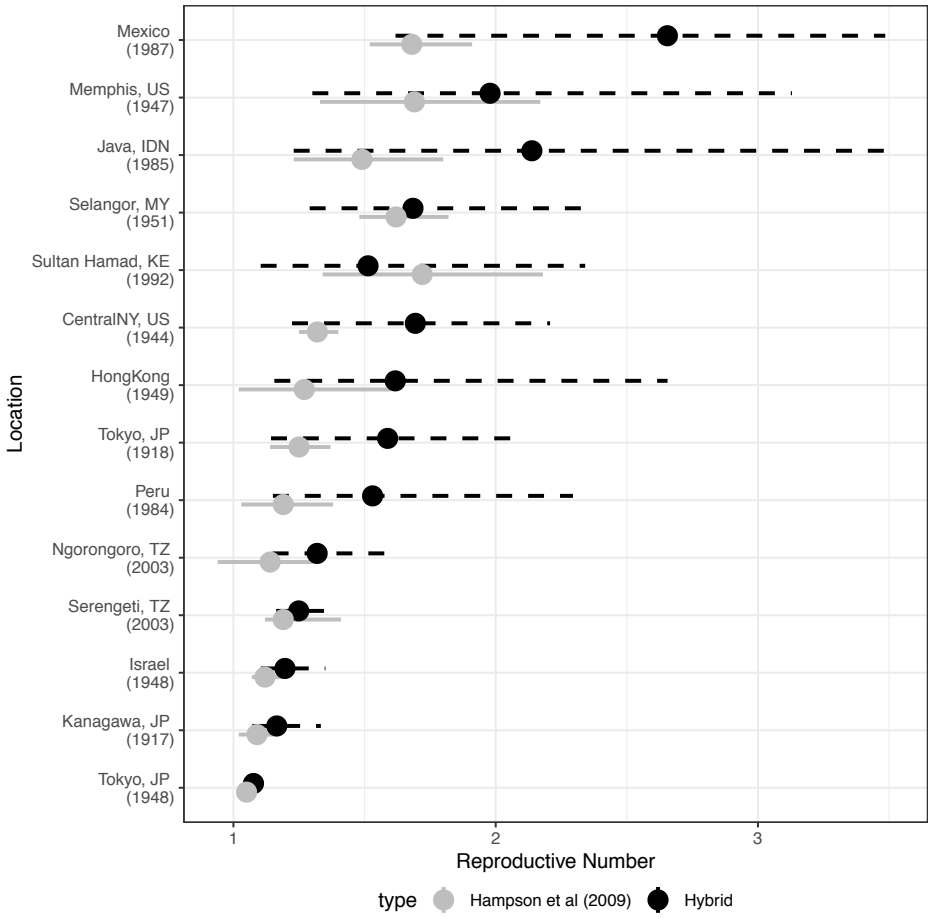

Figure 3: **Reproductive number estimates for global historical outbreaks of rabies, with 95% confidence intervals.** Previous estimates of $\mathcal{R}_0$ (solid line) are shown in gray; $\mathcal{R}_e$ (dash line) estimates from our hybrid approach are shown in black.

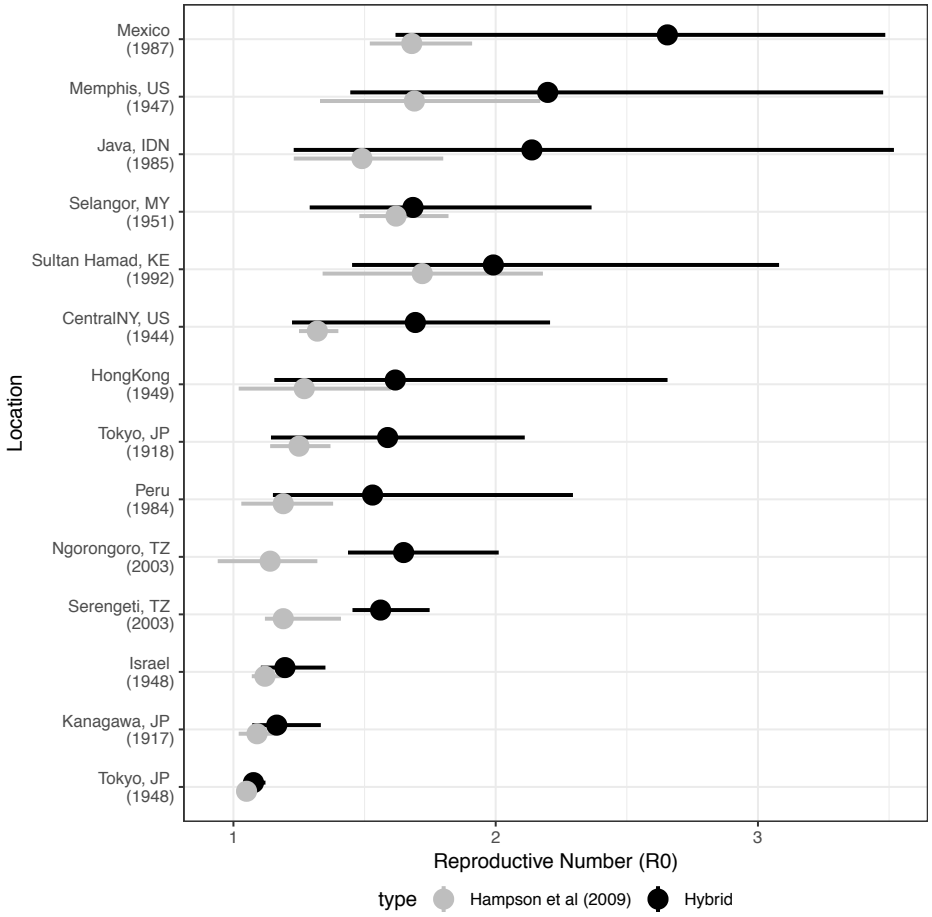

Figure 4: **$\mathcal{R}_0$ estimates for global historical outbreaks, with 95% confidence intervals. Adjusting $\mathcal{R}_e$ estimates for vaccination coverage.**

Of the listed historical outbreaks, four occured in locations with prior rabies vaccination coverage: Memphis in the US (1947), Serengeti in Tanzania (2003), Ngorongoro in Tanzania (2003), and Sultan Hamad in Kenya (1992). The proportion of the dogs thought to be vaccinated in these populations was 10%, 20%, 20%, and 24% respectively. We adjusted our $\mathcal{R}_e$ estimates to obtain $\mathcal{R}_0$ estimates for these four outbreaks (Figure 4).

Discussion

The basic reproductive number $\mathcal{R}_0$ is commonly used to summarize the risk of infectious disease and to inform control measures. Here, we used a relatively simple approach to estimate $\mathcal{R}_0$ by combining initial growth rate (r) estimates from incidence data and generation intervals from contact tracing data. We improved on earlier work by correcting for slowdown in growth in estimating r and by developing a hybrid approach to propagate uncertainty from both r and GI, resulting in higher $\mathcal{R}_0$ estimates with wider confidence intervals.

Re-analysis of these data also allowed us to identify an overlooked fact about rabies generation intervals: observed generation intervals are longer on average than constructed ones, because of within-individual correlations in time distributions and biting behaviour. The unexpected importance of these correlations could have implications for GI-based studies of other infectious diseases. Further investigation of how these correlations affect the overall dynamics of rabies is warranted.

Estimates of $\mathcal{R}_0$ are strongly affected by estimates of the growth rate during the initial phase of the epidemic. The logistic model gives a better approximation of the initial phase of the epidemic resulting in a larger estimate of r compared to the exponential model (Ma et al., 2014). Our estimates of r account for observation error (measurements may not perfectly match reality), but not for process error (the

fundamental stochasticity of the system itself). Thus, there may be more uncertainty in r than we estimate (King et al., 2015), but this is not always true in practice (Li et al., 2018).

Nevertheless, our estimates suggest that rabies $\mathcal{R}_0$ may be larger, and more uncertain than previously thought. This finding may explain some of the formerly unexplained variations in the success of rabies-control programs (e.g., low levels of coverage (30–50%) have been successful in some setings while high coverage 75% was not enough to control rabies in others (Eng et al., 1993)). While exploring why rabies $\mathcal{R}_0$ were low and narrow was our primary goal, we were able to reveal an interesting biological process in rabies generation interval. A mechanistic fitting framework will likely be required to study these patterns in more detail.

References

Abela-Ridder, B., L. Knopf, S. Martin, L. Taylor, G. Torres, and K. De Balogh (2016). 2016: the beginning of the end of rabies? *The Lancet Global Health 4*(11), e780–e781.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of statistical software 76*(1).

Castillo-Neyra, R., A. Toledo, C. Arevalo-Nieto, H. MacDonald, M. De la Puente-Leon, C. Naquira-Velarde, V. A. Paz-Soldan, A. M. Buttenheim, and M. Z. Levy (2019). Socio-spatial heterogeneity in participation in mass dog rabies vaccination campaigns, Arequipa, Peru. *bioRxiv*, 542878.

Champredon, D. and J. Dushoff (2015). Intrinsic and realized generation intervals in infectious-disease transmission. *Proceedings of the Royal Society B: Biological Sciences 282*(1821), 20152026.

Chowell, G. (2017). Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. *Infectious Disease Modelling 2*(3), 379–398.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid monte carlo. *Physics letters B 195*(2), 216–222.

Eng, T., D. Fishbein, H. Talamante, D. Hall, G. Chavez, J. Dobbins, F. Muro, J. Bustos, M. De Los Angeles Ricardy, A. Munguia, et al. (1993). Urban epizootic of rabies

in Mexico: epidemiology and impact of animal bite injuries. *Bulletin of the World Health Organization 71*(5), 615.

Evans, M., J. B. Bailey, F. Lohr, W. Opira, M. Migadde, A. Gibson, I. Handel, B. d. Bronsvoort, R. Mellanby, L. Gamble, et al. (2019). Implementation of high coverage mass rabies vaccination in rural Uganda using predominantly static point methodology. *The Veterinary Journal 249*, 60–66.

Gibson, A. D., S. Mazeri, F. Lohr, D. Mayer, J. L. B. Bailey, R. M. Wallace, I. G. Handel, K. Shervell, M. Barend, R. J. Mellanby, et al. (2018). One million dog vaccinations recorded on mHealth innovation used to direct teams in numerous rabies control campaigns. *PloS one 13*(7), e0200942.

Guerra, F. M., S. Bolotin, G. Lim, J. Heffernan, S. L. Deeks, Y. Li, and N. S. Crowcroft (2017). The basic reproduction number (R0) of measles: a systematic review. *The Lancet Infectious Diseases 17*(12), e420–e428.

Hampson, K., L. Coudeville, T. Lembo, M. Sambo, A. Kieffer, M. Attlan, J. Barrat, J. D. Blanton, D. J. Briggs, S. Cleaveland, et al. (2015). Estimating the global burden of endemic canine rabies. *PLoS neglected tropical diseases 9*(4), e0003709.

Hampson, K., A. Dobson, M. Kaare, J. Dushoff, M. Magoto, E. Sindoya, and S. Cleaveland (2008). Rabies exposures, post-exposure prophylaxis and deaths in a region of endemic canine rabies. *PLoS neglected tropical diseases 2*(11), e339.

Hampson, K., J. Dushoff, S. Cleaveland, D. T. Haydon, M. Kaare, C. Packer, and A. Dobson (2009). Transmission dynamics and prospects for the elimination of canine rabies. *PLoS biology 7*(3), e1000053.

King, A. A., M. Domenech de Cellès, F. M. Magpantay, and P. Rohani (2015). 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.

Kitala, P., J. J. McDERMOTT, P. Coleman, and C. Dye (2002). Comparison of vaccination strategies for the control of dog rabies in Machakos District, Kenya. *Epidemiology & Infection 129*(1), 215–222.

Kurosawa, A., K. Tojinbara, H. Kadowaki, K. Hampson, A. Yamada, and K. Makita (2017). The rise and fall of rabies in Japan: A quantitative history of rabies epidemics in Osaka prefecture, 1914–1933. *PLoS neglected tropical diseases 11*(3), e0005435.

Kwoba, E. N., P. Kitala, L. Ochieng, E. Otiang, R. Ndung'u, G. Wambura, K. Hampson, and S. Thumbi (2019). Dog health and demographic surveillance survey in Western Kenya: Demography and management practices relevant for rabies transmission and control. *AAS Open Research 2*.

Li, M., J. Dushoff, and B. M. Bolker (2018). Fitting mechanistic epidemic models to data: A comparison of simple Markov chain Monte Carlo approaches. *Statistical methods in medical research 27*(7), 1956–1967.

Ma, J., J. Dushoff, B. M. Bolker, and D. J. Earn (2014). Estimating initial epidemic growth rates. *Bulletin of mathematical biology 76*(1), 245–260.

Macdonald, G. (1952). The analysis of equilibrium in malaria. *Tropical diseases bulletin 49*(9), 813.

Mariner, J. C., J. McDermott, J. A. P. Heesterbeek, A. Catley, and P. Roeder (2005, July). A model of lineage-1 and lineage-2 rinderpest virus transmission in pastoral areas of East Africa. *Preventive Veterinary Medicine 69*(3), 245–263.

Mazeri, S., A. D. Gibson, N. Meunier, M. Barend, I. G. Handel, R. J. Mellanby, and L. Gamble (2018). Barriers of attendance to dog rabies static point vaccination clinics in Blantyre, Malawi. *PLoS neglected tropical diseases 12*(1), e0006159.

Minghui, R., M. Stone, M. H. Semedo, and L. Nel (2018). New global strategic plan to eliminate dog-mediated rabies by 2030. *The Lancet Global Health 6*(8), e828–e829.

Mtema, Z., J. Changalucha, S. Cleaveland, M. Elias, H. M. Ferguson, J. E. Halliday, D. T. Haydon, G. Jaswant, R. Kazwala, G. F. Killeen, et al. (2016). Mobile phones as surveillance tools: implementing and evaluating a large-scale intersectoral surveillance system for rabies in Tanzania. *PLoS medicine 13*(4), e1002002.

Rupprecht, C., J. Barrett, D. Briggs, F. Cliquet, A. Fooks, B. Lumlertdacha, F. Meslin, T. Müler, L. Nel, C. Schneider, et al. (2008). Can rabies be eradicated? *Developments in biologicals 131*, 95–121.

Taylor, L. H., K. Hampson, A. Fahrion, B. Abela-Ridder, and L. H. Nel (2017). Difficulties in estimating the human burden of canine rabies. *Acta tropica 165*, 133–140.

Tepsumethanon, V., H. Wilde, and F. X. Meslin (2005). Six criteria for rabies diagnosis in living dogs. *J Med Assoc Thai 88*(3), 419–22.

Wallace, R. M., H. Reses, R. Franka, P. Dilius, N. Fenelon, L. Orciari, M. Etheart, A. Destine, K. Crowdis, J. D. Blanton, et al. (2015). Establishment of a canine rabies burden in Haiti through the implementation of a novel surveillance program. *PLoS neglected tropical diseases 9*(11), e0004245.

Wallinga, J. and M. Lipsitch (2006). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences 274*(1609), 599–604.

Chapter 4: Generation time bias in disease transmission mechanism

In my second of two chapters on canine rabies, I present empirical evidence of the difference in the generation and serial intervals in rabies. In this chapter, I explored the high-resolution contact-tracing rabies database to extract generation and serial intervals and other disease traits for the animals. After finding differences in the generation and serial intervals, I investigate further to explore the mechanism of the discrepancies of these intervals. I exploited the discrepancies by fitting a generalized linear model to disease traits, where I find dogs with longer incubation periods have more secondary cases on average.

The text I present here is draft of a manuscript planned for submission for publication.

Author Contributions

ML performed the statistical analyses with helpful feedback from BMB, JD, and KH; ML wrote the manuscript, and JD revised the first draft of the manuscript.

Acknowledgements

I thank Katie Hampson for sharing the data; Ben Bolker, Jonathan Dushoff, and Katie Hampson for helping me refine and present this work.

Generation time bias in disease transmission mechanism

Michael Li[1*], Katie Hampson[2], Benjamin M. Bolker[1,3,4], Jonathan Dushoff[1,3,4]

* Corresponding author: lim88@mcmaster.ca
1 Department of Biology, McMaster University, Hamilton, Ontario, Canada
2 Institute of BAH&CM, University of Glasgow, Glasgow, UK
3 Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
4 Institute for Infectious Diseases Research, McMaster University, Hamilton, Ontario, Canada

Keywords: incubation period, infectious period, generation interval, serial interval

Abstract

The generation interval distribution links two key quantities in infectious disease dynamics: the exponential rate of epidemic growth r and the basic reproductive number $\mathcal{R}_0$. For many infectious diseases, generation intervals are difficult to observe directly; in these cases, researchers often use the serial interval as a proxy for the generation interval. In fact, the terms "generation interval" and "serial interval" are often (incorrectly) used interchangeably in the literature. Here we explore a high-resolution 20-year contact-tracing dataset for canine rabies, and find sharp differences between the average observed generation and serial intervals: the generation interval is approximately 50% larger than the serial interval — a difference with important effects on estimates of the basic reproductive number. We develop a theoretical framework to show that these differences arise from individual-level correlations between incubation period and biting. These findings are critical for accurate estimation of rabies

transmission parameters, and hold the potential to guide investigations of other diseases.

Introduction

The "generation interval" is the time between the moment a focal (primary) host is infected and the moment that they infect another (secondary) host. The generation-interval distribution links two key quantities that characterize infectious-disease dynamics: the exponential rate of growth (r) and the basic reproductive number ($\mathcal{R}_0$). In particular, the product $\mathcal{R}_0 \approx r\bar{G}$, where $\bar{G}$ is the average generation interval; more precise expressions can be used when the entire generation interval distribution is known (Park et al., 2019).

Generation intervals are often hard to estimate reliably because transmission events are difficult to observe for most diseases. Generation intervals are thus sometimes estimated by summing two quantities: a latent period (the time from infection to infectiousness), and an infectious period; the sum of the average latent and average infectious periods is the average generation interval (Anderson et al., 1991). However, in the previous chapter, I showed that constructed estimated generation intervals and observed generation intervals are different for canine rabies; mean generation intervals are larger than constructed generation intervals. In this chapter, I will examine some issues in estimating mean values of the generation intervals and discrepancies between generation and serial intervals (often as a proxy for generation intervals).

The above formulation of the generation interval as the sum of latent and infectious periods (Anderson et al., 1991) is only true under the assumption of exponential infectious periods. We can generalize this formulation by replacing the infectious period by the infectious-waiting time, which is the time from the onset of infectiousness

to transmission. We thus write the generation interval as:

$$GI = L + F, \tag{1}$$

where L is the latent period and F is the infectious-waiting time.

An alternative way to estimate the generation interval is to use the serial interval; researchers often in fact conflate the two terms (Andreasen et al., 2008; Chunara et al., 2012; Majumder et al., 2016; Vynnycky and Fine, 2000; Wallinga and Lipsitch, 2006; Wallinga and Teunis, 2004; White et al., 2009). The "serial interval" is the time between the onset of signs or symptoms in the primary and secondary host. The serial interval can be constructed by summing two time delays: the symptomatic waiting time of the primary host and an incubation period of the secondary host. Note that the symptomatic waiting time need not always be positive: for some diseases infected individuals can begin to transmit infection before the onset of clinical signs or symptoms.

Intuitively, the serial- and generation-interval distributions describe generations from the same process (but with different focal points), and would be expected to be similar. There are theoretical arguments about circumstances under which these intervals should be the same, or approximately the same. However, these arguments do not always apply, and the distributions are not always similar. In fact, in a disease where clinical symptoms typically come later than infectiousness, it is even possible for the serial interval (but not the generation interval) to be negative. Several studies have reported differences between serial and generation intervals (Cowling et al., 2009; te Beest et al., 2014). To our knowledge, work on how differences between generation and serial intervals affect inferences about disease systems has been very limited.

To examine the differences between generation and serial interval, we construct the intervals using a similar formulation. We construct the generation interval as:

$$GI = I_p + S_p, \tag{2}$$

where I_p and S_p are the incubation periods and symptomatic waiting time of the primary host, and the serial interval as:

$$SI = S_p + I_s, \tag{3}$$

where I_s is the incubation period of the secondary host.

Figure 1 shows our reformulation of the generation interval to match the serial intervals. This view makes clear that the difference between the generation and serial interval is the incubation periods: the generation interval uses the incubation period of the primary host, and the serial interval uses the incubation period of the secondary host. Thus, we expect the two distributions to have the same mean as long as incubation periods are independent of tendency to transmit; and the same variance if incubation periods are also independent of symptomatic waiting times (Svensson, 2007).

In the previous chapter, however, I showed that in canine rabies the observed generation interval compared to estimated constructed generation intervals (Hampson et al., 2009) are *not* the same on average; the incubation periods are *not* independent of tendency to transmit. This means that in this case incubation times of the primary individual in a typical transmission is expected to differ from those for a secondary individual. The intuitive expectation that GIs and SIs will be the same does not hold. In this chapter, I will use canine rabies as a case study to explore how such correlations

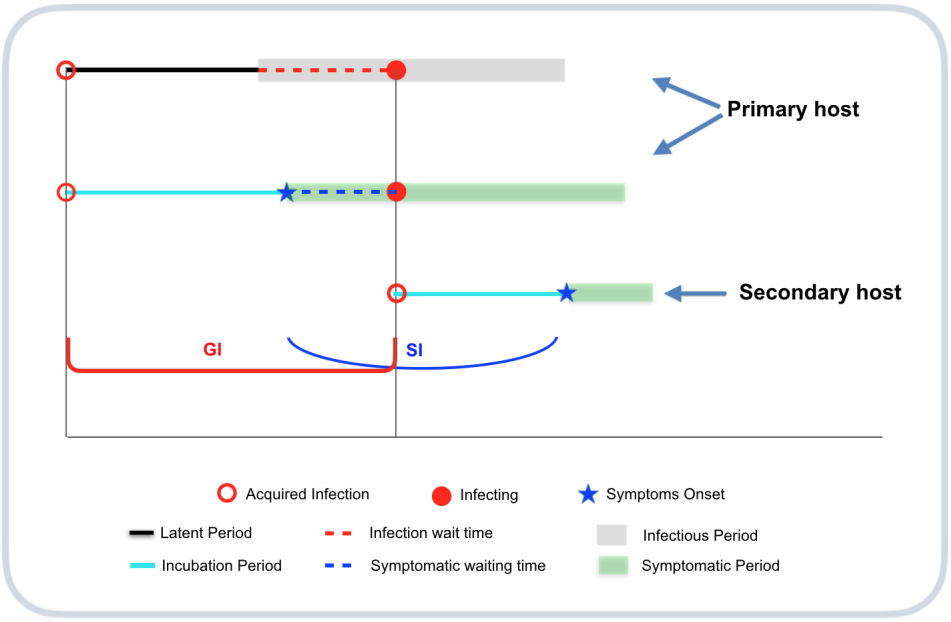

Figure 1: **Generation and serial interval.** Generation intervals start when primary host acquires infection (open red circle); and after virus replication (solid black line), the host becomes infectious; this is the latent period. Once the infectious period (grey block) starts, there is an infectious waiting time (dashed red line) until a susceptible host (solid red circle) is infected. The primary host becomes non-infectious at the end of the infectious period. Another way to construct the generation interval is when the primary host acquires infection; and after virus replication (the incubation period is shown with a light blue solid line) when the host shows clinical signs or symptom onset (blue star) and infects another. There is a symptomatic-waiting time (dashed blue line) until a susceptible host (solid red circle) is infected. Serial intervals start when primary host shows symptoms (blue star); and end when the new secondary host develops symptoms.

affect the relationships between time distributions and tendency to transmit (i.e., number of secondary cases).

Rabies

Canine rabies, spread primarily by domestic dogs, is a preventable disease that causes more than 50,000 estimated annual deaths in humans and economic burden in low- and middle-income countries (LMICs) (Hampson et al., 2015). Rabies transmission chains are relatively easy to record compared to many diseases because infection

events (through biting) are observable and the onset of infectiousness coincides with clinical signs (aggressive behaviours and biting). The transmission cycle begins when an infected animal bites a susceptible host; the virus spreads through the saliva of primary host to the wounds of the secondary host. Once the virus enters the secondary host, it travels through the nerves to the spinal cord and brain; and it takes about 3 to 12 weeks depending on the location of the wound. Clinical signs occur when the virus reaches the brain and infectiousness (when the virus reaches salivary gland) starts at around the same time.

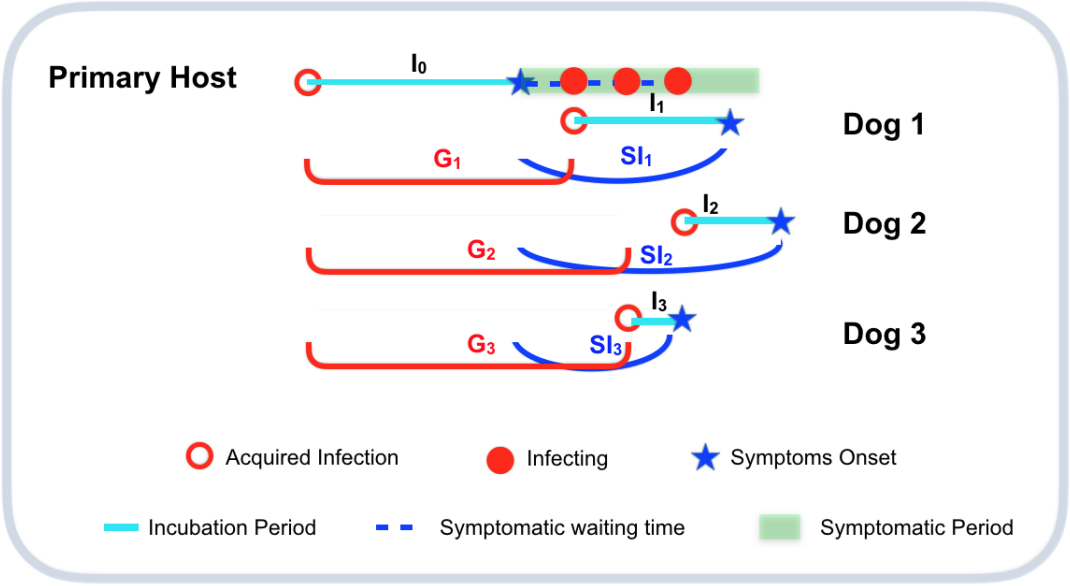

Figure 2: **Generation and serial interval for rabies.** Generation intervals (red intervals) start when the primary host acquires infection (solid red circle) and ends after biting another host. If a single primary host transmits multiple times, the symptomatic waiting times (dashed blue lines) are generally different, but the incubation (I_0) period will be the same. Serial intervals (blue intervals) start when the primary host shows clinical signs (blue star at the end of I_0) and ends when the secondary host show clinical signs. Each secondary case (dog) has its own symptomatic waiting time and incubation period (I_1, I_2, I_3).

In rabies, the incubation and latent periods are well synchronized: clinical signs and infectiousness appear at nearly the same time. Since clinical signs can be directly observed, the incubation period is routinely used as a proxy for the latent period. Figure 2 shows the generation and serial interval of a primary host with multiple transmission for rabies. From Figure 2 we can write the generation intervals as:

$$GI_i = I_0 + S_i \text{ for } i = 1, 2, 3; \tag{4}$$

and the serial intervals as:

$$SI_i = S_i + I_i \text{ for } i = 1, 2, 3. \tag{5}$$

If waiting time is correlated with transmission behaviour, this will affect SIs and GIs in the same way, since both incorporate the (same) waiting time of the primary individual. If incubation time is correlated with transmission behaviour (or with waiting time), however, this will affect GIs, but not SIs (since the incubation period in the SI does not come from the primary individual).

Methods

Zero-inflated negative-binomal

Because infectious periods are highly variable, hard to measure, and do not show clear correlations, we focus for this example on modeling the relationship between incubation period and biting behaviour (specifically, the number of secondary hosts identified). We modeled the number of infectious bites (measured as secondary hosts) as a function of incubation in a GLM framework with a log link and a simple linear response. Thus, we assume that the expected number of bites is exponentially related to the infectious period. To account for (1) a high proportion of dogs without observed

infectious bites and (2) large variation in the number of infectious bites among other dogs, we used a zero-inflated negative binomial model; that is we modeled the outcome variable as being "structurally" zero with a certain probability, and otherwise having a negative-binomial distribution:

$$\begin{aligned} \Pr(x_i = 0) &= p_t + (1 - p_t)\text{NBinom}(0, \mu_t, \theta) \\ \Pr(x_i = k) \quad (k > 0) &= (1 - p_t)\text{NBinom}(k, \mu_t, \theta) \\ \mu_t &= \exp(\beta_0 + \beta_1 t) \\ p_t &= \frac{1}{1 + \exp(-(\beta_0^z + \beta_1^z t))}, \end{aligned} \tag{6}$$

where p_t is the probability of structural zeros when incubation period is t days; β_0 and β_1 are the coefficients of the conditional model; and β_0^z and β_1^z are the coefficients of the zero-inflation model.

To calculate confidence intervals for the effect of incubation period on mean biting, we fixed the intercepts (β_0, β_0^z), and simulated 100 multivariate normal samples of (β_1, β_1^z) using the maximum-likelihood estimated mean and covariance matrix of the zero-inflation model, and used the median and (2.5%, 97.5%) quantiles.

Permutation and Prediction Contour Regions

We used a permutation test to see the effects of correlation between incubation period and biting on generation and serial intervals. We generated 5000 uncorrelated data sets by permuting the bite counts in the original data (i.e., a new bite count for each incubation period). For comparison, we also generated 5000 data sets which include our estimated correlation structure by making random predictions of bite counts from our fitted ZINB model for each incubation period. All incubation periods are then weighted by the simulated number of bite counts.

Data and material

A high-resolution rabies database from an on-going contact-tracing program in Tanzania from January 2002 to present was used for this study. This database contains $\approx 12,000$ specific transmission events documented through retrospective interviews with witnesses to gather information about suspected animals. For each suspected animal (i.e., an animal suspected to have rabies or bitten by a suspected animal), researchers collected: date bitten, clinical sign information, location, and biting history. We restrict our analysis to transmission between domestic dogs, and to cases where we were able to link the transmission events via contact-tracing history (i.e., the infector (primary case) that is responsible for the transmission event of the observed animal (secondary case) also appeared as an observed case of an earlier event in the database). With these restrictions, we obtained 1179 directly observed generation intervals and 1048 directly observed serial intervals.

Results

Time intervals in Rabies

The mean generation interval (32.1 days) is nearly 50% greater than the mean serial interval (21.7 days). This difference in interval lengths occurs both because biters have longer mean incubation periods (26.6 days) than non-biters (19.8 days) and because dogs that bite more have longer incubation periods on average than dogs that bite less. The incubation period of all dogs closely resembles the serial interval distribution, while the weighted incubation period distribution (weighted by number of secondary cases) closely resembles the generation interval distribution.

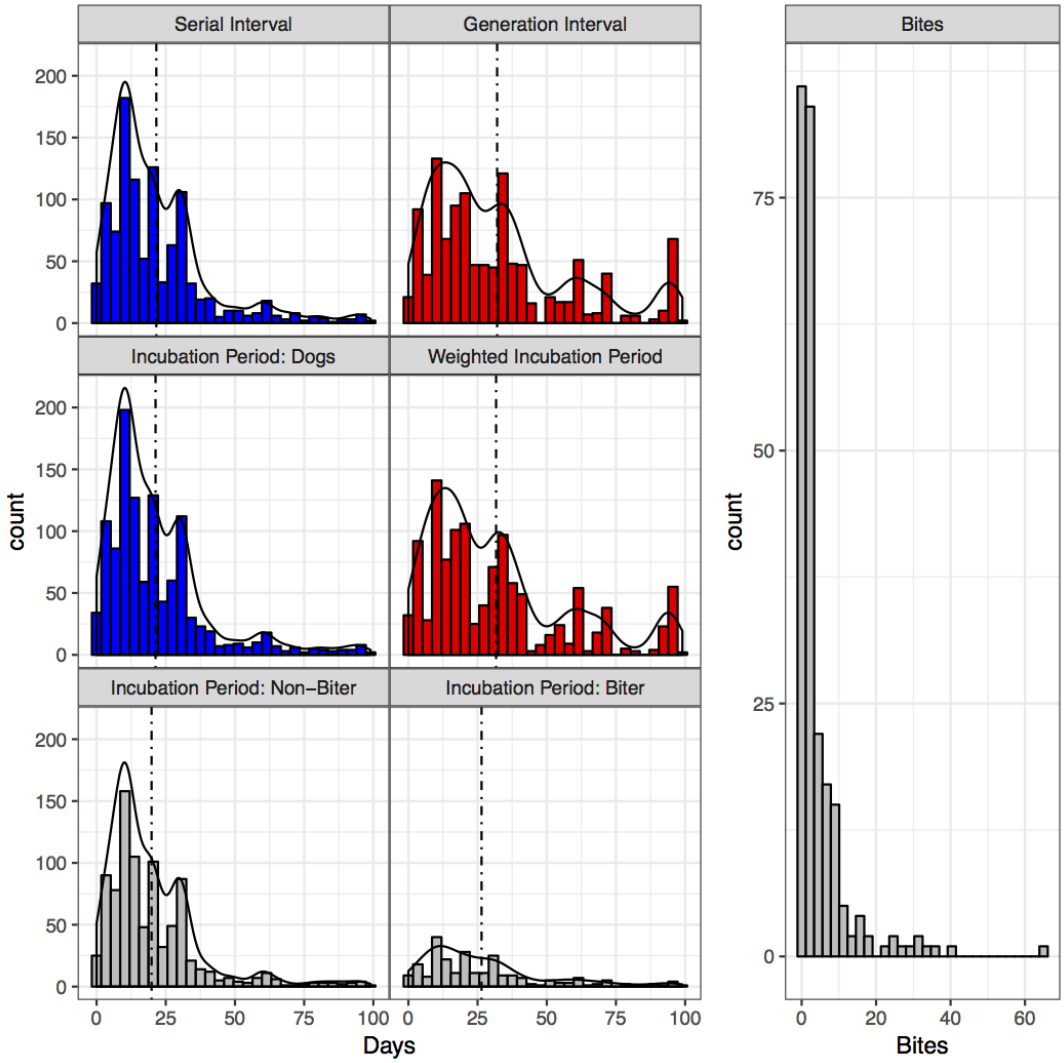

Figure 3: **Rabies interval distributions estimated from contact-tracing data.** First row: directly observed *SI*s (left) are shorter on average than *GI*s (right). Middle row: the observed incubation-period distribution (left) is similar to the *SI* distribution above it, while the same observed incubation periods *weighted by the number of infectious bites* (right) is similar to the *GI* distribution is similar to the same incubation-period distribution. Last row: biters have longer mean incubation periods than non-biters, which is what leads to the lengthening of the *GI*s in the first row. Dashed lines show the means of each time-interval distribution.

Results: Incubation period and secondary infection relationship

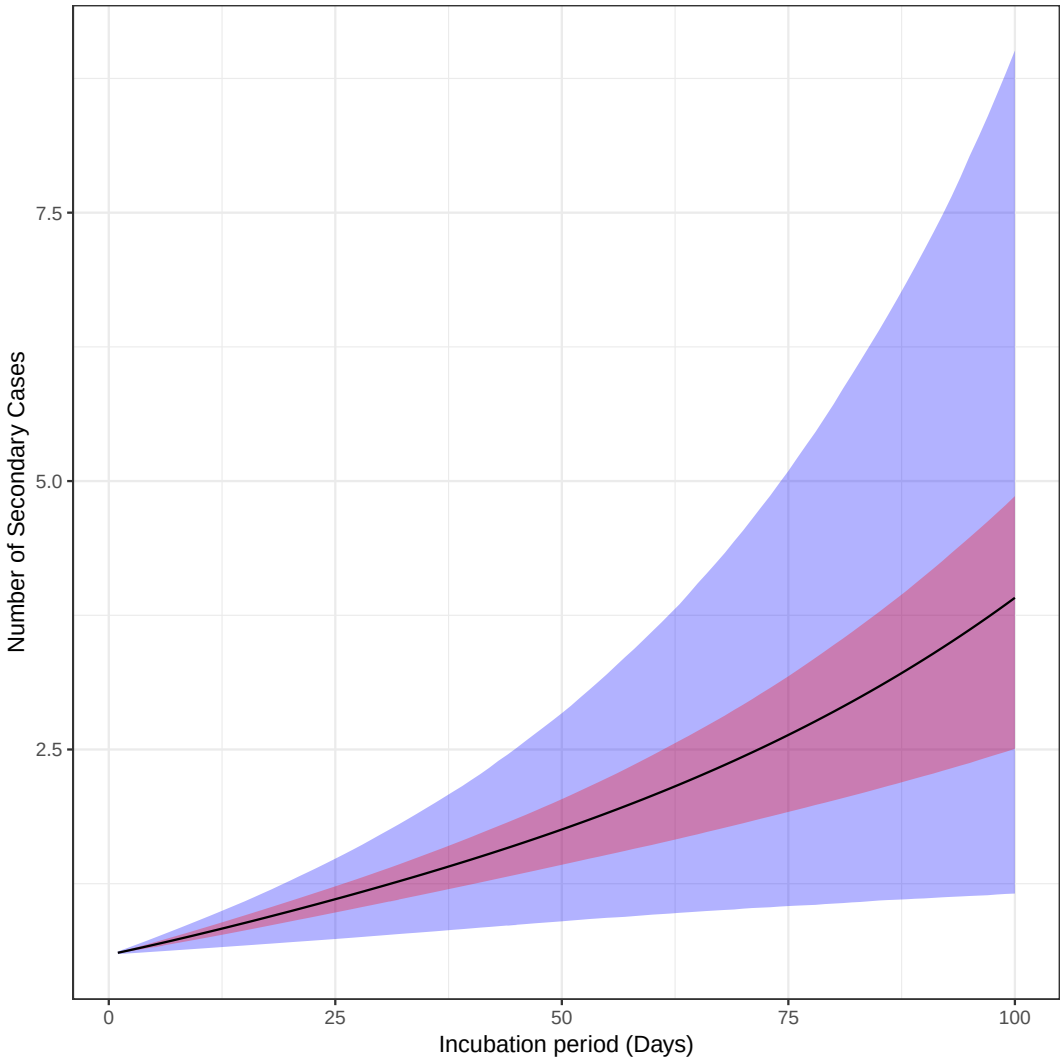

Figure 4: **Effects plot for the zero-inflated negative-binomial model.** The black line shows the expected number of secondary cases as a function of incubation period. The red region is the 50% confidence interval (CI) and the blue region is the 95% CI. Only uncertainty in the slope of the relationship is shown; uncertainty in the intercept is neglected.

We used a GLM to investigate the relationship between incubation period and biting behaviour. Figure 4 shows the inferred relationship with confidence intervals. On average, dogs with longer incubation periods cause more secondary cases.

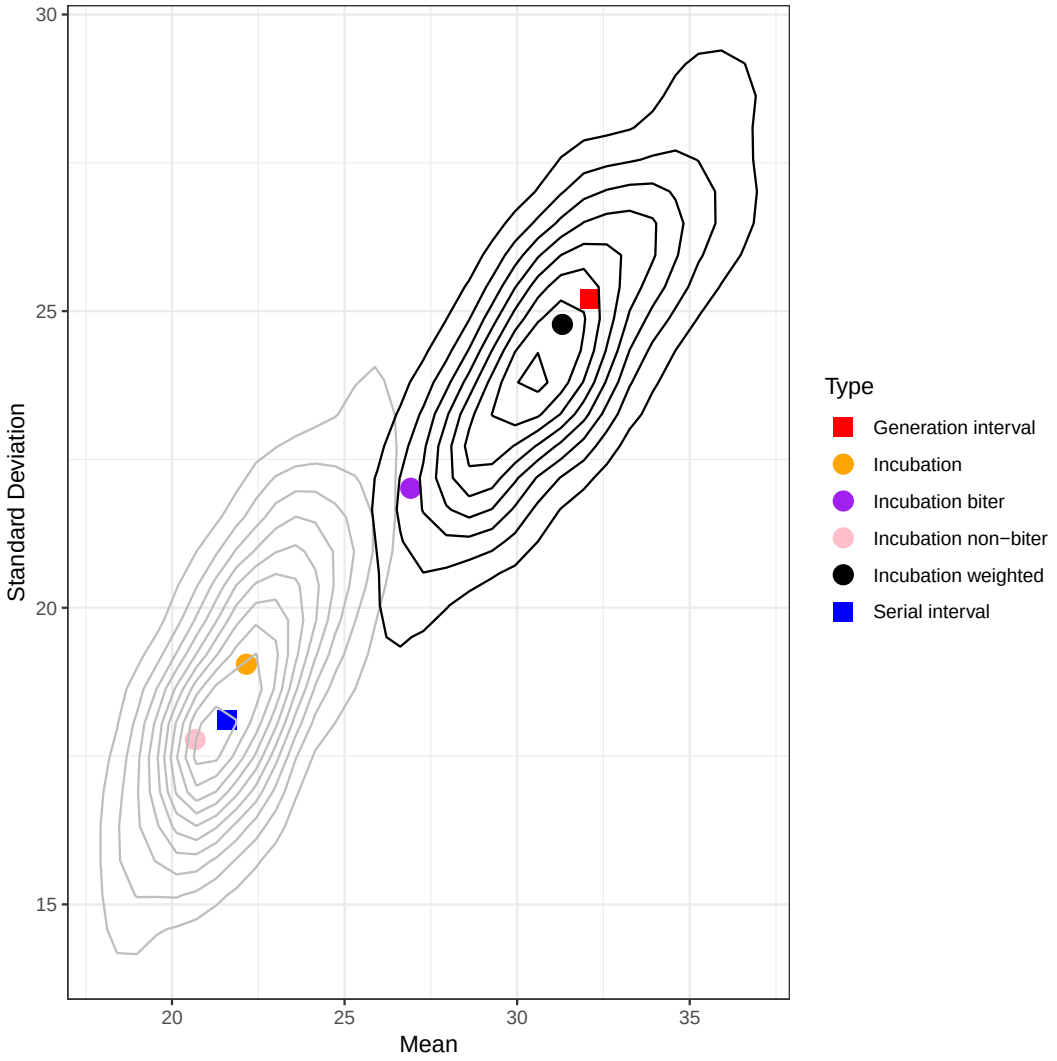

Figure 5: **Correlated and uncorrelated regions.** Prediction regions for the mean and standard deviation of the GI based on uncorrelated (grey) from permutations and correlated (black) zero-inflation negative binomial simulations. Intervals are illustrated with square points and incubation periods with circles.

Figure 5 explores the effects of this observed correlation on simulated intervals. GI distributions from simulated populations with correlations (black contours) have higher mean and standard deviation compared to those from permuted populations without correlations (grey contours). The mean and standard deviation of the generation interval is similar to that for the weighted incubation period, and those of the

serial interval are similar to both the non-biter and overall (unweighted) incubation periods.

Discussion

The link between generation intervals and the reproductive number is an important concept in infectious disease modeling (Park et al., 2019). Reliable estimates of the generation interval distribution are crucial; researchers often overlook important correlation structures. Here, we found that two widely used approaches for estimating generation intervals — summing latent periods and infectious-waiting times (Hampson et al., 2009), or using serial intervals — cannot capture and the observed generation interval distribution in rabies. We used a relatively simple approach to model the relationship between incubation and biting behaviour showed that this relationship explains some of the discrepancy between earlier approaches and observed generation intervals.

In the previous chapter, we estimated r for Serengeti to be 0.23 (95% CI: 0.16, 0.36) per month; the corresponding $\mathcal{R}_e$ estimates (i.e., using $\mathcal{R}_e = \exp(r\bar{G})$ (Park et al., 2019)) estimating $\bar{G}$ from mean serial interval (21.7 days), constructed mean generation interval (24.9 days) and observed generation interval (32.1 days) are 1.18, 1.21, and 1.27 respectively. Although the relative difference between the observed mean serial and generation intervals is large for rabies, the corresponding increase in $\mathcal{R}_e$ estimates is small because r is low.

Reformulating the constructed generation intervals with incubation and symptomatic-waiting times allowed us to compare the difference between generation and serial intervals: generation interval uses the incubation period of the primary host while serial interval uses the incubation period of the secondary host. Compared to other infectious disease systems, rabies is relatively simple conceptually because the latent

and infectious periods are well synchronized. Even in this simple system, where we expected the generation and serial interval to be similar, we find that individual-level correlations instead drive big differences between the two.

References

Anderson, R. M., B. Anderson, and R. M. May (1991). *Infectious diseases of humans: dynamics and control*. Oxford university press.

Andreasen, V., C. Viboud, and L. Simonsen (2008). Epidemiologic characterization of the 1918 influenza pandemic summer wave in Copenhagen: implications for pandemic control strategies. *The Journal of infectious diseases 197*(2), 270–278.

Chunara, R., J. R. Andrews, and J. S. Brownstein (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American journal of tropical medicine and hygiene 86*(1), 39–45.

Cowling, B. J., V. J. Fang, S. Riley, J. M. Peiris, and G. M. Leung (2009). Estimation of the serial interval of influenza. *Epidemiology (Cambridge, Mass.) 20*(3), 344.

Hampson, K., L. Coudeville, T. Lembo, M. Sambo, A. Kieffer, M. Attlan, J. Barrat, J. D. Blanton, D. J. Briggs, S. Cleaveland, P. Costa, C. M. Freuling, E. Hiby, L. Knopf, F. Leanes, F. X. Meslin, A. Metlin, M. E. Miranda, T. Muller, L. H. Nel, S. Recuenco, C. E. Rupprecht, C. Schumacher, L. Taylor, M. A. N. Vigilato, J. Zinsstag, and J. Dushoff (2015). Estimating the global burden of endemic canine rabies. *PLoS Negl Trop Dis 9*, e0003709.

Hampson, K., J. Dushoff, S. Cleaveland, D. T. Haydon, M. Kaare, C. Packer, and A. Dobson (2009). Transmission dynamics and prospects for the elimination of canine rabies. *PLoS biology 7*(3), e1000053.

Majumder, M. S., E. Cohn, D. Fish, and J. S. Brownstein (2016). Estimating a feasible serial interval range for Zika fever. *Bull World Health Organ 10*, 1–6.

Park, S. W., D. Champredon, J. S. Weitz, and J. Dushoff (2019). A practical generation-interval-based approach to inferring the strength of epidemics from their speed. *Epidemics 27*, 12–18.

Svensson, Å. (2007). A note on generation times in epidemic models. *Mathematical biosciences 208*(1), 300–311.

te Beest, D. E., D. Henderson, N. A. van der Maas, S. C. de Greeff, J. Wallinga, F. R. Mooi, and M. van Boven (2014). Estimation of the serial interval of pertussis in Dutch households. *Epidemics 7*, 1–6.

Vynnycky, E. and P. E. Fine (2000). Lifetime risks, incubation period, and serial interval of tuberculosis. *American journal of epidemiology 152*(3), 247–263.

Wallinga, J. and M. Lipsitch (2006). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences 274*(1609), 599–604.

Wallinga, J. and P. Teunis (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of epidemiology 160*(6), 509–516.

White, L. F., J. Wallinga, L. Finelli, C. Reed, S. Riley, M. Lipsitch, and M. Pagano (2009). Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. *Influenza and other respiratory viruses 3*(6), 267–276.

Chapter 5: Reformulating phylogenetic mixed models to improve flexibility and speed

In this chapter, I explored different platforms in modeling trait relationships among different species while account their shared evolutionary history. While a wide range of tools is available for comparative analyses, existing procedures may be either insufficiently flexible or too computationally demanding when analyzing large volumes of data. Using the generalized-linear mixed model framework, I reformulating phylogenetic mixed models to improve flexibility and speed. I demonstrate the method with simulated phylogenies and evolutionary models of varying complexity, as well as real data from several previous studies. This algorithmic approach is general and could be implemented in a wide range of computational platforms, I implemented using the "lme4" and "glmmTMB" R package (the most widely used package for fitting mixed effect models). I also compare our results against existing R packages to explore the limitations of different methods and quantify simulation accuracy and computational efficiency.

The text I present here is a draft of a manuscript planned for submission for publication.

Author Contributions

ML designed the simulation and performed the statistical analyses with helpful feedback from BMB; ML wrote first draft of the manuscript, and BMB revised the manuscript.

Acknowledgements

I thank Morgan Kain for helping me refine and present earlier versions of this work.

Reformulating phylogenetic mixed models to improve flexibility and speed

Michael Li[1*] and Benjamin M. Bolker[1,2,3]

* Corresponding author: Michael Li; lim88@mcmaster.ca

1 Department of Biology, McMaster University, Hamilton, Ontario, Canada
2 Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
3 Institute for Infectious Diseases Research, McMaster University, Hamilton, Ontario, Canada

Abstract

1. Phylogenetic comparative methods (PCM) using phylogenetic regression are a powerful technique to explore relationships among related groups of species traits. However, existing procedures may be either insufficiently flexible or too computationally demanding when analyzing large volumes of data.

2. We propose an alternative formulation of phylogenetic generalized linear mixed models that is mathematically equivalent to previous approaches, but is more flexible in practice. We have implemented this formulation in two R statistical packages (`lme4` and `glmmTMB`).

3. Our reformulation of phylogenetic generalized linear mixed models is computationally efficient, operating orders of magnitude faster than existing methods for fitting phylogenetic mixed models.

4. Our approach can, in principle, be implemented in any platform for generalized mixed models. Our implementation in `lme4` and `glmmTMB` allows users to fit phylogenetic mixed models to a broad range of previously difficult cases (e.g., large data, unbalanced observational designs, complex random effects).

Keywords: phylogenetic comparative methods, phylogenetic correlation, phyloglmm, species–branch matrix

Introduction

Phylogenetic comparative methods (PCMs) are a powerful technique to explore relationships among related groups of species traits. Given a known phylogenetic tree, PCMs explore the relationships among species traits or distributions while taking the underlying evolutionary relationships of the species into account; they can be used to control statistically for phylogenetic relationships, to quantify phylogenetic signal (a measure of the dependence among species responses due to their evolutionary relationships) in trait distributions, or both.

Ever-increasing data collection capabilities (e.g., genomic sequencing, telemetry studies of animal behaviour, or environmental remote sensing), in combination with large-scale synthetic databases of species occurrence and phenotypic traits, are making larger volumes of biological data available over an ever-wider taxonomic range. Researchers use these data to fit complex models describing species occurrence and traits. For example, ecologists have used phylogenetic relationships in multi-species models (Davies et al., 2013; Freckleton et al., 2002; Garland Jr et al., 1992; Ord et al., 2010); more recently they have begun to integrate evolutionary considerations in applied ecological studies addressing biodiversity conservation and the effects of climate change (Lankau et al., 2011; Lavergne et al., 2010; Mace and Purvis, 2008; Santamaría and Mendez, 2012; Winter et al., 2013).

Unlike standard regression-based statistical models, where all of the predictor variables of interest (for example, species traits and environmental factors) are directly observable, PCMs using phylogenetic regression use phylogenetic relationships to estimate the unobserved process of trait evolution (Butler and King, 2004; Felsen-

stein, 1985; Hansen and Bartoszek, 2012). While a wide range of tools is available for phylogenetic regression, existing procedures may be either insufficiently flexible or too computationally demanding when analyzing large volumes of data. In such cases, researchers typically search for ways to simplify their analyses: for example, treating species effects as independent, thus neglecting phylogenetic correlations among species responses (Bunnefeld and Phillimore, 2012); ignoring degrees of relatedness and treating taxon as a strictly hierarchical description (Tella et al., 1999); or neglecting within-species variation (Ord et al., 2010). In this paper, we propose an alternative method for flexibly and efficiently modeling phylogenetic relationships by extending existing software for fitting mixed effect models. This method allows researchers to easily incorporate evolutionary and statistical complexities without sacrificing speed.

Challenges in modeling phylogenetic processes

Standard statistical regression techniques do not allow for correlations among species responses due to their shared evolutionary history. Classic phylogenetic regression uses a statistical model in which phylogenetic correlation in the residuals from a regression between two species-level traits arises because the residual variation in the dependent-variable trait evolves along the branches of the phylogeny according to a Brownian-motion evolutionary model (Felsenstein, 1985). If in addition the residuals are normally distributed and observed without any additional error or within-species variation, Felsenstein's method of phylogenetically independent contrasts (PICS: Felsenstein, 1985; Nicolakakis and Lefebvre, 2000) is sufficient to account for the phylogenetic correlation. More recent approaches – including phylogenetic generalized linear mixed models (PGLMM) (Housworth et al., 2004; Ives and Helmus, 2011), Pagel's λ (Pagel, 1999), and Blomberg's K (Blomberg et al., 2003) — build upon PICs by considering different (non-Gaussian) response distributions and by accounting for evolutionary processes other than Brownian-motion. These

methods partition residual variation into two components: (1) uncorrelated, or independent, residual variation (observation error or tip variation) and (2) phylogenetic signal (evolutionary process error) (Hansen and Bartoszek, 2012; Housworth et al., 2004). If each species' traits are observed more than once, possibly under different conditions, we can potentially distinguish a third level of variation; in this case, phylogenetic variation and tip variation can both be considered part of the evolutionary process error (which we will call tip variation or intercept-level variation) while the among species residual variation is associated with among observation variation within each species. Although many studies include multiple observations per species, phylogenetic analyses rarely take advantage of such information to partition variability more finely. Indeed, many existing methods are restricted to single observations per species, requiring users to collapse multiple observations per species to species mean values.

Classic phylogenetic regressions usually allow the response (trait or distribution) to evolve according to the phylogenetic relationship across species, but the effects of the predictor variables may evolve according to the phylogenetic relationship across species as well. Suppose we have examined a collection of species that came from two groups, and wish to know whether their brain size (Y) is proportional to their body size (X) (Felsenstein's (1985) example using a mixed-effect model. Standard phylogenetic regressions allow for phylogenetic correlations in the intercept of the relationship between body and brain size. However, species within taxonomic groups with similar body sizes may vary in overall brain size, or taxonomic groups may vary in the relationship between brain and body size. Several recent studies have incorporated phylogenetic variation in different ways and looked at species response to phylogenetic variation with changes in environmental factors. For example, Nowakowski et al. (2018) considered phylogenetically correlated slopes in response to habitat conversion when studying the abundance of amphibian species using a Bayesian phylogenetic

Model	Method	Data	Platform
Generalized Linear Model (GLM)	Correlated residual	Single observation	`nlme`:gls, `ape`:pic
	Residual + phylogenetic intercept	Single observation	Pagel's λ Blomberg's k via `nlme`:gls `phylolm`
Generalized Linear Mixed Model (GLMM)	Random effect	Single observation Balanced design	`pez`, `phyr`
		Unrestricted	`lme4`, `glmmTMB`
Bayesian GLMM	Random effect	Balanced design	`MCMCglmm`
		Unrestricted	`brms`

Table 1: List of phylogenetic generalized linear models and R packages.

GLMM, while Li et al. (2017) considered phylogenetically correlated species nested within sites when modeling plant abundance via a phylogenetic GLMM approach. The tools available for extending phylogenetic relationships to predictor variables (or predictor level variation) in a standard frequentist framework are relatively inflexible; thus, many biologist needing to fit random-slopes model have usually turned to more flexible Bayesian approaches, despite their additional computational burden (Bürkner, 2018; Hadfield, 2010). Table 1 summarizes types of platforms, data constraints, and provides model complexities for phylogenetic comparative analysis.

In this paper, we will propose an alternative formulation of phylogenetic regression in a generalized linear mixed modeling framework that is mathematically equivalent to previous approaches, but more flexible. In particular, our new formulation can be implemented in any framework that allows for random effects (such as random intercepts, slopes, and interactions), without the need to implement special correlation structures, by incorporating phylogenetic structures as part of the mean model. We will compare our technique coded in R packages `lme4` and `glmmTMB` with existing R packages (i.e. `nlme` (Pinheiro et al., 2014), `phylolm` (Ho and Ané, 2014), `pez` (Pearse et al., 2015) `phyr` (Li et al, unpublished), `MCMCglmm` (Hadfield, 2010), and `brms` (Bürkner, 2018)), fitting models to data from simulated model that incorpo-

rates phylogenetic variation from both predictors and tips/species as well as residual variation.

Materials and Methods

We generated test data using a simple framework that combines fixed effects, phylogenetic random intercept, slope, and their correlations from a phylogenetic tree, as well as residuals. We then fit the test data from these simulations using our approach implemented using the R packages `lme4` (Bates et al., 2015) and `glmmTMB` (Brooks et al., 2017), using assumptions that match those of the simulation model. We also fit with other platforms, using standard simplifications when necessary for implementation.

Phylogenetic regression

We begin by describing the classic phylogenetic regression in a linear regression setting. Consider a simple linear regression model of observable trait $\mathbf{y}$ as a function of some predictors encoded in a model matrix $\mathbf{X}$, where each species is measured exactly once. The standard phylogenetic regression can be formulated as

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \text{MVN}(0, \sigma^2\mathbf{C}), \end{aligned} \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ response vector; $\mathbf{X}$ is an $n \times m$ model matrix, describing n observations of m predictor variables (phenotypic traits or environmental variables, typically including an intercept column of ones); $\boldsymbol{\beta}$ is an m-vector of coefficients; $\boldsymbol{\epsilon}$ is an $n \times 1$ vector which is assumed to be multivariate normally distributed with mean 0 and variance-covariance matrix given by $\sigma^2\mathbf{C}$ where $\mathbf{C}$ is a $n \times n$ phylogenetic covariance (PC) matrix. The PC matrix is inferred from the topology of the evolutionary tree by quantifying the degree of shared evolution between any pair of taxa (Garamszegi,

2014).

Phylogenetic generalized linear mixed model

Alternatively, one can use the generalized linear mixed effects modeling (GLMM) framework to define a wider range that includes the standard phylogenetic regression as a special case (Lynch, 1991). The generalized linear mixed effect model allows for non-Gaussian responses and uses random effects to flexibly incorporate multiple types of variability. The typical GLMM has the form:

$$\begin{aligned} \mathbf{y} &= \mathcal{D}(g^{-1}(\boldsymbol{\mu}), \phi) \\ \boldsymbol{\mu} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} \\ \mathbf{b} &\sim \text{MVN}(0, \boldsymbol{\Sigma}(\theta)) \end{aligned} \tag{2}$$

where $\mathbf{Z}$ is an $n \times m$ model matrix for the n-dimensional vector-valued m predictor variables; $\mathbf{b}$ (sometimes referred to as the "G-side" effect) representing the conditional modes, is assumed to be multivariate normally distributed with a variance-covariance matrix given by $\boldsymbol{\Sigma}(\theta)$; and ϕ is a scale parameter for the conditional distribution $\mathcal{D}$. When $\mathcal{D}$ is Gaussian, g is the identity function, $\mathbf{Z}$ is the identity matrix, and $\boldsymbol{\Sigma}(\theta) = \sigma^2\mathbf{C}$, (2) reduces to (1).

There are forms of random variation in the mixed model framework. First, random intercepts can allow the response trait to vary independently across groups other than species (e.g., patches, sites, or experiments). Second, random intercepts can also allow the response (trait or distribution) to vary either independently among species (since species represent the tips of the phylogenetic tree) or among species in a phylogenetically correlated way (i.e., species that are closely related tend to have similar responses). The third type of variation that is often neglected is random slopes. Ran-

dom slopes allow fixed effects (the relationship between predictors and responses) to vary among groups. Analogous to phylogenetically correlated variation in intercepts, phylogenetic random slopes can allow the relationship between predictors and responses to vary between species in a phylogenetically correlated way (i.e., similar species will have similar predictor–response relationships).

Reformulating the phylogenetic covariance matrix

Suppose that the evolution follows a Brownian-motion process, i.e., continuous traits evolve independently, following an unbiased, continuous-time random walk, along each branch of the phylogeny. In this case, the phylogenetic variability of a particular species can be written as the sum of the variances of evolutionary changes that occurred on all of the branches in its history. Thus, modeling the evolutionary history of each species with a sequence of independent errors with species–branch matrix $\mathbf{S}$ is equivalent to imposing a correlation $\mathbf{C}$. For example, for the phylogeny in figure 1, the corresponding $\mathbf{S}$ takes the form:

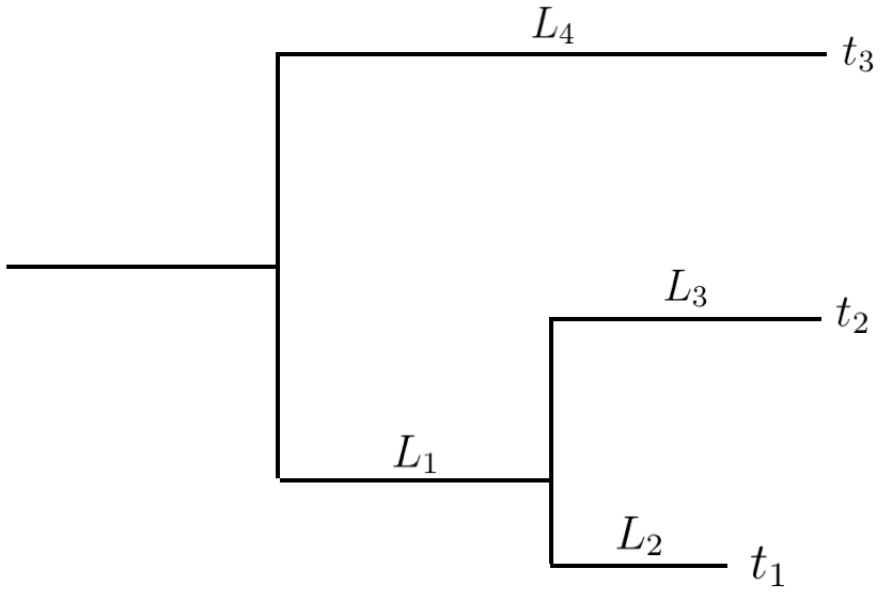

Fig. 1: Three-species phylogenetic tree.

$$\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array} \begin{array}{c} \begin{array}{cccc} L_1 & L_2 & L_3 & L_4 \end{array} \\ \begin{pmatrix} \ell_1 & \ell_2 & 0 & 0 \\ \ell_1 & 0 & \ell_3 & 0 \\ 0 & 0 & 0 & \ell_4 \end{pmatrix} \end{array} \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

The phylogenetic variability corresponding to species 1 is $\ell_1\epsilon_1 + \ell_2\epsilon_2$, where $\ell_i = \sqrt{L_i}$, the square root of the branch length L_i in figure 1, and the ϵ_i are independent Normal variates with zero mean and σ^2 evolutionary variance (i.e. the variance for species 1 is $\mathrm{E}[(\ell_1\epsilon_1 + \ell_2\epsilon_2)^2] = (L_1 + L_2)\sigma^2$).

Constructing the species–branch random effects model matrix

The $\mathbf{S}$ matrix is the product of an $m \times b$ indicator matrix $\mathbf{S}_{ind}$ of branch indices and a vector $\boldsymbol{\ell}$ of square roots of branch lengths:

$$\mathbf{S}_{ind} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad \boldsymbol{\ell} = \begin{bmatrix} \ell_1 \\ \ell_2 \\ \ell_3 \\ \ell_4 \end{bmatrix}.$$

$\mathbf{S}_{ind}$ is a binary (indicator) matrix that describes whether a particular branch occurs in the history of a focal species. $\mathbf{S}\mathbf{S}^T$ gives the variance-covariance matrix of the phylogeny.

In general, the random-effect model matrix $\mathbf{Z}$ for a GLMM can be decomposed into term-wise model matrices $\mathbf{Z}_i$ as described in Bates et al. (2015). Analogous to the procedure described in Bates et al. (2015), the phylogenetic correlated random-effect matrix $\mathbf{Z}_i^C$ is

$$\mathbf{Z}_i^C = (\mathbf{S}^\top \mathbf{J}_i^\top * \mathbf{X}_i^\top)^\top, \tag{3}$$

where $\mathbf{S}$ is the $m \times b$ species–branch matrix; $\mathbf{J}_i$ is the $n_i \times m$ indicator matrix of grouping factors; $\mathbf{X}_i$ is the $n \times p_i$ raw random-effects model matrix; and $*$ is the Khatri-Rao product (Khatri and Rao, 1968) partitioned at the observation level (n).

For example, using the phylogeny above (figure 1), if we begin with a raw model matrix corresponding to a random-slope model,

$$\mathbf{X} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \end{bmatrix}$$

then the term-wise phylogenetic random effects model matrix is,

$$\mathbf{Z}_i^C = (\mathbf{S}^\top \mathbf{J}_i^\top * \mathbf{X}_i^\top)^\top = \left[\left(\begin{bmatrix} \ell_1 & \ell_1 & 0 \\ \ell_2 & 0 & 0 \\ 0 & \ell_3 & 0 \\ 0 & 0 & \ell_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right) * \begin{bmatrix} 1 & 1 & 1 \\ t_1 & t_2 & t_3 \end{bmatrix} \right]^\top \tag{4}$$

$$= \begin{bmatrix} \ell_1 & \ell_1 t_1 & \ell_2 & \ell_2 t_1 & 0 & 0 & 0 & 0 \\ \ell_1 & \ell_1 t_2 & 0 & 0 & \ell_3 & \ell_3 t_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ell_4 & \ell_4 t_3 \end{bmatrix}.$$

Simulation

Single group model

We generated test data based on the random slopes mixed model formulation (2) with a single response variable $\mathbf{y}$ and a single continuous normally distributed predictor variable $\boldsymbol{t}$ for $n = 25$, 50, and 100 species. For simplicity, the response variable $\mathbf{y}$ is conditionally normally distributed (i.e., $\boldsymbol{D}$ is a Gaussian distribution, and g is the identity link function), corresponding to a linear mixed effect model. For the first set of simulations, we simulate one observation per species. Thus, the full simulation model is as follows:

$$\begin{aligned} \mathbf{y} &= (\beta_0 + b_{\text{phy}_{\text{int}}}) + (\beta_1 + b_{\text{phy}_{\text{slope}}})\boldsymbol{t} + \boldsymbol{\epsilon} \\ (b_{\text{phy}_{\text{int}}}, b_{\text{phy}_{\text{slope}}}) &\sim \text{MVN}\left(0, \begin{bmatrix} \Sigma^2_{\text{phy}_{\text{int}}} & \Sigma_{\text{phy}_{\text{int-slope}}} \\ \Sigma_{\text{phy}_{\text{int-slope}}} & \Sigma^2_{\text{phy}_{\text{slope}}} \end{bmatrix}\right) \\ \boldsymbol{\epsilon} &\sim \text{N}(0, \sigma^2_\epsilon). \end{aligned} \tag{5}$$

The model contains two fixed effect parameters (β_0 and β_1), three random effect parameters (phylogenetic random intercept variance $\Sigma^2_{\text{phy}_{int}}$, phylogenetic random slope

variance $\Sigma^2_{\text{phy}_{slope}}$ and covariance between phylogenetic random slope and intercept $\Sigma_{\text{phy}_{\text{int}-\text{slope}}}$) and residual variance (σ^2_ϵ). The covariance between phylogenetic random intercept and slope measures the correlation of phylogenetic dependency variability in regression effect ($b_{\text{phy}_{\text{slope}}}$) and response ($b_{\text{phy}_{\text{int}}}$); i.e. if a positive correlation indicates that similar species have similar relative intercepts and slopes. Predictor-level and intercept-level random effects of species are not applicable in this simulation setting because there is only a single observation per species, so within-species variation cannot be separated from tip variation.

Multi-group model

We extend the simulation model by adding multiple groups where each group has one observation per species. The multi-group model is a generalization of multiple-site models used in community ecology to model phylogenetic attraction (Helmus et al., 2007). The full multi-group model is as follows:

$$
\begin{aligned}
\mathbf{y} &= (\beta_0 + b_{\text{phy}_{\text{int}}} + b_{\text{sp}_{\text{int}}} + b_{\text{group}}) + (\beta_1 + b_{\text{phy}_{\text{slope}}} + b_{\text{sp}_{\text{slope}}})\boldsymbol{t} + b_{\text{sp:group}} + \boldsymbol{\epsilon} \\
(b_{\text{phy}_{\text{int}}}, b_{\text{phy}_{\text{slope}}}) &\sim \text{MVN}\left(0, \begin{bmatrix} \Sigma^2_{\text{phy}_{\text{int}}} & \Sigma_{\text{phy}_{\text{int}-\text{slope}}} \\ \Sigma_{\text{phy}_{\text{int}-\text{slope}}} & \Sigma^2_{\text{phy}_{\text{slope}}} \end{bmatrix}\right) \\
(b_{\text{sp}_{\text{int}}}, b_{\text{sp}_{\text{slope}}}) &\sim \text{MVN}\left(0, \begin{bmatrix} \sigma^2_{\text{sp}_{\text{int}}} & \sigma_{\text{sp}_{\text{int}-\text{slope}}} \\ \sigma_{\text{sp}_{\text{int}-\text{slope}}} & \sigma^2_{\text{sp}_{\text{slope}}} \end{bmatrix}\right) \\
b_{\text{group}} &\sim \text{MVN}(0, \sigma^2_{\text{group}}) \\
b_{\text{sp:group}} &\sim \text{MVN}(0, \text{I}_{\text{group}} \otimes \Sigma^2_{phy}) \\
\boldsymbol{\epsilon} &\sim \text{N}(0, \sigma^2_\epsilon),
\end{aligned}
\tag{6}
$$

where I_{group} is a indicator matrix of groups; and $\otimes$ is the Kronecker product.

The multi-group full simulation model has five additional random effects (predictor-level ($\sigma^2_{\text{sp}_{\text{slope}}}$) and intercept-level ($\sigma^2_{\text{sp}_{\text{int}}}$) random effect of species variance and their

covariance ($\sigma_{\text{sp}_{\text{int}-\text{slope}}}$), random intercept of group (b_{group}) and random intercept of species-group interaction ($b_{\text{sp:group}}$)) compared to the single–group full model. Predictor-level and intercept-level random effects of species are applicable in the multi-group model setting because there are multiple observations per species; thus we can quantify variation among species separately from residual variation. Variance in the intercept of species–group interactions ($b_{\text{sp:group}}$) describes whether the species within a group have more similar responses on average than expected by chance, equivalent to phylogenetic attraction (Helmus et al., 2007).

Platforms

We compare our approach with five other R packages that can fit phylogenetic comparative models: `nlme` (Pinheiro et al., 2014), `phylolm` (Ho and Ané, 2014), `pez` (Pearse et al., 2015), and `brms` (Bürkner, 2018). Phylogenetic generalized least squares (PGLS) (`gls` in `nlme`) is one of the most widely used techniques in phylogenetic comparative analysis; it fits a linear model where the covariance structure between species assumes an evolutionary process on the tree (typically Brownian-motion, but other processes can be used) instead of treating the residual error for each species as independent. Phylogenetic generalized linear models (PGLM) (`phyloglm` in `phylolm`) are a slightly more flexible variation of PGLS that can allow for both phylogenetic and residual variation, as well as non-Gaussian response variables. Both `gls` and `phylolm` can model non-Brownian evolutionary processes and different correlation structures (e.g., Pagel's λ or Blomberg's K), but we restrict our PGLS fits to the simple BM correlation. Neither PGLS nor PGLM can handle random slopes or multiple observations within a species. One of the few packages that currently fit phylogenetic correlations to predictor level variation is `pez` (and very recently `phyr`), which can handle additional random slopes ($\Sigma^2_{\text{phy}_{\text{slope}}}$) and random intercept of species-group interactions ($b_{\text{sp:group}}$) but does not incorporate covariation between phylogenetic random slope-

Package	nlme	phylolm	lme4/glmmTMB	pez	phyr	brms	MCMCglmm
Single Group	X	X	X			X	X
Phylo Intercept	X	X	X			X	X
Phylo Slope			X			X	X
Phylo Slope-Intercept correlation			X			X	X
Residual		X	X			X	X
Multi-group			X	X	X	X	X
Phylo Intercept			X	X	X	X	X
Phylo Slope			X	X	X	X	X
Phylo Slope-intercept correlation			X			X	X
Phylo Species-group interaction			X	X	X	X	
Species intercept			X	X	X	X	X
Species Slope			X	X	X	X	X
Species Slope-intercept correlation			X			X	X
Residual			X	X	X	X	X

Table 2: List of estimable parameters for each R package.

intercept ($\Sigma_{\text{phy}_{\text{int}-\text{slope}}}$). Lastly, Bayesian phylogenetic GLMMs using Markov chain Monte Carlo (MCMC) can handle all of the cases described above. However, MCMC is usually much more computationally expensive for GLMMs compared to platforms using deterministic optimization. `MCMCglmm` (Hadfield and Nakagawa, 2010) is the most widely used Bayesian phylogenetic GLMM, but we will instead use `brms`, which uses a more computationally efficient MCMC technique called Hamiltonian Monte Carlo (HMC)(Duane et al., 1987).

Simulation and evaluations

We simulated 100 phylogenetic trees for each sample size (n = 25, 50, 100 and an additional n = 500 for the multi-group model) to and then modelled responses on each tree using (5, 6). Each realization was fitted using all model variants. All simulation parameters are shown in Figure 2 and Figure 5. Correlation ρ is used in place of covariances in the simulations. Table 2 shows the parameters that are estimable for each platform. We only evaluated the goodness of fit for model fits that passed the convergence tests implemented by the package. For Bayesian GLMM, we included only realizations where we require values of the Gelman-Rubin statistic less than the recommended threshold 1.1. Based on recent concerns about Gelman-

Rubin thresholds (Vats and Knudson, 2018)), we included the additional convergence criterion of effective sample size (ESS) > 1000 for the fixed effect parameters (β_0 and β_1) for each replication (Vehtari et al., 2019). For each replication, we sample two chains starting with 10000 iterations. We first evaluate our estimates by looking at the distribution of the estimated values (maximum likelihood estimates for non-Bayesian platforms and posterior medians for Bayesian platforms) to quantify bias and variance (i.e., quality of the point estimate). We then compute frequentist coverage to assess the quality of the confidence intervals. Coverage refers to the proportion of simulations in which the computed confidence intervals include the true values of parameters. We used 95% Wald confidence intervals for deterministic methods and quantile-based intervals (for Bayesian GLMM) to evaluate coverage. We also compare computational speed between different platforms to evaluate the efficiency of different platforms and methods.

Results

We used our method to reproduce the examples in chapter 11 of Garamszegi (2014) using phylogenetic GLMMs based on `lme4` and `glmmTMB` (for more details see example in supplement). We also used our method and fitted the full model of the dune meadow data recently used with `pez` (Li et al., 2017). `lme4` and `pez` give identical results for fixed and random effect estimates, but our code runs approximately 120 times faster than `pez`. Codes for analyzing the dune meadow data using `lme4` and `pez` are provided in the supplements.

Single Group model simulations

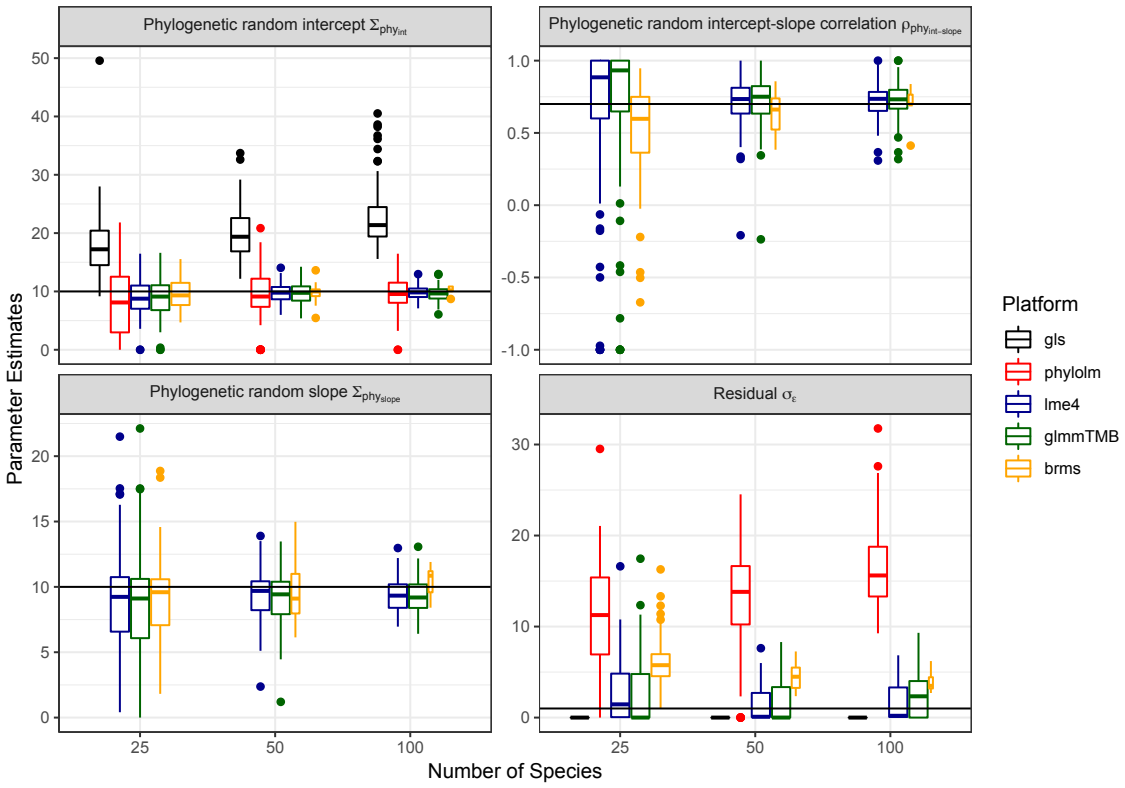

Fig. 2: Comparison of single group model parameter estimates across different R packages in Table 2. Total simulations $N = 100$ for each category. The horizontal line shows the true value of the parameters in the simulation model. Models capable of fitting all parameters (`lme4`, `glmmTMB`, and `brms`) fit well for all parameters.

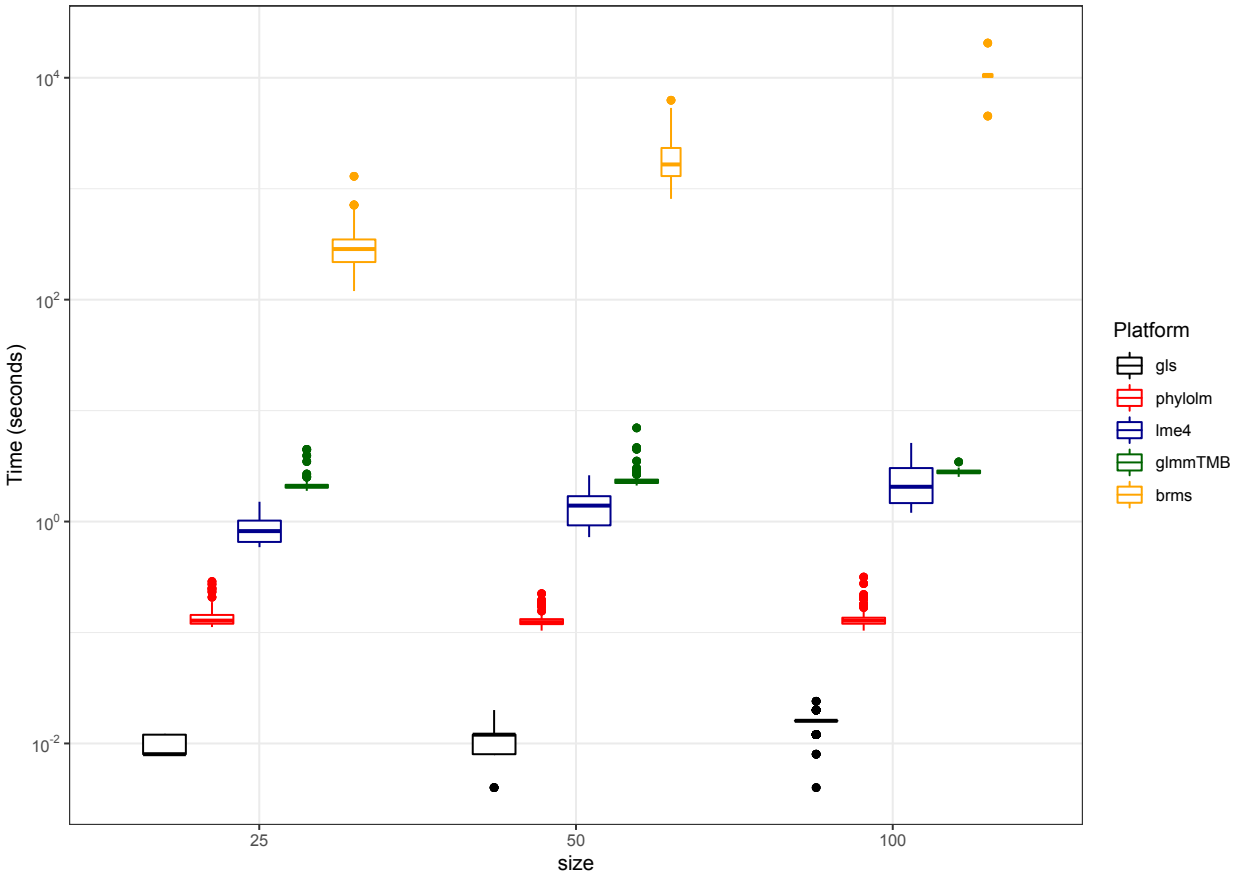

Fig. 3: Comparison of single-group model computational speed.

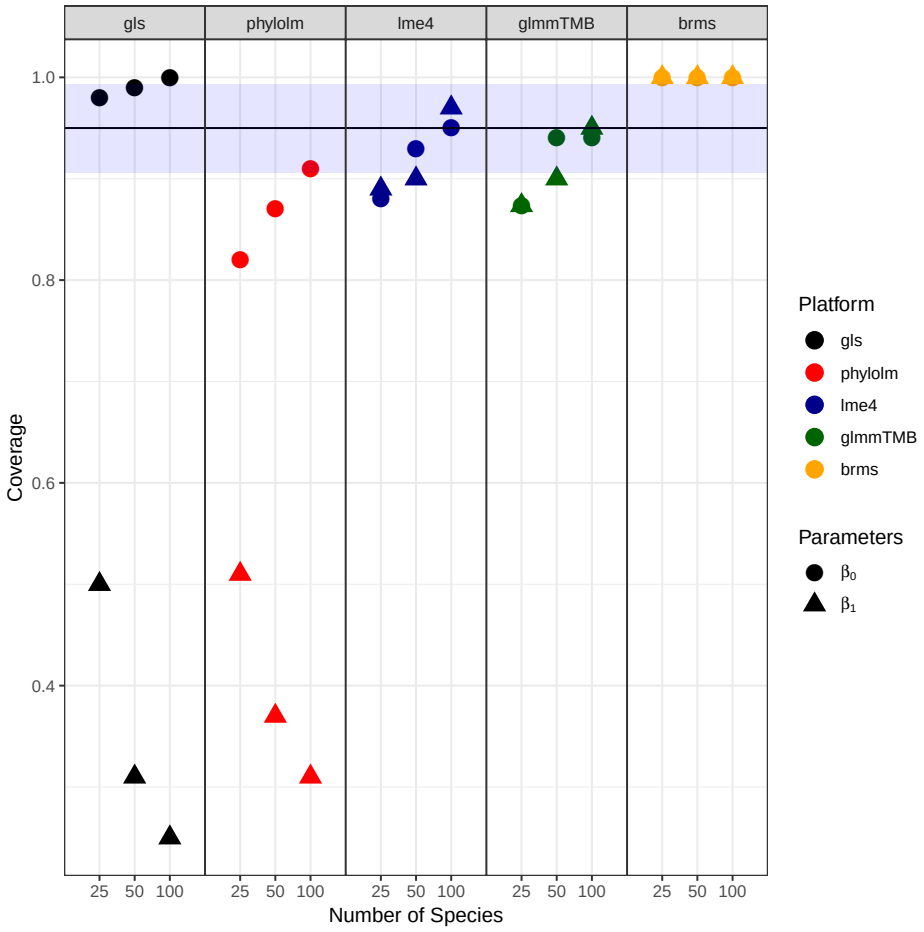

Fig. 4: Comparison of coverage probability for fixed effect parameters. Models matching the simulation model (`lme4`, `glmmTMB` and pkgbrms) have coverage near the nominal value of 0.95. The black line shows the nominal coverage, and the blue ribbon the 95% binomial confidence interval based on 100 simulated fits.

The full fitted model (which matches the simulation model that incorporates phylogenetic intercept, slope, and correlation) provides estimates with low bias (average difference between the estimated parameters and the true simulation parameters) for all parameters. Estimates for fixed effect parameters (β_0 and β_1) approach nominal coverage as the number of species increases for `lme4` and `glmmTMB` but not for other packages. `brms` has higher than nominal coverage (i.e., its confidence intervals are overly conservative) because the prior distributions for the simulation parameters are

centered at the true values (Li et al. (2018) discuss the interaction of informative priors and Bayesian calibration).

In general, models that are insufficiently flexible to match the true simulation model (PGLM and PGLS) will try to fit the data with the parameters available. PGLM (which lacks the phylogenetic slope parameter) provides reasonably good estimates for the phylogenetic intercept standard deviation parameter ($\sigma_{\text{phy}_{\text{int}}}$) but overestimates the residual standard deviation; the estimates for the intercept (β_0) are slightly overconfident (the coverage $\approx 90\%$ with 100 species) and the fixed slope parameter (β_1) has poor coverage ($< 60\%$). PGLS, with only one parameter available, confounds all variation (phylogenetic intercept, slope and residual variation) into the phylogenetic intercept parameter, resulting in overestimating the phylogenetic intercept and over-covering for β_0, and under-covering for β_1.

Multi-group model simulations

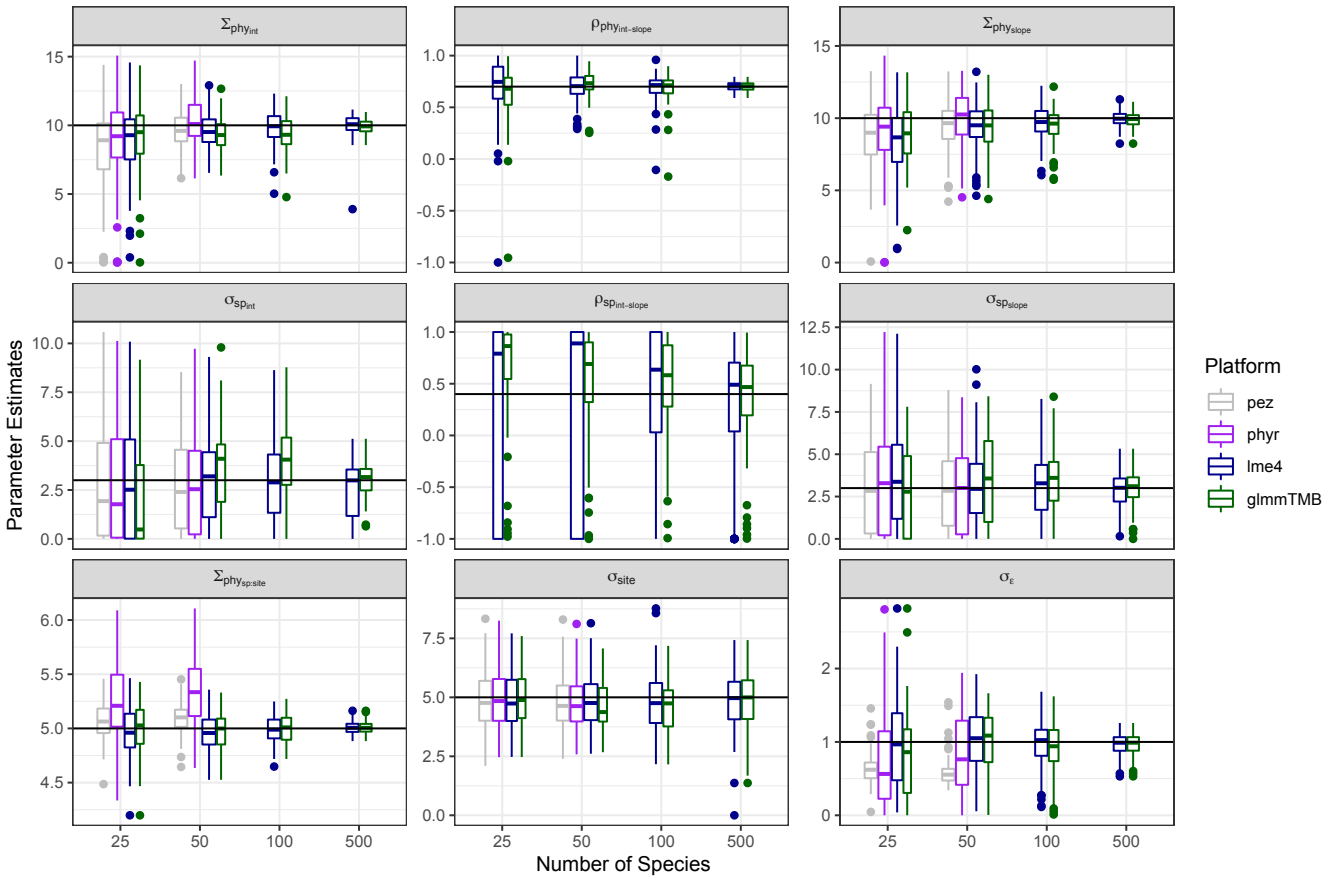

Fig. 5: Comparison of multi-group model parameter estimates. The horizontal line shows the true value of the parameters in the simulation model. Models capable of fitting all parameters (lme4 and glmmTMB) fit well for all parameters. `pez` and `phyr` estimates for n = 100, and 500 are not available because the models did not converge within 30 minutes.

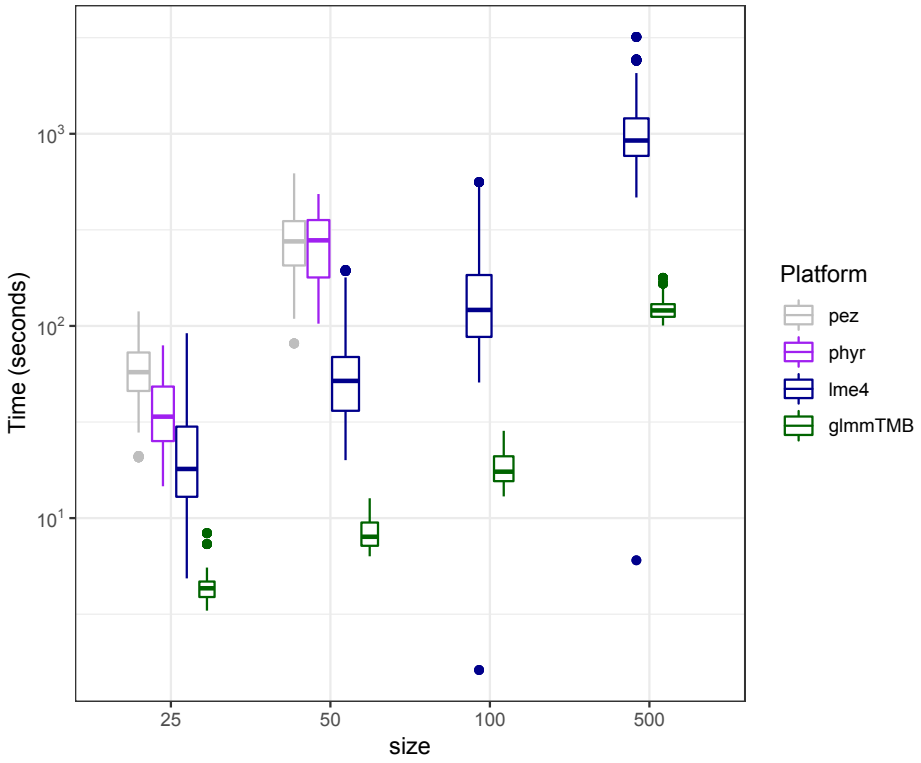

Fig. 6: Comparison of multi-group model computational speed.

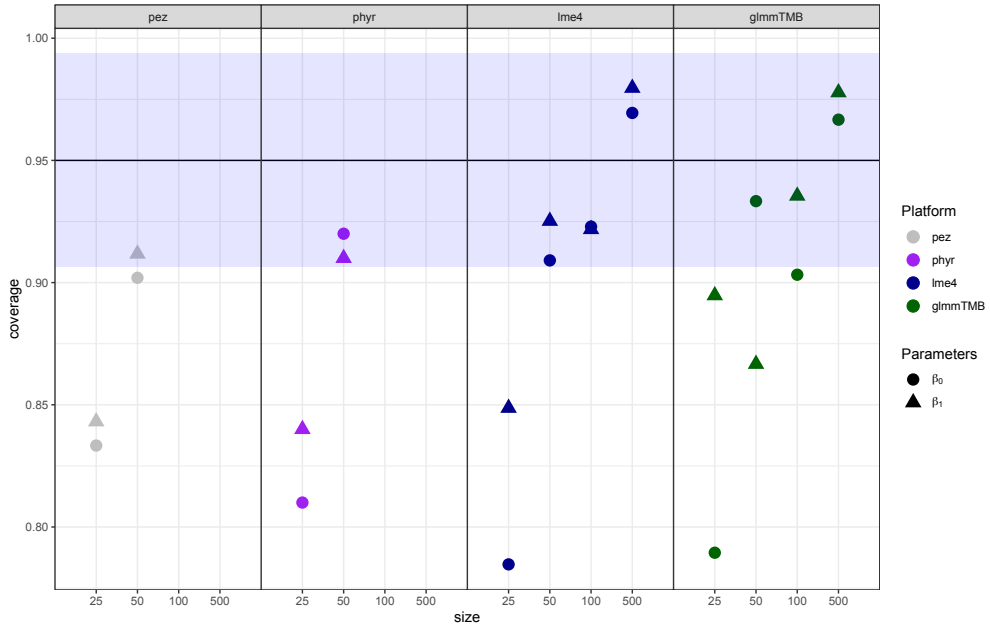

Fig. 7: Comparison of multi-group model coverage.

In contrast, the multi-group model fits are much more similar across platforms (only the more powerful platforms can fit these models at all, and the fitting models are closer to the true simulation model). Similar to the single-group fits, `lme4` and `glmmTMB` match the simulation model well and provide good estimates for all parameters except the correlation ($\sigma_{\mathrm{sp}_{\mathrm{int-slope}}}$) for small numbers of species (i.e. n = 25 and 50). The lack of correlations in `pez` and `phyr`'s statistical models does not appear to have a large effect on the estimates of the remaining parameters in the model but underestimates the residual standard deviation (Figure 5).

Although the parameter estimates are similar across platforms for the multi-group simulation fits, computational efficiency varies enormously across platforms and sample size. For example, the new formulation is implemented in both `lme4` and `glmmTMB`, but `glmmTMB` is almost an order of magnitude faster than `lme4` for the cases studied here. Comparing `glmmTMB` to `pez` and `phyr`, the median time for `glmmTMB` to fit 50 species model is $\approx$ 9 versus $\approx$ 200 seconds for `pez` and `phyr` respectively. `glmmTMB` takes $\approx$ 125 seconds to fit a 500-species model; it was not practical for us to fit 500-species models with `pez` and `phyr`, because computational speed scaled faster than linearly with sample size. `glmmTMB` is almost an order of magnitude faster than `lme4`.

Discussion

We have simulated relatively complex models containing phylogenetic variation in both intercepts and slopes, as well as within-species variation that is quantifiable because we allow multiple observations per species. These models are intrinsically more complex than some simple platforms for phylogenetic regression can handle, which may seem unfair; nevertheless, our models are certainly *less* complex than evolutionary processes occurring in nature. Our results show that models that cannot

match the full "simulation world" perform poorly even for the parameters they do estimate; it is important to understand the limitations of these simpler, commonly used methods.

Even our relatively simple models can incorporate many layers of complexity — e.g. multiple spatial grouping variables as well as correlated phylogenetic variation in the effects of several different traits and environmental variables on a focal trait. In theory, as long as we have enough data and enough computational power, models that can incorporate more of the complexity will always describe a biological system better. However, real applications are always data-constrained. Deciding on a practically appropriate level of model complexity for a given problem and data set is an open and difficult general problem in statistical modeling, not just in phylogenetic studies. Should one use simple models that may be overly conservative or risk overfitting by using more ambitious models? How can one appropriately use the data themselves to choose model complexity (Roberts et al., 2016)? What are the relative costs and benefits of using a step-down procedure starting from the most complex possible model (Barr et al., 2013), choosing simpler models *a priori* (Baayen et al., 2008), or using Bayesian approaches with regularizing priors (Hadfield, 2010)?

Incorporating different levels of variation

In classic GLMMs, random effects are used to handle group (or individual) level variation. In the simplest experimental design with continuous response observations in different levels of a discrete grouping variable, where the response may vary among levels of the grouping variable, fitting random intercepts are the "go-to" method to handle this variation. The random intercept model controls the group effect in the response level by allowing different intercepts for each level of the grouping variable. However, random effects can take more complicated forms as the experimental design becomes more complex. For example, imagine observing another continuous explana-

tory variable in the experimental design above, where the relationship of the response and the new explanatory variable may vary according to the grouping variable. A random slopes model, which allows different slopes (the relationship between continuous variables) for each level of the grouping variable is most appropriate to handle this type of variation. Random slopes models require appropriate observational or experimental designs (i.e., multiple measurements of traits and responses within each evolutionary group) and often require more data overall for reliable estimates, but they are relevant over a wide range of scenarios (Cleasby et al., 2015; Ord et al., 2010; Schielzeth and Forstmeier, 2008). Neglecting random slopes can lead to biased fixed effect estimates with inadequate coverage and type I errors (Schielzeth and Forstmeier, 2008) as shown in the simulations above.

Nevertheless, it is hard to account for all forms of complexities and decide if when it is best to use phylogenetic random effects, simple grouping, or both. Optimal model complexity depends on experimental design and whether the data provides enough signal to estimate these different levels of variation, which can be strongly confounded. For example, for experimental designs with single measurement per species, any method that can account for at least two sources of variation, such as Pagel's λ will be sufficient. However, if multiple observations are available per species, then these simple methods may confound tip variation with residual variation. In this case, multiple observations can summarized to a single measurement (for example, weighted-mean) per species to avoid confounding residual and tip variation. This is equivalent to assuming homoscedasticity using inverse-variance weights for unbalanced datasets. Alternatively, when the within-species variance is actually of interest, accounting for within-species variation (i.e., adding species-level random effects in our example) can automatically handle multiple observations per species.

It may be easier to be conservative to include both (phylogenetic random effects and grouping random effects) of them but simplify the phylogenetic relationships (at

the random slopes level) and think about the random-slopes model in a strictly hierarchical setting (i.e., estimating different slopes for each family, or taxon (Bunnefeld and Phillimore, 2012)) - the PGLMM collapses to a standard random-slopes model.

Another simplifying alternative is to be conservative to include both phylogenetic random effects and grouping random effects but simplify the phylogenetic relationships and think about the random-slopes model in a strictly hierarchical setting (i.e., estimating different slopes for each family, or taxon (Bunnefeld and Phillimore, 2012)). However, this collapses phylogenetic structures to a standard random-slopes model. Users should be aware of two essential questions when fitting random-slope models: How much data do we need in order to practically estimate the random slopes? Are we making a mistake by ignoring random slopes (Schielzeth and Forstmeier, 2008)?

Extension and alternatives

We have presented a range of classical phylogenetic comparative methods (i.e. phylogenetic least squares, linear and mixed models) to fit models that incorporate different levels of phylogenetic variation. Even within this scope, there is additional room for exploration, such as phylogenetic multivariate response models; non-Brownian evolutionary processes such as the Ornstein-Uhlenbeck (OU) model which accounts for both selection and drift processes (Butler and King, 2004)); Bayesian approaches (Hadfield and Nakagawa, 2010); and variable-rate model, where evolutionary parameters vary across the phylogeny. However, the simple approach we developed here offers an efficient way to handle phylogenetic comparative analysis for a wide range of univariate, Brownian-motion evolutionary models. This approach can in principle be combined flexibly with any platforms that supports independent latent variables such as Stan. More importantly, this implementation in `lme4` and `glmmTMB` allows users to fit phylogenetic mixed models to the fullest (large data, unbalanced species observations, complex random effects) and explore new ideas.

Authors' contributions

ML and BMB conceived the ideas and designed methodology; ML and BMB implemented the code in `lme4` and `glmmTMB`; ML ran all simulations; ML and BMB analyzed the results; ML wrote the first draft of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data Availability

All codes are available at DOI:10.5281/zenodo.2639887.

References

Baayen, R. H., D. J. Davidson, and D. M. Bates (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language 59*(4), 390–412.

Barr, D. J., R. Levy, C. Scheepers, and H. J. Tily (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language 68*(3), 255–278.

Bates, D., M. Mächler, B. Bolker, S. Walker, et al. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(i01).

Blomberg, S. P., T. Garland Jr, and A. R. Ives (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution 57*(4), 717–745.

Brooks, M. E., K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Maechler, and B. M. Bolker (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal 9*(2), 378–400.

Bunnefeld, N. and A. B. Phillimore (2012). Island, archipelago and taxon effects: mixed models as a means of dealing with the imperfect design of nature's experiments. *Ecography 35*(1), 15–22.

Butler, M. A. and A. A. King (2004). Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist 164*(6), 683–695.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal 10*(1), 395–411.

Cleasby, I. R., S. Nakagawa, and H. Schielzeth (2015). Quantifying the predictability of behaviour: statistical approaches for the study of between-individual variation in the within-individual variance. *Methods in Ecology and Evolution 6*(1), 27–37.

Davies, T. J., E. M. Wolkovich, N. J. Kraft, N. Salamin, J. M. Allen, T. R. Ault, J. L. Betancourt, K. Bolmgren, E. E. Cleland, B. I. Cook, et al. (2013). Phylogenetic conservatism in plant phenology. *Journal of Ecology 101*(6), 1520–1530.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid Monte Carlo. *Physics Letters B 195*(2), 216–222.

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist 125*(1), 1–15.

Freckleton, R. P., P. H. Harvey, and M. Pagel (2002). Phylogenetic analysis and comparative data: a test and review of evidence. *The American Naturalist 160*(6), 712–726.

Garamszegi, L. Z. (2014). *Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice*. Springer.

Garland Jr, T., P. H. Harvey, and A. R. Ives (1992). Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Systematic biology 41*(1), 18–32.

Hadfield, J. and S. Nakagawa (2010). General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of evolutionary biology 23*(3), 494–508.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software 33*(2), 1–22.

Hansen, T. F. and K. Bartoszek (2012). Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. *Systematic Biology 61*(3), 413–425.

Helmus, M. R., K. Savage, M. W. Diebel, J. T. Maxted, and A. R. Ives (2007). Separating the determinants of phylogenetic community structure. *Ecology letters 10*(10), 917–925.

Ho, L. S. T. and C. Ané (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology 63*, 397–408.

Housworth, E. A., E. P. Martins, and M. Lynch (2004). The phylogenetic mixed model. *The American Naturalist 163*(1), 84–96.

Ives, A. R. and M. R. Helmus (2011). Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs 81*(3), 511–525.

Khatri, C. and C. R. Rao (1968). Solutions to some functional equations and their applications to characterization of probability distributions. *Sankhyā: The Indian Journal of Statistics, Series A*, 167–180.

Lankau, R., P. S. Jørgensen, D. J. Harris, and A. Sih (2011). Incorporating evolutionary principles into environmental management and policy. *Evolutionary Applications 4*(2), 315–325.

Lavergne, S., N. Mouquet, W. Thuiller, and O. Ronce (2010). Biodiversity and climate change: integrating evolutionary and ecological responses of species and communities. *Annual review of ecology, evolution, and systematics 41*, 321–350.

Li, D., A. R. Ives, and D. M. Waller (2017). Can functional traits account for phylogenetic signal in community composition? *New Phytologist 214*(2), 607–618.

Li, M., J. Dushoff, and B. M. Bolker (2018). Fitting mechanistic epidemic models to data: a comparison of simple Markov chain Monte Carlo approaches. *Statistical methods in medical research 27*(7), 1956–1967.

Lynch, M. (1991). Methods for the analysis of comparative data in evolutionary biology. *Evolution 45*(5), 1065–1080.

Mace, G. M. and A. Purvis (2008). Evolutionary biology and practical conservation: bridging a widening gap. *Molecular Ecology 17*(1), 9–19.

Nicolakakis, N. and L. Lefebvre (2000). Forebrain size and innovation rate in European birds: feeding, nesting and confounding variables. *Behaviour 137*(11), 1415–1429.

Nowakowski, A. J., L. O. Frishkoff, M. E. Thompson, T. M. Smith, and B. D. Todd (2018). Phylogenetic homogenization of amphibian assemblages in human-altered habitats across the globe. *Proceedings of the National Academy of Sciences*, 201714891.

Ord, T. J., J. A. Stamps, and J. B. Losos (2010). Adaptation and plasticity of animal communication in fluctuating environments. *Evolution 64*(11), 3134–3148.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature 401*(6756), 877.

Pearse, W. D., M. W. Cadotte, J. Cavender-Bares, A. R. Ives, C. M. Tucker, S. C. Walker, and M. R. Helmus (2015). Pez: Phylogenetics for the environmental sciences. *Bioinformatics 31*(17), 2888–2890.

Pinheiro, J., D. Bates, S. DebRoy, and D. Sarkar (2014). R core team (2014) nlme: linear and nonlinear mixed effects models. r package version 3.1-117. *Available at h ttp://CRAN. R-project. org/package= nlme*.

Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann (2016, December). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*.

Santamaría, L. and P. F. Mendez (2012). Evolution in biodiversity policy–current gaps and future needs. *Evolutionary Applications 5*(2), 202–218.

Schielzeth, H. and W. Forstmeier (2008). Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology 20*(2), 416–420.

Tella, J. L., G. Blanco, M. G. Forero, Á. Gajón, J. A. DONAzar, and F. Hiraldo (1999). Habitat, world geographic range, and embryonic development of hosts explain the prevalence of avian hematozoa at small spatial and phylogenetic scales. *Proceedings of the National Academy of Sciences 96*(4), 1785–1789.

Vats, D. and C. Knudson (2018). Revisiting the Gelman-Rubin diagnostic. *arXiv preprint arXiv:1812.09384*.

Vehtari, A., A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner (2019). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC. *arXiv preprint arXiv:1903.08008*.

Winter, M., V. Devictor, and O. Schweiger (2013). Phylogenetic diversity and nature conservation: where are we? *Trends in Ecology & Evolution 28*(4), 199–204.

Chapter 6: Conclusion

Mathematical and statistical models have proven to be useful in studying infectious disease. Works by Ross (Kermack and McKendrick, 1927) have created a solid mathematical foundation for the epidemiology of infectious disease. These theoretical insights have translated into real-life disease control widely used today and have successful eliminate some of the deadliest diseases in civilization such as smallpox and rinderpest. Mathematical disease models have evolved tremendously in the new technology and information era, capable of capturing realistic aspects of disease epidemics. The increase in power of computation and data integration linked with theoretical methods has allowed for the development of new approaches for disease modeling and these methods bring us closer to real-time disease outbreak analysis.

Principal Findings and Contribution to the Field

In Chapter 2, I show that it is challenging to fit models and make adequate predictions in the early phase of an epidemic outbreak with limited information. Even in a simple simulation setting, where the fitting models are roughly restricted to match the simulation model (i.e., stochastic, discrete-time SIR model), there is still a lot of decision-making when constructing models. For example, what the type of distribution should we use for the transmission and observation processes? What is the best fitting platform? Should we allow for overdispersion in these processes?). When comparing models varying in complexity using several different platforms, I found that it is crucial to allow the model to incorporate uncertainties, and neglecting such uncertainty are likely to be over-confident and less accurate in forecasting. Furthermore, approximating discrete latent state processes with continuous processes increases computational

efficiency without losing accuracy. In response to modeling real emerging epidemic outbreaks, fast models allow more opportunity to try lots of scenarios and do rapid exploration.

In Chapter 3 and 4, I show that rabies dynamics are more complicated and uncertain than previously thought. I developed and applied improved estimation techniques to estimate epidemiological parameters and propagate uncertainties for historical rabies outbreaks around the world. Using the logistic model to estimating initial growth rate r, and propagating uncertainties in both r and generation interval, I found $\mathcal{R}_0$ estimates of historical rabies outbreaks around the world are larger and more uncertain compared to previous estimates(Hampson et al., 2009). The results may explain why rabies persists in these countries and have implications to guide rabies control. In Chapter 4, I used rabies contact tracing data and showed rabies generation intervals and serial intervals have different distributions. This finding led to exploring the relationship between dogs' time distributions and their tendency to transmit where I found these two disease traits are positively correlated and affect rabies transmission.

Lastly, in Chapter 5, I explored a broader evolutionary biology problem, where I developed an alternative method for flexibly and efficiently modeling phylogenetic mixed models. This method allows researchers to model different levels of Brownian-motion evolutionary model in the generalized linear mixed modeling framework. Existing procedures may be either insufficiently flexible or too computationally demanding when analyzing large volumes of data; researchers typically search for ways to simplify their analyses. Using simulations where I incorporate various phylogenetic correlations in the simulation model, refitting models which ignores these structures often leads to bias estimates of the parameters of interest. This improvement offers researchers a flexible way to fit multi-species trait models.

Future direction

This work is primarily motived by the vast amount of new approaches developed in recent years applied to disease modeling applications. At the end of this thesis, two major possible future directions can be extended in the future. Chapter 4 laid out the foundations and blueprints of the generation and serial interval differences, witn an example from rabies. It would be interesting to explore how these results can be extended to other diseases. Rabies is a straightforward disease system where the two pivotal time intervals (i.e., incubation and latent periods) match almost perfectly; this is not the case for many diseases. The second future direction also stemmed from Chapter 4, where I showed correlations in disease traits affects generation intervals and ultimately $\mathcal{R}_0$. The focus in Chapter 3 and 4 is on domestic dogs: it would be interesting to apply this approach to other species that spread canine rabies in this or other systems. In addition, it may be useful to gather data and fit a multi-species disease trait model for rabies using the approach developed in chapter 5.

Bibliography

Hampson, K., J. Dushoff, S. Cleaveland, D. T. Haydon, M. Kaare, C. Packer, and A. Dobson 2009. Transmission dynamics and prospects for the elimination of canine rabies. *PLoS biology 7*(3), e1000053.

Kermack, W. O. and A. G. McKendrick 1927. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character 115*(772), 700–721.

Ma, J., J. Dushoff, B. M. Bolker, and D. J. Earn 2014. Estimating initial epidemic growth rates. *Bulletin of mathematical biology 76*(1), 245–260.

Appendix : Additional figures for Chapter 2

Supplemental figures, originally presented as a supplemental component of the published work presented as Chapter 2.

Supplemental material

In the main text, we present the bias, RMSE, coverage and efficiency plots for aggregated forecast, $\mathcal{R}_0$, MGI, P_{eff}, and P_{rep}. Here, we present plots showing the other parameters (shape G_S and position G_P of the transmission kernel and process and observation overdispersion parameters δ_P and δ_{obs}) and disaggregated forecasts (five forecast steps) that are excluded in the main text. We also add some representative plots of the simulated cases and forecast.

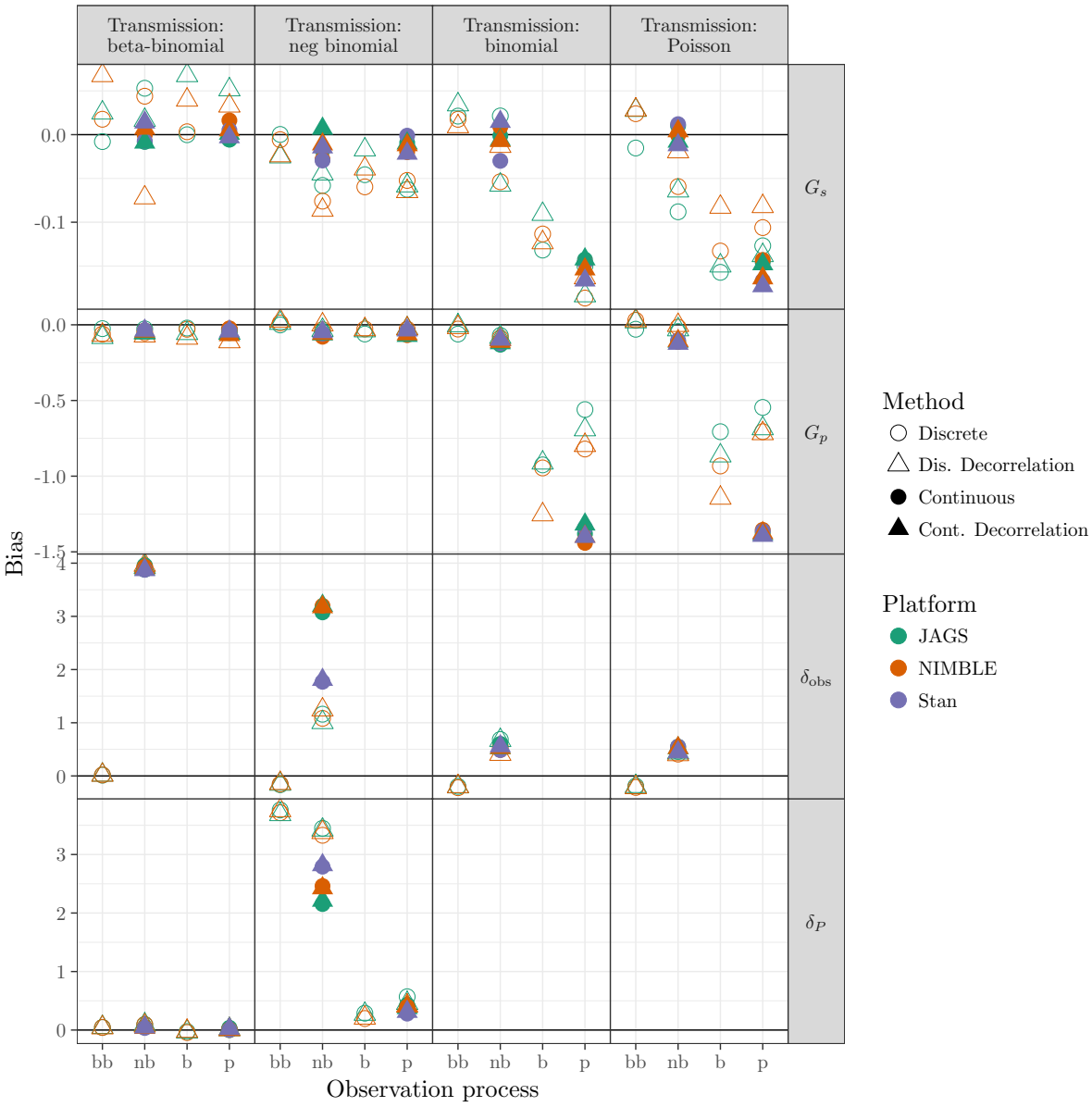

Figure S1. Comparison of bias for G_S (transmission shape), G_P (transmission position), δ_{obs} (observation overdispersion), and δ_P (process overdispersion: more detail given in Sect. 2.2) across different platforms (described in Sect. 2.3.1). Overdispersion parameter δ_P is only applicable in models with dispersion in the transmission process (first and second left column panel) and overdispersion parameter δ_{obs} is only applicable in models with dispersion in the observation process (first and second column within each column panel).

Prepared using sagej.cls

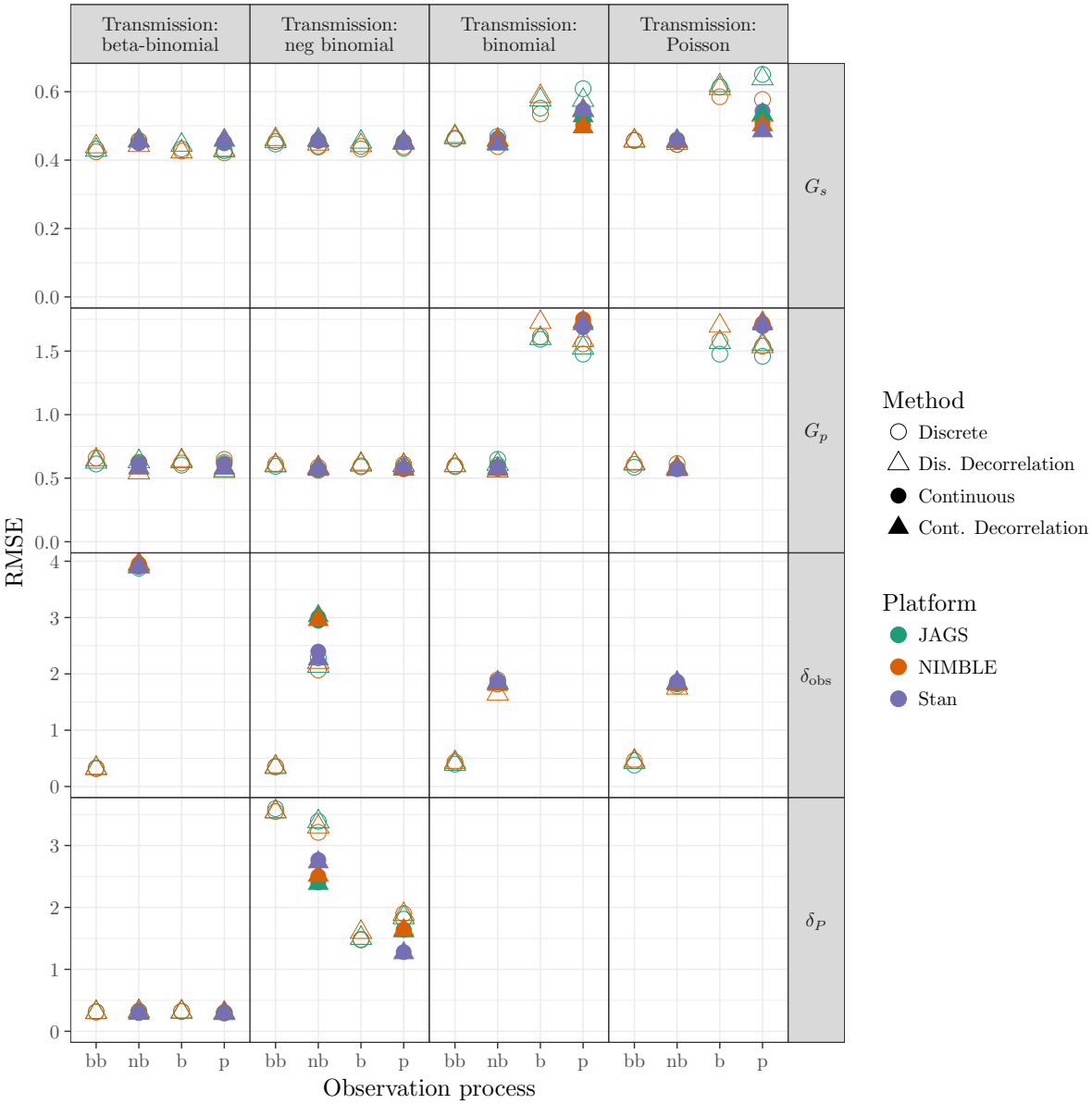

Figure S2. Comparison of RMSE for G_S, G_P, δ_{obs}, and δ_P. See Figure 4 in main text and Figure S1 in appendix for details.

Prepared using sagej.cls

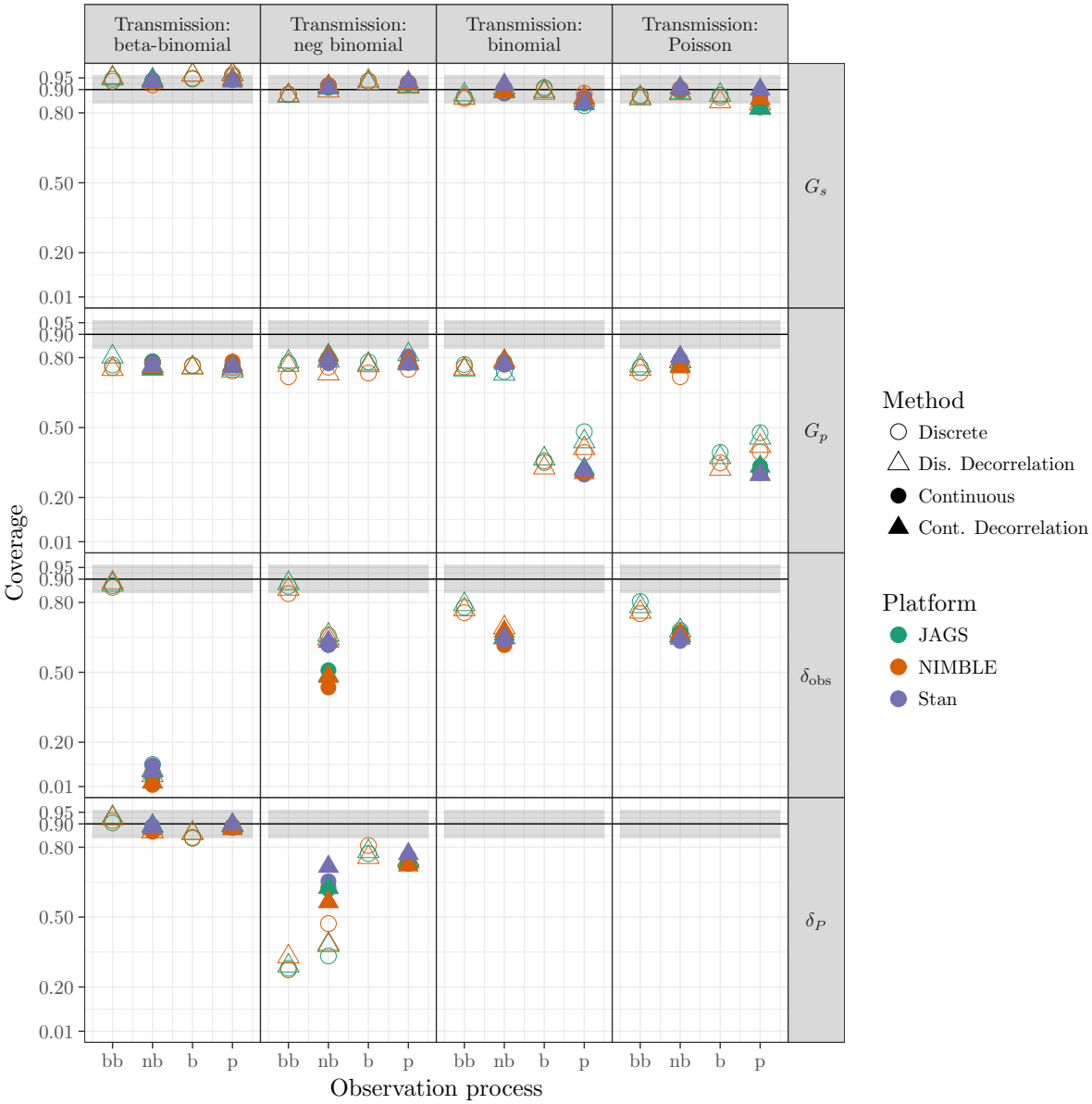

Figure S3. Comparison of coverage for G_S, G_P, δ_{obs}, and δ_P. See Figure 5 in main text and Figure S1 in appendix for details.

Prepared using sagej.cls

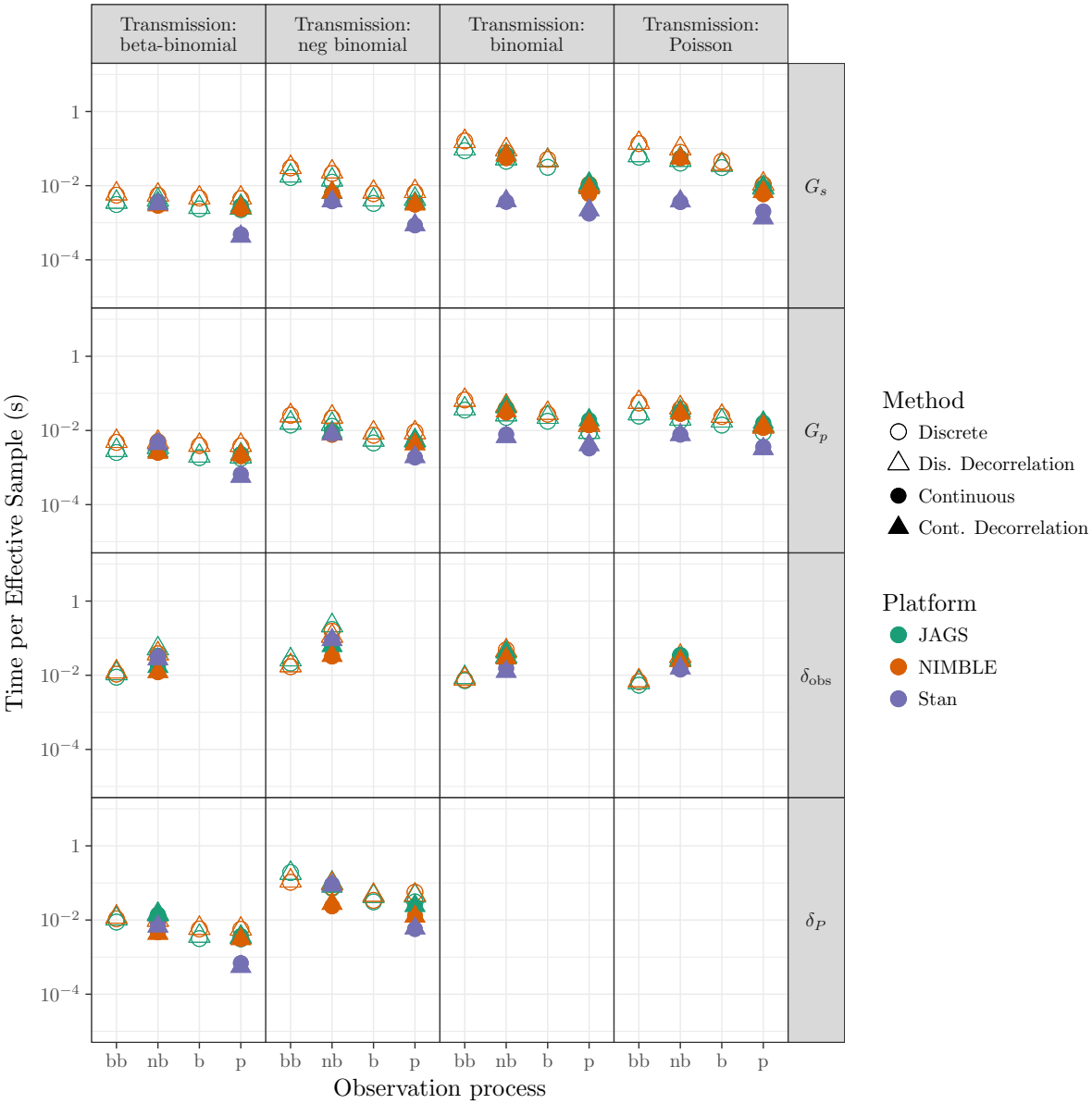

Figure S4. Comparison of coverage for G_S, G_P, δ_{obs}, and δ_P. See Figure 6 in main text and Figure S1 in appendix for details.

Prepared using sagej.cls

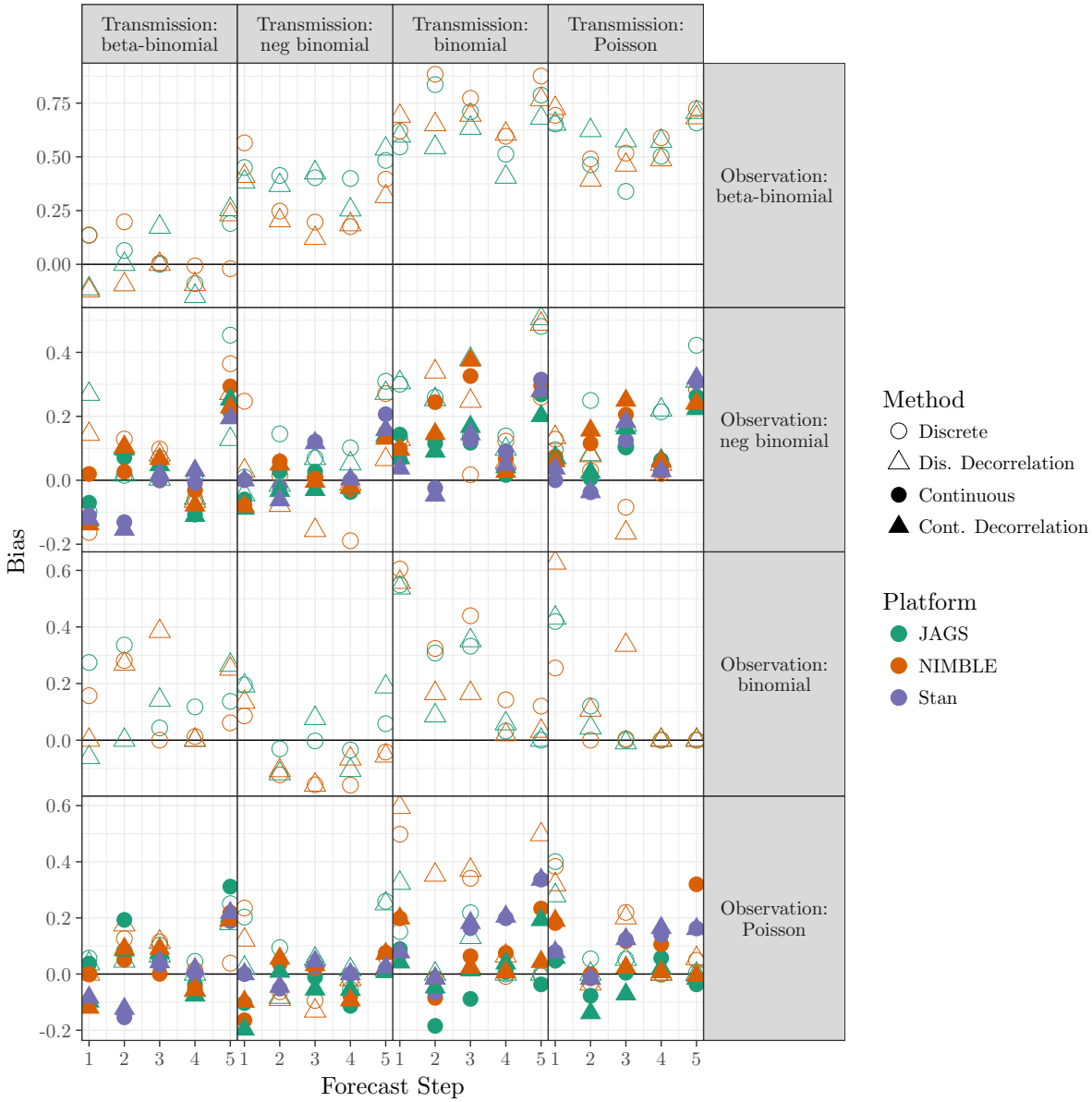

Figure S5. Comparison of bias for five forecast steps (described in Sect. 2.2) across different platforms (described in Sect. 2.3.1).

Prepared using sagej.cls

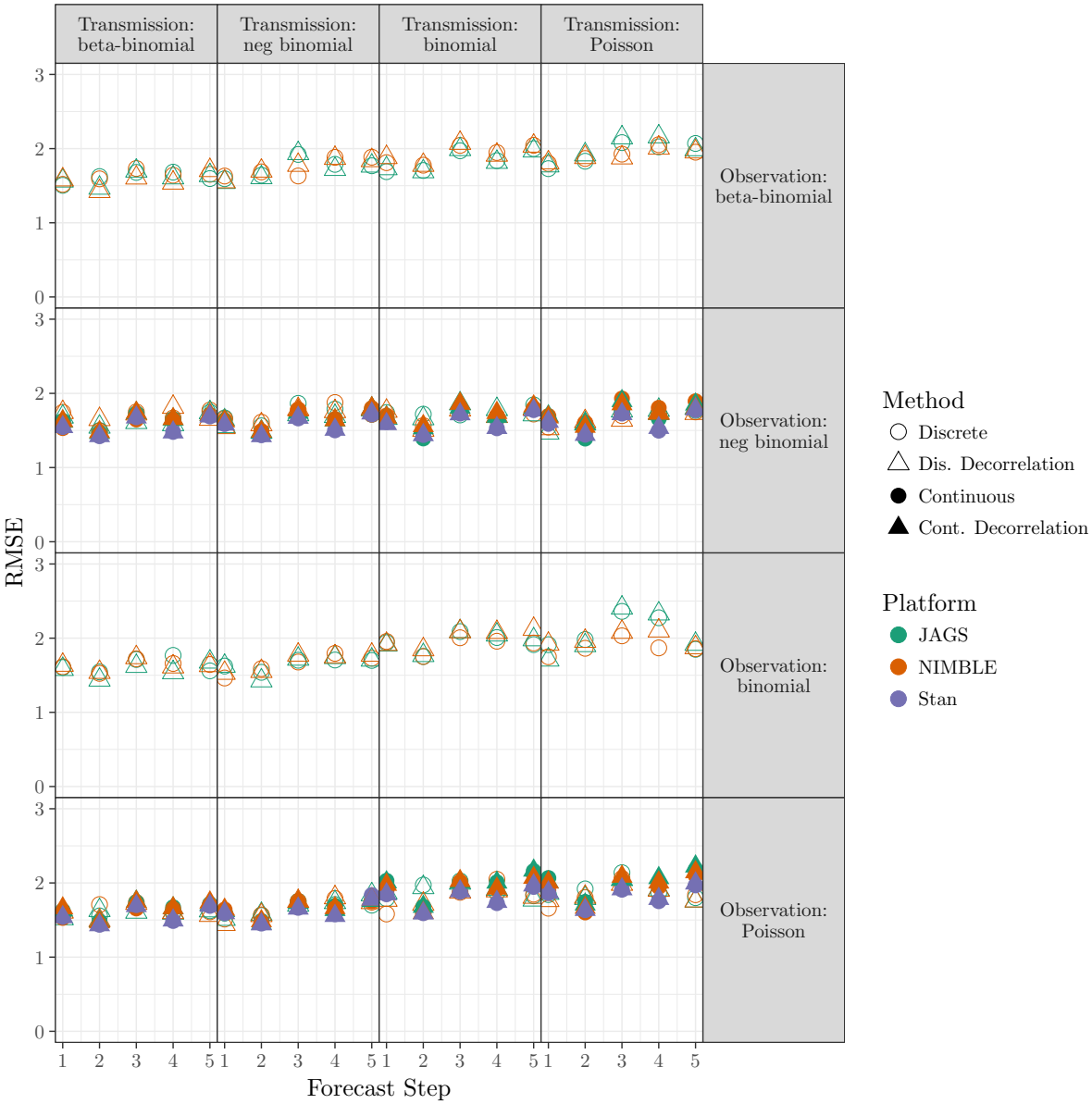

Figure S6. Comparison of RMSE for five forecast steps described in Sect. 2.2 across different platforms described in Sect. 2.3.1. See Figure 4 in the main text for details.

Prepared using sagej.cls

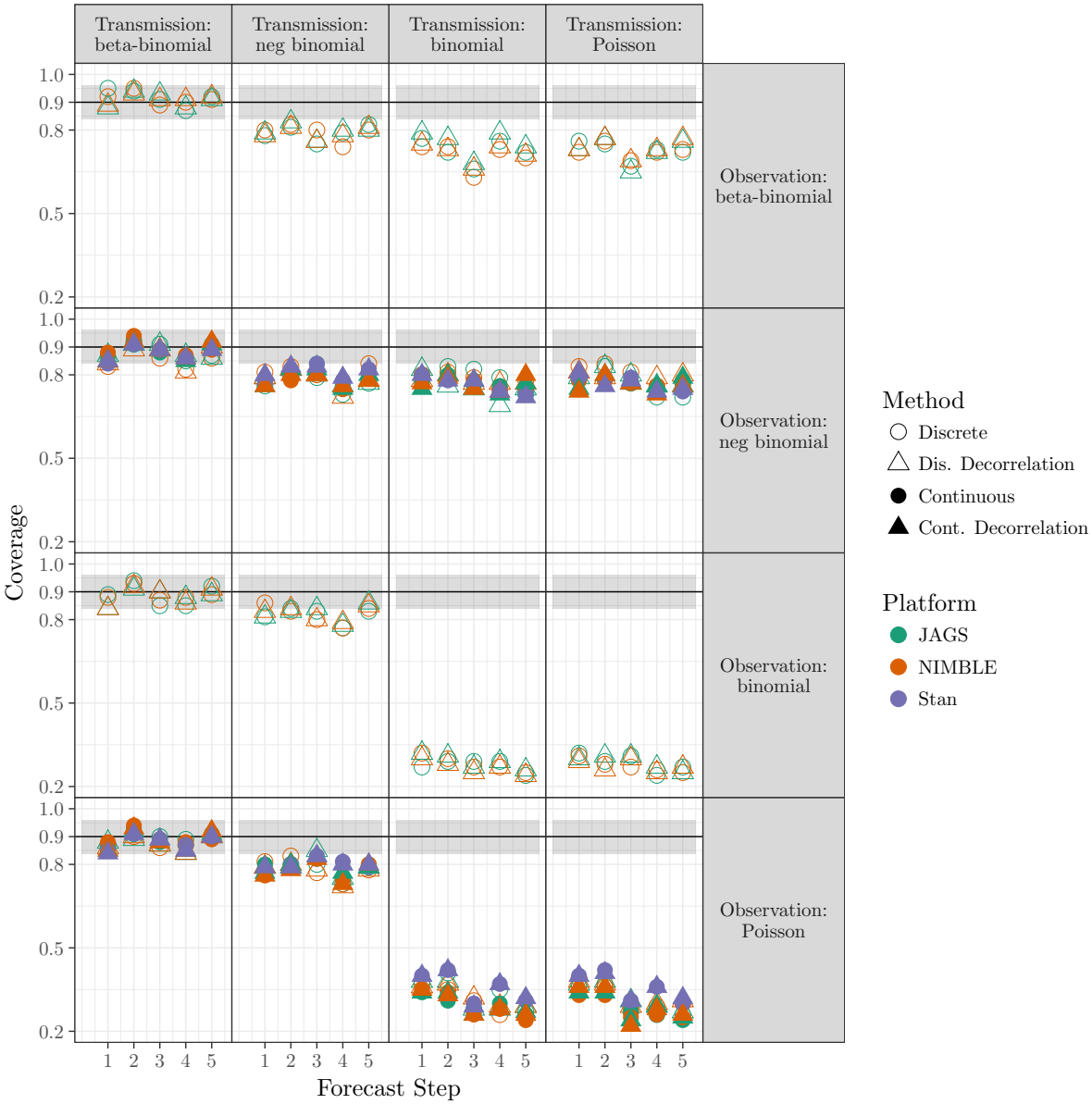

Figure S7. Comparison of coverage for five forecast steps described in Sect. 2.2 across different platforms described in Sect. 2.3.1. See Figure 5 in the main text for details.

Prepared using sagej.cls

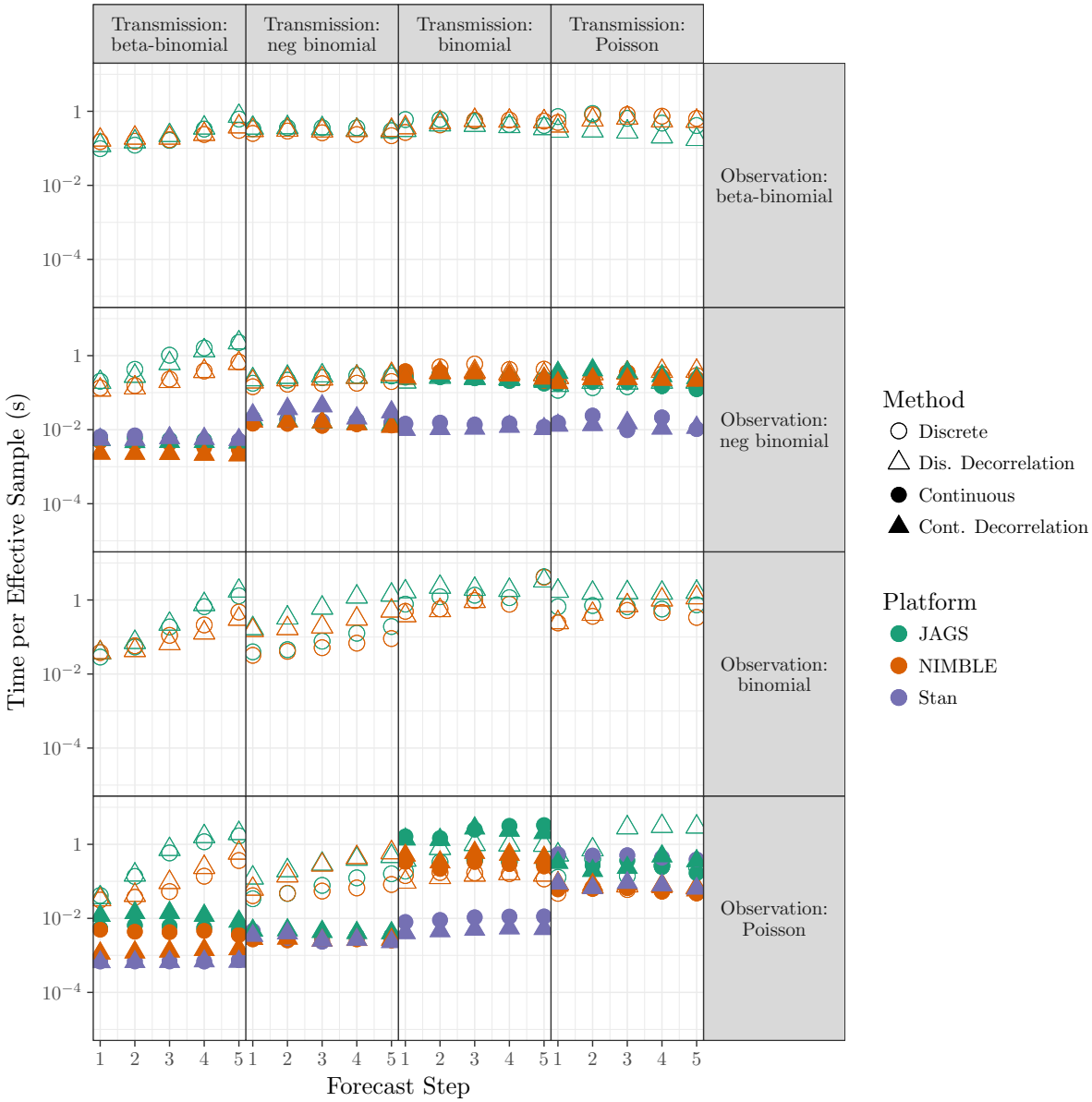

Figure S8. Comparison of sampling efficiency for five forecast steps described in Sect. 2.2 across different platforms described in Sect. 2.3.1. See Figure 6 in the main text for details.

Prepared using sagej.cls

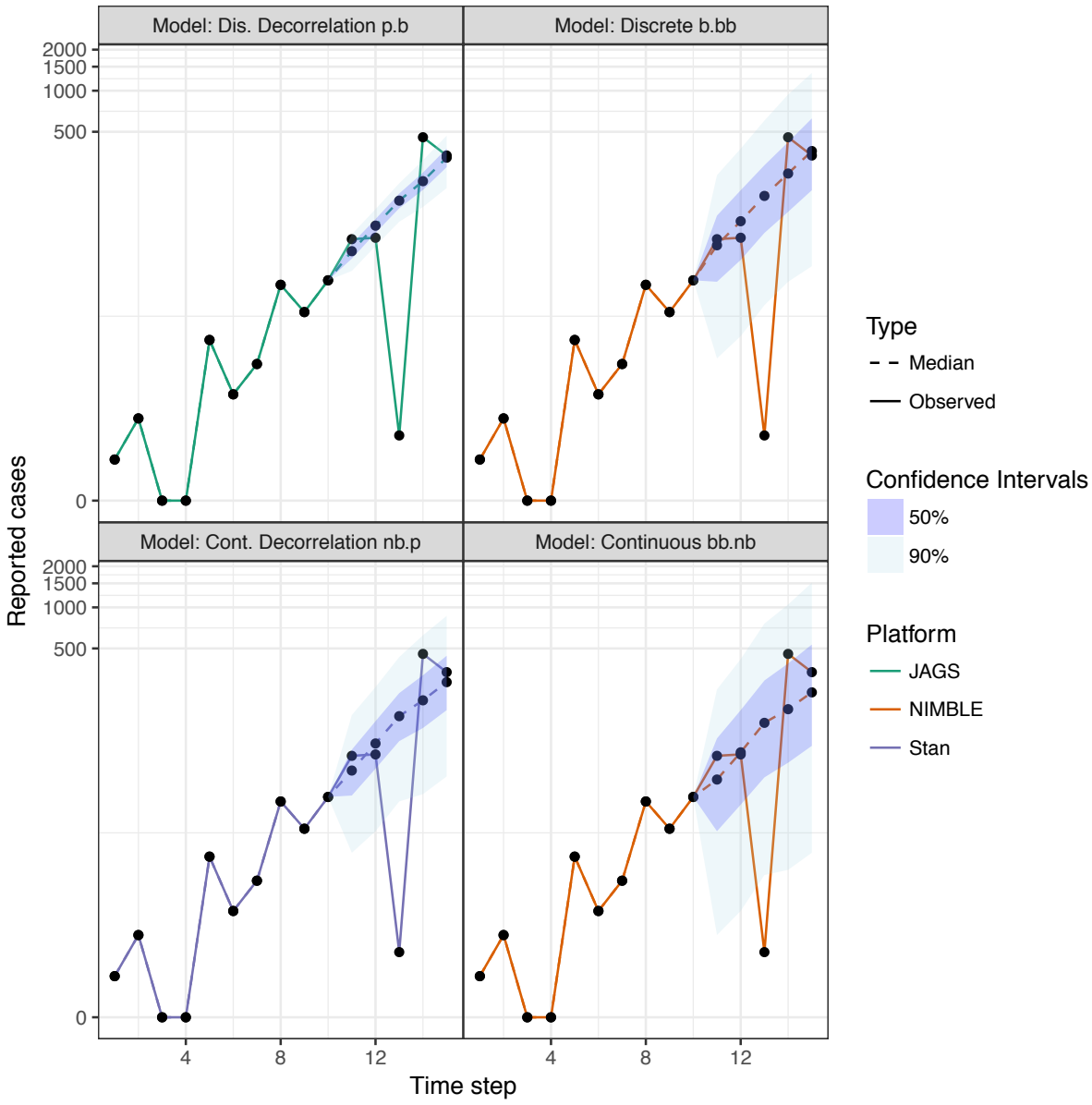

Figure S9. Comparison of forecast using combinations of transmission process, observation process, decorrelation, latent state variables, and platforms described in Sect 2.2 and 2.3.1. Moving from the top to bottom row adds overdispersion in the transmission process (binomial (b) and Poisson (p) to negative-binomial (nb) and beta-binomial (bb)). Moving from left to right adds overdispersion in the observations. Solid line shows the simulated observed cases (15 time steps); dashed line shows the median of the posterior forecast sample with 50% (dark ribbon) and 90% (light ribbon) confidence intervals (last 5 time steps).

Prepared using sagej.cls

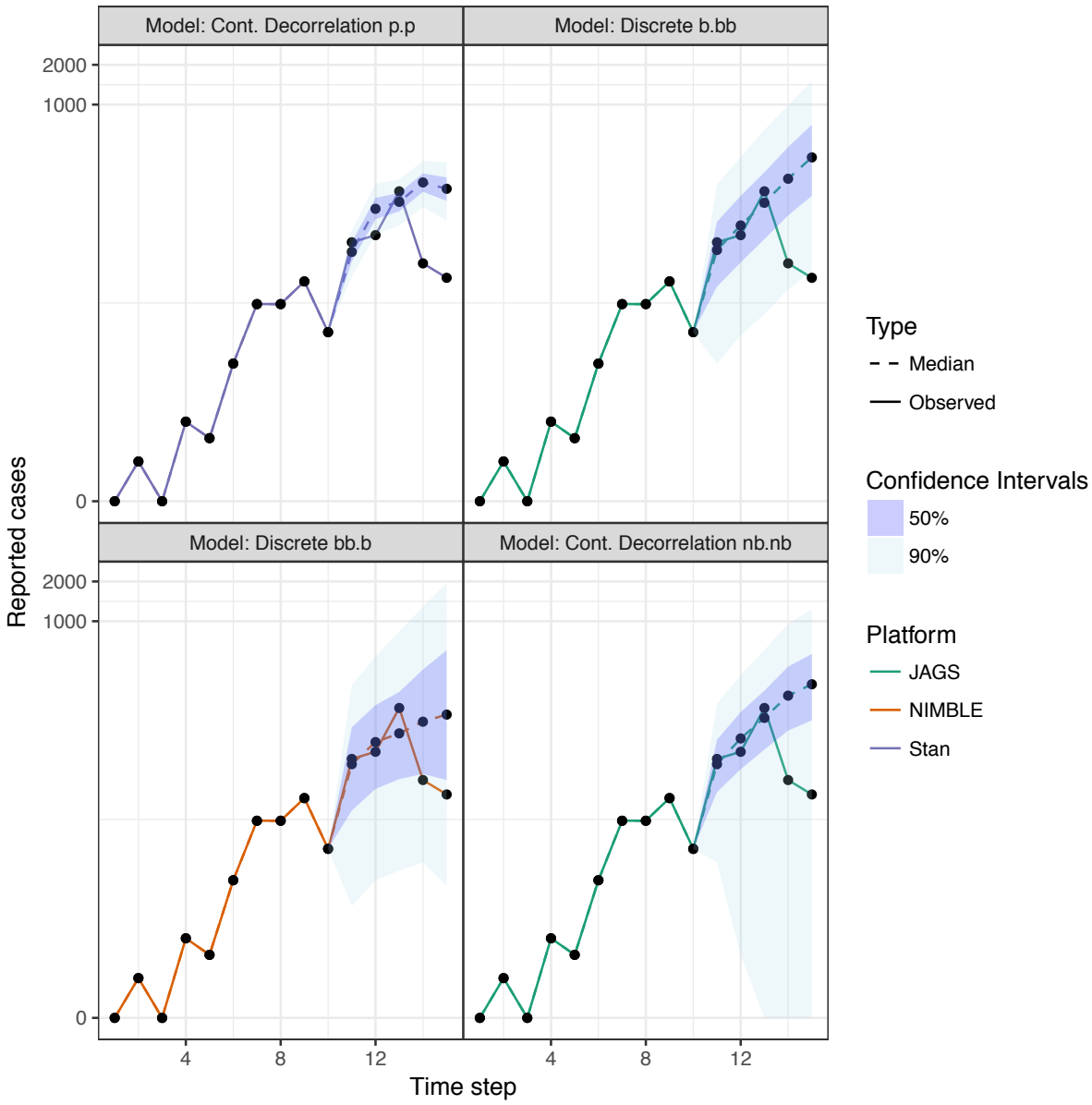

Figure S10. Comparison of forecast using a different set of parameters. See Figure S9 for details.

Prepared using sagej.cls

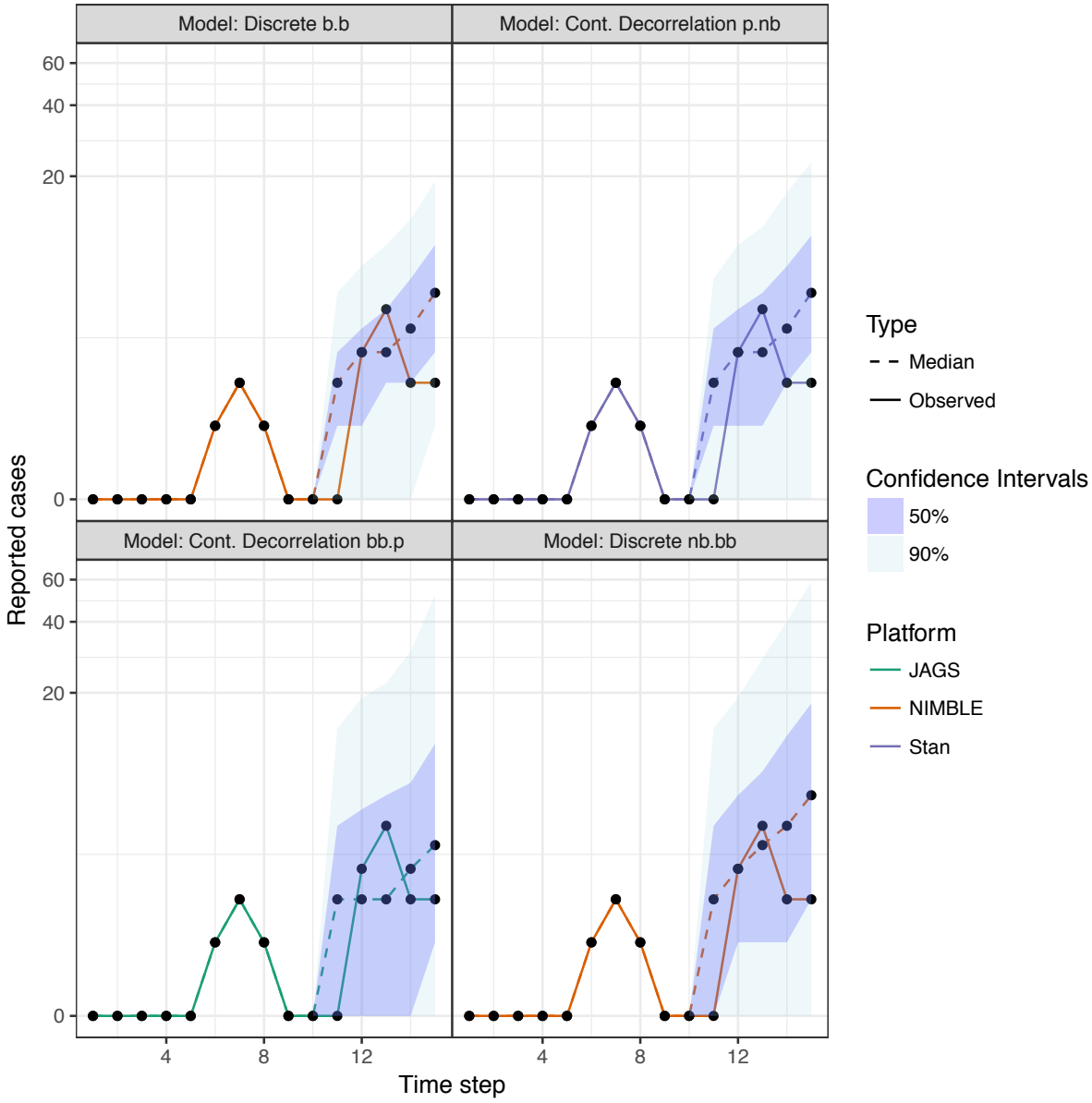

Figure S11. Comparison of forecast of low observed cases. See Figure S9 for details.

Prepared using sagej.cls

Appendix : Additional Tables for Chapter 2

Supplemental tables, originally presented as a supplemental component of the published work presented as Chapter 2.

Tables

Table 1. Simulation model parameters

Parameter	Description	True	Prior
N	Total population size	Fixed at 100,000	NA
ℓ	Maximum length of the generation interval	Fixed at 5 time steps	NA
$\mathcal{R}_0$	Basic reproductive number	3	Gamma(shape=15,rate=5)
P_{eff}	Effective susceptible proportion of the population	0.5	Beta($\frac{B_{size}}{1-P_{\text{eff}}}$, $\frac{B_{size}}{P_{\text{eff}}}$)
P_{rep}	Reporting proportion	0.5	Beta($\frac{B_{size}}{1-P_{\text{rep}}}$, $\frac{B_{size}}{P_{\text{rep}}}$)
G_p	Position parameter for generation interval	0.5	Beta($\frac{2B_{size}}{1-G_p}$, $\frac{2B_{size}}{G_p}$)
G_s	Shape parameter for generation interval	1	Gamma(shape=5,rate=5)
δ_P	Beta Binomial transmission process dispersion	1	Gamma(shape=10,rate=10)
δ_{obs}	Beta-Binomial Observation process dispersion	1	Gamma(shape=10,rate=10)

Table 2. Fitting model parameters

Parameter	Description	True	Prior
N	Total population size	Fixed at 100,000	NA
ℓ	Maximum length of the generation interval	Fixed at 5 time steps	NA
B_{size}	Beta prior size factor	Fixed at 1	NA
$\mathcal{R}_0$	Basic reproductive number	3	Gamma(shape=15,rate=5)
P_{eff}	Effective susceptible proportion of the population	0.5	Beta($\frac{B_{size}}{1-P_{\text{eff}}}$, $\frac{B_{size}}{P_{\text{eff}}}$)
P_{rep}	Reporting proportion	0.5	Beta($\frac{B_{size}}{1-P_{\text{rep}}}$, $\frac{B_{size}}{P_{\text{rep}}}$)
P_{effrep}	Proportion of effective S to I that are observed	$P_{\text{eff}} \times P_{\text{rep}}$	Beta($\frac{B_{size}}{1-P_{\text{effrep}}}$, $\frac{B_{size}}{P_{\text{effrep}}}$)
ρ	Scale splitting factor	0.5	Beta($\frac{B_{size}}{1-\rho}$, $\frac{B_{size}}{\rho}$)
G_p	Position parameter for generation interval	0.5	Beta($\frac{2B_{size}}{1-G_p}$, $\frac{2B_{size}}{G_p}$)
G_s	Shape parameter for generation interval	1	Gamma(shape=5,rate=5)
δ_P	Beta Binomial transmission process dispersion	1	Gamma(shape=10,rate=10)
δ_P (Neg-Binom)	Negative-Binomial Transmission process dispersion	NA	Uniform(min=0,max=100)
δ_{obs}	Beta-Binomial Observation process dispersion	1	Gamma(shape=10,rate=10)
δ_{obs} (Neg-Binom)	Negative-Binomial Transmission process dispersion	NA	Uniform(min=0,max=100)

Appendix : Additional Figures for Chapter 3

Supplemental figures, planned as a supplemental component to the work presented as Chapter 3.

Supplement

In the main text, we presented the $\mathcal{R}_0$ estimates for the historical rabies outbreaks around the world. Here, we present r estimates using two approaches: maximum likelihood estimation and Bayesian HMC. In addition, we made prediction case plots (with 95% prediction regions) using the maximum likelihood estimates of r.

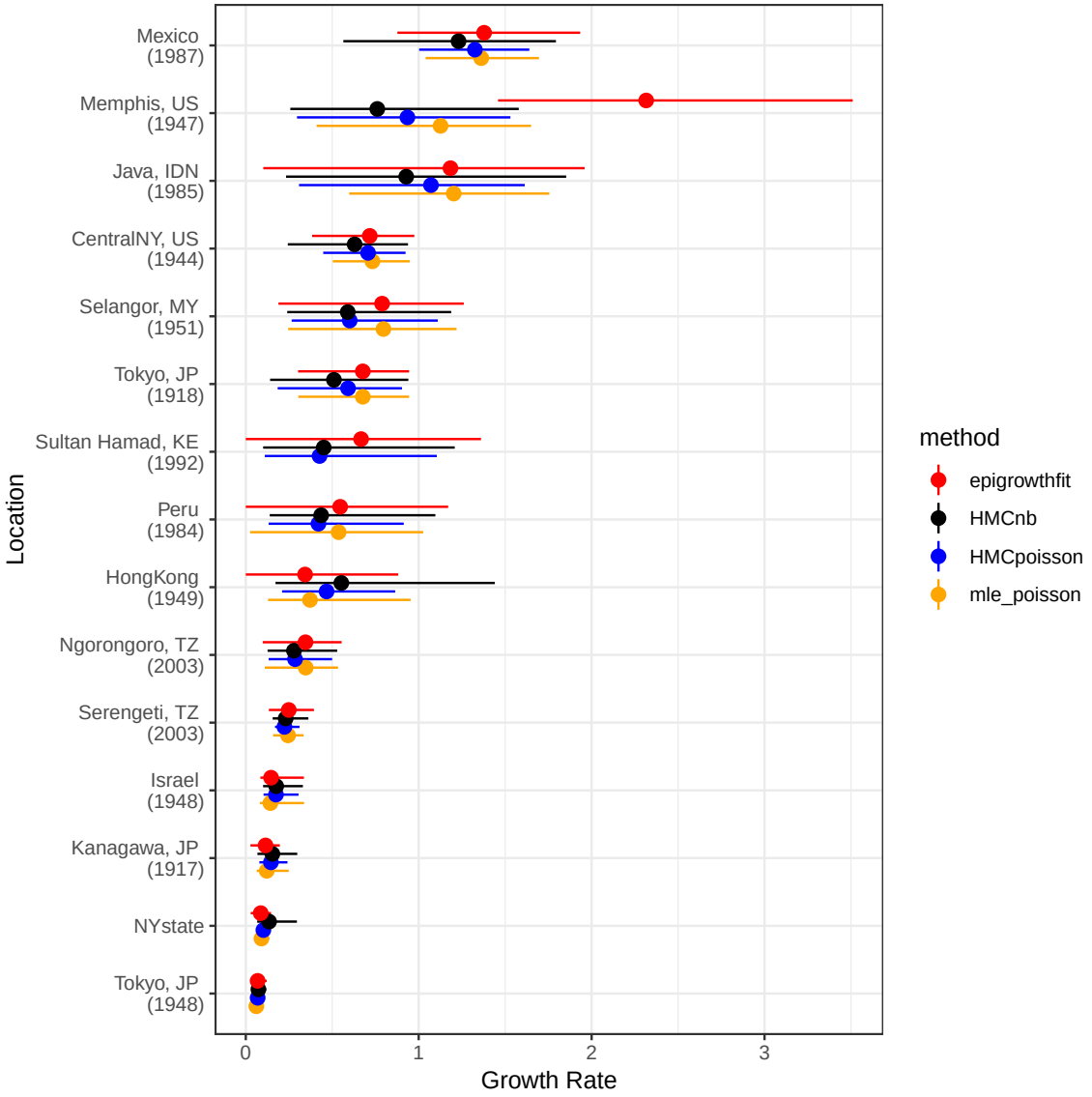

Figure A1: Logistic growth rate fits.

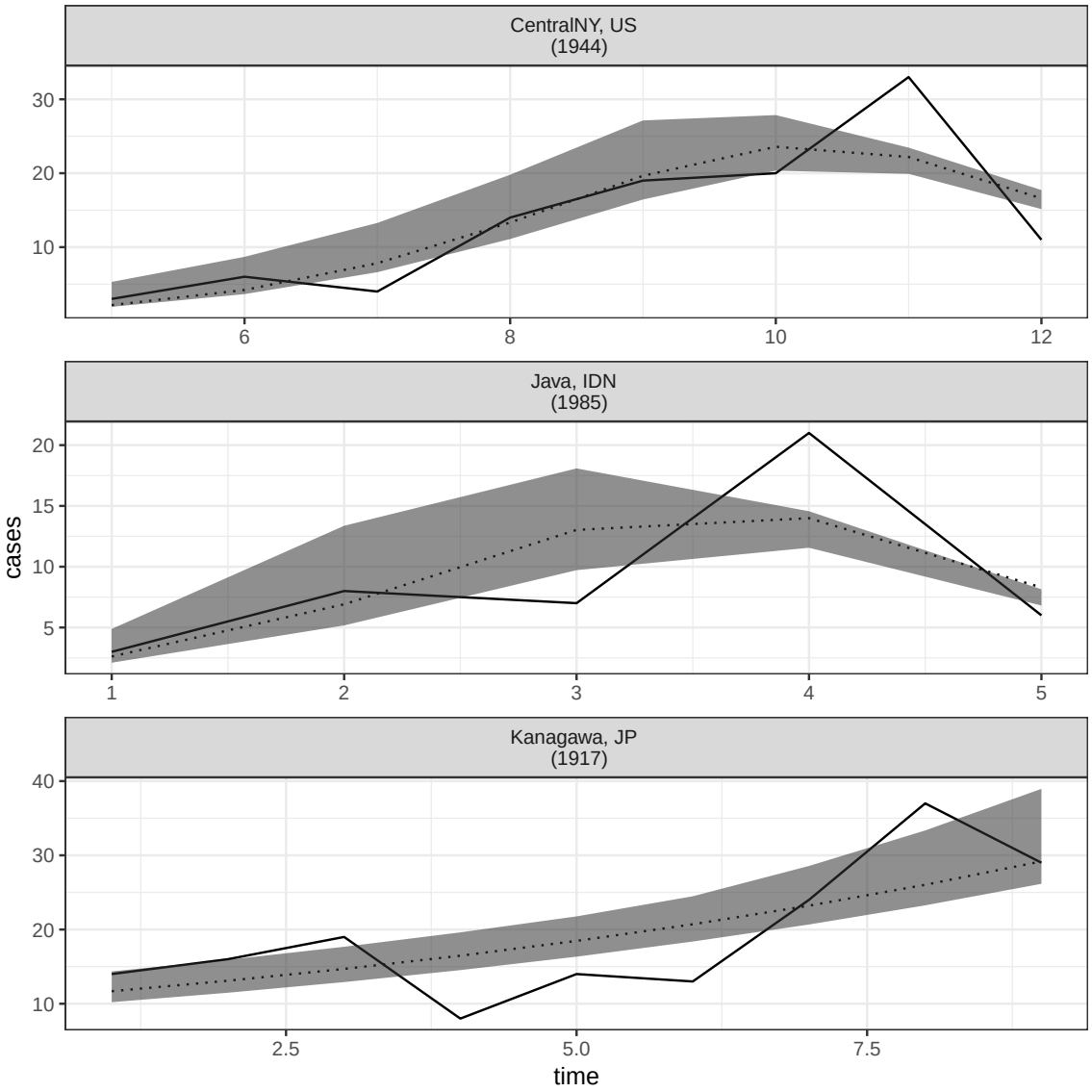

Figure A2: Prediction plots.

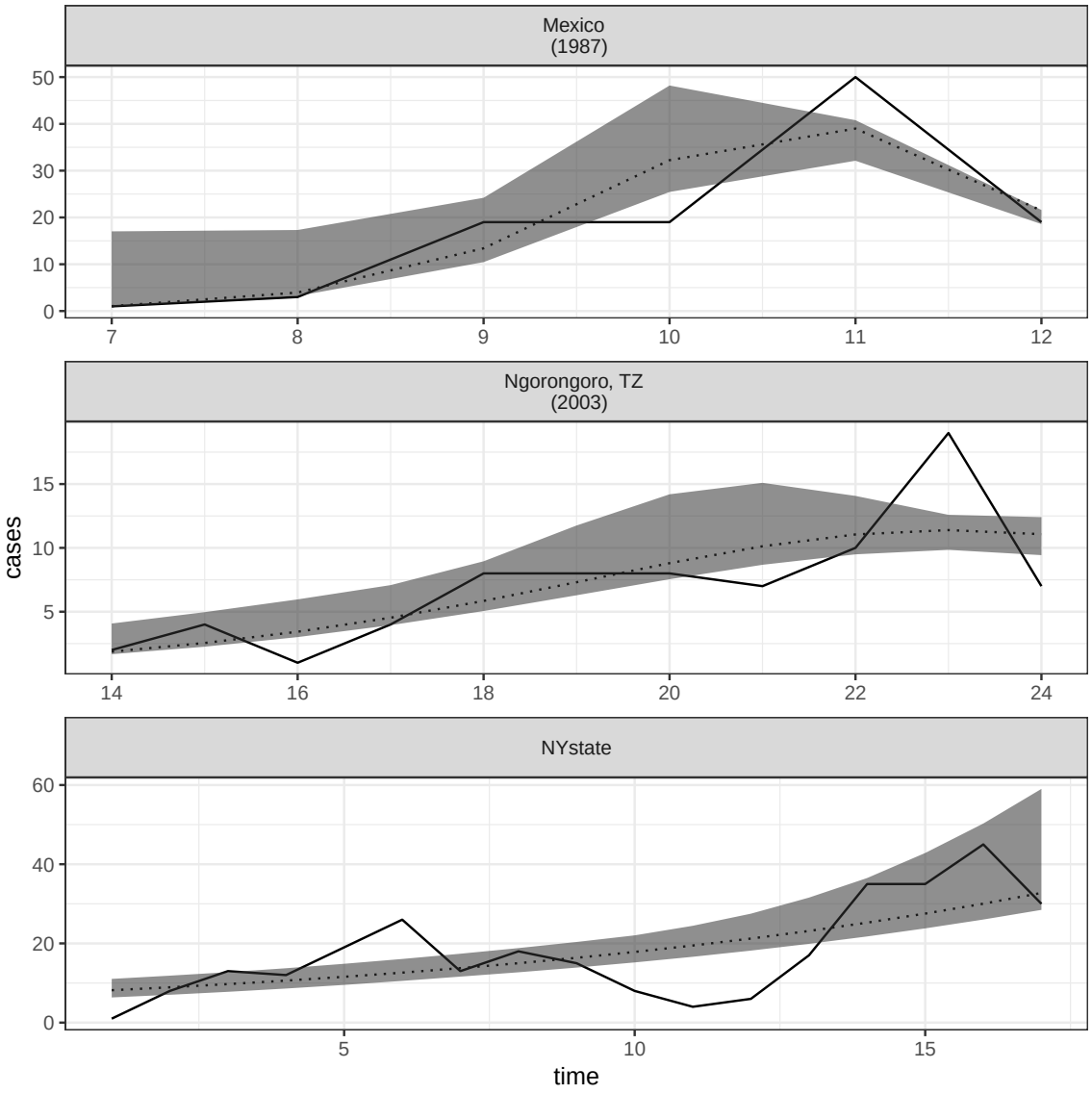

Figure A3: Prediction plots.

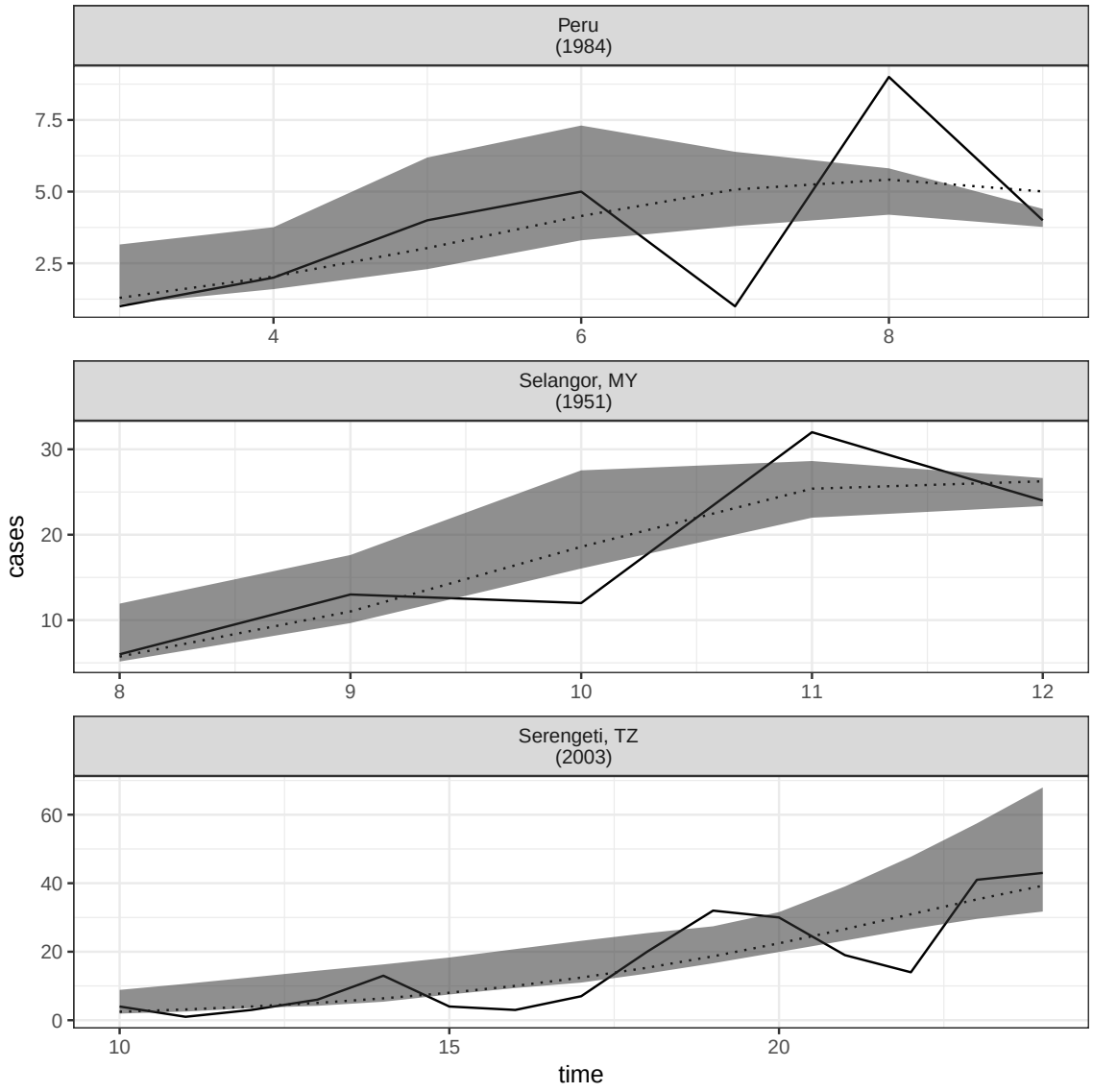

Figure A4: Prediction plots.

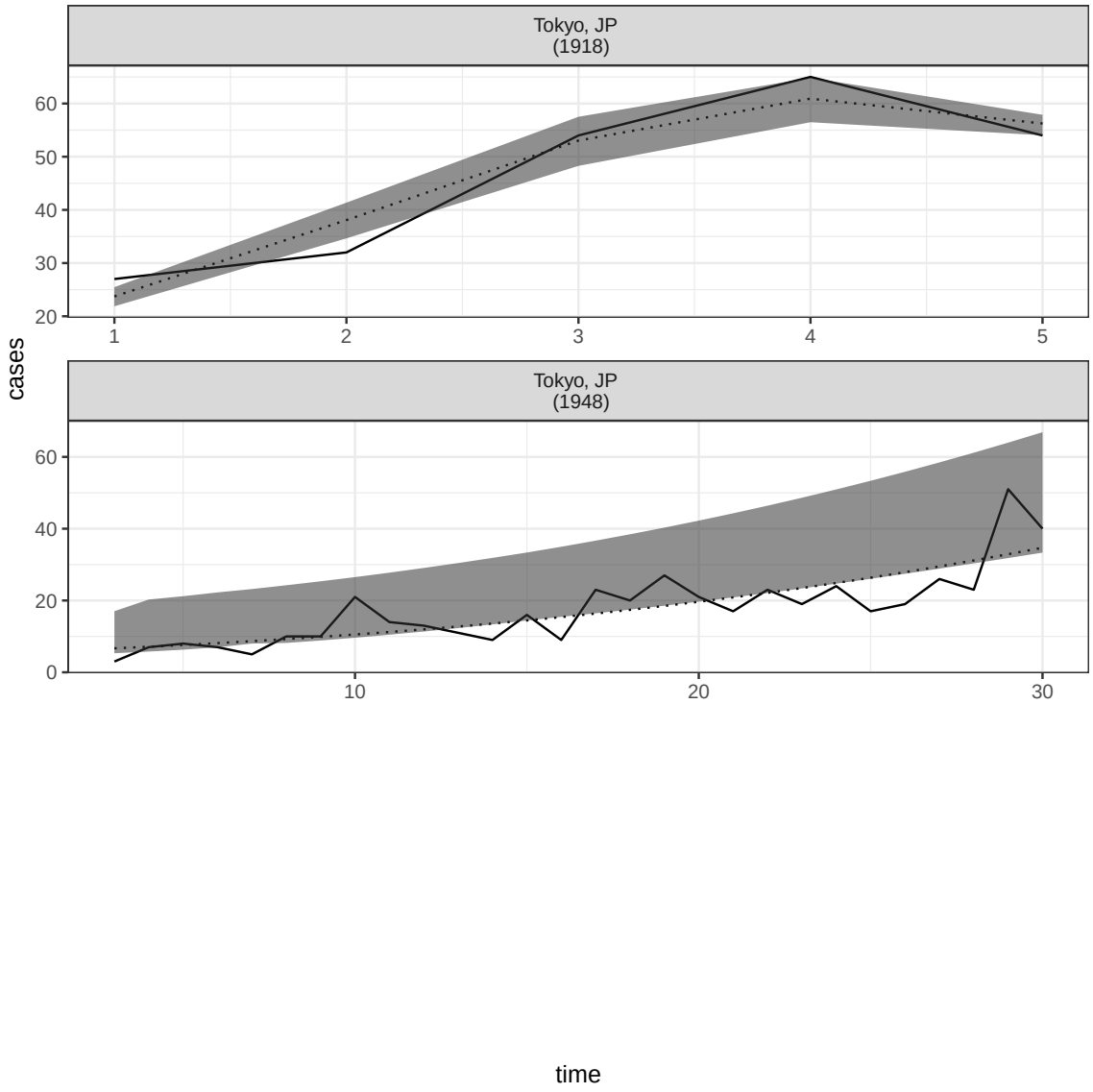

Figure A5: Prediction plots.