Title

Leveraging mechanistic models to characterize the dynamics of zoonotic infectious diseases and assess intervention strategies

Author

Ambrose, Monique R

Publication Date

2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Leveraging mechanistic models to characterize the dynamics of zoonotic infectious diseases and assess intervention strategies

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Biology

by

Monique Renee Ambrose

2018

ABSTRACT OF THE DISSERTATION

Leveraging mechanistic models
to characterize the dynamics of zoonotic infectious diseases
and assess intervention strategies

by

Monique Renee Ambrose
Doctor of Philosophy in Biology
University of California, Los Angeles, 2018
Professor James O. Lloyd-Smith, Chair

Zoonotic diseases, which are caused by pathogens that transmit from animals into humans, are responsible for numerous ongoing public health burdens, such as leptospirosis, rabies, and West Nile virus infections, and are also considered a probable source of future epidemics in humans. Describing and quantifying the transmission dynamics of these pathogens is vital if we wish to assess which of the many known zoonotic pathogens pose a threat to humans and which management strategies would be most effective at minimizing that threat. To conduct these assessments, it is necessary to consider the ecological dynamics and interactions driving zoonotic disease transmission.

A zoonotic pathogen's impact on humans depends not only on transmission dynamics within the human population, including heterogeneities in human contacts and interactions with

endemic human pathogens, but also on disease dynamics within the reservoir and at the human-reservoir interface. Because of the complex ecological interactions driving the spread of zoonotic pathogens, qualitatively and quantitatively characterizing their spread and devising rational management strategies requires combining insights from community ecology, invasion biology, and classical single-host disease ecology with system-specific information about the pathways of transmission within the reservoir, within humans, and between the two. Bringing together these complementary perspectives can shed light on the key processes driving transmission, which is essential for predicting how changes, both purposeful interventions and natural shifts, may alter the system's behavior. In this dissertation, I present three studies that use diverse methods to explore different aspects of zoonotic pathogens' disease dynamics.

In chapter 1, I use a theoretical approach to explore the effects of competition between a zoonotic pathogen and a human-endemic pathogen in the context of a disease eradication program. I use a deterministic compartmental model that tracks spillover and transmission of a zoonotic disease in humans as well as transmission of a partially cross-protective endemic human pathogen to examine how the presence of the zoonotic pathogen can reduce the vaccination coverage necessary to eradicate the human pathogen and how the zoonotic pathogen's prevalence is expected to change during and following a successful eradication campaign. I then use the smallpox-monkeypox system as a case study to assess how the theoretical findings apply in a real-world context.

In chapter 2, I move from theoretical explorations of disease dynamics to using real-world datasets to inform mechanistic models. Zoonotic disease surveillance datasets are valuable sources of information about disease dynamics, but are generally difficult and expensive to obtain and are associated with a variety of data limitations. This chapter develops methods to

extract as much information as possible from these valuable information sources. I develop a model-based inference method that addresses a number of data challenges, including unobserved sources of transmission (both human and zoonotic), limited spatial information, and unknown scope of surveillance, using a spatial model with two levels of mixing. After demonstrating the robustness of the method using simulation studies, I apply the new method to a dataset of human monkeypox cases detected during an active surveillance program from 1982-1986 in the Democratic Republic of the Congo. The results provide estimates of the reproductive number and spillover rate of monkeypox during this surveillance period and suggest that most human-to-human transmission events occur over distances of 30 km or less. Taking advantage of contact-tracing data available for a subset of monkeypox cases, I find that around 80% of contact-traced links could be correctly recovered from transmission trees inferred using only date and location. The results highlight the importance of identifying the appropriate spatial scale of transmission, and show how even imperfect spatiotemporal data can be incorporated into models of zoonotic pathogens to obtain reliable estimates of transmission patterns.

Chapter 3 shifts from examining the dynamics of zoonotic pathogens after they have already spilled into humans to evaluating how interventions in the zoonotic reservoir could help reduce the risk of spillover occurring in the first place. This chapter focuses on evaluating interventions to minimize the risk of spillover of swine-origin influenza A viruses (IAV-S) into humans in the United States. In the past decade, the majority of reported human infections with IAV-S in the United States have been associated with individuals exposed to exhibition swine while attending agricultural shows. Because these exhibition swine make up a largely distinct population within the US swine herd, there is great potential to implement control practices within exhibition swine that could substantially reduce risk of spillover into humans. To

understand the factors that drive influenza prevalence and persistence in US exhibition swine and to evaluate the impact of potential interventions, I develop a network model that characterizes disease spread into and among exhibition swine. The model incorporates key structural information about the system and is informed by a unique surveillance dataset collected from shows in Ohio, Michigan, and Indiana, including IAV-S genomes from more than one hundred infected swine. I use several different approaches based on both epidemiological and sequence data to estimate parameters describing transmission and to evaluate the expected impact of a set of thirty potential interventions on the risk of spillover into humans. Across all approaches, several interventions consistently are found to perform best at reducing projected spillover risk, including requiring participants to take one or two weeks off between shows and implementing strategies to reduce transmission probabilities among swine at shows.

While the studies presented in these chapters range from theoretical explorations of simplified systems to direct comparisons of intervention impacts incorporating messy real-world data and complex system structure, they all pursue the common goal of providing insights relevant for conceptualizing the prominent forces in a system and for using that understanding to inform decisions on control measures in a real-world context.

The dissertation of Monique R. Ambrose is approved.

Kirk Edward Lohmueller

Van Maurice Savage

Marc Adam Suchard

James O. Lloyd-Smith, Committee Chair

University of California, Los Angeles

2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I want to thank my advisor Jamie Lloyd-Smith for his extraordinary mentorship and guidance. I have been continually inspired by the example he sets as a researcher, science communicator, and mentor, and his invaluable insights and advice have helped me grow as a scientist. I could not be more grateful for all of the remarkable opportunities he has provided over the past six years, as well as for his unwavering support and encouragement.

I have benefitted greatly from the insights and advice of numerous faculty and staff who have been generous with their time and attention. I thank the other members of my doctoral committee, Van Savage, Kirk Lohmueller, and Marc Suchard, for their very helpful comments and suggestions on my research and for teaching some of the most valuable courses I have taken. I am grateful to Pam Yeh and Van Savage for their kindness and support throughout my time at UCLA: from providing stimulating academic conversations to sharing invaluable career and work-life balance advice, they have made my time here much richer. I am also grateful for the exceptional mentorship and support of my undergraduate advisor Cherie Briggs, who introduced me to the world of infectious disease modeling research and had a huge impact on my career trajectory.

I have been very fortunate to work with many wonderful collaborators. Anne Rimoin and her students and collaborators have been generous in sharing and discussing surveillance datasets from the Democratic Republic of the Congo, and Andrew Bowman and Jacqueline Nolting have provided invaluable insights into the exhibition-swine system through patiently answering my many questions and sharing a phenomenal dataset. I also thank Adam Kucharski, Martha Nelson, and Cecile Viboud for many fun and intellectually-stimulating discussions and for their technical suggestions and advice.

The students, postdocs, and researchers of the Lloyd-Smith lab group: Benny Borremans, Michael Buhnerkempe, Ana Gomez, Katie Gostic, Angela Guglielmino, Sarah Helman, Ruian Ke, Claude Loverdo, Christian Mason, Riley Mummah, Miran Park, and Katie Prager, made my time at UCLA so much more enjoyable and rewarding. From giving me great feedback and suggestions on practice talks, to joining me for swimming, juggling, and rock climbing adventures, to challenging my endurance during our afternoon planks, they provided highly-valued laughter and perspective.

I would not be who or where I am today without the support and love of my amazing family. They encouraged my interest in biology and the sciences from a young age by allowing me to play with all manner of slimy, scaly, and many-legged creatures and by indulging (with admirable patience) my unending questions and curiosity. I am so grateful for their consistent encouragement and excellent advice throughout my life and for all they have done to help me pursue my passions. Finally, I thank Gabriel for his unwavering support and patience, for his kindness and generosity, and for making me smile every single day.

I have been fortunate to receive financial support during my dissertation from a National Science Foundation Graduate Research Fellowship, a Systems and Integrative Biology Training Grant (NIH), a UCLA Pauley Fellowship, and a UCLA Graduate Dean Scholar's Award. Travel funds from the Infectious Disease Evolution Across Scales Research Coordination Network (IDEAS RCN), the UCLA Gradate Division, and the Department of Ecology and Evolutionary Biology have facilitated collaborations and allowed me to share the results of this work at conferences.

All three chapters have been prepared as manuscripts and are in preparation for submission.

Chapter one is a version of Monique R. Ambrose and James O. Lloyd-Smith. Competition between cross-immunizing human and zoonotic pathogens: implications for disease control and the aftermath of eradication. *In preparation.*

For chapter one, M.R.A was supported by the National Institutes of Health Ruth L. Kirschstein National Research Service Award (T32-GM008185). J.O.L.-S. contributed to the conception of the study, provided feedback on model analyses, and contributed editorial feedback on the manuscript.

Chapter two is a version of Monique R. Ambrose, Adam J. Kucharski, Anne W. Rimoin, and James O. Lloyd-Smith. Quantifying transmission of emerging zoonoses: Using mathematical models to maximize the value of surveillance data. *In preparation.*

For chapter two, M.R.A was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1144087. J.O.L.-S. and A.J.K. contributed to the conception of the study, provided feedback on model structure and analyses, and contributed editorial feedback on multiple drafts of the manuscript. A.J.K. contributed to the development of the model. A.W.R. provided data sources and insights into monkeypox epidemiology and surveillance.

Chapter three is a version of Monique Ambrose, Martha Nelson, Cecile Viboud, Jacqueline Nolting, Andrew Bowman, and James Lloyd-Smith. Evaluating intervention strategies to reduce zoonotic spillover of influenza A viruses from US exhibition swine: a modeling-based analysis. *In preparation.*

For chapter three, M.R.A. was supported by an Edwin W. Pauley Fellowship. J.O.L.-S., M.N., and C.V. contributed to the conception of the study and provided feedback on model structure and analyses. J.O.L.-S. contributed editorial feedback on the manuscript. J.N and A.B. provided data sources, insights into IAV epidemiology and surveillance in U.S. exhibition swine, and feedback on which questions were most relevant to the system.

The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or NSF.

VITA

2008-2012	B.A. in Biology; Minors in Statistical Sciences and Mathematics University of California, Santa Barbara
2012-2014	UCLA Graduate Dean's Scholar University of California, Los Angeles
2012, 2018	Edwin W. Pauley Fellow University of California, Los Angeles
2014, 2015	Teaching Assistant, UCLA Department of Ecology and Evolutionary Biology University of California, Los Angeles
2013-2014	Systems and Integrative Biology Trainee NIH / University of California, Los Angeles
2014-2017	National Science Foundation Graduate Research Fellow, NSF University of California, Los Angeles

PUBLICATIONS

Ambrose, M.R., Nelson, M.I., Viboud, C., Nolting, J.M., Bowman, A.S., and Lloyd-Smith, J.O. Evaluating intervention strategies to reduce zoonotic spillover of influenza A viruses from US exhibition swine: a modeling-based analysis. *In preparation.*

Ambrose, M.R., Lloyd-Smith, J.O. Competition between human and zoonotic pathogens: implications for disease control and the consequences of eradication. *In preparation.*

Lloyd-Smith, J.O., Blumberg, S., Ambrose, M.R., et al. Quantifying the risks of human monkeypox emergence. *In preparation.*

Ambrose, M.R., Kucharski, A.J., Lloyd-Smith, J.O. Inference of a subcritical zoonotic pathogen's transmission dynamics using surveillance data with imperfect spatial information. *In preparation.*

Gostic, K.M., Ambrose, M.R., Worobey, M, Lloyd-Smith, J.O. (2017) Maternal antibodies' role in immunity - Response. *Science*. 355: 705.

Gostic, K.M., Ambrose, M.R., Worobey, M, Lloyd-Smith, J.O. (2016) Potent protection against H5N1 and H7N9 influenza via childhood hemagglutinin imprinting. *Science*. 354: 722-726.

CHAPTER 1:
Competition between cross-immunizing human and zoonotic pathogens: implications for disease control and the aftermath of eradication

1.1 Introduction

Infectious disease management programs often focus on a single target pathogen, ignoring potential interactions with other pathogens and host species. However, pathogens do not exist in isolation: a single host individual may be infected with a multitude of pathogens (many of which may have complicated immune-mediated interactions) and pathogens may move between different host species. The implications of these multi-pathogen, multi-host interactions for disease control remain largely unexplored.

Pathogen community ecology may be of particular importance in the context of pathogen eradication efforts. Following the successful eradication of smallpox and rinderpest, public health agencies and non-profit organizations have become more optimistic about using pathogen eradication to remove public health threats and are currently working to eradicate several other pathogens, such as poliomyelitis, dracunculiasis (guinea worm disease), yaws, and malaria (1–4). While these efforts will assuredly improve human health worldwide, it is important to consider potential unintended consequences of pathogen eradication. From a community ecology perspective, removing an organism from a system may open a 'niche' that can then be filled by competing organisms. Although there has been some debate regarding the validity of niche competition when applied to pathogens (e.g. (5)), the existence of cross-protective immunity within many groups of pathogens suggest that niche replacement is a valid concern. There is already evidence suggesting that Peste des petits ruminants and monkeypox have increased in prevalence following the eradication of their competitors, rinderpest and smallpox, respectively

(6–8). Among the pathogens currently targeted for eradication, the possibility of emergence of related pathogens is already a concern. For example, in a recent study comparing the genomes of human *Plasmodium vivax* and *P. vivax* parasites from other primates, Loy et al. conclude that '*P. vivax* in African apes represents a substantial and genetically diverse parasite reservoir from which future human infections could arise, even if eradication of current human strains were successful' (9).

To investigate how the competition of cross-immunizing pathogens over a limited supply of susceptible individuals may impact pathogen control efforts, we used a deterministic compartmental model to examine the interactions of a human-specific pathogen and a zoonotic spillover pathogen during and following a vaccination campaign. (For clarity, we use the terms 'human pathogen' and 'zoonotic pathogen' in this work; however, these concepts generalize to any system where there is a species-specific pathogen as well as a generalist pathogen that can spill over into that species.) In particular, we focused on several questions relevant to infectious disease control: **1.** Does the added competitive pressure of a zoonotic pathogen that is cross-protective with a human pathogen reduce the level of vaccination required to eradicate the human pathogen? **2.** How will the zoonotic pathogen respond to vaccination pressure? **3.** Following eradication of the human pathogen and cessation of vaccination, how much will the zoonotic pathogen's prevalence increase as a result of competitive release?

As a case study, we applied the model to the smallpox-monkeypox system to examine the current paradigm that monkeypox has invaded the niche space opened by the elimination of smallpox and the cessation of vaccination. This work has implications for estimating the amount of vaccination required to eliminate a pathogen and for interpreting current trends in zoonotic prevalence, and also serves as a reminder that eradicating an endemic pathogen may have

unintended consequences that should be carefully considered when devising management strategies.

1.2 Methods

Model formation

The model expands on the classic continuous-time Susceptible-Infected-Recovered (SIR) model to track the proportion of the host population in each of nine possible states that indicate whether they are currently infected by either pathogen, as well as their infection and vaccination history (see Figure 1.1). Individuals move between states when they are vaccinated and when they are infected by or recover from infection with either the human-specific pathogen or the zoonotic pathogen. The following system of differential equations describes the progression of the system through time:

$$\frac{dS}{dt} = \mu - (\rho_H + \rho_Z + v + \mu) * S$$

$$\frac{dI_H}{dt} = \rho_H * S - (\gamma_H + \mu) * I_H$$

$$\frac{dI_Z}{dt} = \rho_Z * S - (\gamma_Z + \mu) * I_Z$$

$$\frac{dV}{dt} = v * S - ((1-c) * \rho_H + (1-c) * \rho_Z + \mu) * V$$

$$\frac{dR_H}{dt} = \gamma_H * I_H - ((1-c) * \rho_Z + v + \mu) * R_H$$

$$\frac{dR_Z}{dt} = \gamma_Z * I_Z - ((1-c) * \rho_H + v + \mu) * R_Z$$

$$\frac{dI_{HZ}}{dt} = (1-c) * \rho_Z * (R_H + V) - (\gamma_Z + \mu) * I_{HZ}$$

$$\frac{dI_{ZH}}{dt} = (1-c) * \rho_H * (R_Z + V) - (\gamma_H + \mu) * I_{ZH}$$

$$\frac{dR_{HZ}}{dt} = \gamma_H * I_{ZH} + \gamma_Z * I_{HZ} + v * (R_H + R_Z) - \mu * R_{HZ}$$

where $\rho_H = \beta_H (I_H + I_{HZ})$ and $\rho_Z = \beta_Z (I_Z + I_{HZ}) + s_p$ are the rates at which fully susceptible individuals become infected with the human pathogen (via transmission from infectious individuals) and the zoonotic pathogen (via transmission from infectious individuals and spillover). Descriptions of the parameters and the baseline values used in analyses are provided in Table 1.1. Of particular importance, the extent of cross-protective immunity between the pathogens is represented by *c*. When *c*=1, there is complete cross-protective immunity between pathogens, so individuals who recover from infection with one pathogen can no longer be infected by either pathogen. At the other extreme, when *c*=0, individuals who recover from infection with one pathogen experience no protection against future infection with the other pathogen (no cross-protective immunity). Vaccination, which occurs at constant rate *v* in susceptible individuals as well as individuals who have recovered from infection, is assumed to provide the same degree of cross-protective immunity against the human and zoonotic pathogens as the two pathogens provide against one another.

The basic reproduction number, which describes the average number of new infections caused by an infectious individual when the rest of the population is susceptible, is a common metric for describing the transmissibility of a pathogen (10–12). For this system, the basic reproduction number of the human and zoonotic pathogens, respectively, are $R_{0,H} = \beta_H/(\gamma_H + \mu)$ and $R_{0,Z} = \beta_Z/(\gamma_Z + \mu)$, and unless otherwise specified are set at $R_{0,H} = 5$ and $R_{0,Z} = 1.5$. This parameterization corresponds to a system in which the human pathogen is the superior competitor: when there is high enough cross-protective immunity between the pathogens, the

zoonotic pathogen is outcompeted by the human pathogen and is unable to sustain transmission in humans. However, zoonotic infections may still occur because the zoonotic pathogen is repeatedly reintroduced to the human population through spillover events (at rate s_p). In the absence of the human pathogen, the zoonotic pathogen has a high enough basic reproduction number that it would be able to sustain transmission in humans.

Smallpox-monkeypox case study

We simulated smallpox and monkeypox transmission as well as vaccination in the Congo Basin between 1925 and 2010 using a slightly modified version of the model described in section 2.1 (see Appendix for model equations). Instead of using a single parameter, μ, to represent both the per capita birth and per capita death rates, we included separate birth and death rate parameters to reflect the real-world differences in these rates in the Congo Basin. Due to the population growth that results from a higher birth than death rate, we tracked the *number* of individuals in each category, instead of the fraction of the population in each category. The frequency-dependent transmission pattern of the model described in 2.1 was preserved, so that when the birth and death rate parameters are set equal to one another, the two models produce equivalent results.

All parameters used in the smallpox-monkeypox case study were estimated based on demographic and epidemiological studies from the literature. The parameter values used and the references for each value are shown in Table 1.2 and Figure 1.2. These parameter values give a $R_{0,H}$ of 6.9 for smallpox and a $R_{0,Z}$ of 0.8 for monkeypox. The fraction of the population in each infectious state in 1925 was initialized as the equilibrium values under the 1925 vaccination level.

1.3 Results and implications

Effect of competition on the vaccination level needed to eradicate the human pathogen

A classic result in infectious disease modeling is that when a sufficiently large fraction of the population is vaccinated, the pathogen's prevalence decreases to zero (10,22,23). This vaccination-coverage threshold is known as the critical vaccination level, and in the single-pathogen version of the model described above, it is equal to $(1\text{-}1/ R_{0,H})/c$. Our results show that the presence of a cross-protective zoonotic pathogen in the human population reduces the vaccination coverage needed for eradication of the human pathogen below the single-pathogen critical vaccination level. The magnitude of this effect increases as the spillover rate or $R_{0,Z}$ increases (Figure 1.3). Furthermore, because the slope of the critical vaccination versus $R_{0,H}$ curve is steepest at low values of $R_{0,H}$, the presence of a zoonotic pathogen will make the greatest impact on the critical vaccination level for these less-transmissible human pathogens (Figure 1.3). Human pathogens with low $R_{0,H}$ values can be excluded entirely even in the absence of any vaccination and when the zoonotic pathogen has a lower reproduction value, so long as the spillover rate is sufficiently high.

In the context of public health management, calculating the critical vaccination level is an important step in formulating a strategy to control an infectious disease (24–26). The critical vaccination level indicates whether a vaccination-based eradication campaign is feasible and helps optimize the resources expended on control. Because the presence of a zoonotic competitor can substantially decrease the critical vaccination level, public health officials should take the transmission patterns of cross-protective pathogens into account when developing control or eradication strategies.

Effect of vaccination on the zoonotic pathogen

As vaccination pressure increases, the zoonotic pathogen's prevalence declines (Figure 1.4). The magnitude of this effect depends largely on the extent of cross-protective immunity. When cross-protective immunity is weak, the zoonotic pathogen begins at a high prevalence but then declines sharply as vaccination levels are increased. When cross-protective immunity is strong, the zoonotic pathogen's prevalence begins at a lower value, but it experiences little or no decline in prevalence as vaccination levels increase. This phenomenon can be understood if one considers that the susceptible fraction of the population is largely set by the stronger competitor: the human pathogen. The effective reproduction number of the human pathogen ($R_{eff,H}$: defined as the number of new infections caused by an infectious individual in a population that is not necessarily fully susceptible) is determined by the fraction of the population susceptible to infection with that pathogen (S_H), such that $R_{eff,H} = S_H * R_{0,H}$. By definition, at equilibrium the human pathogen's effective reproduction number is one (each infection exactly replaces itself so that its prevalence remains constant). To maintain this equilibrium in the presence of zoonotic spillover and vaccination, the prevalence of the human pathogen will adjust so that the fraction of the population susceptible to the human pathogen at equilibrium remains constant at $1/R_{0,H}$. When there is complete cross-protective immunity, both the human and zoonotic pathogens can only infect completely susceptible individuals, so $S_H = S$. Because the human pathogen's prevalence decreases as vaccination increases to preserve an effective reproduction number of one, the size of the fully-susceptible population stays constant as vaccination increases. Therefore, in the complete cross-protectivity scenario, the zoonotic pathogen sees a constant number of susceptible individuals as vaccination levels increase, up until the human pathogen is eradicated. Once vaccination coverage has reached a high enough level to eradicate the human

pathogen, the effect of additional vaccination increases is no longer partly (or wholly) absorbed by the human pathogen, so the equilibrium zoonotic prevalence declines more sharply as vaccination increases (Figure 1.4).

Understanding the expected response of a zoonotic pathogen to a vaccination campaign will help managers determine whether observed changes in zoonotic prevalence are within the range of anticipated outcomes or whether there is likely some other factor at play that merits further investigation.

Increase in zoonotic pathogen prevalence following eradication of human pathogen

Before eradication of the human pathogen, the zoonotic pathogen competes with the human pathogen and with vaccination for a limited supply of susceptible individuals. After eradication and cessation of vaccination, the zoonotic pathogen no longer shares the susceptible pool and therefore experiences increased equilibrium prevalence. The absolute increase in prevalence is greatest at large $R_{0,Z}$ values because the zoonotic pathogen is able to reach a higher equilibrium infection prevalence. However, the proportional increase in zoonotic prevalence is greatest at intermediate $R_{0,Z}$ values, which balance the higher final prevalence of large $R_{0,Z}$ values with the greater suppression by the human pathogen at lower $R_{0,Z}$ values (Figure 1.5). Furthermore, in systems where the human pathogen has a larger reproduction number or where there is stronger cross-protective immunity, the impact of competition on the zoonotic pathogen is stronger, and therefore the increase in the zoonotic pathogen's prevalence after eradication is greater.

Because the presence of the human pathogen can substantially reduce the zoonotic pathogen's transmission in the human population, estimates of the zoonotic pathogen's

reproduction number that are generated using pre-eradication data may underestimate the true value unless the impact of cross-protective immunity is taken into account. It is worth noting that a zoonotic pathogen that is unable to sustain transmission in humans while competing with the human-specific pathogen may be capable of sustained spread after the human pathogen's eradication. Depending on the anticipated health costs of the zoonotic pathogen's increased prevalence, public health officials may need to consider whether vaccination efforts should be sustained even after eradication of the human-endemic pathogen.

Smallpox-monkeypox case study

To examine whether the findings discussed above, which are based on equilibrium states, are relevant in a real-world context, we ran a dynamic simulation of smallpox and monkeypox transmission in the Congo Basin between 1925 and 2010, covering both pre- and post- smallpox eradication eras. The dynamic smallpox-monkeypox simulation exhibited several of the behaviors reported in the previous sections (Figure 1.6). From 1925 to the late 1960s, the increasing vaccination rate caused only a small decline in monkeypox's simulated prevalence due to the high levels of cross-protective immunity between monkeypox and smallpox (as explained in section 3.2). After smallpox was eradicated but before vaccination was discontinued, monkeypox's prevalence declined more sharply, again, paralleling results from section 3.2. Finally, after smallpox was eradicated and vaccination efforts stopped, monkeypox's simulated prevalence increased substantially as it approached the higher competition-free equilibrium expected in section 3.3 (Figure 1.6). The timescale for monkeypox to asymptote at its competition-free equilibrium is related to the demographic birth and death rates of the population: as older individuals who were infected or vaccinated during the smallpox era are replaced by younger individuals born post-eradication, the system approaches its new

equilibrium state. These results indicate that the behaviors reported in sections 3.2-3.3 have an observable and important impact on real systems.

The results are also of direct relevance to the monkeypox system. The idea that the decline of the vaccinated population has led to an increase in monkeypox prevalence has been proposed previously (7,8). This model provides a theoretical framework to formalize the idea that the relaxation in competition for susceptible individuals led to the increase in monkeypox prevalence. The model was created with the goal of transparency and generality, and thus does not include many important complexities found in the monkeypox system, such as spatial structure of villages, stochasticity, or population structure. These and other complexities not included in this paper's model preclude us from using the simulation output as quantitative predictions for the system. Nonetheless, the qualitative prediction that monkeypox would be expected to increase due to competitive release matches well with the observation and suggests that at least some of the observed increase in monkeypox prevalence is due to competitive release.

Applying these ideas to a heterogeneous world

The present work is a theoretical exploration of a complex situation, and to maximize comprehensibility we have restricted the analysis to a highly homogenous scenario. However, real-world heterogeneities in behavior and risk across the human population are well known to have important impacts on disease dynamics (10,27–32). We expect including spatial heterogeneity in the reproduction number of both pathogens as well as in the spillover rate would have a quantitative effect on our results, though the general qualitative findings should hold. For instance, if spillover risk is higher in remote villages than in cities due to higher rates of human-

animal contact, the presence of the zoonotic pathogen might not have a large effect on the vaccination coverage needed for eradication in cities, but might substantially lower the coverage needed in villages. This effect could be beneficial in the context of an eradication campaign, which may struggle to obtain high vaccine coverage in remote regions. In fact, this type of effect has been proposed to help explain the eradication of the rinderpest virus, where the presence of PPR may have made eradication possible despite vaccination coverage levels in certain regions lower than the expected eradication threshold (6). Furthermore, if the zoonotic pathogen is expected to pose an unacceptable public health burden after eradication of the human-specific pathogen, it may be possible to take advantage of the heterogeneities in spillover to target vaccinations to the areas at highest risk.

Also of relevance to the spillover and disease eradication context, both the human and zoonotic pathogens may have smaller reproduction numbers in remote villages and higher values in densely-packed cities. This means that after eradication of the human-specific pathogen, the increase in the zoonotic pathogen's prevalence if it reaches a city may be substantially higher than in the homogenous case presented here. Furthermore, these heterogeneities in reproduction numbers will also result in different critical vaccination levels required for elimination in different areas. Spatially-structured models, parameterized for the relevant disease system, are needed to explore the implications of these interactions more fully.

While the qualitative patterns from this study will likely hold true in a wide range of situations, the magnitude of effects are likely to be highly dependent on the specific heterogeneities of a particular system. Considering the impact these heterogeneities, in addition to accounting for uncertainty or variability in parameter estimates, will be essential before using the ideas presented here to alter management plans.

Extending to other control scenarios

In this study, we have focused on the relatively simple scenario where vaccination is the only management strategy in place, and the protection provided by vaccination against both the zoonotic and human pathogen is equal (and is the same as the cross-protective immunity between pathogens). These assumptions may be representative of some real-world scenarios, particularly when there is a vaccination available that produces strong, long-lasting immunity. However, additional control practices, such as vector control, bed nets, or drug treatments, will be necessary in many situations to supplement or replace vaccination. In addition, the interventions deployed may have an asymmetric impact on the transmission of the human versus the zoonotic pathogen. A full exploration of the implications of non-immune-mediated or asymmetric control measures is beyond the scope of this work; however, we expect that many of the qualitative patterns discussed here will apply across a wide range of scenarios, so long as the zoonotic and human pathogens are competing over a shared resource of susceptible individuals and the intervention affects the availability of that resource.

Furthermore, here we have focused on systems where the cross-protective immunity between pathogens (and protection from vaccination) is strong and lifelong. The functional relevance of the patterns for management plans will vary between systems, and may become inconsequential in systems where cross-protective immunity is weak or short-lived.

1.4 Conclusions

This work illustrates that a zoonotic competitor can lower both the cost and benefits of eradication: it can decrease the amount of vaccination necessary to eradicate a human pathogen, but its expansion after eradication has the potential to counteract some of the public health

benefits. Applying the model to the monkeypox system indicates that pathogen competition can have real-world consequences and supports the idea that competitive release explains at least some of monkeypox's increase. It is therefore essential to consider the ecological interactions in the broader pathogen community when making decisions about infectious disease control so that consequences of eradication can be anticipated and dealt with appropriately.

1.5 Figures and Tables

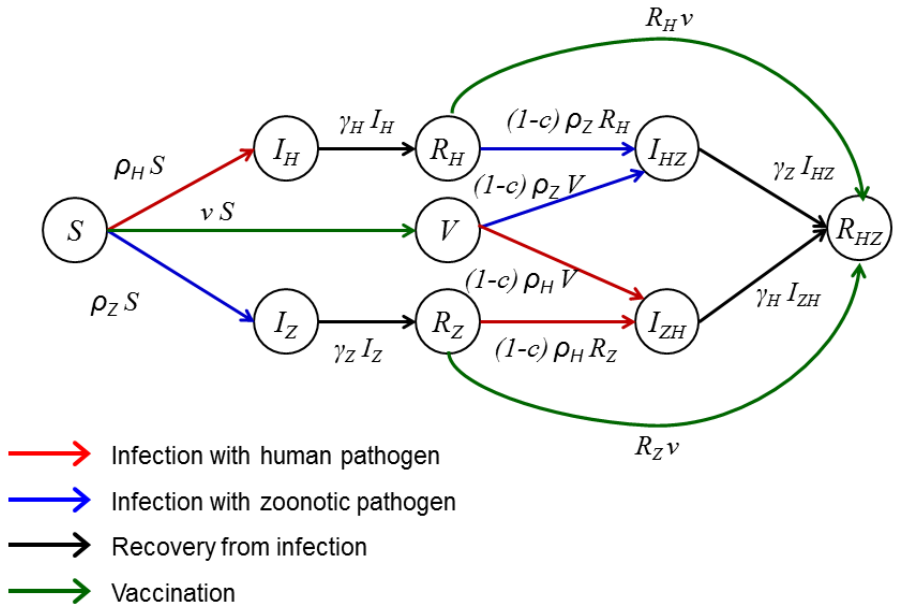


Figure 1.1. Model diagram. Individuals are born into the susceptible state (S). From here, they may be vaccinated (state V) or infected by either the human or zoonotic pathogen (state I_H or I_Z, respectively). After recovering from this first exposure (and moving into state R_H or R_Z), individuals are assumed to be completely immune to the pathogen they were previously infected with and partly protected from infection with the other pathogen. Individuals who have only been vaccinated or infected by one pathogen can be infected by the other pathogen (I_{ZH} if newly infected by the human pathogen or I_{HZ} if newly infected by the zoonotic pathogen). After recovering from this infection (state R_{HZ}), an individual is considered completely immune. Not shown in diagram, all individuals in the population experience a constant per capita death rate μ, and the population size is kept constant by balancing these deaths by births into the susceptible state.

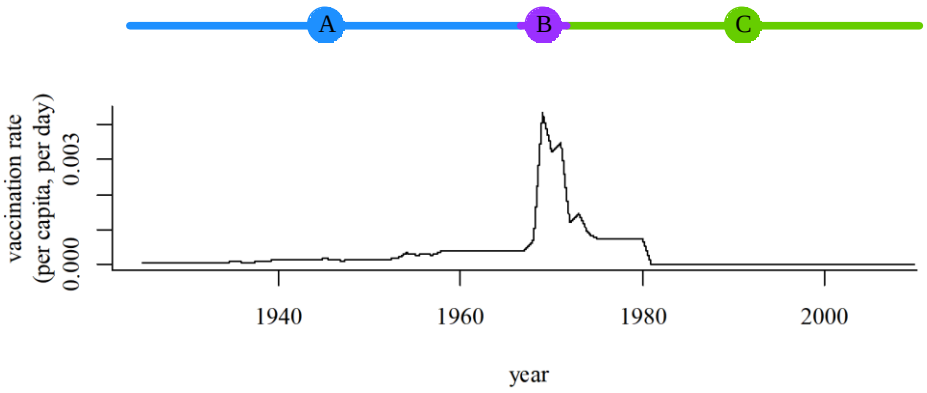


Figure 1.2. Vaccination rates between 1925 and 2010 that were used in the smallpox-monkeypox simulation. The colored bars highlight the periods before (A), during (B), and after (C) smallpox's eradication from the Congo Basin.

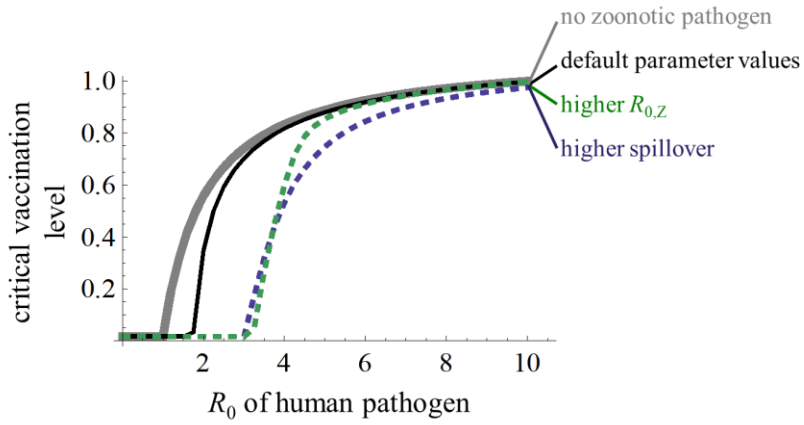


Figure 1.3. Effect of competition on critical vaccination levels. The bold grey line represents the classic curve describing the proportion of the human population that would need to be vaccinated to eradicate the human pathogen (the 'critical vaccination level') versus R_0 of the human pathogen ($R_{0,H}$) without competition. The remaining lines represent the curve under competition with a zoonotic pathogen. The black line shows the baseline parameter values used in this study (s_p = 0.00001 and $R_{0,Z}$=1.5). The green dotted line corresponds to a higher spillover rate (s_p = 1e-04) while the blue dotted line corresponds to a higher zoonotic reproduction number ($R_{0,Z}$ =4). For reference, the two spillover rates result in around 0.4% (for the baseline scenario) and 3.6% (for the high scenario) of the population being exposed to spillover in a given year. All other parameters are as listed in Table 1.1.

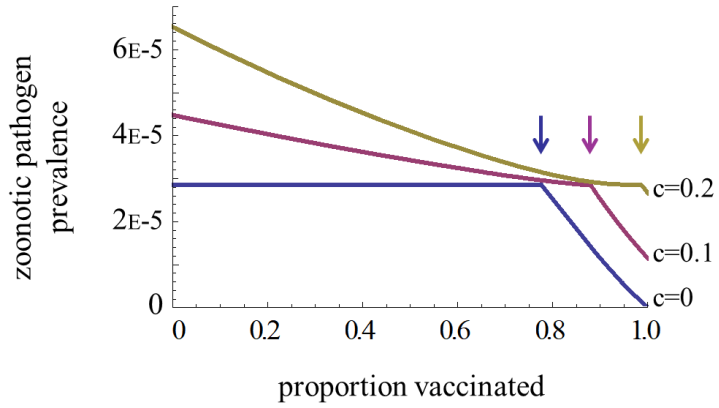


Figure 1.4. Effect of vaccination on zoonotic prevalence. Colors correspond to different levels of cross-protective immunity, with c=1 (blue) indicating complete cross-protection, c=0.9 (magenta) indicating less cross-protection, and c=0.8 (yellow) indicating the least cross-protection. Vertical arrows indicate the vaccination level at which the human pathogen is eradicated.

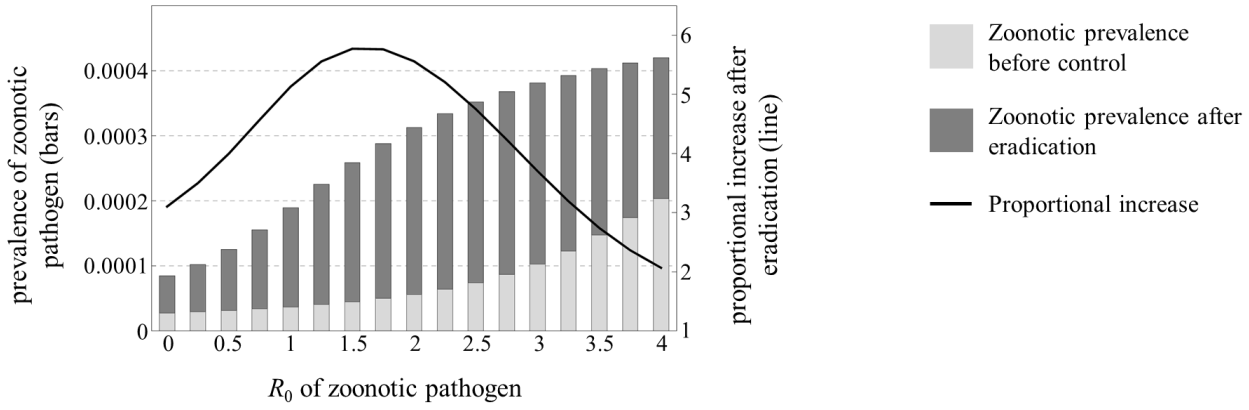


Figure 1.5. Increase in zoonotic prevalence (at equilibrium) after eradication of the human pathogen. Proportional increase of zoonotic pathogen's prevalence after eradication (right axis) is indicated by the line while the bars represent absolute zoonotic prevalence (left axis) before control (light grey bar) and after eradication (dark grey bar) of the human pathogen.

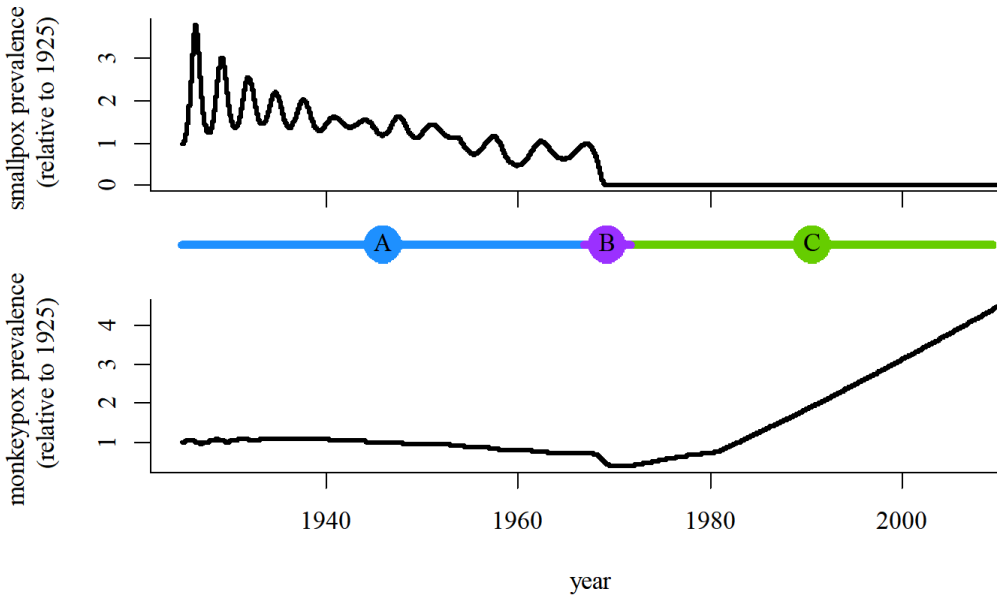


Figure 1.6. Simulation of smallpox-monkeypox dynamics during and following the smallpox eradication campaign. The colored bars indicate the same time periods as in Figure 1.2, which shows the vaccination rate through time. (A – blue bar)Vaccination efforts intensify, resulting in substantial decrease in smallpox. Due to the cross-protective immunity between monkeypox and smallpox, monkeypox experiences only a shallow decrease (relates to Figure 1.4). (B – purple bar) Smallpox has been locally eliminated but vaccination efforts continue. Monkeypox prevalence declines more sharply (relates to Figure 1.4). (C – green bar) Smallpox has been eradicated and vaccination efforts cease. As individuals protected by vaccination or smallpox exposure are replaced with fully susceptible individuals, monkeypox's prevalence increases (relates to Figure 1.5).

Table 1.1. Parameter descriptions and baseline values used in analyses.

Parameter	Description	Default Value
μ	per capita birth and death rate	5.5E-5 individuals per day
β_H	transmission parameter for the human pathogen	0.50 [individuals * day]$^{-1}$
β_Z	transmission parameter for the zoonotic pathogen among humans	0.15 [individuals * day]$^{-1}$
γ_H	recovery rate from the human pathogen	0.1 day $^{-1}$
γ_Z	recovery rate from the zoonotic pathogen	0.1 day $^{-1}$
s_p	spillover rate of the zoonotic pathogen	1E-5 day $^{-1}$
v	per capita vaccination rate	0 day $^{-1}$
c	protection from infection due to cross-protective immunity	0.9

Table 1.2. Parameter values used in the smallpox-monkeypox simulation, along with references.

Parameter	Value	Reference
μ_b	1.3E-4	(13)
μ_d	6.0E-5	(13)
β_H	0.430	(14,15)
β_Z	0.051	(16,17)
γ_H	0.0625	(14,18,19)
γ_Z	0.0625	(14,20)
s_p	4.66e-07	(16,20)
$v(t)$	See Figure 1.2	(7,14,16–18,21)
c	0.8	(7)

1.6 Appendix

For the model used in smallpox-monkeypox analysis, we are tracking the number of individuals in each state, rather than the fraction of individuals as described in the main text. In addition, the model allows for different birth and death rates, resulting in changes in the population size through time:

$$\frac{dS}{dt} = \mu_b - (\rho_H + \rho_Z + v + \mu_d) * S$$

$$\frac{dI_H}{dt} = \rho_H * S - (\gamma_H + \mu_d) * I_H$$

$$\frac{dI_Z}{dt} = \rho_Z * S - (\gamma_Z + \mu_d) * I_Z$$

$$\frac{dV}{dt} = v * S - ((1-c) * \rho_H + (1-c) * \rho_Z + \mu_d) * V$$

$$\frac{dR_H}{dt} = \gamma_H * I_H - ((1-c) * \rho_Z + v + \mu_d) * R_H$$

$$\frac{dR_Z}{dt} = \gamma_Z * I_Z - ((1-c) * \rho_H + v + \mu_d) * R_Z$$

$$\frac{dI_{HZ}}{dt} = (1-c) * \rho_Z * (R_H + V) - (\gamma_Z + \mu_d) * I_{HZ}$$

$$\frac{dI_{ZH}}{dt} = (1-c) * \rho_H * (R_Z + V) - (\gamma_H + \mu_d) * I_{ZH}$$

$$\frac{dR_{HZ}}{dt} = \gamma_H * I_{ZH} + \gamma_Z * I_{HZ} + v * (R_H + R_Z) - \mu_d * R_{HZ}$$

where $\rho_H = \beta_H (I_H + I_{HZ})/N$ and $\rho_Z = \beta_Z (I_Z + I_{HZ})/N + s_p$ and N is the population size. Notice that, like in the equations from the main text, transmission occurs in a frequency-dependent way (based on the fraction of the population in the infectious category). Because I_H,

I_Z, I_{ZH}, and I_{HZ} are now in terms of the *number* of individuals, we divide them by N to get the fraction of the population in each infectious class.

1.7 References

1. Tanner M, de Savingy D. WHO | Malaria eradication back on the table. WHO [Internet]. 2008 [cited 2018 Sep 11]; Available from: http://www.who.int/bulletin/volumes/86/2/07-050633/en/#.W5coziTucZs.mendeley

2. Aylward RB, Birmingham M. The human story. BMJ Br Med J [Internet]. 2005;331(7527):1261–2. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1289331/

3. Asiedu K, Fitzpatrick C, Jannin J. Eradication of Yaws: Historical Efforts and Achieving WHO's 2020 Target. PLoS Negl Trop Dis [Internet]. 2014;8(9):e3016. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4177727/

4. Bill and Melinda Gates Foundation. Malaria Strategy Overview [Internet]. 2018 [cited 2018 Nov 9]. Available from: https://www.gatesfoundation.org/What-We-Do/Global-Health/Malaria

5. Fenner F, Hall A, Dowdle W. What is eradication? In: Dowdle W, Hopkins D, editors. The Eradication of Infectious Diseases. Chichester, UK: Wiley; 1998. p. 3–18.

6. Roeder PL. The animal story. BMJ [Internet]. 2005 Nov 24;331(7527):1262 LP-1264. Available from: http://www.bmj.com/content/331/7527/1262.abstract

7. Rimoin AW, Mulembakani PM, Johnston SC, Smith JOL, Kisalu NK, Kinkela TL, et al. Major increase in human monkeypox incidence 30 years after smallpox vaccination campaigns cease in the Democratic Republic of Congo. Proc Natl Acad Sci [Internet]. 2010; Available from: http://www.pnas.org/content/early/2010/08/24/1005769107.abstract

8. Lloyd-Smith JO. Vacated niches, competitive release and the community ecology of pathogen eradication. Philos Trans R Soc Lond B Biol Sci [Internet]. 2013;368(1623):20120150. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3720048&tool=pmcentrez&rendertype=abstract

9. Loy DE, Plenderleith LJ, Sundararaman SA, Liu W, Gruszczyk J, Chen Y-J, et al. Evolutionary history of human Plasmodium vivax revealed by genome-wide analyses of related ape parasites. Proc Natl Acad Sci [Internet]. 2018 Sep 4;115(36):E8450 LP-E8459. Available from: http://www.pnas.org/content/115/36/E8450.abstract

10. Anderson RM, May RM. Infectious diseases of humans : dynamics and control. Oxford. New York: Oxford University Press; 1991.

11. Heesterbeek JAP, Dietz K. The concept of Ro in epidemic theory. Stat Neerl [Internet]. 1996 Sep 11;50(1):89–110. Available from: https://doi.org/10.1111/j.1467-9574.1996.tb01482.x

12. Heffernan JM, Smith RJ, Wahl LM. Perspectives on the basic reproductive ratio. J R Soc Interface. 2005;2.

13. United Nations DESA / Population Division. World Population Prospects 2017 [Internet]. 2017. Available from: https://esa.un.org/unpd/wpp/DataQuery/

14. Eichner M, Dietz K. Transmission Potential of Smallpox: Estimates Based on Detailed Data from an Outbreak. Am J Epidemiol [Internet]. 2003 Jul 15;158(2):110–7. Available from: http://dx.doi.org/10.1093/aje/kwg103

15. Gani R, Leach S. Transmission potential of smallpox in contemporary populations. Nature [Internet]. 2001 Dec 13;414(6865):748–51. Available from: http://dx.doi.org/10.1038/414748a

16. Fine PEM, Jezek Z, Grab B, Dixon H. The Transmission Potential of Monkeypox Virus in Human Populations. Int J Epidemiol [Internet]. 1988;17(3):643–50. Available from: http://ije.oxfordjournals.org/content/17/3/643.abstract

17. Lloyd-Smith JO, et al. Quantifying the risks of human monkeypox emergence in the aftermath of smallpox eradication. Prep.

18. Fenner F, Henderson DA, Arita I, Jezek Z, Ladnyi ID. Smallpox and its Eradication. World Health Organization; 1988.

19. Meltzer MI, Damon I, LeDuc JW, Millar JD. Modeling potential responses to smallpox as a bioterrorist weapon. Emerg Infect Dis [Internet]. 2001;7(6):959–69. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2631899/

20. Ježek Z, Fenner F. Human monkeypox [Internet]. S Karger Ag; 1988. Available from: http://books.google.com/books?id=fyupMQEACAAJ

21. Schneider WH. Smallpox in Africa during Colonial Rule. Med Hist [Internet]. 2009 Apr;53(2):193–227. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2668906/

22. Smith CEG. Prospects for the Control of Infectious Disease. Proc R Soc Med [Internet]. 1970 Nov;63(11 Pt 2):1181–90. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1811055/

23. Fine P, Eames K, Heymann DL. “Herd Immunity”: A Rough Guide. Clin Infect Dis [Internet]. 2011 Apr 1;52(7):911–6. Available from: http://dx.doi.org/10.1093/cid/cir007

24. Elbasha EH, Dasbach EJ, Insinga RP. Model for Assessing Human Papillomavirus Vaccination Strategies. Emerg Infect Dis [Internet]. 2007;13(1):28–41. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2725801/

25. Orenstein WA, Gay NJ. The Theory of Measles Elimination: Implications for the Design of Elimination Strategies. J Infect Dis [Internet]. 2004;189(Supplement 1):S27–35. Available from: http://jid.oxfordjournals.org/content/189/Supplement_1/S27.abstract

26. Anderson RM, May RM. Directly transmitted infections diseases: control by vaccination. Science (80-) [Internet]. 1982 Feb 26;215(4536):1053 LP-1060. Available from: http://science.sciencemag.org/content/215/4536/1053.abstract

27. Arita I, Wickett J, Fenner F. Impact of Population Density on Immunization Programmes. J Hyg (Lond) [Internet]. 1986;96(3):459–66. Available from: http://www.jstor.org/stable/3863139

28. Woolhouse MEJ, Dye C, Etard J-F, Smith T, Charlwood JD, Garnett GP, et al. Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. Proc Natl Acad Sci [Internet]. 1997 Jan 7;94(1):338 LP-342. Available from: http://www.pnas.org/content/94/1/338.abstract

29. Grenfell, Bolker. Cities and villages: infection hierarchies in a measles metapopulation. Ecol Lett [Internet]. 1998 Jul 1;1(1):63–70. Available from: http://dx.doi.org/10.1046/j.1461-0248.1998.00016.x

30. Keeling MJ, Woolhouse MEJ, Shaw DJ, Matthews L, Chase-Topping M, Haydon DT, et al. Dynamics of the 2001 UK Foot and Mouth Epidemic: Stochastic Dispersal in a Heterogeneous Landscape. Science (80-) [Internet]. 2001;294(5543):813–7. Available from: http://www.sciencemag.org/content/294/5543/813.abstract

31. Sebastian J. Schreiber, James O. Lloyd-Smith. Invasion Dynamics in Spatially Heterogeneous Environments. Am Nat [Internet]. 2009;174(4):490–505. Available from: http://www.jstor.org/stable/10.1086/605405

32. Keeling MJ, Rohani P. Modeling Infectious Diseases in Humans and Animals. Princeton Univ. Press; 2008. 366 p.

CHAPTER 2:
Quantifying transmission of emerging zoonoses: Using mathematical models to maximize the value of surveillance data

2.1 Introduction

Many recent infectious disease threats have been caused by pathogens with zoonotic origins, including Ebola, pandemic H1N1 influenza, and SARS- and MERS- Coronaviruses, and zoonotic pathogens are expected to be a primary source of future emerging infectious diseases (1–8). By definition, zoonotic pathogens can transmit from animals to humans; those also capable of human-to-human transmission are of particular public health concern (5,9). Infectious disease surveillance serves a crucial role for detecting and gathering information on zoonotic pathogens: data obtained through surveillance are often the primary resource available for informing public health management decisions (10). Developing methods that improve our ability to infer information about a pathogen's transmission dynamics from available surveillance data is therefore an essential frontier for understanding and ultimately combating these pathogens (11,12).

For zoonoses, three epidemiological measures are crucial for summarizing transmission dynamics and informing risk assessments. The first of these is the spillover rate, which indicates how frequently the pathogen is transmitted from the animal reservoir into humans and helps inform the total expected disease incidence (13). The second measure describes the pathogen's potential for further spread once in the human population and is commonly assessed using the reproductive number (R), which gives the average number of secondary human cases caused by an infectious individual (14,15). Values of R greater than one indicate that the pathogen is capable of sustained (i.e. 'supercritical') transmission in humans. Pathogens with subcritical

transmission (*R* less than one but greater than zero) can cause limited chains of transmission in humans after a zoonotic introduction, and they pose a risk of acquiring ability for supercritical transmission via evolutionary or environmental change (2,5,16). The third epidemiological measure is the distance over which human-to-human transmission occurs, which informs how the disease will spread spatially and the risk of it being introduced into new populations. Combined, these three measures can help evaluate the current public health threat posed by the pathogen, the risk of future emergence, and the most effective approaches for disease management.

Estimating epidemiological measures is a challenging task in any pathogen system, and the unique properties of zoonotic diseases can exacerbate these difficulties. Infectious disease surveillance often records temporal information and certain aspects of spatial information about human cases, but the underlying transmission events are seldom observed. In a zoonotic system, this means that an observed human infection could have been caused by a previous human case or by zoonotic spillover. Without intensive contact tracing, or sequence data in the case of fast-evolving pathogens, quantifying the relative contribution of zoonotic versus human-to-human transmission is a major challenge; identifying the source of infection for specific individuals is an even bigger one.

Epidemiological analyses are often hindered by data truncation and unknown denominators (17,18). In many disease surveillance systems, the total set of localities under surveillance (i.e. those that would appear in the dataset if a case occurred there) can be separated into ‘observed localities,’ which appear in the dataset because they reported one or more cases, and ‘silent localities,’ which have no cases during the surveillance period and therefore do not appear in the dataset. This form of truncation, where localities with zero cases are absent from

the dataset, obscures the true scope of the surveillance effort. Without knowledge of the total number of localities under observation (the 'unknown denominator'), accurately estimating the spillover rate and probability of human-to-human transmission between localities is not straightforward. Simply disregarding these silent localities in the analysis is the functional equivalent of selectively removing zeros from the dataset and can lead to problematic inference biases.

Complicating inference efforts further is the fact that surveillance datasets often report the geographic location of cases only at a coarse resolution, obscuring information about a transmission process that occurs on a much finer scale (19–21). Precise spatial information is often absent from historic datasets and data collected in remote or low-resource areas, replaced by the names of the locality and broader administrative units where the case occurred. For example, only the village name and the region and country to which the village belongs may be recorded in a dataset. Furthermore, linking a village name to spatial coordinates is often impossible when maps of the region do not exist or only unofficial local names are used. Although collecting exact spatial coordinates has become more practical in contemporary disease surveillance, privacy and confidentiality concerns can arise in both human and agricultural contexts when data contains high-resolution spatial information (19,20,22–25), leading to data being reported in a non-localized manner. Methods that can use this inexact spatial information are especially needed for zoonotic diseases, where any additional information about the proximity of human cases to one another can improve the power to distinguish between human-to-human transmission and zoonotic spillover.

Despite these challenges, a series of research efforts have expanded our ability to estimate the transmission properties of zoonotic pathogens from case onset data. A key set of

methods revolve around inferring *R* from the sizes of case clusters (a cluster is defined as a group of cases that occur in close spatiotemporal proximity to one another) or from the proportion of observed cases that were infected by zoonotic spillover (16,26–30). However, these approaches either require detailed case investigations to determine whether a case was infected by a zoonotic or human source or assume that each cluster is caused by one single spillover event followed by human-to-human transmission. A likelihood-based approach for estimating *R* for human-to-human transmission using only symptom onset dates of cases was introduced by Wallinga and Teunis (31). This method was extended to apply to zoonotic systems by Lo Iacono et al. (32), but the extension requires that chains of exclusively human-to-human transmission can be identified, and is thus not applicable to many zoonotic surveillance systems where human and zoonotic transmissions are intermixed. A different approach was taken by White and Pagano (33), who introduced a different likelihood-based method that compares the observed number of cases on each day with the expected number, as calculated using the number and timing of previous cases. Though the White and Pagano approach was only applicable to human-to-human transmission, it was expanded by Kucharski et al. (34) to work in zoonotic spillover systems in scenarios where a control measure, implemented at a known point in time, causes an abrupt reduction in spillover. A related approach that requires knowledge of the human and animal reservoir population sizes was also explored in Lo Iacono et al. (35). Crucially, however, none of these methods incorporate information about the spatial location of cases to improve inference power or to estimate patterns of spatial spread. Spatial data is a powerful tool in transmission inference in single-species studies (e.g. (36–39)), but has largely been excluded from analyses of zoonotic transmission, which often implicitly assume homogenous mixing across the study area or that

human-to-human transmission can only occur within a locality. One recent exception to this is the analysis by Cauchemez et al. (40), which includes transmission at several spatial levels.

In this work, we present model-based inference methods that allow us to infer *R*, the spillover rate, and properties of spatial spread among humans from surveillance datasets with non-localized spatial information and an unknown total number of surveilled localities. Our approach builds on methods introduced by White and Pagano (33) and Kucharski et al. (34), but allows continuous spillover throughout the surveillance period and makes use of available spatial information on case location. While the method could be readily adjusted to incorporate more precise geographic information should it be available, in this study we focus on the more challenging scenario in which only the names of the locality and broader administrative units where a case occurred are known. To make use of this form of non-localized spatial data, our model considers two scales of spatial mixing and transmission (Figure 2.1A), reminiscent of the 'epidemics with two levels of mixing' structure utilized in Ball et al. (41) and Demiris and O'Neill (42). The first mixing level is the locality in which the case occurred, such as a village or a farm, conceptualized as a group of individuals geographically separated from other localities. We assume that individuals within the same locality have more frequent contact with one another than with individuals from other localities, and therefore that infection is more likely to be transmitted within a locality. However, the total number of localities under surveillance is unknown because only localities with one or more cases appear in the dataset (the 'unknown denominator' problem discussed above). We refer to the second spatial level as the 'broader contact zone.' It describes a collection of localities that all occur within the same administrative unit and likely share some amount of human movement. When multiple types of administrative units of different sizes are reported in the dataset (e.g., districts, regions, provinces, etc.), the

ideal choice for broader contact zone is the smallest administrative unit that contains inter-locality human-to-human transmission events. If this scale is not known *a priori*, inferring the appropriate scale of administrative unit is necessary.

We tested the method against a variety of datasets simulated using different epidemiological parameters, offspring distributions for human-to-human transmission, and spatial transmission kernels. To assess the performance of the method, we compared the estimated and true values for epidemiological measures such as the reproductive number and spillover rate, and also examined how well the method was able to estimate the probable transmission source of each case. When silent localities were not accounted for, substantial biases arose in zoonotic spillover rate estimates. However, a modified method that accounts for these silent localities was successful in a wide range of circumstances. We therefore applied this 'corrected-denominator method' to a dataset on human monkeypox cases from an active surveillance effort conducted in the Democratic Republic of the Congo (formerly Zaire) in the 1980s (43) (Figure 2.2). We repeated the analyses for four different assumptions about the appropriate spatial scale to use to represent the 'broader contact zone' over which human-to-human transmissions take place and selected the preferred option using the deviance information criterion (DIC) method for model comparison. In the monkeypox dataset, contact-tracing data are available for a subset of the cases, providing a rare opportunity to compare inferred transmission sources with those suggested by epidemiological investigation. In addition, some localities were associated with known GPS coordinates, enabling us to estimate the spatial transmission kernel in greater detail. As such, our monkeypox analysis yielded estimates of R and the spillover rate, as well as insights into the spatial scale of human transmission of monkeypox.

2.2 Results

Overview of the approach

We first validated the inference framework using a simulation study, then applied the validated method to a dataset on human monkeypox cases to estimate key epidemiological parameters and the spatial scale of transmission. To generate simulated test datasets and perform parameter inference, we used a mathematical model of the zoonotic pathogen's transmission into and among humans. The model tracks the number of human cases that occur in each locality on each day; infections can arise from spillover from the zoonotic reservoir or from human-to-human transmission (Figure 2.1B). Three key parameters govern the behavior of the system. The spillover rate (λ_z) describes the average number of human cases caused by animal-to-human transmission ('primary cases') in each locality per day. The reproductive number of the pathogen (R) determines the average number of ('secondary') cases caused by each infected human. And the spatial dispersal of the pathogen is controlled by the fraction of cases arising from human-to-human transmission that occur in the same locality as the source case (σ) and the rules governing inter-locality transmission events. Two spatial scales of transmission are included in the model: within the locality of the case and between localities in the same broader contact zone. Using this model (described further in Methods 4.1) and values for the three parameters, the likelihood of observing $N_{t,v}$ cases on each day t and locality v can be calculated. Markov chain Monte Carlo (MCMC) methods were used to infer posterior parameter distributions (and hence parameter estimates) for a given dataset of cases.

Robustness of model-based inference method

Basic method (assumes the total number of localities under surveillance is known)

To assess the accuracy and precision of our method's estimates of spillover and transmission parameters, we simulated datasets with known parameter values and compared these true values with the inferred values. We investigated a range of R and λ_z values in the neighborhood of values previously estimated for monkeypox (16,44), with R ranging from 0.2 to 0.6 and λ_z ranging from 0.0001 to 0.0007 expected spillover events per locality per day (λ_z values correspond to 59 to 415 expected spillover events in the five year simulation period, across all localities). Transmission events between humans had a probability σ=0.75 of occurring within a locality and otherwise were equally likely between any localities in the same broader contact zone. We were interested in seeing how well the inference methods are able to use the spatial-temporal arrangement of cases to estimate the true parameter values.

Across 125 simulations (25 simulations for each of five parameter sets), estimated values clustered around the true parameter values. The true value for R was included in the 95% credible interval (CI) 119 times (95.2%) and for λ_z was included 121 times (96.8%) (Figure 2.3A). On average, the posterior mean estimate of R differed from the true value by 8.6%; the analogous percent errors for λ_z and σ estimates were 6.3% and 7.0%, respectively (Table S2.1).

However, this method assumes that the true number of localities under surveillance is known. In real-world situations, 'silent' localities that experience zero cases often do not appear in the dataset, resulting in an unknown true number of surveilled localities. We investigated possible biases in parameter estimates that could arise from assuming that the number of localities that reported one or more cases represents the total number of localities under surveillance. To do so, we used the same set of simulated datasets as described above, but

removed knowledge about the number of silent localities. In these datasets, silent localities make up between 21% and 85% of all localities under surveillance, with the proportion driven primarily by the spillover rate. Estimates for the reproductive number *R* were not strongly impacted (95.2% of the 95% CIs contained the true value with an average percent error of 8.4%), but the spillover rate λ_z was consistently overestimated (Figure 2.3B). The true value for λ_z was contained in none of the simulations' 95% CIs and the posterior mean had an average percent error of 153% (Table S2.1).

To further investigate the effect of this data truncation (whereby localities with zero cases do not appear in the dataset), we performed inference assuming that the observed localities represented all, 1/2, or 1/5 of the total localities under surveillance. While this assumption had a relatively small impact on the estimated *R*, it greatly impacted the inferred λ_z (which is measured as the number of spillover events *per locality* per day and is therefore strongly affected by changes in the assumed number of localities) (Figure S2.1). Assuming that a larger fraction of surveilled localities appear in the dataset resulted in substantially higher estimated spillover rates.

Corrected-denominator method (conditions on the locality observation process)

Because the total number of localities assumed to be under surveillance has a substantial impact on parameter estimates, we developed a modified version of the likelihood function that accounts for localities that were under surveillance but never observed in the dataset. This approach calculates the likelihood of the observed dataset conditional on the fact that only localities with one or more cases are included (details on the modified likelihood function can be found in Methods 4.2 and Appendix).

We tested the performance of the corrected-denominator method against simulated datasets, looking at the same parameter sets as in the first section. The inferred parameter values cluster well with their corresponding true values (Figure 2.3C): mean percent error in *R* estimates was 8.4% and in λ_z estimates was 14.0%. Across the 125 simulations, the true parameter value was included in the 95% CI 116 times (92.8%) for *R* and 117 times (93.6%) for λ_z (Table S2.1).

Because an estimate of the true number of localities under surveillance would help determine the size of the population that could be detected for a given system, we assessed how well we could approximate this value. Given the number of localities with one or more cases and the mean parameter estimates, it is possible to calculate the expected total number of localities under surveillance (see Appendix). Estimates of the true number of localities calculated for the simulated datasets center on the correct value (Figure S2.2). The magnitude of estimate error is driven by the spillover rate, which largely determines the proportion of localities that are observed by surveillance. The mean percent error across simulations with spillover rate of 0.0001, 0.00036, and 0.0007 were 25.4%, 7.9%, and 2.4%, respectively, while simulations with spillover rates of 0.004 and above almost always recorded at least one case in each locality and therefore tended to estimate the exact true number of localities.

Inferring the sources of transmission events

We investigated how well sampled transmission trees recovered the source of individual cases as well as higher-order measures, such as the fraction of cases originating from zoonotic, within-locality, and between-locality transmission. We tested our method using 125 simulated datasets, with 25 datasets simulated for each of five sets of true parameter values (these are the same datasets as discussed above, simulated with *R* between 0.2 and 0.6 and spillover rate

between 0.0001 and 0.0007). Two hundred plausible transmission trees were sampled for each simulated dataset.

When comparing the overall fraction of cases attributed to each source type (zoonotic versus within-locality versus between-locality transmission), the sampled transmission trees closely match the true transmission patterns (Figure 2.4). On average, the difference between the true fraction of cases caused by zoonotic spillover and the fraction inferred in a tree was 0.022 (standard deviation 0.018), the difference for within-locality transmission was 0.006 (standard deviation 0.005), and the difference for between-locality transmission was 0.022 (standard deviation 0.018).

The success at recovering individual transmission links was high overall but varied slightly depending on the true parameters underlying the simulation (Figure S2.3). On average, sampled transmission trees inferred 85.9% of all sources correctly. Better performance was observed for lower spillover rates and lower *R*, presumably due to the fewer opportunities for misattribution of cases. Some transmission links were more likely to be captured than others: on average 90.9% and 90.1% of sampled trees correctly inferred links with zoonotic and within-locality sources, respectively, but only 36.8% of trees correctly identified the source of between-locality transmission events.

Epidemiological insights into monkeypox

Applying the corrected-denominator method to 1980s monkeypox surveillance data

Between 1982 and 1986, the active monkeypox surveillance program in the Democratic Republic of the Congo detected 331 human cases in 171 localities (43). For each human case, we know the name of the locality as well as the district or administrative subregion (henceforth

referred to simply as ‘district’) and region to which it belongs. However, the total number of localities that would have been detected by surveillance had they experienced a case is unknown. We therefore used the corrected-denominator method to generate estimates under four different assumptions about which administrative unit most suitably represents the broader contact zone. The country-level, region-level, and district-level models correspond to progressively smaller choices of broader contact zone, while the locality-level model assumes that all instances of human-to-human transmission occur within a locality. We anticipate that assuming an excessively large broader contact zone could result in overestimating R and underestimating λ_z if too many spurious human-to-human transmission events are inferred from pairs of cases that just happen to occur within a generation-time interval of one another, while assuming an inappropriately small broader contact zone could result in the opposite parameter biases if the model is unable to detect actual incidents of human-to-human transmission because the cases occur in different (assumed) broader contact zones.

In the monkeypox analysis, the size of the administrative unit used as the broader contact zone has a strong effect on the resulting parameter estimates (Figure 2.5A). When larger administrative units are assumed to represent the broader contact zone, a given pair of cases is more likely to belong to the same broader contact zone, giving the model more opportunities to infer inter-locality human-to-human transmission events and resulting in larger estimated reproductive number R and a smaller spillover rate λ_z. Mean values of the posterior distribution of R range from 0.29 when transmission is assumed to occur only within localities to 0.52 when transmission is assumed to occur among all localities in the country (Table 2.1).

We used the mean parameter estimates obtained using each of the four broader contact zone assumptions to generate estimates of the expected total number of localities under

surveillance. While only 171 localities were observed in the dataset, estimates of the total number of surveilled localities ranged from 337 (using the locality-level model) to 408 (using the country-level model). The district-level and region-level models generated similar estimates of 351 and 366 total localities, respectively.

Insights into how underlying assumptions drive monkeypox estimates

We investigated how different assumptions about the true number of localities and the spatial scale of human-to-human transmission would affect the parameter estimates for the monkeypox system. To explore how the presence of silent localities affects results, we repeated the analysis using the basic method (which does not account for silent localities) under the assumption that the localities observed in the monkeypox dataset represent all, 1/2, and 1/5 of the total number of localities that were under surveillance. Furthermore, for each of these assumptions about the total number of localities under surveillance, we repeated the analysis using the four different choices of broader contact zone to determine how the assumed spatial scales of inter-locality transmission impacted inference results.

Both the choice of broader contact zone and the assumed total number of localities have a large impact on estimates of R and λ_z (Figure 2.5B). As noted above, models assuming smaller broader contact zones allow fewer opportunities for human-to-human transmissions to be inferred, and these models estimate substantially lower R values and correspondingly higher spillover rates. In contrast, assuming that a smaller fraction of surveilled localities were observed leads to slightly higher estimates of R and substantially lower estimates of λ_z because the presence of many silent localities effectively 'dilutes' the observed number of spillover events *per locality* per day and drives the estimate lower. Estimates of R are most strongly affected by the choice of broader contact zone, while estimates of λ_z are most strongly impacted by assumed

fraction of localities observed. For all assumptions of broader contact zone and total number of localities, the means of the parameters' posterior distributions fall along the line

$$R = 1 - \frac{V*T*\lambda_z}{N} \quad , \qquad (1)$$

where *V* is the true number of localities, *T* is the number of days over which surveillance occurred, and *N* is to total number of cases in the monkeypox dataset. This relationship arises because the expected number of total cases is equal to the expected number of spillover events ($V * T * \lambda_z$) multiplied by the total number of human cases expected to occur from each spillover event ($1/(1-R)$ for $0<R<1$). Each assumption about the total number of localities under surveillance corresponds to a separate line along which parameter estimates fall (Figure 2.5B). The position of the parameter estimates along this line depends on the spatio-temporal distribution of the *N* cases and the assumed spatial scale of human-to-human transmission.

District-level broader contact zone preferred in model comparisons

To assess which broader contact zone assumption is most appropriate for the monkeypox system, we used the deviance information criterion (DIC) to perform model comparisons for the corrected-denominator method as well as for each assumption about the number of surveilled localities. For the corrected-denominator method, the district-level model had the best DIC score, followed by the region and country-level models (Table 2.1). The locality-level model received a much larger DIC value, indicating that the data strongly support models that allow transmission between localities. Similarly, for each of the three assumptions about the true number of surveilled localities, the district-scale model performed best in DIC model comparisons (Table 2.1).

Inferring the sources and distances of transmission events

We used the district-level corrected-denominator method to sample 20,000 transmission trees for the monkeypox dataset. The sampled transmission trees attributed an average of 60.8% (standard deviation of 2.2%) of cases to zoonotic spillover, 28.5% (standard deviation of 0.9%) to within-locality human-to-human transmission, and 10.7% (standard deviation of 2.1%) to between-locality human-to-human transmission. For comparison, the results using the three other broader contact zone assumptions are shown in Figure S2.4A. Each model's trees include a similar number of within-locality human-to-human transmission events, but increasing the spatial scale of the broader contact zone increases the number of inferred between-locality transmission events.

To characterize the distance range over which inter-locality transmission occurs, we focused on links in the sampled transmission trees that occurred between cases with known GPS coordinates (280 out of 331 monkeypox cases had recorded GPS coordinates). The number of transmission events in each sampled tree that occurred over a certain distance was then compared to the number of transmission events expected to occur over each distance if transmission between all localities in a broader contact zone was equally likely (see Methods 4.3 for how this 'null distribution' was calculated).

For all models allowing inter-locality transmission, more transmission events were inferred to occur across ≤ 30 kilometers than expected based on the null distribution (Figure 2.6, Figure S2.4B). For each inferred transmission tree, a binomial test was used to examine whether more transmissions were inferred to occur over ≤ 30 kilometers than expected based on the null distribution of transmission distances. Out of 20,000 sampled trees for each model, p-values of less than 0.1 were obtained in 93% of the district, 72% of the region, and 81% of the country-

level models' trees. The median p-values for these three models were 0.007, 0.030, and 0.012, respectively (Figure S2.5 shows the full distributions of p-values obtained across all sampled trees).

Comparison of sampled transmission trees with contact-tracing data

Contact-tracing, where the contacts of a case were recorded and follow-up investigations determined whether or not the contacts had become infected, was done for a subset of monkeypox cases. Instances where a contact developed an infection are presumed to be instances of human-to-human transmission. For each of these epidemiologically contact-traced links, we looked at how frequently the sampled transmission trees for each model captured the transmission link.

Of the 53 case pairs linked through contact tracing, an average of 79.5% (standard deviation of 4.2%) were recovered in each of the district model's sampled transmission trees (Figure 2.7A). The highest success was seen for pairs of epidemiologically-linked cases whose dates of symptom onset were between 7 and 25 days apart (Figure 2.7B). Although it is generally believed that the generation interval for human-to-human transmission of monkeypox is between 7 and 23 days (43,45), several case pairs that occurred more than 23 days apart were epidemiologically linked. It is possible that these links, which were often missed in the sampled transmission trees, are not true instances of human-to-human transmission. Cases that occurred in different localities were also less likely to be linked in a sampled transmission tree, though even for these inter-locality pairs, the district-level model tended to perform better than the other three models (Figure S2.6) The four models had similar success at recovering within-locality links. In all models, when a link was incorrectly inferred, it frequently was inferred to originate from zoonotic spillover instead. Although the district model had the highest success at

recovering contact-traced links, the sampled trees from all models recovered an average of >76% of contact pairs.

Comparison of the transmission tree generated using only contact-tracing data with the trees created using the district-level and locality-level models highlights how much our perception of the transmission dynamics depends on assumptions about spatial spread (Figure 2.8). Most of the within-locality transmission links detected through epidemiological contact-tracing appear in the locality-level model's tree, though the locality-level tree suggests substantially more human-to-human transmission events than captured in the contact-tracing tree. However, the locality-level tree misses all inter-locality links. The district-level model's tree captures most of the links indicated by the locality-level tree, and also suggests that inter-locality transmission is occurring, though it has low power to determine exactly which case pairs are linked through inter-locality transmission.

Sensitivity analyses

We conducted a variety of sensitivity analysis tests using simulated datasets to assess how robust the method was over a range of parameter values and assumption violations (full descriptions are provided in the appendix). The method continued to perform well even at very high spillover rates (Figure S2.7, Table S2.3) and when the offspring distribution used in simulations differed from the one assumed in the inference (Figure S2.8, Table S2.4). In some situations, assuming a larger broader contact zone than the one used for simulations could lead to an overestimation of R and an underestimation of λ_z (Tables S2.5, S2.6). This outcome is consistent with what was observed in the monkeypox analysis where assuming a larger spatial scale for the broader contact zone corresponded to a higher estimate of R and a smaller estimate

of the spillover rate (Figure 2.5). When simulations were run with highly structured, non-homogeneous spillover, substantial biases in the inference results occurred because this spillover process gives rise to clusters of primary cases that the model mistakes as arising from human-to-human transmission (Figure S2.9).

2.3 Discussion

Principal findings

In this work, we developed and tested a method to infer fundamental epidemiological parameters and transmission patterns for zoonotic pathogens from epidemiological surveillance data with aggregated spatial information. When tested against simulated datasets, the method successfully recovered estimates of R and spillover rate close to the true values and also inferred the fraction of cases resulting from zoonotic, within-locality, and between-locality sources with a high degree of accuracy. The 'unknown denominator problem' that occurs when the total number of localities under surveillance is unknown can cause large biases in parameter estimates, so we modified the inference method to account for this observational process and enable unbiased estimation in the presence of this common data gap.

We applied the method to a rich surveillance dataset of human monkeypox in the Congo basin from the 1980s and found that human-to-human transmission of monkeypox between localities plays an important role in the pathogen's spread. Of the four assumptions we tested for the spatial scale of the broader contact zone, the district-level model was best supported by DIC model comparisons and validation with contact-tracing. In addition, the signal of elevated inter-locality transmission occurring over ≤ 30 kilometers suggests that most inter-locality transmissions occur in a relatively small neighborhood, consistent with the limited transportation

infrastructure in the DRC. This further corroborates that the district-level model, which is the smallest spatial aggregation scale that still permits inter-locality transmission, is likely the most appropriate choice for capturing inter-locality transmission patterns of human monkeypox.

The district-level model estimates a reproductive number for human monkeypox of 0.38 (0.31-0.45 95% CI). This value is slightly higher than previous estimates of *R* for the 1980s DRC monkeypox dataset, which was estimated as 0.30 (90% CI 0.22-0.40) in Blumberg and Lloyd-Smith (16), as 0.32 (90% CI 0.22-0.40) in Lloyd-Smith et al. (46), and as 0.28 in Jezek et al. (44). There are several explanations for the higher estimate we obtained. The previous studies may have underestimated the reproductive number, particularly if contact-tracing or cluster formation methods were liable to miss transmissions that occurred between localities. Indeed, the estimate obtained using the locality-level model ($R = 0.29$) closely matches previous estimates. It is also possible that the district-level model may overestimate the amount of human-to-human transmission in the same way that the region- and country-level models picked up a higher signal of human-to-human transmission than the district-level model due to their larger broader contact zone sizes. The size of the DRC's districts and administrative subregions used for the district-level model vary in size, but average around fifteen thousand square kilometers, or around one hundred forty kilometers across, encompassing a much greater distance than most human-to-human transmission events likely occur over. We therefore expect that the true value of *R* is bounded by the estimates of the locality-level and the district-level models.

In addition to providing an estimate of monkeypox's reproductive number, the methods give insight into the frequency of spillover and the spatial scale of human-to-human transmission. The district-level model estimates a mean spillover rate of around 0.11 spillover events per locality per year, which corresponds to roughly one spillover event every nine years in

each locality. It also estimated that around 70% of human-to-human transmissions occur within a locality. This finding contrasts with the assumption that human-to-human transmission occurs within a locality, which is commonly used to generate transmission clusters, and suggests that estimates generated using that assumption may substantially underestimate the amount of human-to-human transmission occurring in the system. The importance of inter-locality contacts has been reported for the neighboring country of Uganda, where a survey by le Polain de Waroux et al. (47) on rural movement and social contact patterns indicated that 12% of social contacts occurred outside participants' village of residence.

Among human monkeypox cases with recorded geographical coordinates, a clear signal emerged of higher rates of human-to-human transmission between localities ≤ 30 kilometers apart. This pattern seems reasonable given the infrastructure and general difficulty of transportation in the more remote regions of the DRC. It also suggests a similar pattern of movement as found in the le Polain de Waroux et al. (47) survey. Their analyses indicate that 90% of people who traveled outside their village of residence remained within 12 km.

Spatial scale of transmission and aggregated spatial data

The potential biases introduced when analyzing data reported at a course spatial scale have been explored in a wide range of contexts (48–50), yet the implications of using this type of spatial information to infer the transmission dynamics of an infectious disease is not obvious. When spatial information is only reported at the level of large spatial zones like districts, regions, or countries, no finer-scale information is available to inform which human cases transmitted infection to one another between different localities. Here we explored how the size of these spatial zones would affect inference for the monkeypox system by repeating the analysis using

spatial information at the district, region, or country resolution. The large differences in parameter estimates generated under different broader contact zone assumptions in the monkeypox analysis illustrates how sensitive inference results can be to the spatial scale assumed for human-to-human transmission, and suggests that reporting spatial data at too large a scale or ignoring inter-locality transmissions can lead to substantial estimate biases.

In the context of monkeypox in the DRC, analysis of simulations using the exact geographic coordinates reported for 80% of localities in the monkeypox surveillance dataset replicated the increasing estimates of R and decreasing estimates of spillover rate as the spatial aggregation scale increased (Tables S2.5, S2.6). However, the magnitude of the effect in simulated datasets was smaller than in the monkeypox analysis. This could be a result of the particular assumptions about inter-locality transmission patterns used in the simulations, but it also opens the question of whether outside large-scale factors such as seasonality or fluctuations in surveillance effort might induce temporal autocorrelation among unlinked human cases, giving rise to temporal clustering of cases that the model interprets as human-to-human transmission.

This analysis serves to emphasize the importance of selecting an appropriate spatial scale and using caution when interpreting results obtained using spatially aggregated data. Many methods implicitly assume a certain scale of spatial transmission, often ignoring the possibility of longer-range transmissions, so careful consideration of whether that scale is appropriate for the system is essential.

In general, recording precise spatial locations of cases is vital for increasing the inferential power of modeling analyses. Developing methods that maintain spatial information without risking a breach in confidentiality is a nontrivial challenge, but progress has already been

made in generating possible solutions such as geographic masking or the verified neighbor approach (51,52).

Model assumptions and future directions

In this work, we assumed that the spillover rate was homogenous through time and space, but more complex disease dynamics in the reservoir or spatiotemporal heterogeneity in animal-human contacts may cause nontrivial deviations from this assumption in real-world systems. Of particular concern is the possibility that outbreaks in the reservoir could cause periods of amplified local spillover, which could create a clustering pattern of human cases potentially indistinguishable from human-to-human transmission. Without information about disease dynamics in the reservoir, accounting for this heterogeneous spillover will be challenging, but certain types of pathogen dynamics, such as seasonal epidemics or expanding wave-fronts of infection, could be incorporated into the model structure.

Similarly, spatially and temporally variable surveillance intensity could potentially mimic the signal of human-to-human transmission clusters and result in overestimates of the reproductive number. Future surveillance programs could help mitigate this challenge by recording a measure of surveillance effort undertaken at each location and time.

This work assumes that R is constant across all localities; however, to obtain a full picture of pathogen emergence risk, it may be necessary to consider the heterogeneity in transmission intensity among different human populations, as well as the interplay between where R is highest versus where spillover tends to occur (53). In some zoonotic systems, for instance, spillover predominantly occurs into remote villages and towns that are in close proximity to forested regions. However, we generally expect these villages to have lower levels of human-to-human

transmission than the more densely-packed cities (54–56). A pathogen may even be incapable of supercritical spread until it reaches such a city. Therefore, to assess the probability a pathogen will successfully emerge and to determine which populations to target with control measures, it may be necessary to establish not only the spillover rate and R across different populations, but also the rate of dispersal of the pathogen between those populations (53).

Several assumptions may need to be modified when applying this method to other zoonotic systems. Because we assume that the source of human-to-human transmission events will show symptoms before the recipient, the likelihood function can treat human cases as occurring independently conditional on preceding cases. For zoonotic diseases in which infected individuals frequently transmit the pathogen before showing symptoms (or when asymptomatic cases contribute non-negligibly to transmission), the likelihood expression would need to be modified substantially, and the lack of independence between cases might make a simulation-based inference approach necessary.

We assume that sufficiently few infections occur relative to the population size that depletion of susceptible individuals does not affect transmission dynamics. While appropriate when there are few human infections or in the early stages of invasion, this assumption could bias estimates if applied in a system with sufficiently high levels of human infection or where transmission occurs primarily among highly clustered contacts, such as individuals within a household. We also note that in the monkeypox example we are estimating the *effective* reproductive number, which takes into account existing population immunity. If the goal instead were to establish the basic reproductive number (the reproductive number for the pathogen in a fully susceptible human population), accounting for past exposure to the pathogen or other cross-immunizing pathogens or vaccines would be necessary.

The current methods assume that all human cases that occur during the surveillance period inside the surveillance area are observed. This assumption is plausible for the analysis of the 1980s monkeypox dataset, given the unusually high resources and experience level of this surveillance effort in the aftermath of the smallpox eradication program and the use of serology to detect missed cases retrospectively (43). However, most zoonotic surveillance systems operate with limited resources and have a much lower detection rate. Ignoring unobserved cases will lead to underestimation of the spillover rate, while the effect on estimation of R will depend on the nature of the surveillance program. For instance, in the chain-size analyses of Ferguson et al. (28) and Blumberg and Lloyd-Smith (16), R is underestimated when the detection probability of each case is independent of one another or when right-censoring occurs but overestimated when a detected case triggers a retrospective investigation that detects all cases in that transmission chain.

Conclusions

This work expands our ability to assess and quantify important zoonotic pathogen traits from commonly available epidemiological surveillance data, even in the absence of exact spatial information or a complete count of localities under surveillance. We anticipate that these methods will have greatest value in the common circumstance when the source of cases, particularly whether a case came from an animal or human source, cannot be readily established. In such situations, the ability to infer the pathogen's reproductive number, spillover rate, and spatial spread patterns from available surveillance data, will greatly enhance our understanding of the pathogen's behavior and could provide valuable insights to help guide surveillance design and outbreak response.

2.4 Methods

Model

In broad terms, the model describes the probability of observing a set of symptom onset times and locations of human cases given the timing and location of previous cases and parameters that underlie the transmission process. Human infections can arise from either animal-to-human transmission ('zoonotic spillover') or human-to-human transmission (Figure 2.1B). Human-to-human contact occurs more frequently within a locality than between localities, but can still occur between localities that belong to the same broader contact zone (Figure 2.1A).

All sources of infection are assumed to generate new cases independently of one another. The number of human cases that become symptomatic on each day in each locality caused by zoonotic spillover is assumed to follow a Poisson distribution with mean λ_z. For simplicity and because reservoir disease dynamics are rarely well characterized, we assume the Poisson process is homogenous through time and across localities, but this assumption could be modified for a system where more information is available about the reservoir dynamics (e.g., (34)). New infections can also arise from contact with infected humans. The number of new infections that become symptomatic on day *t* in locality *v* caused by an infectious individual who became symptomatic on day *s* in locality *w* is assumed to be a Poisson-distributed random variable with mean $\lambda_{\{s,w\},\{t,v\}}$.

Aggregating cases caused by all sources of infection (both human and zoonotic), the total number of new cases on day *t* in locality *v* is a Poisson-distributed random variable with mean

$$\mu_{t,v} = \sum_{s=1}^{t-1} \sum_{w=1}^{\mathcal{V}} [N_{s,w} * \lambda_{\{s,w\},\{t,v\}}] + \lambda_z \quad , \qquad (2)$$

where $\mathcal{V}$ is the number of localities under surveillance and $N_{s,w}$ is the number of cases with symptom onset on day *s* in locality *w*.

The mean of the Poisson random variable describing human-to-human transmission, $\lambda_{\{s,w\},\{t,v\}}$, depends on the reproductive number of the pathogen in humans, the generation time distribution, and the coupling between localities:

$$\lambda_{\{s,w\},\{t,v\}} = R * g(t-s) * H(v,w) \ , \quad (3)$$

where *R* is the reproductive number of the pathogen; *g*(*t*-*s*) is the generation time distribution, which gives the probability that a secondary case becomes symptomatic *t*-*s* days after the index case shows symptoms; and *H*(*v*,*w*) describes the amount of transmission between localities *v* and *w* and takes values between zero (if no transmission can occur between localities *v* and *w*) and one (if all cases arising from an infected individual in locality *v* arise in locality *w*). The generation time *g*(*t*-*s*) is assumed to follow a negative binomial distribution. For this study, we used a mean of 16 days and a dispersion parameter of 728.7 (calculated by fitting a negative binomial distribution to observed generation interval counts for smallpox presented in Fig. 2b of (57)), which is consistent with previous estimates of the generation time for both smallpox and monkeypox (43,45,57,58).

The factor that describes the amount of transmission that occurs between localities *v* and *w* (*H*(*v*,*w*)) could reflect Euclidean distance, travel time, inclusion in different spatial zones, or any other available measurement. To accommodate the imperfect spatial information available for many zoonotic surveillance systems, this study focused on developing methods for the situation when only a locality name and an aggregated spatial zone (such as district or country) is reported for cases, rather than an exact position. We assume that inter-locality transmission

occurs only among localities within the same broader contact zone (Figure 2.1A). Because transmission will be greater within a locality than between localities, a proportion σ of secondary cases are assumed to occur in the same locality as the source case and a proportion (1- σ) of secondary cases are assumed to occur amongst the outside localities that are within the same broader contact zone as the source case. This outside transmission is assumed to be divided equally among all localities within the index case's broader contact zone:

$$H(v,w) = \begin{cases} 0\ , & Z_v \neq Z_w \\ \sigma\ , & v = w \\ \frac{(1-\sigma)}{(\mathcal{V}_v - 1)}\ , & Z_v = Z_w, v \neq w \end{cases} ,$$

where Z_v indicates the broader contact zone of locality v and $\mathcal{V}_v$ is the total number of localities in the broader contact zone of locality v. For a given locality v, the sum of $H(v,w)$ across all w equals one. To observe the effect of assuming different broader contact zones, the monkeypox case study was repeated under four different assumptions about the spatial scale of human-to-human transmission: locality, district, region, and country-level.

Model inference

Likelihood function

Using the model described above, a likelihood function was used to evaluate a parameter set ($\theta = \{R, \lambda_z, \sigma\}$) given the data ($D = N_{t,v}$ cases observed on each day t and locality v):

$$\mathcal{L}(\theta|D) = \prod_{t=1}^{T} \prod_{v=1}^{V} \frac{e^{\mu_{t,v}}\ {\mu_{t,v}}^{N_{t,v}}}{N_{t,v}!} ,$$

where T is the number of days surveillance was conducted and V is the total number of localities under surveillance.

While this approach works well when the total number of surveilled localities is known (see Figure 2.3A), localities often only appear in the dataset if they have reported cases; as a result we may not know the total number of localities under surveillance. Ignoring localities with zero cases can lead to biased parameter estimates (see Figure 2.3B). We explored several alternative approaches to account for these silent localities; the preferred approach rescales the likelihood function to reflect that localities with zero cases are not included in the data. Several approximations are made in this approach to estimate unknown parameters and improve computational tractability. The details of the derivation for the model are given in the appendix, and the final likelihood function is:

$$\mathcal{L}(\theta|D) = \prod_{w=1}^{W} \frac{\prod_{t=1}^{T} \frac{e^{\mu_{t,w}} \mu_{t,w}^{N_{t,w}}}{N_{t,w}!}}{\left(1-e^{-\lambda_z T-\left(\frac{R\,T\,\lambda_z\,(1-\sigma)(E[V]-1)}{E[V]-1-R(E[V]-2+\sigma)}\right)}\right)},$$

where W is the number of observed localities (localities with one or more cases) and $E[V]$ is the expected number of localities given the parameter values and the number of observed localities.

Parameter estimation

Markov chain Monte Carlo (MCMC) was used to obtain the posterior distributions of the model parameters. The fraction of transmissions occurring within a locality (σ) and the reproductive number (R) were given uniform priors on zero to one. The expected number of spillover events per locality per day (λ_z) was given a uniform prior with a lower bound of zero and an upper bound selected to be far above the converged posterior distribution (ranging from 0.0075 to 1, see Figure S2.10 for comparison of spillover priors and posterior distributions).

The chains were run for 100,000 steps, with a burn-in of 20,000. They satisfied visual inspection for convergence. In addition, the Gelman and Rubin multiple sequence diagnostic was evaluated for three parallel chains from each of the models for the monkeypox dataset (59). The Gelman-Rubin potential scale reduction values were less than 1.00033 across all models, indicating that the chains have converged close to the target distribution (60).

DIC model comparisons:

For the monkeypox dataset, four assumptions about the choice of broader contact zone were compared using the deviance information criterion (DIC). This approach combines a complexity measure, used to capture the effective number of parameters in each model, with a measure of fit in order to perform model comparisons. Models are rewarded for better 'goodness-of-fit' to the data and penalized for increasing model complexity. Similarly to the well-known Akaike information criterion (AIC) model comparisons, models with smaller DIC values are preferred. As a rule of thumb, a difference between models' scores of four or more generally indicates that the model with the larger value is 'considerably less' well supported by the empirical evidence; similarly, a difference of ten or more indicates 'essentially no' empirical support (61). The values necessary to calculate the DIC can be readily obtained from the MCMC output (62).

Transmission tree reconstruction

The origin of cases (zoonotic spillover, intra-locality human-to-human transmission, or inter-locality human-to-human transmission) and the distances of inter-locality human-to-human transmission events (when case localities are known) can be established given a particular transmission tree. To gain estimates of these measures, trees were sampled based on the model

and the parameter posterior distributions. From the MCMC output (representing draws from the posterior distribution), d_1 sets of parameter estimates were drawn to create d_1 transmission-probability matrices (***P***). The entry P_{ij} describes the probability that individual *i* was infected by individual *j*. The diagonal values of the matrix represent the probability a case originated from zoonotic spillover. For a case *i* observed to occur on day *t* in locality *v*, the probability that case *j* was the source of case *i* (P_{ij}) was taken to be the proportion of $\mu_{t,v}$ (the expected total number of cases on that day and locality; defined in equation 2) contributed by case *j*. By sampling d_2 transmission trees from each of these transmission-probability matrices, we were able to calculate the proportion of cases that resulted from spillover, within-locality transmission, and between-locality transmission in each sampled tree. When testing the method against 125 simulated datasets, 200 sampled transmission trees were generated for each datasets, with d_1 =20 and d_2 =10. For the monkeypox dataset, 20000 transmission trees were generated with d_1 =200 and d_2 =100.

For inferred inter-locality human-to-human transmission events in the monkeypox dataset, if the GPS coordinates were known for both localities in a transmission pair, the transmission distance was calculated using the *gdist* function in the R package Imap (63). The 'null distribution,' used for comparing the number of inferred inter-locality transmission events with the number expected to occur if spatial location played no role in transmission, was calculated by pooling all cases for which locality GPS coordinates are known, sampling all inter-locality pairs permitted by the model, and recording the distance between the localities in each pair.

Simulation of test datasets

To test the effectiveness of the methods, datasets with known parameter values were simulated using the model explained above. Simulations were run over 1825 days (approximately 5 years) and 325 surveilled localities. The localities were assumed to be partitioned across thirty districts and six regions, with the distribution of localities across districts and regions similar to that observed for the monkeypox dataset. The generation time interval (the number of days between symptom onset of the source and recipient cases) was assumed to follow a negative binomial distribution with a mean of 16 days and a dispersion parameter of 728.7 (as described above), with a maximum generation time interval of 40 days. A number of parameter sets, as well as different underlying model structures, were used for simulations (Table S2.2). Simulation parameters were chosen to approximate the monkeypox dataset, with σ set at 0.75, R ranging from 0.2 to 0.6, and λ_z ranging from 0.0001 to 0.1. Unless otherwise specified, simulations were performed assuming the district-level model. Details on the models used for sensitivity analyses that use the exact spatial location of cases or allow highly structured and non-homogenous spillover patterns are provided in the appendix.

Monkeypox data

Data on human monkeypox cases in the Democratic Republic of the Congo (DRC), formerly 'Zaire,' were collected as part of an intensive surveillance program supported by the World Health Organization. During the peak surveillance period, between 1982 and 1986 (64), data on 331 cases of laboratory-confirmed human monkeypox were recorded (see Figure 2.2) (43). As part of field investigations, mobile teams visited the locality of a monkeypox case to collect information about the case, such as the date of fever and rash onset (for this study, the

symptom onset date was taken to be the fever onset date; if the date of onset was not recorded, the rash onset date was used instead), as well as to identify individuals who had had close contact with the case (44,65). If one of these contacts developed monkeypox within 7 to 21 days of first exposure, the presumptive source case was recorded (43,65).

Between 1982 and 1986, human monkeypox cases were observed in 171 distinct localities, distributed among 30 districts and administrative subregions (simply referred to as ‘districts’) and 6 regions. The total number of localities that could have been detected by surveillance is unknown. Of the 171 observed localities, GPS coordinates are available for 136 localities (which corresponds to 280 out of 331 cases). The district, region, and country of a locality were always recorded.

2.5 Figures and Tables

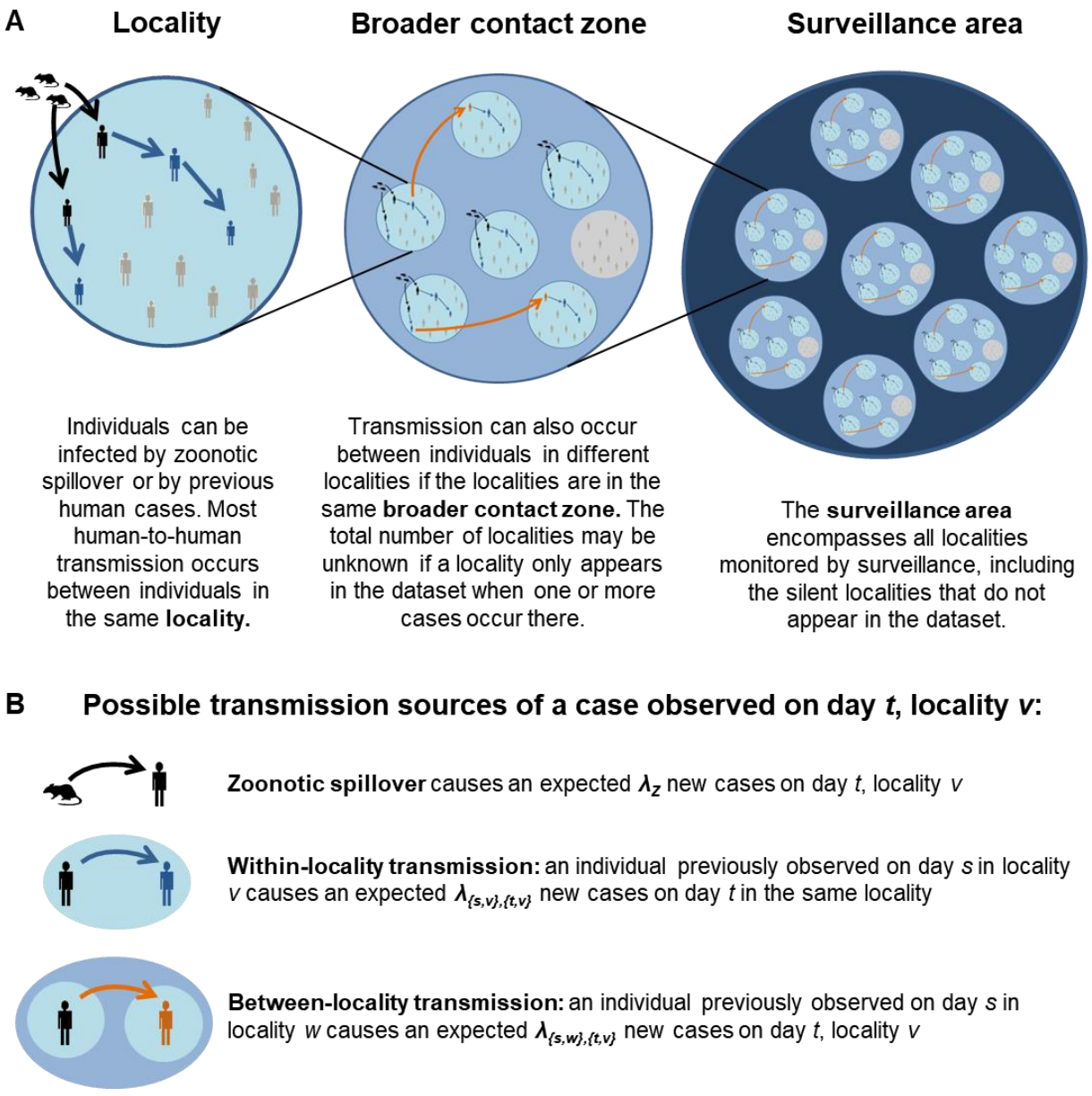


Figure 2.1. Model schematic. **A.** The schematic illustrates the spatial scales considered in the model and the types of transmission that occurs at different scales. Human cases are represented in black if they were infected by zoonotic spillover, blue if they were infected by within-locality human-to-human transmission, and orange if infected by between-locality human-to-human transmission. Individuals who are not infected are colored gray and do not appear in the surveillance dataset. Similarly, if zero individuals in a locality are infected, that 'silent locality' does not appear in the dataset (represented by the gray locality in the broader contact zone). **B.** The possible sources of human infection, which in aggregate determine the number of new

infections on day *t*, locality *v*. The number of cases arising from spillover and human-to-human transmissions follow Poisson distributions with means λ_Z and $\lambda_{\{s,w\},\{t,v\}}$, respectively.

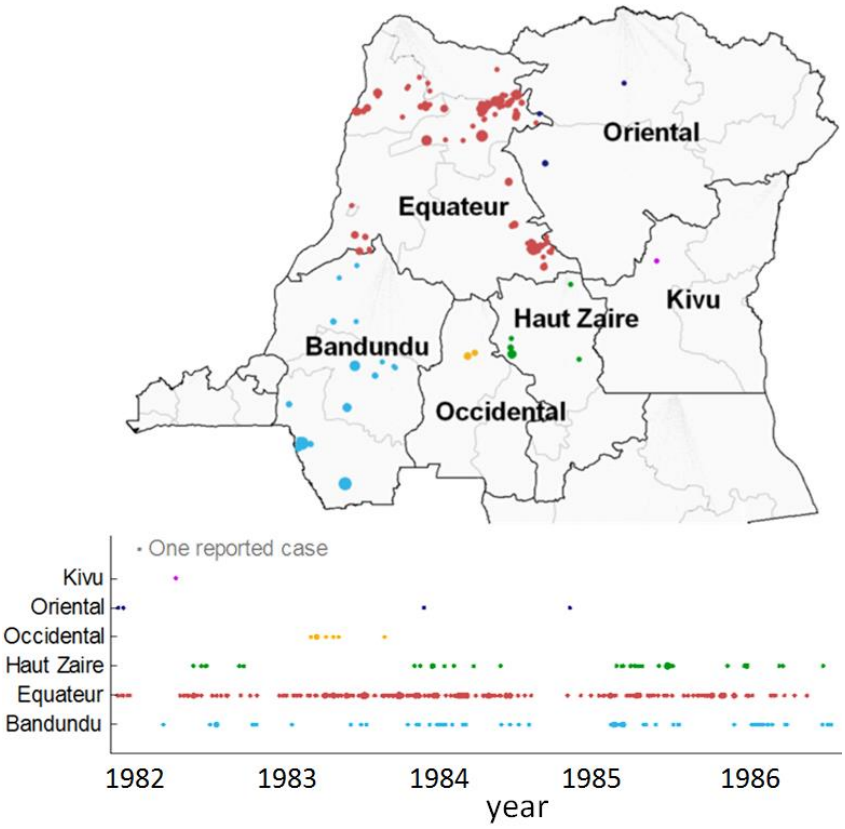


Figure 2.2. Map and time-series showing locations and dates of human monkeypox cases. The size of points on the map indicate the number of cases and the color of points corresponds to the region in which the cases occurred. Dark lines indicate region boundaries while light lines indicate the official boundaries for districts (though in the monkeypox surveillance dataset these may be further divided into administrative subregions).

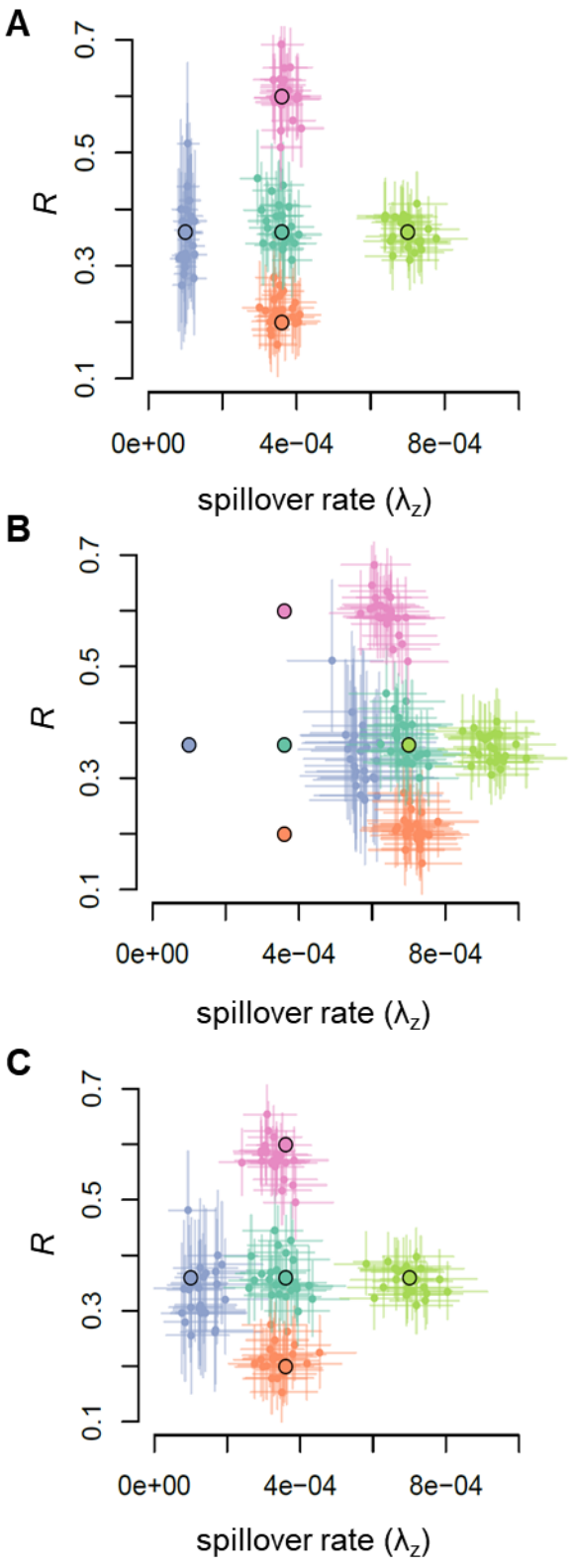

A
R
0.1
0.3
0.5
0.7
0e+00
4e−04
8e−04
spillover rate (λ_z)
B
R
0.1
0.3
0.5
0.7
0e+00
4e−04
8e−04
spillover rate (λ_z)
C
R
0.1
0.3
0.5
0.7
0e+00
4e−04
8e−04
spillover rate (λ_z)

Figure 2.3. Comparison of true and inferred parameter values in simulation study. Within each color, large points outlined in black indicate the true parameter set and smaller points indicate the inferred parameter values from simulated datasets (lines show the 95% credible interval). Inferences were performed **A)** when the true number of localities under surveillance was known, **B)** when the true number was unknown and it was assumed that the number of observed localities was the total number of localities, and **C)** when the true number of localities was unknown and the corrected denominator method was used to control for the locality observation process.

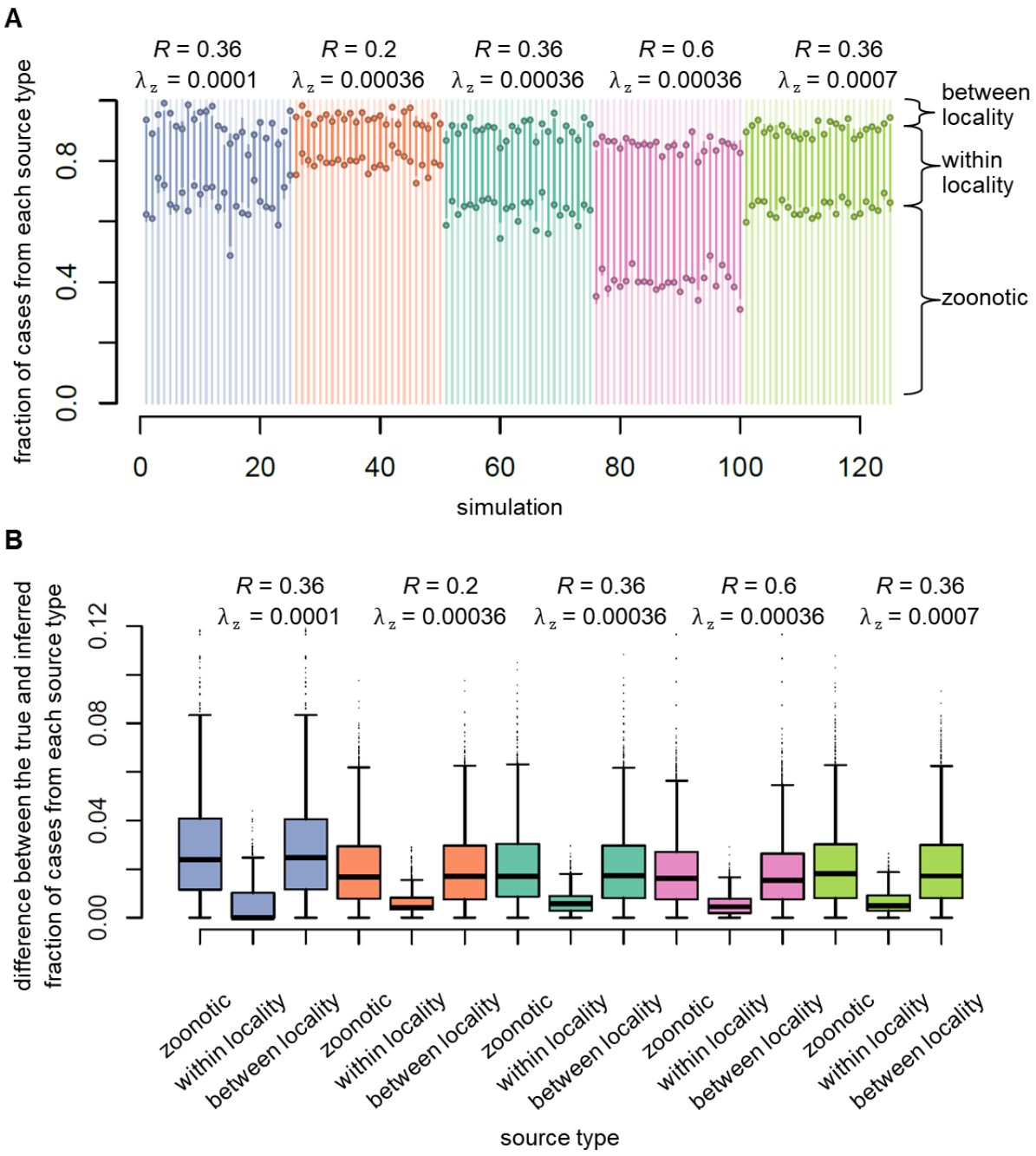


Figure 2.4. Comparison of the true and inferred fraction of transmissions from each source type. For each of five parameter sets, 25 datasets were simulated and 200 transmission trees were sampled for each of these simulated datasets. **A.** Stacked bars show the true fraction of transmissions from zoonotic (bottom bar, medium-darkness), within-locality (middle bar, light color), and between-locality (top bar, darkest color). Points on the bars indicate the inferred values. **B.** Box plots summarize the error in the inferred fraction of cases originating from each source type The error size is small across all parameter sets, especially for within-locality

human-to-human transmission. The upper whisker was calculated as min(max(x), Q_3+1.5*IQR) and the lower whisker was calculated as max(min(x),Q_1-1.5*IQR).

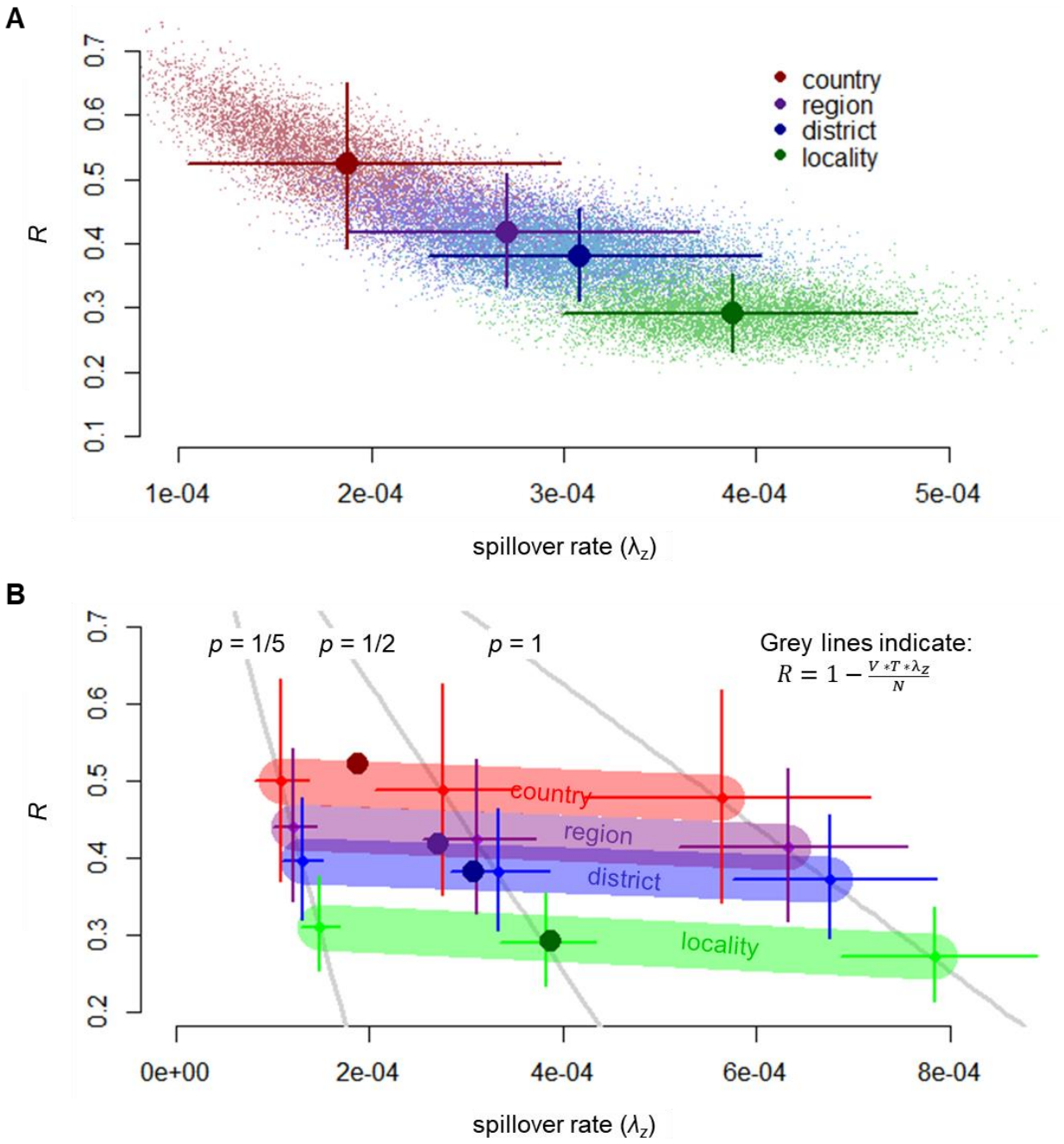


Figure 2.5. Assumptions about the total number of localities under surveillance and the broader contact zone affect parameter estimates for the monkeypox dataset. Estimates and 95% CIs for the reproductive number (R) and the spillover rate (λ_z) of the monkeypox dataset are shown for each of the four choices of spatial scale for the broader contact zone (locality = green, district = blue, region = purple, country = red). **A.** Inference performed using the corrected denominator method that accounts for silent localities. Light background dots are draws from the posterior, larger dots designate the mean value, and bars indicate the 95% CI. **B.** Inference performed assuming that the fraction of localities with one or more monkeypox cases (p) is 1/5, 1/2, or 1.

For each assumption about the total number of localities, parameter estimates fall roughly along the line $\boldsymbol{R = 1 - \frac{V*T*\lambda_z}{N}}$ (indicated by grey lines); their position along this line depends on the spatial model used. Note that the slope of each line is proportional to -1/*p* because *V* = (number of observed localities) / *p*. Dots represent the mean posterior estimates and bars indicate the 95% CI. The four darker dots show the mean estimates from panel **A.**

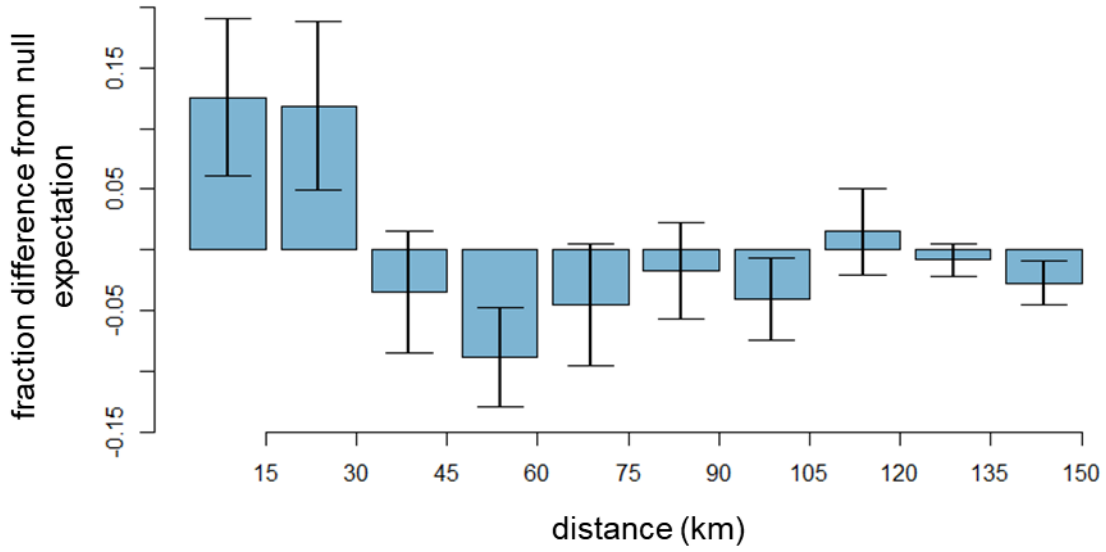


Figure 2.6. Distance of inferred inter-locality human-to-human transmission events. Bars show the difference between the proportion of inter-locality human-to-human transmissions inferred to occur over a given distance by the district model and the proportion expected based on the spatial distribution of localities (the ‘null expectation’).

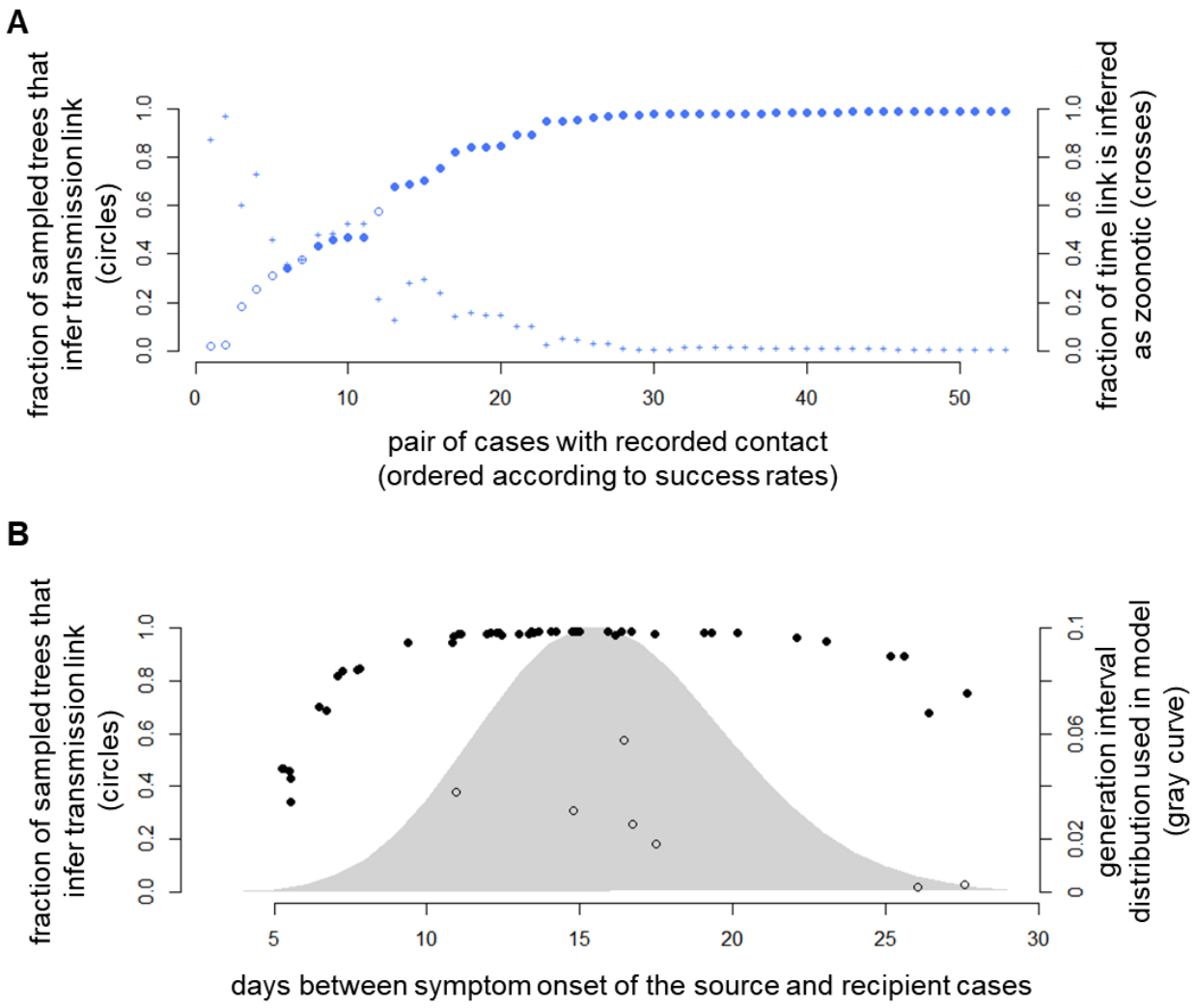


Figure 2.7. Comparison of epidemiologically contact-traced links with sampled transmission trees. **A.** Circles (left axis) show the fraction of sampled trees that infer the epidemiologically-traced source. Open circles represent inter-locality links while closed circles represent intra-locality links. Crosses (right axis) indicate the probability that a link is instead inferred to have a zoonotic source. Results are shown for the model assuming the district-level broader contact zone. Links are sorted from lowest to highest success. **B.** The fraction of sampled transmission trees that recover a contact-traced link is influenced by the number of days between the symptom onset of source and recipient cases. Circles (left axis) show how often a given link was inferred as a function of the generation interval while the gray curve (right axis) shows the probability density for the generation interval assumed by the model.

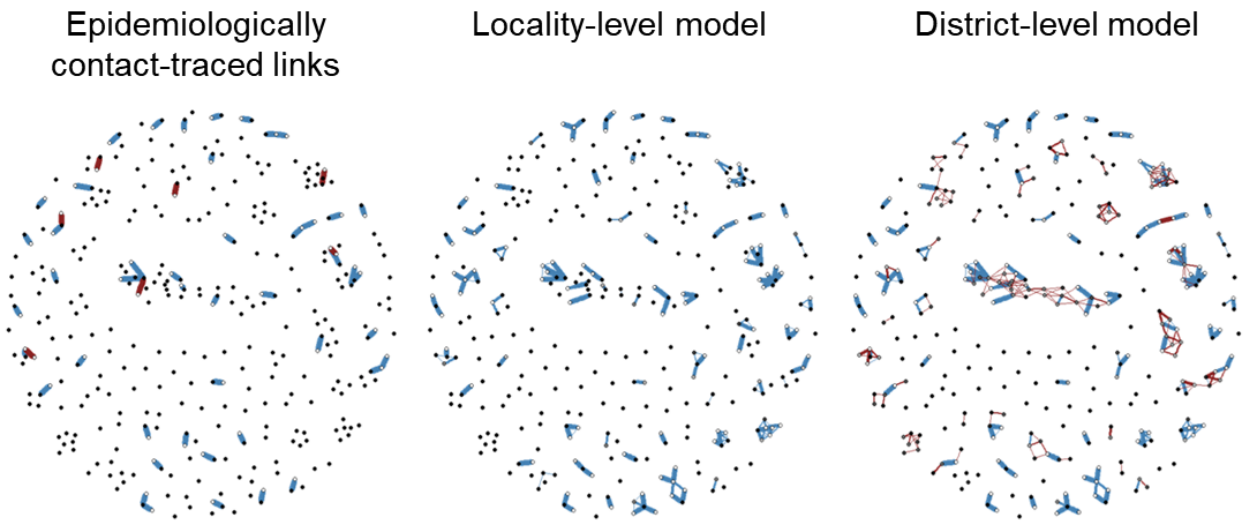


Figure 2.8. Comparison of monkeypox transmission trees created from contact-tracing, the locality-level model, and the district-level model. Points represent cases and edges indicate inferred transmission links between cases. Edge thickness corresponds to the frequency with which a given transmission link was inferred while edge color indicates whether a pair of linked cases occurred within the same (blue) or different (red) localities. The darkness of a point's fill indicates how frequently the case was inferred to have a zoonotic source, so transmission links often go from black points (cases caused by zoonotic spillover) to white points (cases infected by a human source).

Table 2.1. District model performs best for the monkeypox dataset in DIC model comparisons. Parameter inference for the monkeypox dataset was performed using four different approaches for dealing with the silent locality problem: the corrected denominator method (which conditions on the observation process for localities under surveillance) and three assumptions about the fraction of localities under surveillance that were observed. For each of these approaches, inference was repeated using four choices for the broader contact zone and the DIC was calculated. Parameter estimates and ΔDIC values are shown. The model with lowest ΔDIC is preferred and is shown in bold text.

Approach for dealing with silent localities	**Model**	**ΔDIC**	**mean *R***	**mean λ_z**	**mean σ**
Corrected denominator method	Locality	23.11	0.290	0.000387	1
	District	**0.0**	**0.381**	**0.000309**	**0.696**
	Region	5.88	0.418	0.000271	0.622
	Country	5.82	0.522	0.000188	0.464
Assume all surveilled localities were observed	Locality	21.98	0.272	0.000785	1
	District	**0.0**	**0.372**	**0.000676**	**0.717**
	Region	6.25	0.413	0.000633	0.656
	Country	10.92	0.479	0.000564	0.568
Assume 1/2 of surveilled localities were observed	Locality	17.06	0.290	0.000382	1
	District	**0.0**	**0.381**	**0.000334**	**0.756**
	Region	3.12	0.424	0.000311	0.684
	Country	6.79	0.488	0.000276	0.598
Assume 1/5 of surveilled localities were observed	Locality	15.05	0.310	0.000148	1
	District	**0.0**	**0.395**	**0.000130**	**0.777**
	Region	2.01	0.439	0.000121	0.704
	Country	5.34	0.500	0.000108	0.622

2.6 Appendix

Corrected denominator method: Derivation for the conditional likelihood function

The model described in the main text tells us that the number of new human cases on day *t* in locality *v* follows a Poisson distribution with mean

$$\mu_{t,v} = \sum_{s=1}^{t-1} \sum_{w=1}^{\mathcal{V}} [N_{s,w} * \lambda_{\{s,w\},\{t,v\}}] + \lambda_z \quad , \qquad (1)$$

which represents the sum of the expected numbers of cases caused by spillover and all previous human cases (Table S2.7 provides a description of parameters). Based on this model, the likelihood of a set of parameters ($\theta = \{R, \lambda_z, \sigma\}$) given surveillance data ($D = N_{t,v}$ cases observed on each day *t* and locality *v*) is:

$$\mathcal{L}(\theta|D) = \prod_{t=1}^{T} \prod_{v=1}^{V} \frac{e^{\mu_{t,v}} \mu_{t,v}^{N_{t,v}}}{N_{t,v}!} \quad . \qquad (2)$$

A challenge in applying this likelihood function to surveillance data arises when the total number of localities under surveillance, *V*, is unknown. Instead, we observe *W* localities that have one or more observed cases. If we re-arrange the product functions in the likelihood function, it becomes more apparent that we are taking the product of the likelihood for each locality:

$$\mathcal{L}(\theta|D) = \prod_{v=1}^{V} \prod_{t=1}^{T} \frac{e^{\mu_{t,v}} \mu_{t,v}^{N_{t,v}}}{N_{t,v}!} \quad . \qquad (3)$$

However, because we only observe localities with one or more cases in the surveillance data, we need that conditioning to be reflected in the likelihood. In other words, we now want to express the likelihood of a particular time-series of cases in a locality *conditional on that locality having one or more cases.* This can be done for each locality by multiplying its component of the likelihood by the inverse of the probability (q) of having one or more cases:

$$\mathcal{L}(\theta|D) = \prod_{w=1}^{W} \frac{\prod_{t=1}^{T} \frac{e^{\mu_{t,w}} \mu_{t,w}^{N_{t,w}}}{N_{t,w}!}}{q} \quad . \tag{4}$$

It is now necessary to calculate the probability a surveilled locality experiences one or more cases. This probability is equivalent to one minus the probability of no cases occurring at a locality during the surveillance period. The following section explains how the probability of zero cases occurring at a given locality (here denoted *p*) is calculated.

For zero cases to occur in a locality, there must be no zoonotic spillover into that locality as well as no human-to-human transmission from an outside locality. The zoonotic component is relatively straightforward to calculate, as it is simply the probability of zero spillover events on each of the *T* days (which equals $e^{-\lambda_z T}$). The probability of no transmission from an outside human source is a bit more complicated and can be broken down by the generation of the outside case to avoid double-counting. The generation of a case indicates how many human-to-human transmission events occurred leading to the case. We refer to cases resulting from zoonotic spillover as primary cases. Individuals infected by primary cases are second generation cases, individuals infected by second generation cases are third generation cases, etc. For there to be no cases in a locality, no transmission may have occurred into that locality from outside cases in any generation:

$$P(\text{no transmission from cases in other localities})$$

$$= P(\text{no transmission from primary cases})$$

$$* P\left(\text{no transmission from second generation cases} \middle| \text{no transmission from primary cases}\right)$$

$$* P\left(\text{no transmission from third generation cases} \middle| \text{no transmission from primary or second generation cases}\right)$$

$$* \ldots$$

The number of cases caused by a given case (of any generation) in the target locality is described by a Poisson distribution with expected value equal to $R\frac{(1-\sigma)}{(\mathcal{V}_w-1)}$, where V_w is the number of localities within the target locality's broader contact zone. Because each case transmits disease independently of one another (conditioned on the previous cases), the probability that no generation *i* cases cause infections in the target locality is $e^{-R\frac{(1-\sigma)}{(\mathcal{V}_w-1)}n_i}$, where n_i is the total number of i^{th} generation cases within the broader contact zone (given knowledge that none of the cases from previous generations transmitted to the target locality). Incorporating this information, the probability of observing zero cases in a locality (p) becomes:

$$p = e^{-\lambda_z T} * \prod_{i=1}^{\infty} e^{-R\frac{(1-\sigma)}{(\mathcal{V}_w-1)}n_i}$$

$$= e^{-\lambda_z T} * e^{-R\frac{(1-\sigma)}{(\mathcal{V}_w-1)}\sum_{i=1}^{\infty} n_i}. \quad (5)$$

We next need to calculate estimates for the expected values of each of the n_i. The expected number of primary cases in the entire broader contact zone (given that no spillover events occurred into the target locality) is the expected number of spillover events per locality (λ_z) multiplied by the number of localities under consideration ($\mathcal{V}_w - 1$), multiplied by the number of surveillance days (T). For subsequent case generations, we can calculate the expected number of cases in generation i+1 as the number of cases caused by the i^{th} generation in their own localities plus those caused in the $\mathcal{V}_w - 2$ other possible localities (there are $\mathcal{V}_w - 2$ other possible localities because the case's current locality and the target locality have already been counted):

$$\mathbb{E}[n_{i+1}] = n_i \left(R\sigma + R\sum_{v=1}^{\mathcal{V}_w-2}\frac{(1-\sigma)}{(\mathcal{V}_w-1)}\right)$$

$$= R\, n_i \frac{(\mathcal{V}_w+\sigma-2)}{(\mathcal{V}_w-1)}. \quad (6)$$

If we approximate the values of n_i with $\mathbb{E}[n_i]$, we get

$$\mathbb{E}[n_{i+1}] \approx \lambda_z T(\mathcal{V}_w - 1)\left[R\, \frac{(\mathcal{V}_w+\sigma-2)}{(\mathcal{V}_w-1)}\right]^i. \quad (7)$$

Returning to our estimation of p, we can approximate n_i values with $\mathbb{E}[n_i]$ and get

$$p \approx\ e^{-\lambda_z T} * e^{-R\frac{(1-\sigma)}{(\mathcal{V}_w-1)} \sum_{i=1}^{\infty} \mathbb{E}[n_i]}$$

$$=\ e^{-\lambda_z T} * e^{-R\frac{(1-\sigma)}{(\mathcal{V}_w-1)} \sum_{j=0}^{\infty} \lambda_z T(\mathcal{V}_w-1)\left[R\, \frac{(\mathcal{V}_w+\sigma-2)}{(\mathcal{V}_w-1)}\right]^j}$$

$$=\ e^{-\lambda_z T} * e^{-R\frac{(1-\sigma)}{(\mathcal{V}_w-1)} * \frac{\lambda_z T(\mathcal{V}_w-1)}{1-R\frac{(\mathcal{V}_w+\sigma-2)}{(\mathcal{V}_w-1)}}}$$

$$=\ e^{-\lambda_z T} * e^{\frac{-R\,\lambda_z T(1-\sigma)(\mathcal{V}_w-1)}{\mathcal{V}_w-1-R\,(\mathcal{V}_w+\sigma-2)}}$$

$$=\ e^{-\lambda_z T-\frac{R\,\lambda_z T(1-\sigma)(\mathcal{V}_w-1)}{\mathcal{V}_w-1-R\,(\mathcal{V}_w+\sigma-2)}}. \quad (8)$$

With some additional algebraic simplification, we can insert this value in the original equation:

$$\mathcal{L}(\theta|D) = \prod_{w=1}^{W} \frac{\prod_{t=1}^{T} \frac{e^{\mu_{t,w}}\, {\mu_{t,w}}^{N_{t,w}}}{N_{t,w}!}}{1-e^{-\lambda_z T-\frac{R\,\lambda_z T(1-\sigma)(\mathcal{V}_w-1)}{\mathcal{V}_w-1-R\,(\mathcal{V}_w+\sigma-2)}}} \quad . \quad (9)$$

This expression still includes the parameter $\mathcal{V}_w$, though fortunately the sensitivity of results to the value of this parameter is relatively low. We therefore approximate $\mathcal{V}_w$ using the expected number of localities under surveillance in the broader contact zone. This calculation is explained in the following section.

Estimating total number of localities under surveillance

We wish to use the estimated parameter values for R, λ_z, and σ in conjunction with the number of observed localities in a broader contact zone (W_w) to estimate the total number of localities under surveillance in that broader contact zone (V_w). If we let q be the probability a locality is observed (has one or more cases during the surveillance period), then we expect $V_w * q \approx W_w$. From the section above, we approximate $q = 1\text{-}p$ as:

$$q \approx 1 - e^{-\lambda_z T - \frac{R\,\lambda_z T(1-\sigma)(\mathcal{V}_W - 1)}{\mathcal{V}_W - 1 - R\,(\mathcal{V}_W + \sigma - 2)}}. \quad (10)$$

So we estimate V_w as the value that satisfies the equation:

$$0 = \mathcal{V}_w \left(1 - e^{-\lambda_z T - \frac{R\,\lambda_z T(1-\sigma)(\mathcal{V}_W - 1)}{\mathcal{V}_W - 1 - R\,(\mathcal{V}_W + \sigma - 2)}}\right) - W_w. \quad (11)$$

Simulation methods

Simulations with exact spatial locations

Although the model assumes that inter-locality transmission with a broader contact zone is equal between all locality pairs, we expect that the actual amount of shared transmission between two localities is strongly influenced by the distance between those localities. We conducted two simulations using localities with set geographic locations and inter-locality transmissions depending on the spatial relationship of the localities. We took the 178 GPS records available from monkeypox surveillance in the DRC during the 1980s and simulated transmission across localities with the same coordinates and the same district and region boundaries. Two types of inter-locality transmission rules were explored. In the first of these, inter-locality transmissions were assumed to occur equally into a source locality's five closest

neighbors. In the second set of simulations, inter-locality transmissions from a source locality were assumed to occur equally among all outside localities within 30 km of the source locality.

Simulations with highly structured and non-homogeneous spillover patterns

To illustrate how highly structured and non-homogeneous spillover could bias parameter estimates, we simulated an extreme case of a zoonotic epidemic traveling through time and space. We imagined that disease dynamics in the reservoir would occur in a single location for 25 days before moving to a new spot, in an extreme form of a traveling zoonotic epidemic. For each 25 day period, three localities (selected to be in the same district when possible) would be selected to experience all of the spillover in the entire system. Aside from this extreme spillover pattern, the simulation followed the district-level model.

Sensitivity analyses

Sensitivity of parameter inference to elevated or heterogeneous spillover

To test whether a high rate of spillover would inundate the system with so many cases that the temporal clustering patterns resulting from human-to-human transmission could be obscured, we simulated datasets with spillover rates up to 0.1. This value corresponds with an expected 59,312.5 spillover events during the five year simulation, which corresponds to an average of 36.5 per year in each locality. At this rate of spillover, there is an average of only ten days between spillover events, a shorter period than the mean generation time for human-to-human transmission events, which was sixteen days. Across the range of spillover rates tested, the method did very well at both point estimates and capturing the true parameter values within the 95% CI (an average of 94.3% of CIs included the true value of R and 94.9% included the true value of λ_z; Figure S2.7, Table S2.3). As the spillover rate increased from 0.0001 to 0.1,

estimates of R tended to improve (posterior means closer to true value and smaller CIs). While the absolute error on estimates of λ_z increased as spillover rate increased, the relative error tended to decrease. As such, it appears that elevated spillover rates, far from obscuring patterns, may actually correspond with improved estimates, presumably due to the increased inference power resulting from a larger number of cases.

Spillover is unlikely to occur homogeneously through time and space in real-world settings. As an illustration of the potential effect this occurrence could have on parameter estimates, we simulated an extreme case (see 'Simulations with highly structured and non-homogeneous spillover patterns,' above) where spillover occurs into three localities at a time. The parameter inference results for this situation were strongly biased (Figure S2.10).

Sensitivity of parameter inference to offspring distribution assumptions

The model used in this study assumes that the number of new cases caused by an infectious individual follows a Poisson distribution, but previous work suggests that the offspring distribution is often better characterized by a negative binomial distribution, which allows for a greater amount of variation between individuals (Lloyd-Smith et al. 2005). We simulated datasets using a negative binomial offspring distribution (using a dispersion parameter k=0.58 in accordance with previous estimates for monkeypox from Lloyd-Smith et al. 2005) and examined how well our inference method, which assumes a Poisson offspring distribution, estimated the true parameter values. Estimates for these datasets were only marginally less accurate than estimates for datasets generated with a Poisson offspring distribution (with an average percent error of 10.9% as opposed to 8.2% for R and of 11.6% as opposed to 10.4% for spillover rate estimates) (Figure S2.8, Table S2.4). As such, there are unlikely to be strong biases introduced

from a mis-specified offspring distribution for the monkeypox dataset, though this bias could increase if applied to pathogens with more extreme transmission variance.

Sensitivity of parameter inference to broader contact zone assumption

To examine how assuming different broader contact zones would affect inference results, we compared parameter estimates obtained under three choices of broader contact zones for data simulated under two inter-locality transmission rules. We simulated disease spread in a system where localities were placed in the same arrangement as seen in 178 localities with GPS coordinates included in the monkeypox surveillance system, district and region arrangement were the same as in the 1980s surveillance, and human-to-human transmission could occur either between a locality and its five closest neighbors or between localities located within 30 km of one another. Inference results again showed increasing estimates of R and decreasing estimates of spillover rate as the size of the assumed broader contact zone increased (Tables S2.5, S2.6).

2.7 Appendix Figures and Tables

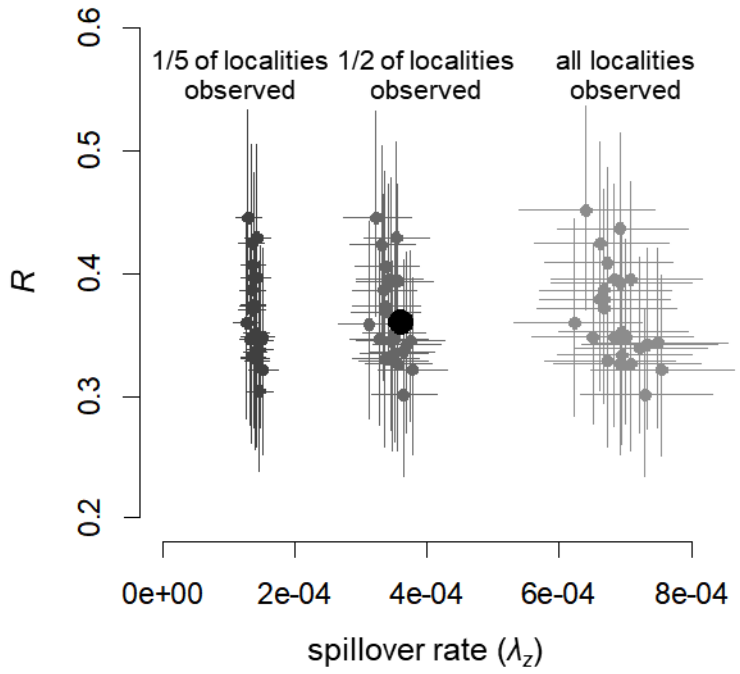


Figure S2.1. Effect of assumed fraction of localities observed on parameter estimates. The true parameter values are indicated by a large black dot and while smaller points indicate the inferred values from 25 simulated datasets (lines show the 95% credible interval). For each dataset, inference was performed assuming that 1/5, 1/2, and all of the localities under surveillance were observed. For these simulations, the true percentage of localities observed ranged from 46% to 57%, with a mean of 52%.

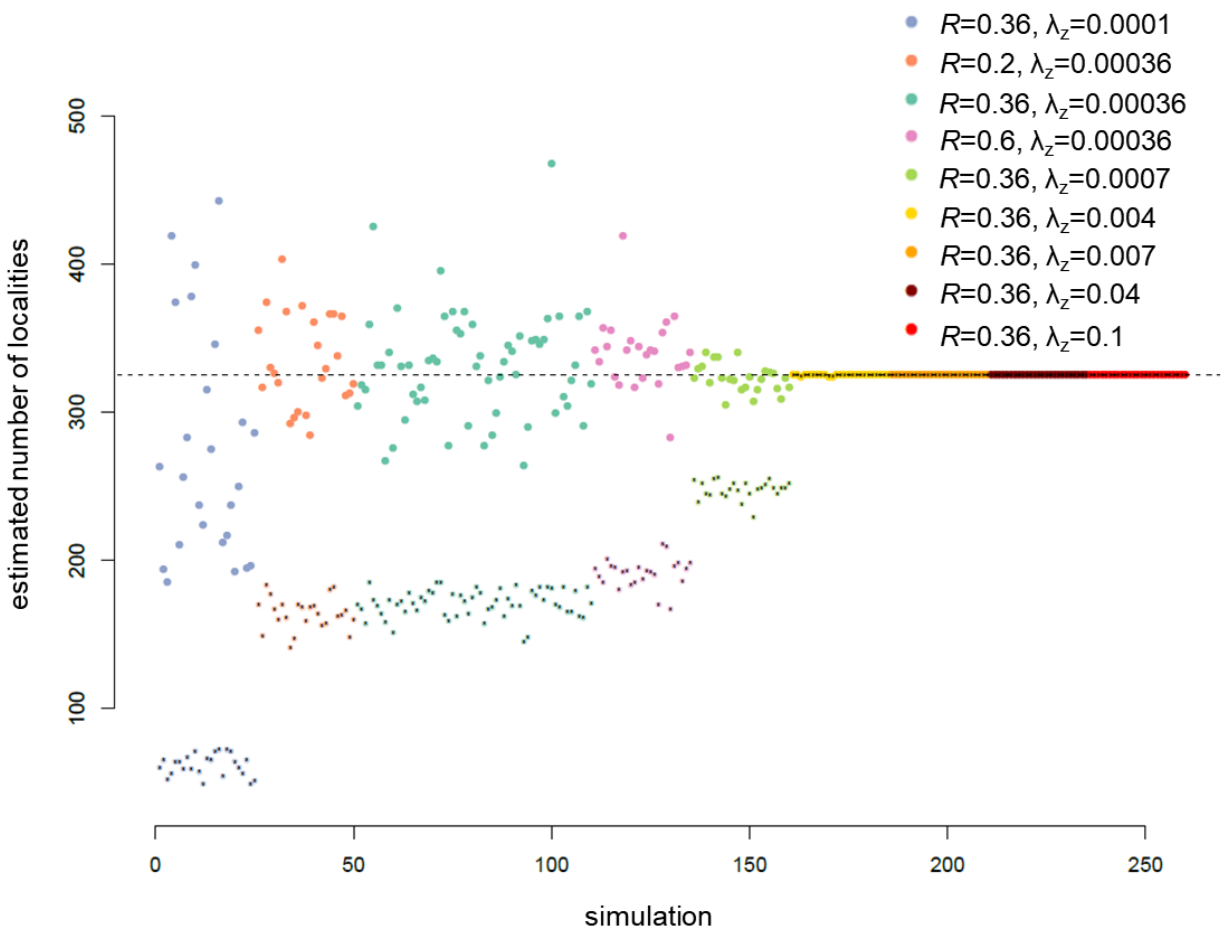


Figure S2.2. Estimated number of localities under surveillance (calculated given the number of observed localities and the estimated parameter values). Large colored dots indicate the estimated number of localities under surveillance for each simulated dataset while the smaller dots show the number of localities observed in the dataset. The true number of localities is represented by the horizontal dashed line. Each color corresponds to a different parameter set used for simulations.

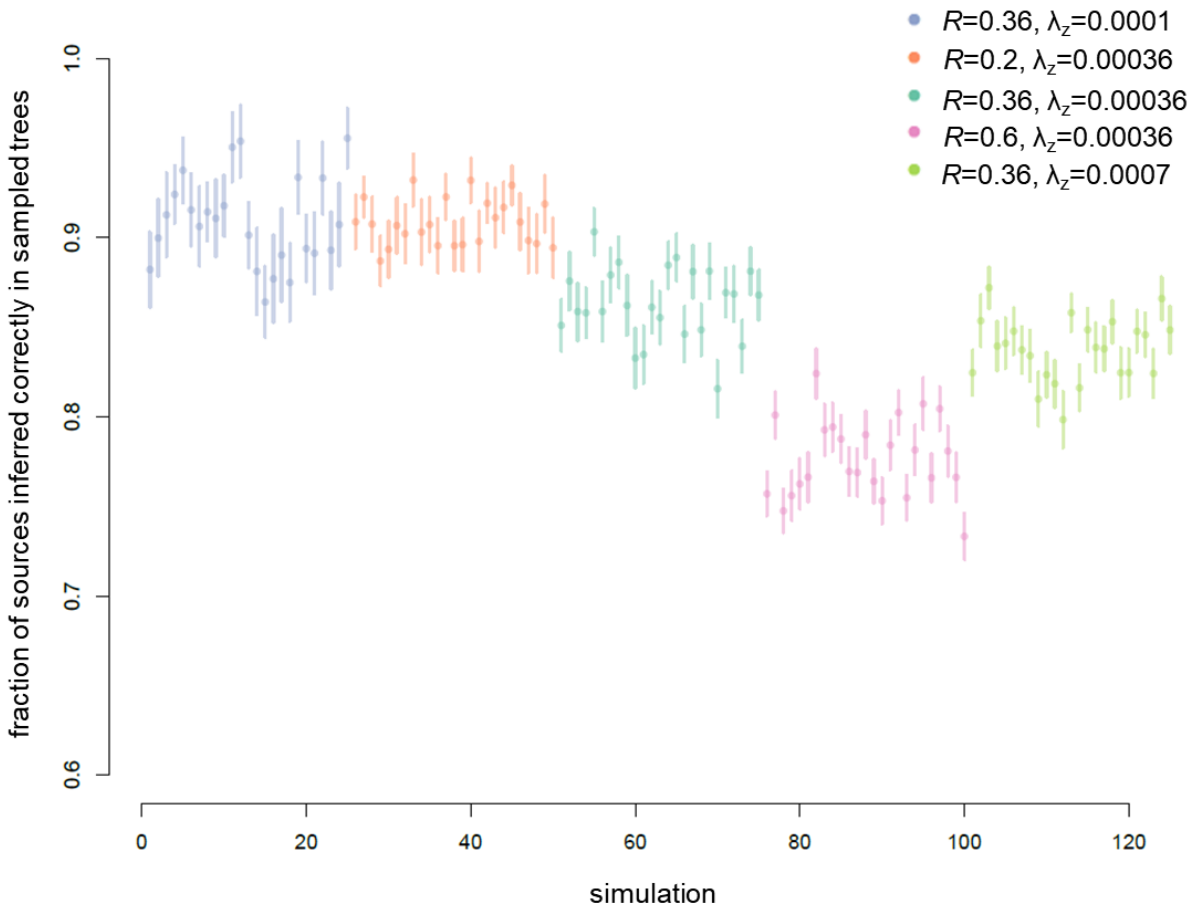


Figure S2.3. Accuracy of inferred transmission trees at inferring the correct source of cases. For each simulated dataset (25 simulations for each of 5 parameter sets), 200 transmission trees were drawn. Points show the mean fraction of cases inferred correctly in a sampled transmission tree and bars indicate the standard deviation.

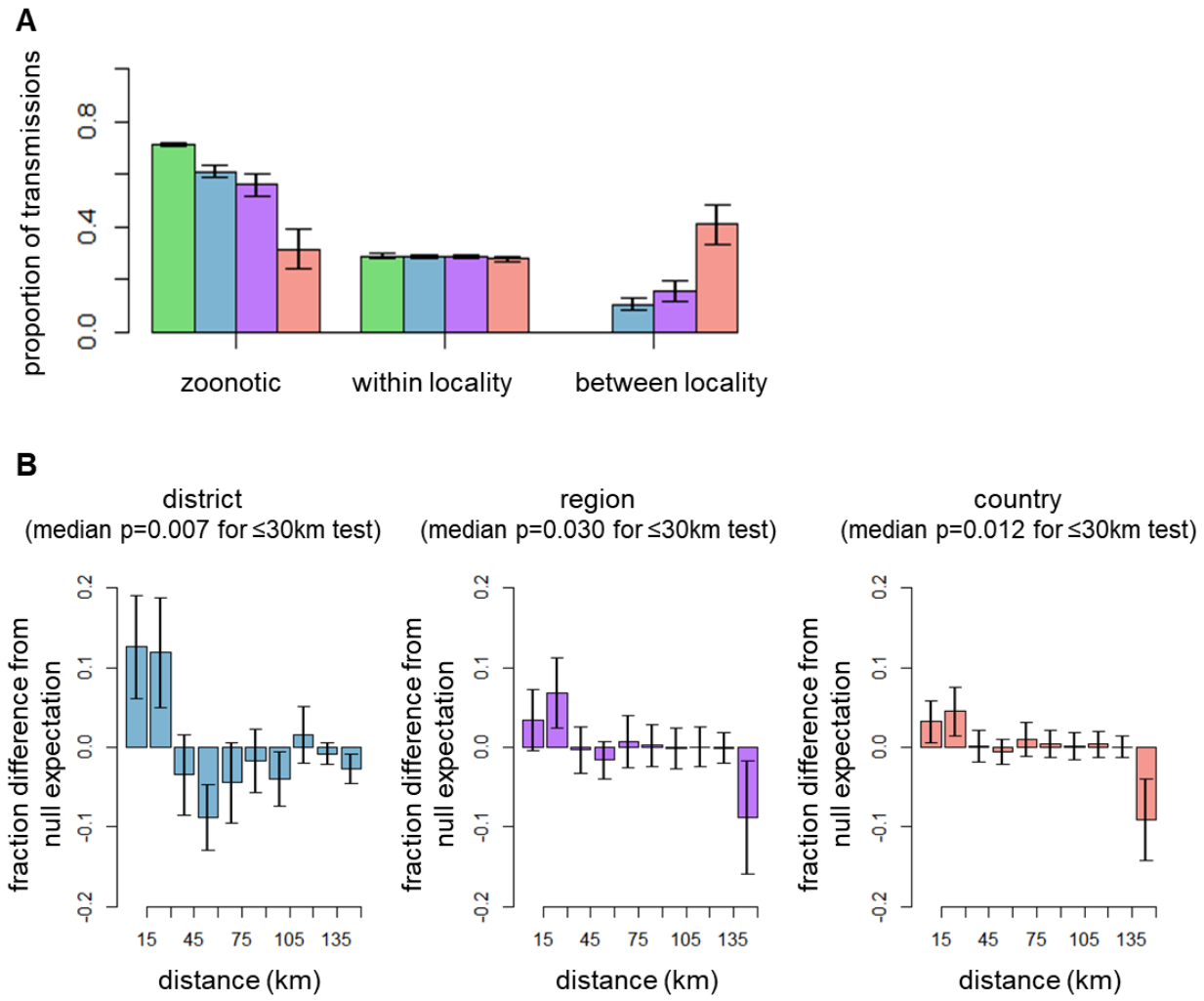


Figure S2.4. Inferred sources of monkeypox cases. **A.** The fraction of cases inferred to have originated from each source using each of the four spatial models (locality-green, district-blue, region-purple, country-red). **B.** Difference in the proportion of inter-locality human-to-human transmissions inferred by the models to occur over a given transmission distance versus expected based on the spatial distribution of localities. The p-values indicate the probability of observing as many or more transmissions over distances of ≤ 30 kilometers based on the null model (i.e. assuming distance plays no role in determining which localities are linked by inferred transmission events). The median p-value of sampled transmission trees is given, and the full distribution of p-values can be seen in Figure S2.5.

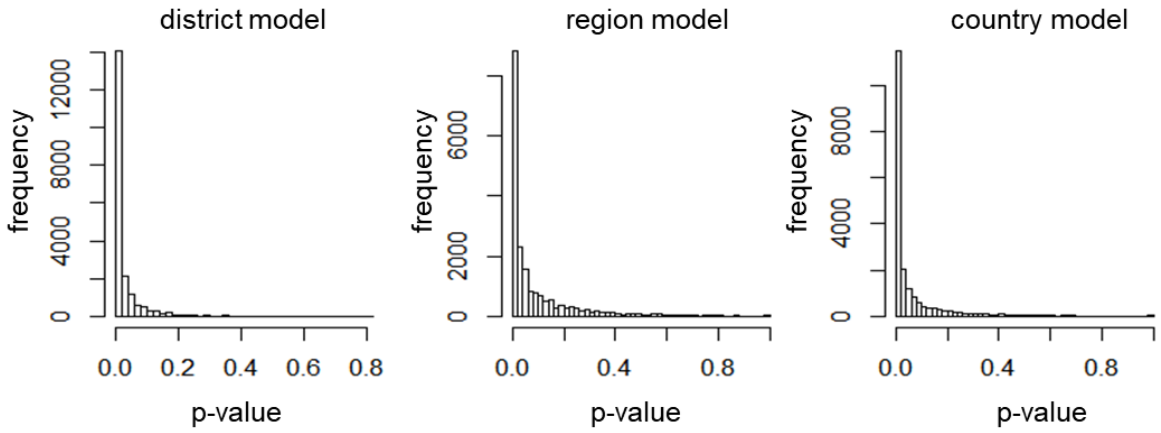


Figure S2.5. The distribution of p-values obtained across sampled transmission trees. P-values obtained from a binomial test examining whether the number of transmission events inferred to occur across thirty or fewer kilometers is greater than that expected based on the null distribution. Each p-value corresponds to a sampled transmission tree.

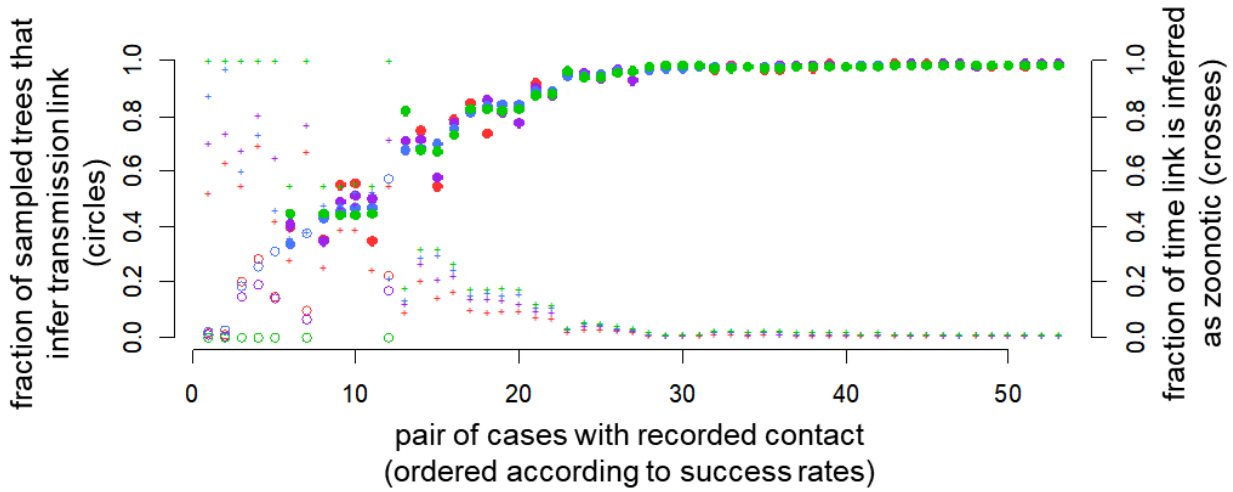


Figure S2.6. Comparison of epidemiologically contact-traced links with sampled transmission trees. Circles (left axis) show the fraction of sampled trees that infer the epidemiologically-traced source. Open circles represent inter-locality links while closed circles represent intra-locality links. Bars (right axis) indicate the probability that a link is instead inferred to have a zoonotic source. Results are shown for models that use the country-level (red), region-level (purple), district-level (blue), and locality-level (green) broader contact zones. Links are sorted from lowest to highest success in the district model.

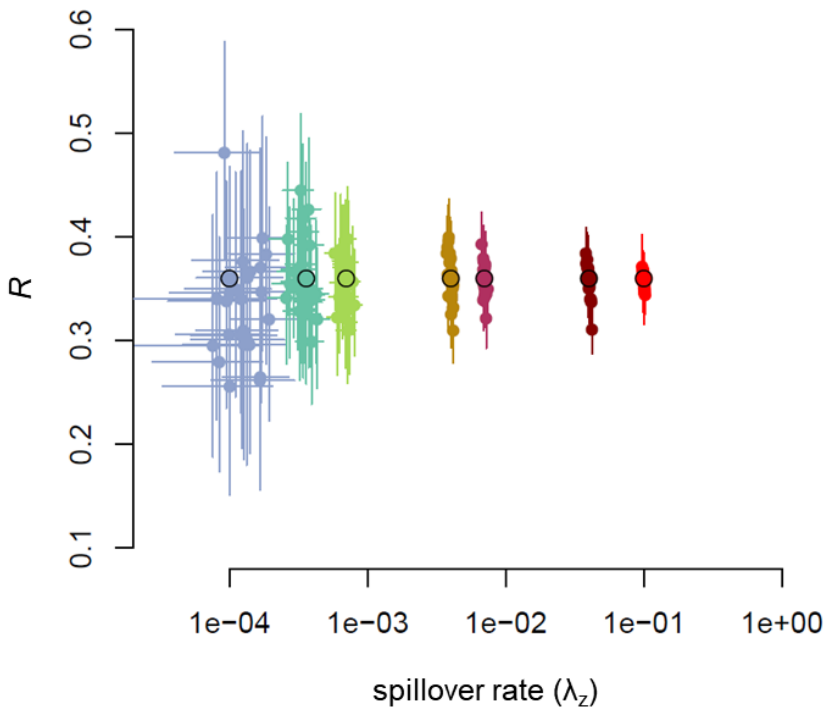


Figure S2.7. Effect of increasing spillover rate on parameter estimate success. Within each color, large points outlined in black indicate the true parameter set and smaller points indicate the inferred parameter values from 25 simulated datasets (lines show the 95% credible interval). Warmer colors correspond with higher spillover rates. Note the log-scale x-axis.

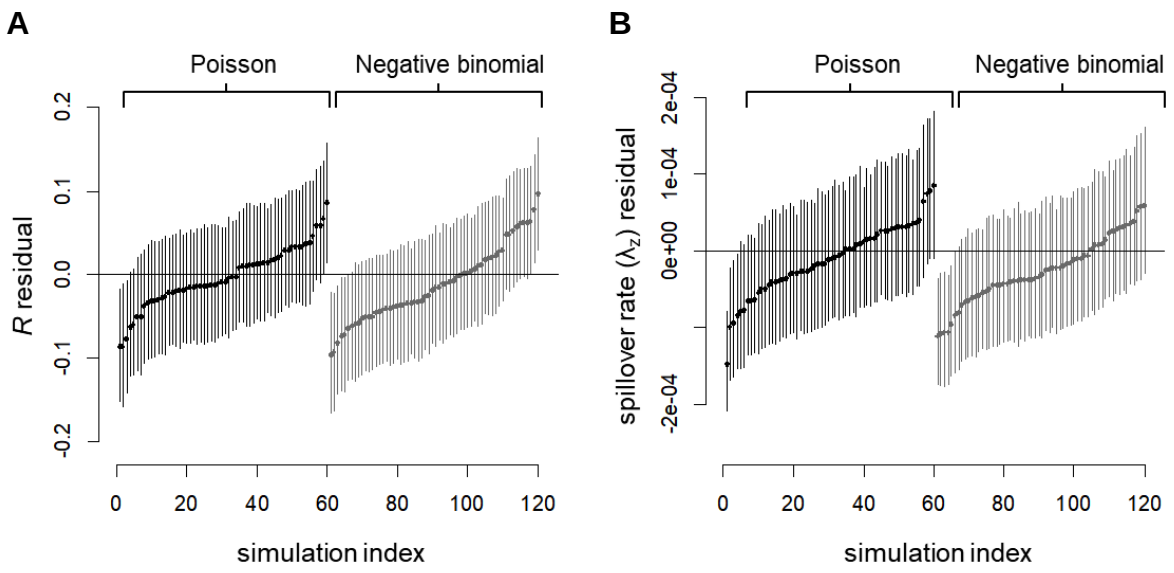


Figure S2.8. Parameter estimate residuals for data simulated using a negative binomial versus Poisson offspring distribution. Because the inference method assumes a Poisson offspring distribution, we compared the inference successes for datasets simulated assuming a Poisson offspring distribution versus datasets simulated assuming a negative binomial offspring distribution. The residuals in parameter estimates for 25 simulations are shown for **A)** the reproductive number and **B)** the spillover rate.

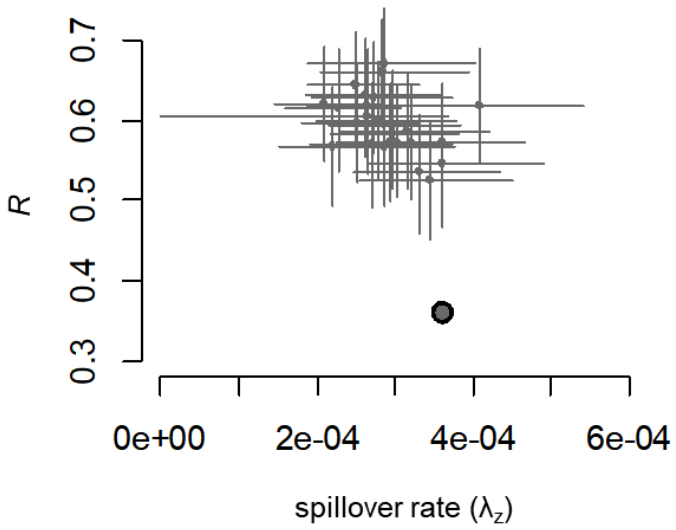


Figure S2.9. Strongly heterogeneous spillover causes bias in parameter estimates. The true parameter value is indicated by the large dot while smaller points indicate the inferred values from 25 simulated datasets (lines show the 95% credible interval). Simulations were conducted to mimic pockets of zoonotic disease moving through the reservoir population. To capture the idea that, at any given time, only a small subset of localities might be experiencing high levels of spillover while the rest of the localities experienced no spillover, the simulations assumed that every 25 days a new set of three localities experienced the full force of spillover for the entire system. This gave rise to clusters of primary cases, which tend to be misclassified as human-to-human transmission events by our inference approach, which assumes homogeneous spillover rates.

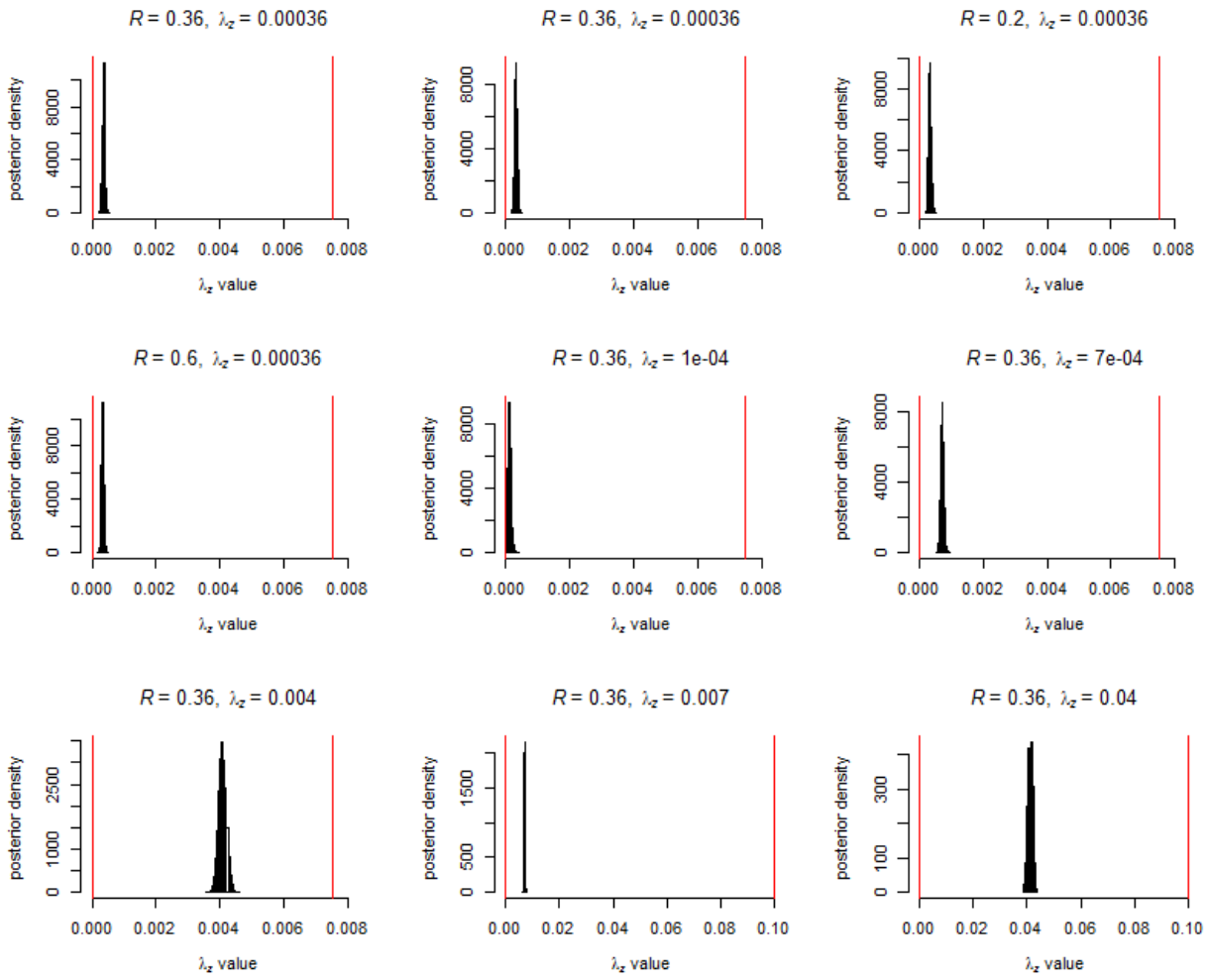


Figure S2.10. Comparison of prior and posterior distributions for spillover rate λ_z. Black bars represent posterior distribution while red lines mark limits of the uniform prior distribution. One representative simulation is shown for each of the nine parameter sets. Notice that the posterior distribution is always relatively far from upper bound of the prior.

Table S2.1. Comparison of inference method success over the same simulated datasets.

	R			λ_z			σ		
Inference Approach	Fraction of 95%CIs include true value	Average error size	Average percent error	Fraction of 95%CIs include true value	Average error size	Average percent error	Fraction of 95%CIs include true value	Average error size	Average percent error
True number of localities known	95.2% (119/125)	0.0293	8.6%	96.8% (121/125)	1.99E-05	6.3%	96.0% (120/125)	0.0522	7.0%
Assume all localities are observed	95.2% (119/125)	0.0288	8.4%	0.0% (0/125)	3.30E-04	153.0%	94.4% (118/125)	0.0575	7.7%
Corrected denominator method (account for silent localities)	92.8% (116/125)	0.0298	8.4%	93.6% (117/125)	3.59E-05	14.0%	88.0% (110/125)	0.0665	8.9%

Table S2.2. Simulated datasets.

Inter-locality transmission rule	Offspring distribution	True R	True λ_z	# datasets simulated
Broader contact zone: district-level	Poisson	0.36	0.00036	60
Broader contact zone: district-level	Poisson	0.2	0.00036	25
Broader contact zone: district-level	Poisson	0.6	0.00036	25
Broader contact zone: district-level	Poisson	0.36	0.0001	25
Broader contact zone: district-level	Poisson	0.36	0.0007	25
Broader contact zone: district-level	Poisson	0.36	0.004	25
Broader contact zone: district-level	Poisson	0.95	0.007	25
Broader contact zone: district-level	Poisson	0.36	0.04	25
Broader contact zone: district-level	Poisson	0.36	0.1	25
Broader contact zone: district-level	NBinom (k=0.58)	0.36	0.00036	60
Broader contact zone: district-level	Poisson	0.01	0.00036 (intensity heterogeneous through time and space)	25
Localities have same spatial coordinates as recorded for DRC monkeypox localities, inter-locality transmission with closest 5 neighbors	Poisson	0.36	0.00036	25
Localities have same spatial coordinates as recorded for DRC monkeypox localities, inter-locality transmission with neighbors within 30 km	Poisson	0.36	0.00036	25

Table S2.3. Success of the corrected denominator inference method for datasets simulated with increasing spillover rates.

	R				λ_z				σ			
True λ_z value	Fraction of 95% CIs include true value	Average error size	Average percent error	Average width of CI	Fraction of 95% CIs include true value	Average error size	Average percent error	Average width of CI	Fraction of 95% CIs include true value	Average error size	Average percent error	Average width of CI
0.0001	96% (24/25)	0.0458	12.7%	0.226	96% (24/25)	3.54 E-05	35.4%	1.75 E-04	88% (22/25)	0.1094	14.58%	0.395
0.00036	92% (23/25)	0.0279	7.8%	0.137	88% (22/25)	3.68 E-05	10.2%	1.72 E-04	84% (21/25)	0.0587	7.83%	0.239
0.0007	100% (25/25)	0.0213	5.9%	0.108	96% (24/25)	3.85 E-05	5.5%	2.06 E-04	88% (22/25)	0.0493	6.57%	0.187
0.004	88% (22/25)	0.0173	4.8%	0.068	96% (24/25)	1.04 E-04	2.6%	4.71 E-04	100% (25/25)	0.0255	3.39%	0.122
0.007	92% (23/25)	0.0121	3.4%	0.060	96% (24/25)	1.27 E-04	1.8%	7.05 E-04	92% (23/25)	0.0261	3.49%	0.113
0.04	92% (23/25)	0.0121	3.4%	0.050	92% (23/25)	7.50 E-04	1.9%	3.15 E-03	96% (24/25)	0.0215	2.87%	0.100
0.1	100% (25/25)	0.0071	2.0%	0.038	100% (25/25)	1.12 E-03	1.1%	5.90 E-03	100% (25/25)	0.0134	1.79%	0.082

Table S2.4. Success of the corrected denominator inference method for datasets simulated with different offspring distributions.

	R			λ_z			σ		
Offspring distribution	Fraction of 95% CIs include true value	Average error size	Average percent error	Fraction of 95% CIs include true value	Average error size	Average percent error	Fraction of 95% CIs include true value	Average error size	Average percent error
Poisson	91.7% (55/60)	0.0289	8.0%	93.3% (56/60)	3.76E-05	10.4%	83.3% (50/60)	0.0649	8.7%
Negative binomial (k=0.58)	86.7% (52/60)	0.0393	10.9%	90.0% (54/60)	4.18E-05	11.6%	91.7% (55/60)	0.0555	7.4%

Table S2.5. Comparison of parameter estimates inferred using models of increasing spatial scale – data simulated using the 'nearest five neighbors' inter-locality transmission rule where localities take the same GPS coordinates as in the DRC monkeypox surveillance dataset (true R is 0.36, true spillover rate is 0.00036; mean parameter estimates from inference on 25 simulated datasets)

Model used for inference	mean R	mean λ_z
District	0.314	0.000346
Region	0.323	0.000343
Country	0.354	0.000328

Table S2.6. Comparison of parameter estimates inferred using models of increasing spatial scale – data simulated assuming inter-locality transmission can occur between any localities located within 30 km of one another, where localities take the same GPS coordinates as in the DRC monkeypox surveillance dataset (true R is 0.36, true spillover rate is 0.00036; mean parameter estimates from inference on 25 simulated datasets)

Model used for inference	mean R	mean λ_z
District	0.348	0.000385
Region	0.357	0.000355
Country	0.379	0.000334

Table S2.7. Parameter descriptions.

Symbol	Description
$\mu_{t,v}$	Expected number of cases observed on day *t*, in locality *v*
$N_{t,v}$	Actual number of cases observed on day *t*, in locality *v*
N	Actual number of cases observed across all localities over the course of surveillance
V	Total number of localities under surveillance
V_w	Total number of localities under surveillance in the broader contact zone of locality *w*
W	Number of localities with one or more cases (the number of localities that appear in the surveillance dataset)
W_w	Number of localities with one or more cases in the broader contact zone of locality *w*
T	Duration of surveillance: number of days surveillance was conducted
λ_z	Spillover rate: the expected number of spillover events per day in a given locality
$\lambda_{\{s,w\},\{t,v\}}$	The expected number of new infections that become symptomatic on day *t* in locality *v* caused by an infectious individual who became symptomatic on day *s* in locality *w*
R	Reproductive number: the average number of secondary cases caused by an infectious individual
σ	Within-locality transmission proportion: the fraction of cases arising from human-to-human transmission that occur in the same locality as the source case

2.8 References

1. Morse SS. Factors in the Emergence of Infectious Diseases. Emerg Infect Dis J [Internet]. 1995;1(1):7. Available from: http://wwwnc.cdc.gov/eid/article/1/1/95-0102

2. Woolhouse MEJ, Gowtage-Sequeria S. Host Range and Emerging and Reemerging Pathogens. Emerg Infect Dis [Internet]. 2005 Dec;11(12):1842–7. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367654/

3. Xu R-H, He J-F, Evans MR, Peng G-W, Field HE, Yu D-W, et al. Epidemiologic Clues to SARS Origin in China. Emerg Infect Dis [Internet]. 2004 Jun;10(6):1030–7. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3323155/

4. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. Nature [Internet]. 2008 Feb 21;451(7181):990–3. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5960580/

5. Lloyd-Smith JO, George D, Pepin KM, Pitzer VE, Pulliam JRC, Dobson AP, et al. Epidemic dynamics at the human-animal interface. Science [Internet]. 2009 Dec 4;326(5958):1362–7. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3891603/

6. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science (80-) [Internet]. 2014 Sep 11;345(6202):1369–72. Available from: http://science.sciencemag.org/content/345/6202/1369.abstract

7. Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P. Host and viral traits predict zoonotic spillover from mammals. Nature [Internet]. 2017 Jun 29;546(7660):646–50. Available from: http://dx.doi.org/10.1038/nature22975

8. Memish ZA, Cotten M, Meyer B, Watson SJ, Alsahafi AJ, Al Rabeeah AA, et al. Human Infection with MERS Coronavirus after Exposure to Infected Camels, Saudi Arabia, 2013. Emerg Infect Dis [Internet]. 2014 Jun;20(6):1012–5. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4036761/

9. Woolhouse M, Gaunt E. Ecological Origins of Novel Human Pathogens. Crit Rev Microbiol [Internet]. 2007 Jan 1;33(4):231–42. Available from: https://doi.org/10.1080/10408410701647560

10. National Research Council (US) Committee on Achieving Sustainable Global Capacity for Surveillance and Response to Emerging Diseases of Zoonotic Origin; Keusch GT, Pappaioanou M, Gonzalez MC, et al. editors. Sustaining Global Surveillance and Response to Emerging Zoonotic Diseases. Washingt Natl Acad Press [Internet]. 2009;

Available from: https://www.ncbi.nlm.nih.gov/books/NBK215317/

11. Heesterbeek H, Anderson R, Andreasen V, Bansal S, De Angelis D, Dye C, et al. Modeling infectious disease dynamics in the complex landscape of global health. Science [Internet]. 2015 Mar 13;347(6227):aaa4339-aaa4339. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4445966/

12. Lloyd-Smith JO, Funk S, McLean AR, Riley S, Wood JLN. Nine challenges in modelling the emergence of novel pathogens. Epidemics [Internet]. 2015;10:35–9. Available from: http://www.sciencedirect.com/science/article/pii/S1755436514000504

13. Plowright RK, Parrish CR, McCallum H, Hudson PJ, Ko AI, Graham AL, et al. Pathways to zoonotic spillover. Nat Rev Microbiol. 2017;15(8):502–10.

14. Anderson RM, May RM. Infectious diseases of humans : dynamics and control. Oxford. New York: Oxford University Press; 1991.

15. Keeling MJ, Rohani P. Modeling Infectious Diseases in Humans and Animals. Princeton Univ. Press; 2008. 366 p.

16. Blumberg S, Lloyd-Smith JO. Inference of <italic>R</italic>$_0$ and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. PLoS Comput Biol [Internet]. 2013;9(5):e1002993. Available from: http://dx.doi.org/10.1371%2Fjournal.pcbi.1002993

17. Kravitz HM. Denominator Difficulties. Wiley StatsRef Stat Ref Online. 2014;

18. Tatem AJ. Mapping the denominator: spatial demography in the measurement of progress. Int Health [Internet]. 2014 Sep 14;6(3):153–5. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4161992/

19. Ozonoff A, Jeffery C, Manjourides J, White LF, Pagano M. Effect of spatial resolution on cluster detection: a simulation study. Int J Health Geogr [Internet]. 2007;6(1):52. Available from: https://doi.org/10.1186/1476-072X-6-52

20. Zhang Z, Manjourides J, Cohen T, Hu Y, Jiang Q. Spatial measurement errors in the field of spatial epidemiology. Int J Health Geogr [Internet]. 2016;15(1):21. Available from: https://doi.org/10.1186/s12942-016-0049-5

21. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature [Internet]. 2017 Apr 20;544(7650):309–15. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5712493/

22. Curtis AJ, Mills JW, Leitner M. Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina. Int J Health Geogr [Internet]. 2006;5(1):44. Available from: https://doi.org/10.1186/1476-072X-5-44

23. National Research C. Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data. Stern PC, Gutman MP, editors. Washington, DC: National Academies Press; 2007.

24. Gutmann M, Witkowski K, Colyer C, O'Rourke JM, McNally J. Providing Spatial Data for Secondary Analysis: Issues and Current Practices relating to Confidentiality. Popul Res Policy Rev [Internet]. 2008;27(6):639–65. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2600804/

25. de Montjoye Y-A, Hidalgo CA, Verleysen M, Blondel VD. Unique in the Crowd: The privacy bounds of human mobility. 2013 Mar 25;3:1376. Available from: http://dx.doi.org/10.1038/srep01376

26. De Serres G, Gay NJ, Farrington CP. Epidemiology of Transmissible Diseases after Elimination. Am J Epidemiol [Internet]. 2000 Jun 1;151(11):1039–48. Available from: http://dx.doi.org/10.1093/oxfordjournals.aje.a010145

27. Jansen VAA, Stollenwerk N, Jensen HJ, Ramsay ME, Edmunds WJ, Rhodes CJ. Measles Outbreaks in a Population with Declining Vaccine Uptake. Science (80-) [Internet]. 2003 Aug 7;301(5634):804 LP-804. Available from: http://science.sciencemag.org/content/301/5634/804.abstract

28. Ferguson NM, Fraser C, Donnelly CA, Ghani AC, Anderson RM. Public Health Risk from the Avian H5N1 Influenza Epidemic. Science (80-) [Internet]. 2004;304(5673):968–9. Available from: http://www.sciencemag.org/content/304/5673/968.short

29. Cauchemez S, Epperson S, Biggerstaff M, Swerdlow D, Finelli L, Ferguson NM. Using Routine Surveillance Data to Estimate the Epidemic Potential of Emerging Zoonoses: Application to the Emergence of US Swine Origin Influenza A H3N2v Virus. PLoS Med [Internet]. 2013 Mar 5;10(3):e1001399. Available from: http://dx.doi.org/10.1371%2Fjournal.pmed.1001399

30. Blumberg S, Lloyd-Smith JO. Comparing methods for estimating R(0) from the size distribution of subcritical transmission chains. Epidemics [Internet]. 2013 Sep 3;5(3):10.1016/j.epidem.2013.05.002. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3821076/

31. Wallinga J, Teunis P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. Am J Epidemiol [Internet]. 2004 Sep 15;160(6):509–16. Available from:

http://aje.oxfordjournals.org/content/160/6/509.abstract

32. Lo Iacono G, Cunningham AA, Fichet-Calvet E, Garry RF, Grant DS, Khan SH, et al. Using Modelling to Disentangle the Relative Contributions of Zoonotic and Anthroponotic Transmission: The Case of Lassa Fever. PLoS Negl Trop Dis [Internet]. 2015 Jan 8;9(1):e3398. Available from: http://dx.doi.org/10.1371%2Fjournal.pntd.0003398

33. White LF, Pagano M. A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. Stat Med [Internet]. 2008 Jul 20;27(16):2999–3016. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3951165/

34. Kucharski A, Mills H, Pinsent A, Fraser C, Van Kerkhove M, Donnelly C, et al. Distinguishing Between Reservoir Exposure and Human-to-Human Transmission for Emerging Pathogens Using Case Onset Data. PLoS Curr Outbreaks. 2014;

35. Lo Iacono G, Cunningham AA, Fichet-Calvet E, Garry RF, Grant DS, Leach M, et al. A Unified Framework for the Infection Dynamics of Zoonotic Spillover and Spread. Foley J, editor. PLoS Negl Trop Dis [Internet]. 2016 Sep 2;10(9):e0004957. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5010258/

36. Keeling MJ, Woolhouse MEJ, Shaw DJ, Matthews L, Chase-Topping M, Haydon DT, et al. Dynamics of the 2001 UK Foot and Mouth Epidemic: Stochastic Dispersal in a Heterogeneous Landscape. Science (80-) [Internet]. 2001;294(5543):813–7. Available from: http://www.sciencemag.org/content/294/5543/813.abstract

37. Höhle M, Jørgensen E, O'Neill PD. Inference in disease transmission experiments by using stochastic epidemic models. J R Stat Soc Ser C (Applied Stat [Internet]. 2005;54(2):349–66. Available from: http:https://doi.org/10.1111/j.1467-9876.2005.00488.x

38. Boender GJ, Hagenaars TJ, Bouma A, Nodelijk G, Elbers ARW, de Jong MCM, et al. Risk Maps for the Spread of Highly Pathogenic Avian Influenza in Poultry. PLOS Comput Biol [Internet]. 2007 Apr 20;3(4):e71. Available from: https://doi.org/10.1371/journal.pcbi.0030071

39. Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc R Soc B Biol Sci [Internet]. 2012;279(1728):444–50. Available from: http://rspb.royalsocietypublishing.org/content/279/1728/444.abstract

40. Cauchemez S, Nouvellet P, Cori A, Jombart T, Garske T, Clapham H, et al. Unraveling the drivers of MERS-CoV transmission. Proc Natl Acad Sci [Internet]. 2016 Aug

9;113(32):9081–6. Available from: http://www.pnas.org/content/113/32/9081.abstract

41. Ball F, Mollison D, Scalia-Tomba G. Epidemics with Two Levels of Mixing. Ann Appl Probab [Internet]. 1997;7(1):46–89. Available from: http://www.jstor.org/stable/2245132

42. DEMIRIS N, O'NEILL PD. Bayesian inference for epidemics with two levels of mixing. Scand J Stat [Internet]. 2005;32(2):265–80. Available from: http:https://doi.org/10.1111/j.1467-9469.2005.00420.x

43. Ježek Z, Fenner F. Human monkeypox [Internet]. S Karger Ag; 1988. Available from: http://books.google.com/books?id=fyupMQEACAAJ

44. Ježek Z, Grab B, Szczeniowski M V, Paluku KM, Mutombo M. Human monkeypox: secondary attack rates. Bull World Health Organ [Internet]. 1988;66(4):465–70. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2491159/

45. Fenner F, Henderson DA, Arita I, Jezek Z, Ladnyi ID. Smallpox and its Eradication. World Health Organization; 1988.

46. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature [Internet]. 2005;438(November):355–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16292310

47. le Polain de Waroux O, Cohuet S, Ndazima D, Kucharski AJ, Juan-Giner A, Flasche S, et al. Characteristics of human encounters and social mixing patterns relevant to infectious diseases spread by close contact: a survey in Southwest Uganda. BMC Infect Dis [Internet]. 2018 Apr 11;18:172. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5896105/

48. Clark WA V., Avery KL. The effects of data aggregation in statistical analysis. Geogr Anal. 1976;8:428–38.

49. Openshaw S, Taylor PJ. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. Stat Appl Spat Sci. 1979;127–44.

50. Beale L, Abellan JJ, Hodgson S, Jarup L. Methodologic issues and approaches to spatial epidemiology. Environ Health Perspect. 2008;116:1105–10.

51. Richter W. The verified neighbor approach to geoprivacy: An improved method for geographic masking. J Expo Sci Environ Epidemiol [Internet]. 2017 Sep 20;28:109. Available from: http://dx.doi.org/10.1038/jes.2017.17

52. Zandbergen PA. Ensuring Confidentiality of Geocoded Health Data: Assessing

Geographic Masking Strategies for Individual-Level Data. Adv Med [Internet]. 2014;2014. Available from: https://doi.org/10.1155/2014/567049

53. Sebastian J. Schreiber, James O. Lloyd-Smith. Invasion Dynamics in Spatially Heterogeneous Environments. Am Nat [Internet]. 2009;174(4):490–505. Available from: http://www.jstor.org/stable/10.1086/605405

54. Arita I, Wickett J, Fenner F. Impact of Population Density on Immunization Programmes. J Hyg (Lond) [Internet]. 1986;96(3):459–66. Available from: http://www.jstor.org/stable/3863139

55. Grenfell, Bolker. Cities and villages: infection hierarchies in a measles metapopulation. Ecol Lett [Internet]. 1998 Jul 1;1(1):63–70. Available from: http://dx.doi.org/10.1046/j.1461-0248.1998.00016.x

56. Neiderud C-J. How urbanization affects the epidemiology of emerging infectious diseases. Infect Ecol Epidemiol [Internet]. 2015 Jun 24;5:10.3402/iee.v5.27060. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4481042/

57. NISHIURA H, EICHNER M. Infectiousness of smallpox relative to disease age: estimates based on transmission network and incubation period. Epidemiol Infect [Internet]. 2007 Oct 7;135(7):1145–50. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2870668/

58. Fine PEM. The Interval between Successive Cases of an Infectious Disease. Am J Epidemiol [Internet]. 2003 Dec 1;158(11):1039–47. Available from: http://dx.doi.org/10.1093/aje/kwg251

59. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Stat Sci [Internet]. 1992;7(4):457–72. Available from: https://projecteuclid.org:443/euclid.ss/1177011136

60. Brooks SP, Gelman A. General Methods for Monitoring Convergence of Iterative Simulations. J Comput Graph Stat [Internet]. 1998 Dec 1;7(4):434–55. Available from: https://www.tandfonline.com/doi/abs/10.1080/10618600.1998.10474787

61. Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd ed. New York: Springer-Verlag New York, Inc; 2002.

62. Spiegelhalter D, Best N, P Carlin B. Bayesian Deviance, the Effective Number of Parameters, and the Comparison of Arbitrarily Complex Models. Vol. 64, Journal of Royal Statistical Society. 1998.

63. Wallace JR. Imap: Interactive Mapping [Internet]. 2012. p. R package version 1.32. Available from: https://cran.r-project.org/package=Imap

64. Anonymous. The current status of human monkeypox: Memorandum from a WHO meeting. Bull World Health Organ [Internet]. 1984;62(5):703–13. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2536211/pdf/bullwho00094-0031.pdf

65. Jezek Z, Marennikova SS, Mutumbo M, Nakano JH, Paluku KM, Szczeniowski M. Human Monkeypox: A Study of 2,510 Contacts of 214 Patients. Vol. 154, The Journal of infectious diseases. 1986. 551-555 p.

CHAPTER 3:
Evaluating intervention strategies to reduce zoonotic spillover of influenza A viruses from US exhibition swine: a modeling-based analysis

3.1 Introduction

Following the 2009 swine-origin influenza pandemic, there has been growing concern that swine may serve as an important source of other influenza A viruses (IAVs) capable of causing future pandemics in humans (1–9). The growing global domestic swine population (10) supports a large diversity of IAVs (2,5,6,11), most of which have never circulated in humans. In addition, recent studies have concluded that the types and distributions of the receptors for influenza A viruses in swine and human respiratory tracts are very similar (12–14), and it is known that there have been numerous occasions of transmission of human IAVs into swine (4,6,15–17), suggesting that there may be a relatively low barrier to host-jumps between these species. This idea is further supported by the fact that there have been more than 400 documented human infections with IAVs acquired from swine since 2011 in the United States alone (18). There are substantial concerns that spillover of IAVs from swine to humans may lead to sustained human-to-human transmission; these concerns are heightened by the fact that there have already been documented cases of limited human-to-human transmission of swine-origin H3N2 IAVs and by the large diversity of IAVs being given repeated opportunities to invade the human population (19–21).

The majority of the documented instances of swine-to-human spillover of IAVs in the United States can be linked to contact with show pigs at agricultural exhibitions, where millions of individuals from the general public come into close proximity with swine each year (22). Previous surveillance studies have detected IAVs in exhibition swine at around 20%-30% of

sampled agricultural fairs, with the prevalence of IAVs in swine at these shows often reaching high levels, sometimes even over 75% (23–26). Characterizing the transmission dynamics of IAV in exhibition swine is therefore a priority if we want to design and implement effective interventions to reduce the frequency of swine-to-human transmission events in the United States.

Despite their central role in IAV transmission to humans, show pigs make up only around 1.5% of the US swine herd and form a largely distinct group from commercial swine (23,27–29). They are often raised alone or in small groups on family farms or backyards and brought to compete at agricultural fairs, such as county and state fairs (21,29). Generally only swine from a particular county or state are permitted to compete in that county or state's fair, so many swine exhibitors will also bring their pigs to regional 'jackpot' or national swine shows, which are open to individuals from any county. Because show pigs reach their peak competitive condition at around six months of age, most swine exhibitors buy piglets several months in advance of their county's show and sell their pigs for pork products at the end of a show season, leaving their farms vacant of swine for part of the year (29).

Due to the strong seasonality of shows and when exhibition swine are present on home farms, it is unlikely that IAV transmission could be sustained in exhibition swine year-round; instead it is likely introduced from commercial swine over the course of the show season. Phylogenetic analyses using data from active IAV surveillance among exhibition swine between 2009-2013 suggest that a diversity of IAVs are introduced annually from US commercial swine into exhibition swine (21,27). The transmission pathway of IAVs from commercial to exhibition swine is not yet well understood, but it may occur through show pigs housed on the same

premises as commercial swine or through fomites (like boots or shared equipment) that may carry IAVs back to a home farm or to a show.

When an IAV outbreak occurs at a swine show, it can rapidly infect a large number of pigs from many different home farms (23–26). The viruses can spread even more broadly in the system if infected exhibition swine carry the infection between shows. However, the relative importance of repeated spillover versus onward transmission between shows, and the role that different types of shows play in sustaining IAV transmission within the show, is not yet well understood. For instance, because county fairs only allow resident pigs to compete, national and jackpot shows could act as bridges that allow IAVs to spread between counties. Clarifying the processes driving the disease dynamics is crucial if we want to devise and evaluate management plans to minimize the risk of spillover into humans.

However, there are numerous challenges associated with quantifying disease dynamics and assessing interventions in livestock systems. The transmission of infectious diseases in livestock is often heavily reliant on the movement patterns of animals transported between localities (30,31), but details of these movement patterns are often undocumented and unknown (21,32–34). Privacy is often a major concern in the U.S. agricultural industry, so even the locations where animals are kept and the sizes of farms are frequently unavailable (35). Additionally, few livestock diseases have compulsory reporting (36) or large-scale, structured surveillance programs. As a consequence, information about a pathogen's spread and prevalence is often patchy and incomplete (34,37). The result is a system where local structure can have an important impact on transmission dynamics, but where only disjoint pieces of information are available to inform that complexity (34). Mathematical models have served a valuable role in numerous livestock infectious disease systems by pulling together available information about a

complex system into a single framework to infer information about disease spread and test potential interventions (36–44).

In the current work, we use a network model to combine data from active surveillance of IAV at swine exhibitions in Indiana, Michigan, and (Figure 3.1) with insights about the swine show system's structure to evaluate which management strategies are likely to be most effective at reducing the risk of IAV spillover into humans. Several disparate estimation approaches were used to fit the parameters in the network model, including approaches based on epidemiological as well as sequence data, to promote robust results. We first tested our method using simulated datasets from twenty-one different true parameter sets covering a broad range of possible transmission behaviors. We examined how well the parameters estimated based on these simulated datasets matched the true values, as well as whether our method of assessing interventions reproduced the same intervention recommendations as the true parameter set. We then repeated this process for each of the three major HA lineages detected in show pigs in 2016: H3-2010, H3-2000, and H1-δ1 to estimate the effectiveness of thirty different potential interventions at reducing IAV in swine at county and state fairs.

3.2 Overview of the approach

To compare the expected impact of different potential interventions, we used data on IAV circulation at more than 120 shows in 2016 to fit the parameters of a model capturing the spread dynamics of IAVs in the swine show system. The fitted model was then used to assess the extent to which tested interventions reduced the fraction of county and state fairs with IAV outbreaks in attending swine, since these are the events where there is maximal contact with the general public.

Given that surveillance of IAV in show pigs was conducted at numerous swine shows but never at home farms and the expectation that shows are the primary avenue through which influenza viruses spread between exhibition pigs from different farms, we represented the transmission dynamics of IAV in the exhibition swine system as a network model where each node corresponds to a swine show (Figure S3.1). Nodes can take values of zero (if the show does not have an IAV outbreak among its swine) or one (if there is an IAV outbreak). The weight of edges between nodes represents the probability of indirect transmission from one show to another (via home farms that send pigs to both shows). A set of eight parameters quantifies IAV transmission probabilities: three parameters describe the probability of spillover into shows (λ_c for county and state shows, λ_j for jackpot shows, and λ_n for national shows), two parameters describe the probability a pig is infected by the end of an IAV-positive show (γ_c for county, state, and national shows and γ_j for jackpot shows), two parameters describe the probability that an infected pig starts an outbreak at a show (β_c for county, state, and national shows and β_j for jackpot shows), and one parameter describes the probability that an infected pig infects a susceptible pig on a given day on their home farm (ρ). Because pigs are generally housed on-site for several days at county, state, and national shows, but they may remain on their trailer except while competing at jackpot shows, the values of parameters describing probabilities of transmission at jackpot shows were allowed to differ from the parameters describing transmission at other types of shows. Parameters are summarized in Table 3.1.

To ensure that our findings were robust to model assumptions, we used several different methods to generate the estimated parameter sets. From a dataset of IAV in exhibition swine (either a real dataset from active IAV surveillance on one of the three HA lineages or a dataset simulated using known parameter values), we applied four principal parameter-estimation

methods to identify parameter sets expected to best describe disease transmission in the system. In the first two methods, we used epidemiological data on which shows were determined to be IAV positive (first row of Figure 3.1), along with the likelihood function based on the network model, to estimate parameters' posterior distributions using MCMC. We then either sampled parameter sets directly from the posterior (parameter-estimation method 'EP': Epidemiological data & Posterior distribution) or selected one hundred unique parameter sets that generated the highest likelihoods in the MCMC chain (parameter-estimation method 'EL': Epidemiological data & Likelihood).

The other two parameter-estimation methods used the phylogenetic tree formed from samples collected from both commercial swine and exhibition swine in 2016. In particular, we compared the size-distribution of show-only IAV clades in observed trees (third row of Figure 3.1) with those simulated over a grid of 6^4 parameter sets. We defined a show-only IAV clade as a monophyletic group of IAVs that were all collected from swine shows. The simulations used to generate the size-distribution of show-only IAV clades included both the observation process of IAV samples as well as the transmission process of IAV within the swine-show network (Figure S3.2). All analyses were repeated with three different tip-sampling assumptions about how IAV strains were sampled from the commercial swine IAV reservoir: a 'moderate-assortativity' assumption represents our expectation for the system and was used for the primary analyses, and two bounding assumptions ('completely-random' and 'high-assortativity') were used to test whether results were robust to the tip-sampling assumption used. For each tip-sampling assumption, two parameter-estimation methods were used: one based on summary statistics describing the size-distribution of show-only IAV clades (parameter-estimation method 'CS': Clade-size distribution data & Summary statistics) and one based on estimated likelihood

distributions that were fitted using the simulated clade-size distributions (parameter-estimation method ‘CL’: Clade-size distribution data & Likelihood). The parameter-estimation methods are described fully in the Methods section and summarized in Table S3.1.

A number of strategies have been proposed to help mitigate the risk of spillover from exhibition swine into humans, such as shortening the duration of county fairs, requiring downtime between attending shows, or improving biosecurity, but the potential impacts of each have not been formally assessed or compared (22,23) We are particularly interested in understanding the relative effectiveness of a collection of thirty different potential interventions that span reducing transmission at shows, on home farms, or from commercial swine; changing the timing of shows; removing certain shows; or requiring farms to wait for a certain period between attending shows (Table S3.2). Because each parameter-estimation method yields a cloud of plausible parameter sets, we used a collection of 1000 parameter sets from each parameter-estimation method to simulate disease spread under thirty different intervention scenarios as well as the no-intervention scenario (giving a total of thirty-one tested scenarios). We compared the simulation results to assess robust patterns in the extent to which each intervention is expected to reduce the fraction of IAV-positive county and state shows. The tested intervention scenarios are described in Table S3.2.

3.3 Results

Results from simulated test datasets

Parameter-estimation methods produce precision and accuracy in some parameter estimates, but are unable to precisely identify other parameter values

To assess how effectively parameter values could be inferred using each of the four different parameter-estimation methods, we tested the methods using twenty-one simulations with true parameter values selected to span the range of plausible parameter values (Table S3.3). All four methods generated fairly precise and accurate estimates for the probability of IAV spillover into county/state and jackpot shows (Table S3.4, see Figure S3.3 for two example parameter sets). The average error in the estimated probability of spillover per show was 0.04 across the twenty-one simulated datasets (where true values spanned from 0.001 to 0.9) and the four parameter-estimation methods. There was not evidence of a bias toward over-estimating or under-estimating the spillover parameter values ($p = 0.59$ for a one-sample, two-sided t-test). However, for the remaining parameter values, the parameter-estimation methods could only identify broad ranges of parameter space rather than precise estimates for each parameter. The mean absolute error between the estimated and true parameter values across all simulated datasets and parameter-estimation methods was 0.20 (true parameter values spanned 0.001 and 0.95). Here, estimated values for these parameters tended to overestimate the true parameter values ($p < 2E\text{-}16$ for a one-sample, two-sided t-test). There was a strong correlational structure between estimated parameters, so rather than identifying each individual parameter value, the methods identified parameter sets that were most consistent with the data. Significant correlations occurred between many pairs of parameters within parameter sets, especially between the parameters for spillover and transmission within a certain show type (e.g., between

λ_j and β_j) and between the at-show transmission probabilities (e.g. between γ_c and β_c and between γ_c and γ_j).

The accuracy of parameter estimates was similar across the twenty-one simulated datasets (Table S3.5). The absolute error for estimates of the county/state and jackpot spillover probabilities was lowest for parameter sets with small spillover probabilities (from less than 0.01 for λ1), and higher as spillover probability increased (to greater than 0.15 for λ6), though this trend disappears when considering the relative error (Table S3.6).

There was also similar performance across the four parameter-estimation methods. The methods that use data on the size-distribution of show-only IAV clades (CS and CL) performed marginally better than the methods based on the epidemiological data (EP and EL), with mean absolute error sizes of 0.19 compared to 0.21 (Table S3.4). However, each of the test datasets was simulated using a parameter set from the 6^4 parameter sets used in the clade-size distribution grid search, while the parameters underlying transmission in a real-world context would not necessarily match perfectly with one of the parameter sets tested. In addition, the tip-sampling assumption used to generate the clade-size distribution parameter estimates (see Methods) exactly matched the tip-sampling assumption used to generate the test datasets. When parameters were estimated based on incorrectly-specified tip-sampling assumptions, the CS and CL error sizes were slightly higher, with mean error sizes of 0.06 instead of 0.04 for the county/state and jackpot spillover probabilities and mean error sizes of 0.24 instead of 0.22 for the remaining parameters (Table S3.4). Estimates using the completely-random tip-sampling assumption (methods CSr and CLr) underestimated county/state and jackpot spillover probabilities by an average of 0.07 ($p < 0.001$) while the high-assortativity tip-sampling

assumption (methods CSh and CLh) overestimated these probabilities by an average of 0.03 ($p <$ 1E-4).

Broadly, all estimation methods achieved good accuracy on spillover parameters, with mean errors for the county/state and jackpot spillover probabilities falling between 0.02 and 0.07, and moderate but consistent accuracy on other parameters, with mean errors for the non-spillover parameters falling between 0.17 and 0.31 (Table S3.4).

Impact of interventions varies based on the underlying parameter values, but patterns emerge

The expected effectiveness of each intervention at reducing the fraction of IAV-positive county and state fairs varies depending on the parameters that govern the transmission dynamics in the system (Figure 3.2). For instance, when the parameters associated with at-show transmission probabilities (γ and β) are larger, removing jackpot and national shows (interventions 2-4), requiring downtime between shows (interventions 9-11, 13-16, 30-31), and reducing transmission at shows (interventions 24-26, 30-31) are all highly effective at reducing the fraction of county and state shows that are IAV positive. However, when the at-show transmission parameters are lower or when spillover probabilities (λ) are high, those interventions are less effective and instead interventions that reduce spillover probabilities (interventions 17 and 28), even only to 90% of the original values, cause the greatest expected reduction of the fraction of county and state fairs that are IAV positive (Figure 3.2).

While the predicted magnitude of an intervention's impact and the relative effectiveness of different interventions vary, certain interventions were frequently found to be effective across a wide range of parameter values, particularly intervention numbers 9, 13, 14, 23, 24, and 31, which correspond to interventions that reduce spillover and at-show transmission probabilities by

50%, require downtime before shows, and combine downtime before some shows with a minor decrease in transmission probabilities at shows (Figure 3.2).

The expected impacts of interventions estimated from test datasets closely approximates the impacts calculated using true parameter values

When calculating the expected impacts of interventions in a real-world setting, we do not know the true parameters underlying disease spread and instead must use parameter sets estimated from observed data. To assess how well the intervention impacts predicted by estimated parameter sets would agree with impacts predicted by true parameter values, we compared the expected reduction in the fraction of IAV-positive county and state fairs relative to the no-intervention scenario calculated using estimated parameter values with the reductions calculated using true parameter values. Despite the inability of all four parameter-estimation methods to obtain precise estimates of all but the county/state and jackpot spillover parameter values, the expected impacts of different interventions were largely consistent between estimated and true parameter values (Table S3.7; also see Figure S3.4 for results with two example parameter sets). The average error size in these values across all tested interventions and simulated datasets was less than 0.10 for each parameter-estimation method (Table S3.7). The clade-size-distribution methods (CS and CL) had the best performance with mean errors of 0.05.

There was a slight trend toward underestimating the effect of interventions, with the estimated effect-size of interventions an average of 0.013 below the true effect size across all interventions, parameter sets, and parameter-estimation methods ($p < 2\text{E-}16$ for a one-sample, two-sided t-test). However, the direction and significance of bias was not consistent when the results from different parameter-estimation methods were considered separately. Estimates based on the EP method did not have evidence of bias ($p = 0.41$), estimates based on the EL and CL

methods slightly underestimated the impact of interventions with a mean error of less than 0.06 ($p < 0.001$), and estimates based on the CS method slightly overestimated the impact of interventions with a mean error of 0.009 ($p < 0.002$). In addition, the directions and magnitudes of bias differed between intervention scenarios, with the largest bias observed for the 'remove all jackpot shows' intervention (intervention number 2), where average impact calculated across all parameter sets and parameter-estimation methods underestimated the true effectiveness by 0.058 ($p < 0.003$).

When the tip-sampling assumption was misspecified (the assumption used to generate the simulated dataset was different from the assumption used to generate parameter estimates), the average error size was 0.08, with the largest mean error size of 0.10 for the CL method assuming completely random tip-sampling (Table S3.8). Estimates from the completely-random tip-sampling assumption methods (CSr and CLr) underestimated the impact of interventions by an average of 0.07 ($p < 2E\text{-}16$), and estimates from the high-assortativity tip-sampling assumption methods (CSh and CLh) overestimated the impact of interventions by an average of 0.04 ($p < 7E\text{-}12$).

Multiplicative effects and antagonism when implementing multiple interventions simultaneously

Three of our intervention scenarios (interventions 29, 30, and 31) involved simultaneous combinations of two other interventions. We examined the interactions of these interventions, comparing the effects of a total of 21 (parameter sets used to generate simulated datasets) * 9 (ways to choose parameter estimates, including using the true parameter values) * 1000 (simulations run for each combination of parameter set / parameter-estimation method) = 189,000 simulations for each intervention. In one of the tested combined-intervention scenarios,

when both the parameters describing transmission at shows and the parameters describing spillover were reduced to 90% of their original values, the effect of the two interventions appeared multiplicative (with a difference of less than 0.005 between the mean effect of the combined interventions and the product of the means of the two individual effects). However, in the other two situations, when a 90% reduction in parameters describing transmission at shows was paired with one or two weeks of downtime before 2/3 of county and state fairs, the interaction appeared slightly antagonistic, although the difference between the mean effect of the combined interventions and the product of the means of the two individual effects was less than 0.03 and the mean effect of the combined intervention was greater than the mean effect of either intervention alone.

Results from the 2016 surveillance datasets

Similar parameter values estimated for the three HA lineages

Parameter estimates were generated using the 2016 surveillance datasets for each of the three HA lineages that appeared in more than two swine shows (H3-2010, H3-2000, and H1-δ1), yielding results that repeated many of the themes seen with the simulated datasets. For all three HA lineages, the parameter-estimation methods showed a strong preference for a relatively narrow range of county and jackpot spillover probabilities (Table S3.8). The mean estimates for λ_j among the parameter sets selected for intervention simulations were similar across HA-lineages, ranging from 0.07 to 0.08. The mean estimate for λ_c was highest for the H3-2010 lineage (0.08), followed by the H3-2000 lineage (0.06), and smallest for the H1-δ1 lineage (0.04). As was seen with the simulated test datasets, there was high uncertainty in estimates of the non-spillover parameters (Table S3.8).

The parameter sets selected using alternative tip-sampling assumptions differed (Table S3.8), but mean parameter estimates were largely consistent across assumptions, with an average difference of 0.06 between the primary tip-sampling assumption used in this analysis and each of the two other tested assumptions. Among the spillover estimates, the mean difference between the different assumptions was 0.04. As was seen for the simulated datasets, the completely-random tip-sampling assumption methods (CSr and CLr) produced estimates of the spillover probabilities that were an average of 0.05 smaller than estimates generated using the moderate-assortativity assumption, and the high-assortativity tip-sampling assumption methods (CSh and CLh) produced estimates of the spillover probabilities that were 0.04 larger than estimates generated using the moderate-assortativity assumption.

The individual parameter estimates, however, may not be as important as the parameter set taken as a whole for determining the behavior of the system. Certain parameter values tend to pair with one another, indicating a correlation structure where there may be the potential for parameters to compensate for one another to produce the same behavior. For example, in Figure 3.3, which shows tanglegrams of parameter sets selected under three different parameter-estimation methods for all three HA lineages, parts **E**, **F**, and **H** show that lower values of β tend to occur in the same parameter sets as higher values of γ. Across the 12,000 parameter sets sampled using the four parameter-estimation methods for the three HA lineages, values of γ_c above its median paired with values of β_c below its median or vice versa in 60% of parameter sets. Nearly identical patterns were observed for γ_j and β_j. The Pearson correlation coefficients describing the relationship between paired values of γ_c and β_c and between paired values of γ_j and β_j were -0.1 ($p < 2E\text{-}16$). As was seen in the simulated test datasets, significant correlation between parameter values within a parameter set were found for many parameter pairs,

especially between the parameters for spillover and transmission within a certain show type (e.g., between λ_c and γ_c and between λ_j and β_j), between spillover and home-farm transmission probabilities (e.g., between λ_c and ρ) and between the at-show transmission probabilities (e.g. between γ_c and β_c and between γ_c and γ_j).

Several interventions identified as having the largest expected impact in the exhibition swine system

The magnitude of the expected impact of interventions varied strongly, particularly across different methods used to generate parameter sets within a HA-lineage. For H3-2010, the two methods based on the epidemiological data (EP and EL) led to predictions of smaller impacts than the predictions based on clade-size-distribution methods (CS and CL) (Figure 3.4A, Figure S3.5), while the opposite pattern held for H1-δ1 (Figure 3.4A). Averaging across parameter-estimation methods, the largest reductions in the expected fraction of IAV-positive fairs were seen for H3-2010, which also had the largest number of IAV-positive shows in the 2016 dataset. The top intervention reduced the fraction of IAV-positive county and state fairs to below 40% of the no-intervention scenario. But even for the other two HA lineages, the top interventions are expected to reduce the fraction of shows infected to 55-65% of the fraction infected without intervention.

Yet despite the differences in predicted magnitude, the qualitative patterns of which interventions are expected to be most impactful were largely consistent across HA lineages and parameter-generation methods (Figure 3.4). The most impactful interventions included those that required downtime before shows and those that reduced spillover or at-show transmission probabilities. Requiring one or two weeks downtime before all shows (interventions 9 and 13), requiring two weeks downtime before county and state shows (intervention 14), and reducing

transmission at all shows or only at county and state shows (interventions 24 and 25) were found in the top ten interventions for all parameter-estimation methods and all HA lineages.

When the results from all parameter-estimation methods were averaged within each HA lineage, these same interventions were in the top ten interventions for each HA lineage, along with removing all jackpot shows (intervention 2), one week downtime before county/state fairs (intervention 10), reducing spillover probabilities (intervention 23) and combining two-weeks downtime for 2/3 of county/state fairs with a 90% reduction in at-show transmission (intervention 31). Each of these interventions is expected to reduce the fraction of IAV-positive county and state fairs to below 75% of the no-intervention scenario, and the 'two weeks downtime required before all shows' scenario is expected to reduce the fraction to below 60% across all HA lineages.

The intervention scenarios where jackpot or national shows were removed (interventions 2, 3, and 4) were often included in the top ten most effective interventions. Removing all eighty jackpot shows (intervention 2) or all four national shows (intervention 4) reduced the fraction of IAV-positive county and state shows to 69% or 85% of the no-intervention scenario, respectively, when averaged across all HA lineages and parameter-estimation methods.

Moving jackpot shows earlier in the season (interventions 5, 6, and 7) was never selected as one of the top ten interventions, though the average fraction of IAV positive county and state shows was decreased by 10-20% when jackpot shows were moved 2 or 4 weeks earlier (interventions 6 and 7) or when national shows were moved 4 weeks earlier (intervention 8).

Interventions that required farms to take downtime between shows (interventions 9-16) were often among the most effective interventions. Taking two weeks downtime was universally more effective than taking a single week downtime, reducing the fraction of IAV positive fairs

by an additional 10% on average. Only taking downtime after the four national shows (interventions 12 and 16) was less effective than the other downtime interventions tested, but those other interventions require the participation of a minimum of 172 shows. The interventions requiring farms to take one or two weeks downtime after national shows (interventions 12 and 16) reduced the fraction of IAV-positive county and state fairs to 93% and 82% of the no-intervention scenario, respectively

Reducing transmission at national shows or on home farms (interventions 21, 22, 27, and 28) was not effective at reducing the expected fraction of IAV-positive county and state fairs. Even when transmission parameters were reduced to 50% of their original values, the mean fraction of IAV-positive shows was reduced by less than 4% relative to the no-intervention scenario.

The sensitivity of the results to the tip-sampling assumption used to generate CS and CL parameter estimates was assessed by comparing the results obtained using the 'moderate assortativity' assumption with results obtained using completely-random and highly-assortative tip selection assumptions. The average difference between tip-selection assumptions in the expected reduction of IAV-positive fairs was 0.07, but differences as high as 0.51 were observed for certain interventions (Figure S3.6). However, across all tip-sampling assumptions, parameter-estimation methods, and HA-lineages, six interventions were consistently selected in the top ten most effective interventions, namely interventions 9, 13, 14, 24, 25, and 31, which involve downtime before shows and reducing transmission at shows. These are the same interventions as were identified above as the most impactful across all HA lineages.

3.4 Discussion

Because exhibition swine have a role as a type of 'intermediate host' allowing IAVs circulating in US swine substantially more opportunities to infect susceptible humans, implementing interventions that target this special group of swine has the potential to efficiently reduce the risk of spillover into humans. In this work, we used several different approaches to estimate the expected impact of thirty interventions on reducing the fraction of county and state fairs with IAV outbreaks in exhibition swine. The predicted impact of interventions was remarkably consistent across all methods, despite the distinct data sources and methods used to generate estimates, and despite challenges in identifying precise parameter values.

To estimate the impact of interventions, it was necessary to estimate parameter sets that describe IAV spillover from commercial swine and transmission among exhibition swine. For both simulated test datasets and real-world datasets, the parameter values describing spillover probabilities into county/state and jackpot shows were identified precisely, while the remaining parameter values exhibited a strong correlation structure, allowing values of parameter pairs to trade off with one another. This correlation structure, combined with the two well-identified spillover parameters and the structure imposed by the network model, meant that even though precise estimates of all individual parameters could not be obtained, estimated parameter sets predicted similar intervention impacts. Furthermore, for the test datasets simulated using known true parameter values, the expected intervention effectiveness calculated using estimated parameter sets was similar to that calculated using the true parameter values.

While this work indicates which interventions would be most effective if carried out, it does not evaluate the practicality of actually implementing interventions. Requiring downtime between shows, reducing transmission at shows and spillover into shows, and removing jackpot

or national shows were all indicated as being effective interventions, but not all of these are realistic options for real-world implementation. For instance, national and jackpot shows are highly valued and present many educational, training, cultural, and networking opportunities; removing them is therefore not a viable management policy. However, the expected impact of removing these shows sheds light on the role they play in contributing to infection at county and state fairs. Interestingly, even though our results suggest that these shows serve as an important bridge for IAVs to transmit between swine in different counties, removing all jackpot or all national shows would not be expected to reduce the fraction of IAV-positive county and state shows by more than 31%.

Intervention scenarios that involve reducing transmission probabilities help highlight locales where such policies would be most impactful, and thereby aid in prioritizing limited resources. For instance, reducing spillover probabilities or at-show transmission probabilities were substantially more effective at decreasing the fraction of IAV-positive fairs than reducing transmission probabilities on home farms. Based on this finding, we would expect that biosecurity measures which minimize transmission opportunities at shows, such as frequent sanitation of intake equipment or limiting the time swine are permitted to spend at a show (and thereby reducing the size of outbreaks) would be more effective than reducing swine-to-swine contact on home farms. However, because the current model did not explicitly include within-show transmission dynamics between swine, additional modeling and experimental work will be needed to quantify how much specific measures would affect transmission probabilities.

It is likely that the most practical and effective approach will involve multiple control measures implemented simultaneously. The present work investigated three examples of mixed strategies and found both multiplicative and slightly antagonistic interactive effects; however a

more thorough examination of different combined interventions is merited. Where possible, combining multiplicative or even synergistic interventions will yield the most cost-effective strategy.

Like all model-based analyses, the accuracy of this study's results depends on the model framework capturing the key transmission processes driving disease spread among exhibition swine. In particular, the present study assumes that the only way IAV can spread between pigs from different home farms is via transmission at a show. However, it is likely that there are some opportunities for farm-to-farm transmission outside of the show context, such as through open houses, swine sales, or contaminated equipment. It seems improbable that these transmission routes, likely between neighboring or local farms, would be an important driver in the spread of IAVs at the broader scales represented by exhibition swine, but studies that quantified the interactions between different farms would indicate whether this transmission route may be worth including in future studies.

The model used in the present analysis did not track the immune status of swine throughout the season. We expect that the lack of saturation of IAV in the system (only around 20-30% of tested shows are IAV positive (24,25)), the short lifespans of the pigs, and the large number of shared farms connecting shows would diminish the sensitivity of IAV dynamics to immunity from previous exposures. Nonetheless, immunity acquired from past IAV exposures could potentially change patterns of disease spread, especially given the high variance in the number of shows farms participate in during the season. While most swine exhibitors attend 1-4 shows during a season, a few exhibitors report bringing pigs to more than 20 shows. The potential of pigs from these farms to spread IAV widely depends on whether exposure early in the season protects them from infection later. Challenge experiments have suggested that there is

at least short-term protection against infection and virus replication within and, to a lesser degree, between IAV subtypes in swine (45–49), suggesting that this effect may be worth including in future work. Similarly, the use and efficacy of influenza vaccination in exhibition swine is currently not well documented but could be worth quantifying and including in future models.

Conclusion

While many measures implemented to reduce the risk of infectious disease spillover into humans are focused on the animal-human interface or on pathogen spread in humans after spillover, this work highlights the benefits of also considering interventions aimed at transmission in the reservoir to minimize spillover risk. The results we have presented in this work reinforce the value of combining data from surveillance in the reservoir and information about system structure with mathematical models to highlight key processes and evaluate potential interventions. A key requirement for this type of work is high-quality surveillance data, yet such information is difficult to obtain and unavailable for many systems. Conducting surveillance within the zoonotic reservoir and on developing mathematical modeling approaches to translate surveillance datasets into insights about routes of transmission would be a very worthwhile investment in many systems, as the resulting mechanistic understanding could reveal opportunities for efficient and effective prevention of spillover.

3.5 Methods

Data

Overview

Data on the presence or absence of IAV in exhibition pigs at 118 shows in Ohio, Michigan, and Indiana, as well as at four national shows in other states, were obtained from active IAV surveillance in 2016 (described in (26,50)). In total, exhibition swine from 96 county fairs, 2 state fairs, 4 national shows, and 20 jackpot shows were sampled. At each of these shows, between 20 and 600 (most commonly 20) nasal swabs or nasal wipes were collected from exhibition pigs and were tested for IAV using real-time reverse transcription PCR. For shows with one or more IAV positive samples, a subset of samples (generally one or two samples) were sequenced. The hemagglutinin (HA) protein from each sequenced sample was classified as belonging to H1δ1, H3-2010, H3-2000, or another HA lineage. This study made use of sequences publicly available on the National Center for Biotechnology Information (NCBI) Influenza Virus Resource (51).

Date and location of shows

In the Midwestern states, most county fairs take place between July and September, with state fairs in late July through early September. Jackpot shows tend to be held earlier in the season, with most occurring on weekends from May to July. There are also a few large national shows held throughout the year that attract pigs from all over the United States (52). For shows included in the 2016 active IAV surveillance, the exact GPS coordinates of each show was recorded. However, we also found the dates and counties of a number of shows that were not sampled in 2016 based on information reported on state agricultural fair websites, jackpot show

advertisements posted online, and 4-H schedules (data available upon request). For these unsampled shows, we used the centroid GPS coordinates for the relevant county to represent the show's location.

Number of pigs attending each show

When the number of pigs attending a show was reported as part of the 2016 surveillance dataset, exact numbers were used. When the numbers of attending swine was not reported for a 2016 county fair, the number reported for that county fair in 2013 in (29) was used. When the numbers of attending swine were unknown for a jackpot show, we used the average number of swine reported across the jackpot shows with reported swine numbers in 2016 because we do not expect there was a bias toward sampling jackpot shows of larger or smaller than average size.

Distribution of home-farm sizes and number of pigs brought to a show from the same home farm

The distribution of home-farm sizes and the distribution of the number of pigs brought to a show from the same home farm were both calculated using the responses of survey participants (29). Across nine swine exhibitions, a total of 428 surveys were collected from participants. The probability distribution for the number of swine brought to a show from the same home farm was taken as the exact distribution reported by survey participants. The distribution of the number of pigs on each home farm was also calculated from survey results, but because a few individuals housed their exhibition swine on the same premises as commercial swine, we only included the home-farm sizes for participants who reported a total of 15 or fewer swine on the premises.

Phylogenetic trees

Phylogenetic trees were obtained from the NCBI Influenza Virus Database (51), for the H1δ1, H3-2010, and H3-2000 HA lineages from among all US swine IAV sequences with collection dates in 2016. Phylogenetic trees created using the single linkage/nearest neighbor clustering algorithm and F84 nucleotide distances were downloaded from NCBI (51).

Creating augmented datasets

We expect that we recorded the date and location of all county and state fairs from the three states included in the study area but only around 70% of all jackpot shows from the three states. In addition, the number of farms in each county, and which farms attend which shows, are both unknown. However, all of these pieces of information are needed to inform the structure of the network model used in analyses. We therefore created "augmented datasets" using the time and location of known shows (see "Date and Location of Shows" section, above), the number of swine present at each show (see "Number of pigs attending each show" section, above), and survey information (29). First, we estimated the number of farms in each county as the number of show pigs that attend their county's fair (see "Number of pigs attending each show" section, above) divided by the estimated average number of pigs that attend a show from the same home farm (see "Distribution of home-farm sizes and number of pigs brought to a show from the same home farm" section, above). Next, we assigned which farms attended each show. For county shows, only farms within the same county are permitted to attend. For state shows, we give every farm in the relevant state an equal probability of being selected to attend the state show. For national shows, each farm in the three states is given an equal probability of being selected to attend, so long as the farm is not already scheduled to attend another show at the same time. For each jackpot show, the attending farms are drawn based on 1) which farms are not already

attending a concurrent show and 2) the distance between the centroids of the farm's county and of the show's county. To inform the fraction of farms that should be selected at different distances from the jackpot show, we used data on the GPS coordinates of home farms of pigs sampled at jackpot shows as part of the 2016 active surveillance. The fraction of home farms that reported being in each of six distance bins (0-25, 26-50, 51-100, 101-150, 151-200, and >200 miles) away from the sampled jackpot shows in the 2016 dataset was used to inform the number of farms to draw from each of these distance bins in the augmented dataset. If, for a particular jackpot show, there were insufficient farms available from within one of the distance bins, the remainder was drawn from the next-largest distance bin.

While we believe we found the dates and locations for the county and state shows in Ohio, Michigan, and Indiana, we expect that we may have missed several jackpot shows. We are aware of the timing and location of 56 jackpot shows, 20 of which were sampled as part of the active IAV surveillance. Based on the estimation that around 25% of all jackpot shows in Ohio, Michigan, and Indiana were included in the active IAV surveillance program (giving a total estimated 80 shows for that region), we estimated that we missed around 24 jackpot shows. To account for this, we added 24 extra 'augmented' jackpot shows to the augmented dataset, whose counties were chosen at random. The timing of each augmented show was drawn from a list of the dates of known jackpot show, with a 1, 2, or 3 week jitter either earlier or later in the season (with uniform probability for each jitter value). Other properties of the augmented jackpot shows, such as the number of participating swine, were selected at random from the observed jackpot shows with known values.

The augmented datasets were used 1) when creating simulated test datasets used to evaluate the method's performance (see "Simulations using the network model" section), 2)

during inference of parameters using likelihood from network model (explained in "Network model" section), and 3) for the simulations used to evaluate the expected impact of different intervention scenarios (see "Simulations using the network model" section). Different augmented datasets were generated for each of these three uses.

Network model

Because our data come from IAV surveillance at swine shows (and we lack farm-level sampling) and because shows are believed to allow influenza to spread between pigs from different home farms, we modeled disease spread in the system using a network model where the nodes are shows and the directional edges represent the probability of IAV transmission from one show to another. A node takes the value one (and is called 'IAV positive') if there is an IAV outbreak in pigs at that show or zero (and is called 'IAV negative') if there is not an IAV outbreak. A graphical depiction of the network model is shown in Figure S3.1.

A show can become IAV positive if spillover of IAV from commercial swine or from shows outside the modeled system occurs into pigs attending that show or if a show from earlier in the season 'transmits' IAV to the show. In this system, for one show to 'transmit' to another several events must occur: first, a pig must be infected while attending the initial show. That pig may either be brought directly to another show, or it may return to its home farm where it may infect other pigs from that home farm. Finally, an infected pig from that farm, either the original pig or a pig infected as a result of the original pig's infection, must be brought to the second show and must initiate an IAV outbreak there.

A set of eight parameters describe the probabilities associated with these transmission processes (Table 3.1). The probability of spillover occurring into a show depends on the type of

show, where there is probability λ_c that an IAV outbreak is started by spillover at a county or state show, probability λ_j at a jackpot show, and probability λ_n at a national show. The edge weight between show *a* and show *b* ($\omega_{a \to b}$), describing the probability of transmission from show *a* to show *b*, depends on the number of farms that send pigs to both shows ($\eta_{a,b}$) as well as the probability $\kappa(d_{a,b}, \xi_a, \xi_b)$ that pigs from a given shared farm will bring the infection from the first show (of show-type ξ_a) to the second show (of show-type ξ_b) $d_{a,b}$ days later and start an outbreak. The probability $\kappa(d_{a,b}, \xi_a, \xi_b)$ is informed by a set of five transmission parameters and is calculated using a simulation approach, as explained in Methods: "Shared-farm transmission simulations" below. The edge weight from show *a* to *b* is calculated as one minus the probability that none of the shared farms (farms that bring pigs to both show *a* and show *b*) transferred infection from show *a* to show *b*:

$$\omega_{a \to b} = 1 - \left[1 - \kappa\left(d_{a,b}, \xi_a, \xi_b\right)\right]^{\eta_{a,b}} \quad .$$

The probability that show *b* of show-type ξ_b is IAV positive (I_b=1) is equal to one minus the probability it was not infected from spillover nor from an IAV positive ancestor node (a show that occurred earlier in the season and is connected to show *b* with edge weights greater than zero):

$$P(I_b = 1 | A_b, \theta) = 1 - (1 - \lambda_{\xi_b}) * \prod_{a \in A_b} (1 - \mathbb{I}_{\{I_a = 1\}} \omega_{a \to b})$$

where A_b is the set of ancestor nodes of show *b*, and θ is the set of eight parameters that describe transmission probabilities.

The probability that a node takes value one or zero is conditionally independent of other nodes' values, given the values of the node's ancestors. Therefore, the overall likelihood for a set of parameters θ given the observed positive/negative status of all shows $D = \{I_1, I_2, I_3, \ldots, I_n\}$ is:

$$\mathcal{L}(\theta|D) = \prod_{b=1}^{n}\left[\mathbb{I}_{\{I_b=1\}}P(I_b = 1|A_b,\theta) + \mathbb{I}_{\{I_b=0\}}\left(1 - P(I_b = 1|A_b,\theta)\right)\right].$$

We assume no direct farm-to-farm transmission outside of shows. Although our model assumes that pigs on a farm can only be infected while attending an infected show, functionally, spillover could occur either into a farm before it brings pigs to a show or into pigs while they attend the show.

Shared-farm transmission simulations

To inform the probability $\kappa(d_{a,b}, \xi_a, \xi_b)$ that a farm that sends pigs first to IAV-positive show *a* and then to show *b* will start an outbreak at show *b*, we ran stochastic simulations to integrate over variability in the number of pigs on each farm and the number of pigs selected to attend shows, as well as to incorporate stochasticity in transmission dynamics arising from the small numbers of pigs on home farms. As above, $d_{a,b}$ is the number of days between the end of show *a* and the start of show *b*, ξ_a is the show-type of show *a* (county/state, jackpot, or national), and ξ_b is the show-type of show *b*.

We ran 100 simulations each time the probability of a farm transmitting IAV from one show to another was calculated for a new parameter set. In each simulation, we drew a farm's size and the number of pigs sent to each show that farm attended from two distributions described in "Distribution of home-farm sizes and number of pigs brought to a show from the same home farm," above. For each pig sent to show *a*, we drew random variables to determine whether or not that pig was infected at show *a* (with probability $\gamma_{\xi a}$ for each pig, which takes different values depending on whether or not show *a* was a jackpot show). If infected, we drew a second random variable giving the day of infection, where the probability distribution for different days of infection depended on the show type and when during the show a pig was

expected to have been infected. We assumed that pigs returning from jackpot shows could be on their first (with probability 0.33) or second (with probability 0.67) day of infection. Pigs returning from county, state, or national shows could be on their first through sixth day of infection, reflecting the longer length of these shows, with the highest probabilities associated with infection later in the show. This progression of increasing probability later in the show is meant to capture the growing epidemic size during the course of the show. The exact values used for the probabilities an infected pig was on its first through sixth day of infection when it returned home from the show were, in order: 0.38, 0.32, 0.16, 0.08, 0.04, and 0.02.

We then simulated transmission between swine on a home farm using a stochastic model that tracked the infectious state of each pig on the farm. All pigs started as susceptible except for pigs that were infected at show a. The probability that a given infectious pig transmitted the infection to a given susceptible pig on a particular day was equal to ρ. Each infected pig progressed through three stages of infection: the incubation stage, the infectious stage, and finally the recovered stage. The number of days pigs spent in each stage was drawn independently for each individual. The incubation period was drawn from a truncated normal distribution with mean 2.83 days and standard deviation 1.14 days. The infectious period was drawn from a truncated normal distribution with mean 4.5 days and standard deviation 1.07 days (values taken from (53) and similar to values used or reported in (45,54–58)). Both distributions were truncated (and renormalized) such that the minimum number of days in each stage was one and the maximum number of days was ten.

After recovery from infection, pigs on a farm were considered immune to future infection during that simulation. However, the immune status of pigs on each farm was not tracked through time, so each farm was assumed fully susceptible at the beginning of a within-farm IAV-

transmission simulation. In other words, the model did not track which farms had already been exposed to IAV earlier in the season. We assumed that pigs on a farm could remain infectious for at most 40 days from a single IAV introduction.

To establish whether any pigs from the farm transmitted infection to show *b*, we first randomly selected pigs from the home farm to attend show *b*. Each infected pig selected for attending show *b* had probability β_{ξ_b} of starting an outbreak at show *b*, where β_{ξ_b} takes different values depending on whether or not *b* is a jackpot show.

In summary, the probability $\kappa(d_{a,b}, \xi_a, \xi_b)$ is calculated using five parameters: the parameters γ_c and γ_j give the probability that a pig attending an IAV positive show becomes infected by the end of the show for county, state, or national shows and for jackpot shows, respectively. The parameter ρ describes the infectiousness of swine on the home farm. And the probability that an infected pig starts an outbreak at a show is given by β_c and β_j for county, state, or national shows and for jackpot shows, respectively. Based on the values for these parameters, simulations of the entire process: from pigs being taken home from show *a*, to transmission on the home farm, to pigs being brought to show *b*, are used to generate a single value, $\kappa(d_{a,b}, \xi_a, \xi_b)$, giving the probability that a given shared farm transmitted infection from show *a* to show *b*.

Selecting parameter sets to use in intervention simulations

For each dataset (either one of the 3 HA lineages or one of 21 test datasets simulated – see "Evaluating success of method based on test datasets simulated with known true parameter values"), we explored eight ways of generating parameter sets to use in intervention simulations. The first two approaches to generate the set of 1000 parameter sets $\Omega = \{\theta_1, \theta_2, \theta_3, \ldots \theta_{1000}\}$ used for intervention simulations are based on data on the positive/negative status of each sampled

show while the last six approaches are based on the observed clade-size distribution of IAV samples taken from show pigs. Table S3.1 provides a summary of the data sources and methods used to generate each of the eight sets of parameter sets Ω.

Parameter inference based on IAV positive/negative show statuses

Inference of parameter posteriors using MCMC

Using the IAV positive/negative status for each sampled show, the augmented dataset, and the likelihood function described in the 'Network model' section, we used MCMC to arrive at a posterior distribution for the parameter set θ. Because the IAV positive/negative status of unsampled shows is unknown, these values were treated as nuisance parameters in the inference. Chains were run for 35,000 steps, and the first third of each chain was discarded. The priors for all eight parameters in θ were set as uniform over the range [0,1].

Sampling 1000 parameter sets

To obtain the 1000 parameter sets used in the intervention simulations, two approaches were taken. In the first approach, the 1000 parameter sets were sampled from the posterior (from the output of the MCMC chain). In the second approach, the 100 unique parameter sets that were associated with the highest recorded likelihoods from the MCMC chain were selected and each was repeated ten times to give rise to a total of 1000 parameter sets.

Parameter set selection based on size-distribution of show-only IAV clades in phylogenetic tree

Our goal for this section of the analysis was to assess which parameter sets yield phylogenetic trees that are consistent with the trees observed in the real-world datasets. In

particular, we sought to determine which parameter sets reproduced similar size-distributions of show-only IAV clades.

Tips in the real-world IAV phylogenetic tree (obtained from NCBI Influenza Virus Database, as described in the "Phylogenetic trees" section, above) came both from IAV samples collected from commercial swine in 2016 in the United States, as well from IAV samples collected from pigs at sampled shows in Ohio, Michigan, and Indiana in 2016. We defined a show-only IAV clade as a monophyletic group whose tips all came from swine shows. In other words, in the phylogenetic tree, all descendants of the common ancestor of a show-only IAV clade must be IAV samples collected from swine shows. The only exception to this occurred when tips from both commercial swine IAV samples and from exhibition swine IAV samples all had zero branch lengths. In these cases, the show-swine tips with branches of zero length from one another were considered to be a show-clade.

To compare the observed size-distribution of show-only IAV clades with the size-distribution expected under different parameter sets, it is necessary to account for both the transmission and the observation processes. Figure S3.2 shows a schematic of the sampling and transmission processes we believe gave rise to the phylogenetic tree formed from the NCBI database of *observed* commercial- and show-pig IAV samples. We conceptualized the system as follows: from the full, mostly-unobserved true phylogenetic tree of all IAVs in commercial swine, only a subset of tips were actually sampled and appear on the NCBI Influenza Virus Database. The tips that were sampled may not have been chosen completely at random from the set of all possible tips. For example, commercial farms from certain geographic areas may have been more likely to submit their samples to the USDA, which is a major contributor of IAV samples from commercial swine on the NCBI database. Also from the full phylogenetic tree of

all IAVs in commercial swine, only a small subset of tips was responsible for spillover into exhibition swine. Again, the tips associated with these spillover events were unlikely to be independent from one another, especially considering that IAV in commercial swine occurs all over the United States while we focused on shows only from three states. After spillover into exhibition swine, an IAV may have spread between shows; we expect that this behavior was governed by the set of transmission parameters and the structure of the system (such as the time between shows and the number of farms that attended a pair of shows). Among all potentially-IAV-positive shows, 122 were sampled during the 2016 active surveillance and viruses from those shows were submitted to the NCBI database.

Simulating the size-distribution of show-only IAV clades for different parameter sets

For a given candidate parameter set, we ran 100 simulations of the observation and transmission processes that give rise to size-distributions of show-only IAV clades, enabling comparisons with the clade-size distributions from real-world data. In each simulation, we began by drawing the full 'true' phylogenetic tree to represent all IAV tips for commercial swine using the rTree() function from the Analysis of Phylogenetics and Evolution (ape) package in R (59). We assumed that the diversity of IAVs from one HA lineage in commercial swine in 2016 was represented by 10,000 tips. We then considered three different assumptions for how tips should be selected for commercial sampling and for spillover into exhibition swine. For the 'completely-random' tip-sampling assumption, we assumed that tips for both commercial and spillover are selected at random independently of one another from among all 10,000 commercial tips. For the second ('moderate-assortativity') and third ('high-assortativity') tip-sampling assumptions, we assumed that certain clades were more likely than others to be sampled or to spill over into exhibition swine. To capture this idea, we used an algorithm for choosing a subset of clades that

involves randomly selecting a tip from the full phylogenetic tree, finding the ancestor of that tip some fraction *F* of the way to the root ancestor, and then adding all tips from that clade to the collection of tips from which sampling or spillover is assumed to occur. This process was repeated until there were at least 1000 tips from which sampling or spillover could occur. Clades were selected independently for commercial sampling and for spillover (the same clade could be selected for both within the same simulation). In the second tip-sampling assumption, the value of *F* used for both commercial sampling and spillover was $F=1/3$, which results in sampling occurring from around 25 separate clades. In the third tip-sampling assumption, $F=1/3$ for commercial sampling but $F=3/5$ for spillover, which results in spillover occurring from an average of around 3.5 clades. The complete analysis was repeated using each of these three tip-sampling assumptions.

The number of spillover events that occurred during a given simulation depended on the spillover parameters. Following spillover of IAV into show pigs, the network model (described in the "Network model" section) was used to simulate the transmission of IAV between shows in the system. We assumed that the genetic sequences of IAV viruses taken from shows that transmitted to one another would form a clade in the full phylogenetic tree (formed from all commercial IAV tips and all show IAV tips). To reflect the real-world observation process of IAV from swine shows, only tips from shows that were part of the 2016 active surveillance were considered 'observed.'

Using only the sampled commercial and show tips, we constructed an 'observed' phylogenetic tree, similar to how tips in the tree obtained from NCBI consist of *observed* samples rather than all possible IAV tips. We obtained the size-distribution of show-only IAV clades directly from this tree. In summary, for a given candidate parameter set, each of the 100

simulations of the transmission and observation processes led to one instance of a clade-size distribution.

Because the process to generate the 100 simulated clade-size distributions for each parameter set is time-consuming, and our network model is described by eight parameters, a high-resolution search over the full parameter space was impractical. Instead, we first split the eight individual parameters into four parameter groups: $\{\lambda_c, \lambda_j, \lambda_n\}$, $\{\gamma_c, \gamma_j\}$, $\{\beta_c, \beta_j\}$, and $\{\rho\}$, where probabilities across different show types are grouped together. Six different sets of values were tested for each parameter group (see Table S3.3), with a full factorial design, giving a total of 6^4 parameter sets, with 100 simulations run for each parameter set.

Sampling 1000 parameter sets

To obtain the 1000 parameter sets used in the intervention simulations, two approaches were taken for each of the three tip-sampling assumptions. In the first approach, we fit a binomial distribution to the number of clades observed in each of the 100 simulations for a given parameter set and fit a negative binomial distribution to the clade-size distribution across those 100 simulations. These two fitted distributions were used to estimate the likelihood that a particular parameter set gave rise to the observed real-world clade numbers and clade-size distributions. The 1000 parameter sets used for the intervention simulations were sampled (with replacement) from among all 6^4 possible parameter sets, with the probability of selecting a particular parameter set proportional to the likelihood estimated for that parameter set.

The second approach to generate 1000 parameter sets for intervention simulations used a set of summary statistics calculated for all 100 of a parameter set's simulation outputs as well as for the real-world clade-size distribution. The statistics included in this test were 1) the number of show tips in the observed phylogenetic tree, 2) the number of show-only clades in the

observed phylogenetic tree, 3) the average show-only clade size, 4) the number of show-only clades containing two or more tips, 5) the number of show-only clades containing five or more tips, 6) the number of show-only clades containing ten or more tips, 7) the number of show-only clades containing fifteen or more tips, and 8) the maximum show-only clade size observed in phylogenetic tree. For a parameter set to be selected, the distribution formed by calculating a statistic for each of a parameter set's 100 simulations had to include the real-world value within the 90% CI for all eight statistics tested. Because fewer than 1000 parameter sets satisfied this criteria for every real-world (and test) dataset explored, the accepted parameter sets were repeated as necessary to make up the 1000 parameter sets used in intervention simulations. In two of the test datasets used to evaluate the method, none of the 6^4 parameter sets satisfied the 90% CI criteria for all eight tested statistics. In this case, the 95% CI was used in place of the 90% CI. If there still were no parameter sets that satisfied that criterion (as was the case for one test dataset), we accepted parameter sets that captured the true values within the 95% CI for four or more statistics.

Simulations of intervention scenarios using the network model

When estimating the expected impact of thirty-one intervention scenarios, a collection of 1000 parameter sets $\Omega = \{\theta_1, \theta_2, \theta_3, \ldots \theta_{1000}\}$ were selected in several different ways (see 'Selecting parameter sets to use in intervention simulations,' above). Each parameter set $\theta_i = \{\lambda_c, \lambda_j \lambda_n, \gamma_c, \gamma_j, \beta_c, \beta_j, \rho\}$ in Ω was used to run one simulation for all 31 intervention scenarios, giving a total of 1000 simulations under each intervention scenario. The network model was used to simulate IAV transmission into and among shows.

Among the 31 intervention scenarios (described in Table S3.2), one scenario was 'no intervention,' 13 scenarios involved reducing some of the transmission parameters, 3 involved removing a subset of shows, 4 involved changing the timing of shows, 8 involved requiring farms to take a period of time off between attending different shows, and 2 involved both a reduction in transmission parameters and requiring farms to take down time between attending certain shows. The 'no intervention' scenario and the 13 transmission-reduction scenarios used the same augmented dataset, while the remaining scenarios required drawing new augmented datasets that reflected the changed system structure.

Evaluating the success of the method using test datasets simulated with known parameter values

To assess how reliable we expected the results from this analysis to be at indicating the impact of different potential interventions, we used a set of twenty-one test datasets simulated with known parameter values that represent a wide range of transmission behaviors. The parameter sets used to generate test datasets were taken from Table S3.3. One simulation was run using value 2 for the spillover parameter group and value 3 for all other parameter groups. The remaining twenty simulations were run holding three parameter groups at the same values used for the first simulation while varying the fourth parameter group through the remaining five parameter sets indicated in Table S3.3.

For each simulated dataset, we sampled from the simulation output to mimic the real-world surveillance process and repeated the same analysis steps that were used for each HA lineage from the real-world dataset. Like in the HA-lineage analysis, we used eight methods to generate the 1000 parameter set estimates Ω and ran thirty-one intervention simulations for each parameter set. We then used the known 'true' parameter values (which were originally used to

generate each of the twenty-one simulated datasets) to simulate the effect of the thirty-one interventions. We compared the 'true' results (from simulations using 'true' parameter values) with the results of intervention simulations based on each set of estimated parameter values. For each pair of simulated test dataset and method to generate parameter estimates from that test dataset, we calculated the difference in the expected fraction of IAV positive county and state shows (relative to the no-intervention scenario) between the 'true' and 'estimated' results across intervention scenarios as well as the Pearson's correlation coefficient.

3.6 Figures and Tables

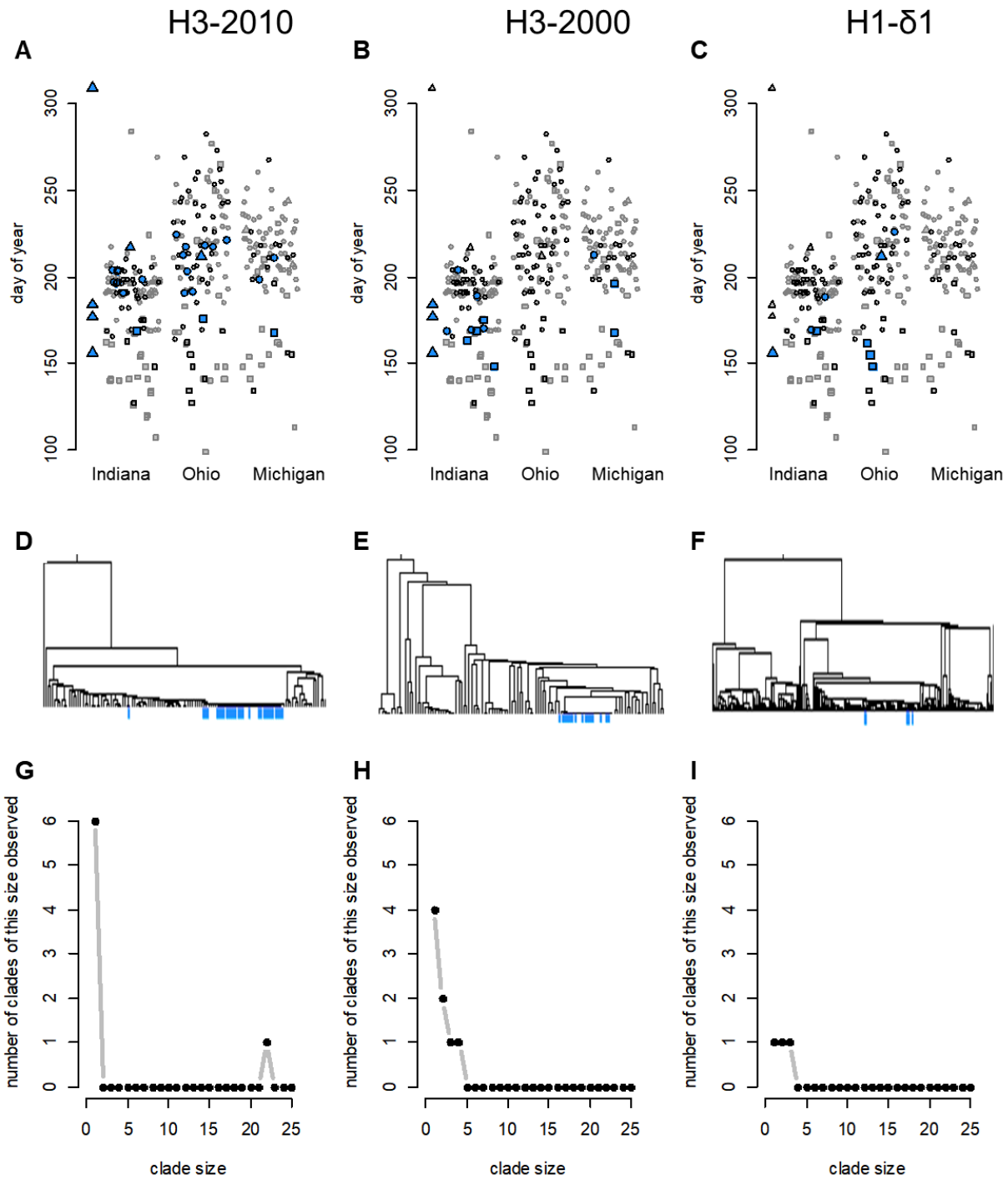
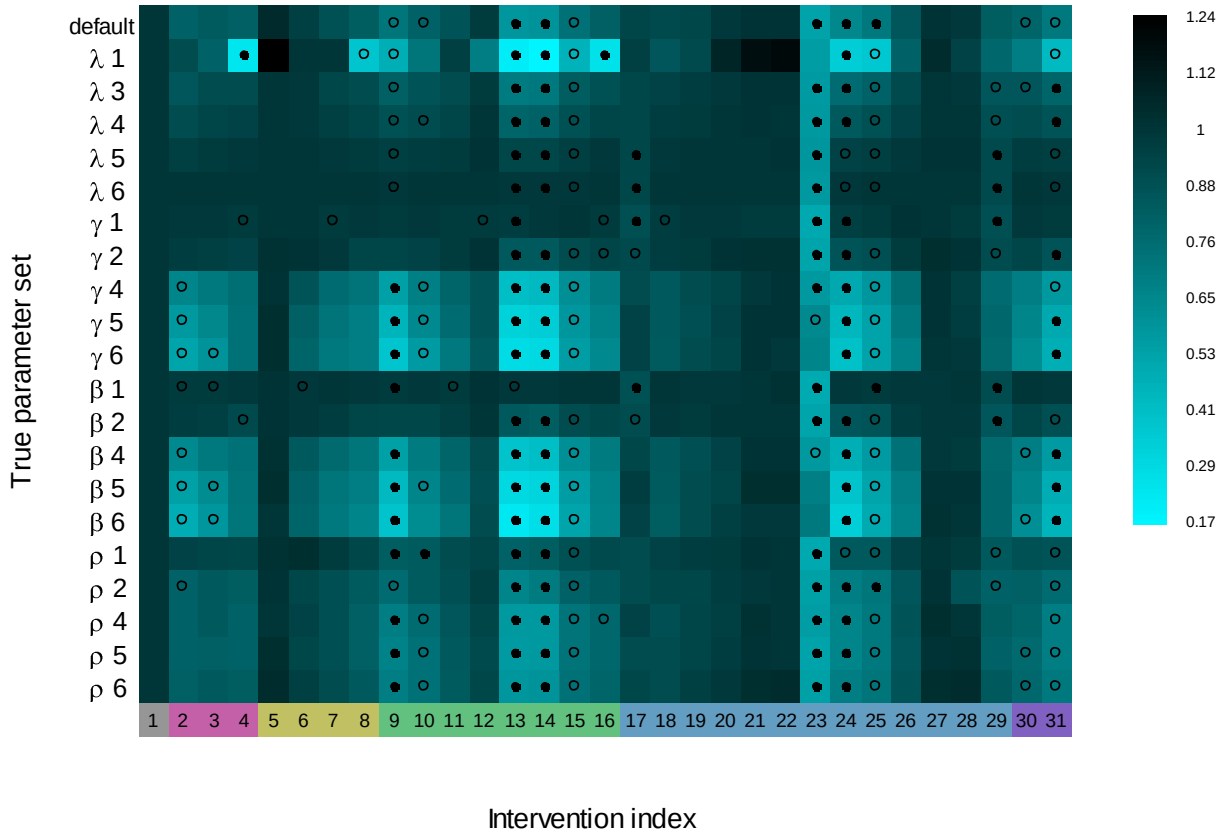


Figure 3.1. Data available for each HA lineage (columns) from the 2016 active surveillance of IAV in exhibition swine in Indiana, Ohio, and Michigan. The first row (**A-C**) displays epidemiological information: the timing, location, and types of shows (circles are county fairs,

squares are jackpot shows, and triangles are state and national shows). Along the x-axis, shows are grouped by state and then sorted by longitude within a state. Shows that were sampled in 2016 are plotted with a black outline; those that tested positive for IAV of a particular HA lineage are plotted as larger shapes with blue fill. The second row (**D-F**) shows the phylogenetic tree for each HA lineage based on sequences available on the NCBI Influenza Virus Database. Tips that come from active surveillance of exhibition swine in 2016 are indicated with blue (a maximum of one tip in each HA lineage is included per show). The third row (**G-I**) displays the show-only clade size distribution for each HA lineage.



Intervention types:

1: No intervention
2-4: Remove a subset of shows
5-8. Change the timing of some shows
9-16. Require farms to take downtime between attending shows
17-29: Reduce transmission probabilities
30-31: Combine downtime and reduced transmission probabilities

Figure 3.2. The expected fraction of IAV-positive county and state shows under different intervention scenarios relative to no-intervention is indicated by the color of a grid cell where lighter values indicate a more effective intervention. Each column corresponds to a tested intervention (Table S3.2). Each of the twenty-one rows corresponds to a different parameter set used to simulate disease spread (Table S3.3). These twenty-one parameter sets were obtained by partitioning the eight parameters in a parameter set into four parameter groups. Three parameter groups were held at default values while the remaining parameter group was cycled through from the smallest to largest of five non-default values. Parameter sets are named according to the parameter group that takes non-default values, so λ1 indicates that the smallest spillover values were used in the simulation while λ6 indicates that the largest spillover values were used. The top five interventions for each row are indicated by filled black circles and the next five best interventions are indicated by open circles.

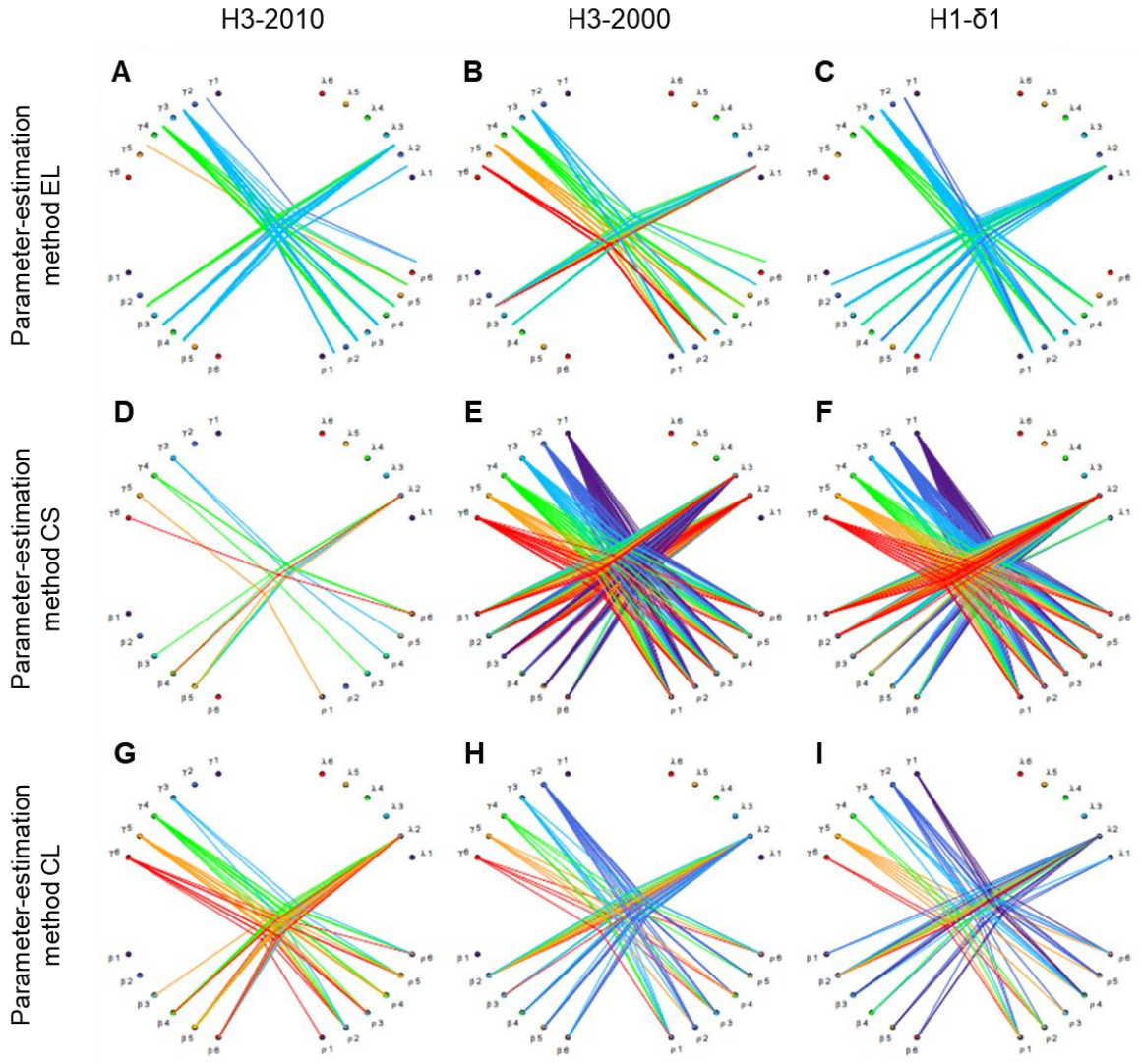


Figure 3.3. Tanglegrams of parameter sets used in intervention simulations; shown for all three HA lineages (columns) and for three of the parameter-estimate generating methods (rows). In each tanglegram, there are four parameter groups, one for the three spillover parameters (λ_c, λ_j, and λ_n), one for the two probabilities a pig was infected at a show (γ_c and γ_j), one for the two probabilities a pig started an outbreak at a show (β_c and β_j), and one parameter describing transmission on home farms (ρ). The values used in each parameter group are given in Table S3.3. Each estimated parameter set is represented in the tanglegram by four lines radiating out from a centroid point. Each line points to one of the six parameter sets from each parameter group. Lines are colored according to the γ value. Tanglegrams in the first row (**A-C**) show parameter sets generated using the EL parameter-estimation method (Table S3.1). Because parameter values within each parameter group are estimated individually, the endpoints of lines were calculated as the weighted average of each parameter in that group. So the location of the line in the λ group was determined by (λ_c* n_c + λ_j* n_j + λ_n* n_n) / (n_c + n_j + n_n) w here n_i is the number of shows of type *i*. Tanglegrams in the second (**D-F**) and third (**G-I**) rows show the

parameter sets estimated using the CS and CL parameter-estimation methods, respectively (Table S3.1). Here, parameters within a parameter group were not estimated independently (a grid search over the same 6^4 parameter sets represented in the tanglegrams was used to generate parameter estimates) so lines point directly to one of the six values within each parameter group.

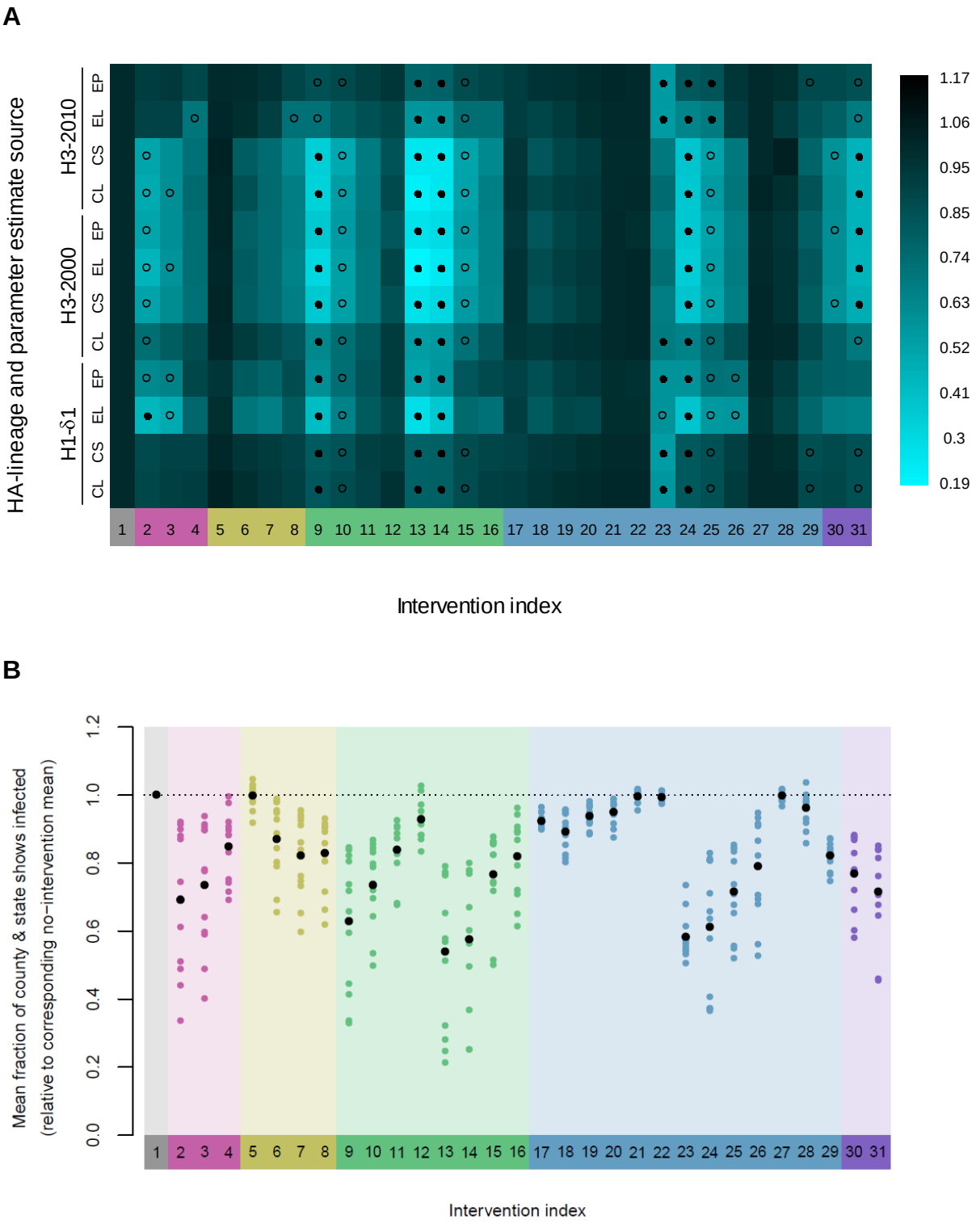


Figure 3.4. Expected impact of intervention scenarios. **A.** The expected fraction of IAV-positive county and state shows under different intervention scenarios relative to no-intervention is indicated by the color of a grid cell where lighter values indicate a more effective intervention.

Each row corresponds to a HA-lineage / parameter-estimate method pair. For each HA-lineage, the results from four parameter-estimation methods are shown (Table S3.1). Each column corresponds to a tested intervention. The top five interventions for each row are indicated by filled black circles and the next five best interventions are indicated by open circles. **B.** Average fraction of county and state shows IAV positive relative to the no-intervention mean. Colored points correspond to combinations of HA-lineage and the four methods used to generate parameter estimates (the height of each point corresponds to that HA-lineage / estimation method's color in subplot **A**). The black dots indicate the mean fraction for each intervention across all three HA lineages and all four parameter-estimate-sources.

Table 3.1. Description of model parameters.

Parameter	Description
λ_c	Probability spillover of an IAV occurs into a county or state fair from commercial swine or from exhibition swine outside Indiana, Michigan, and Ohio
λ_j	Probability spillover of an IAV occurs into a jackpot show from commercial swine or from exhibition swine outside Indiana, Michigan, and Ohio
λ_n	Probability spillover of an IAV occurs into a national show from commercial swine or from exhibition swine outside Indiana, Michigan, and Ohio
γ_c	Probability a pig will be infected with an IAV while attending a county, state, or national show where there is an IAV outbreak
γ_j	Probability a pig will be infected with an IAV while attending a jackpot show where there is an IAV outbreak
β_c	Probability an infected pig will start an IAV outbreak at a county, state, or national show
β_j	Probability an infected pig will start an IAV outbreak at a jackpot show
ρ	Probability that, while on the home farm, a given infected pig infects a given susceptible pig on a given day

3.7 Appendix Figures and Tables

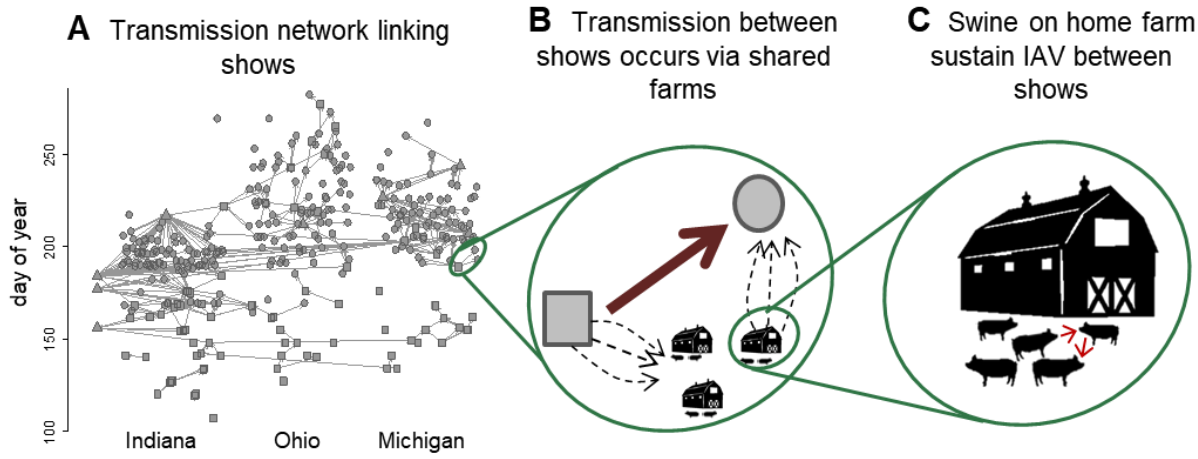


Figure S3.1. Schematic of the network model used to represent the transmission of IAV in the exhibition swine system. **A.** In the network model, each node corresponds to a swine show (circles are county fairs, squares are jackpot shows, and triangles are state and national shows) and the weight of an edge between a pair of shows indicates the probability of indirect transmission between the shows. **B.** Transmission between shows occurs if pigs from a shared farm bring infection from one show to the next. The probability of indirect transmission between two shows depends on the number of farms that bring pigs to both shows, the number of days between shows, and parameters that describe the probability a pig gets infected on a farm, spreads the infection at the home farm, and starts an outbreak at a show. **C.** The probability that a given shared farm brings IAV from one show to another is estimated using stochastic simulations using an SEIR-type model, with dynamics depending on the number of swine on the home farm.

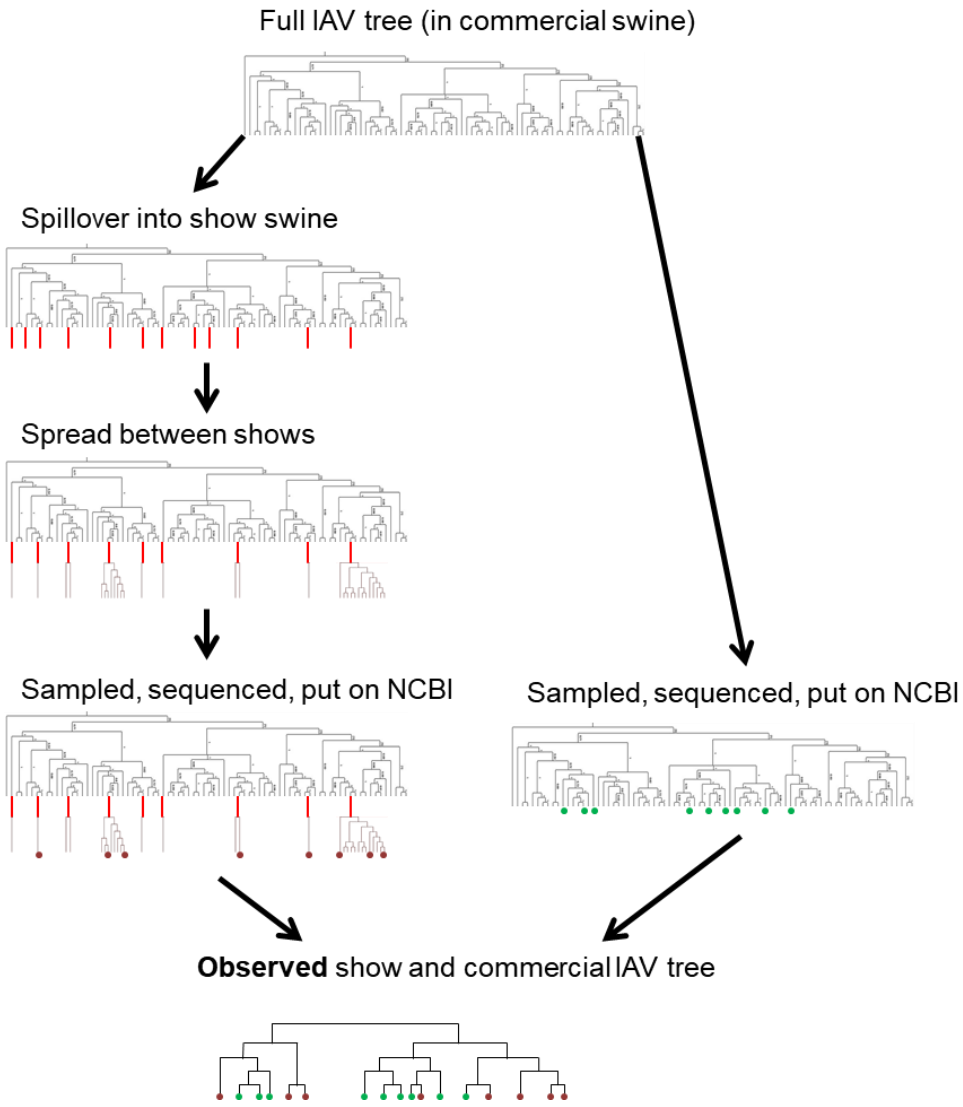


Figure S3.2. Diagram of the transmission and sampling processes that yield the observed sequences found on the NCBI influenza database. From a full tree of all IAV tips in commercial swine, a small subset of viruses spill into exhibition pigs at shows. From each spillover event, multiple shows may be infected if the virus is transmitted between shows via their shared farms. Only a subset of IAV-positive shows is under active surveillance, and sequences from these shows may appear on the NCBI database (indicated by purple dots). Similarly, only a small subset of tips from the commercial IAV tree is sampled and appears on NCBI (indicated by green dots). The tree created using the observed tips will have a different structure that the full, largely-unobserved tree.

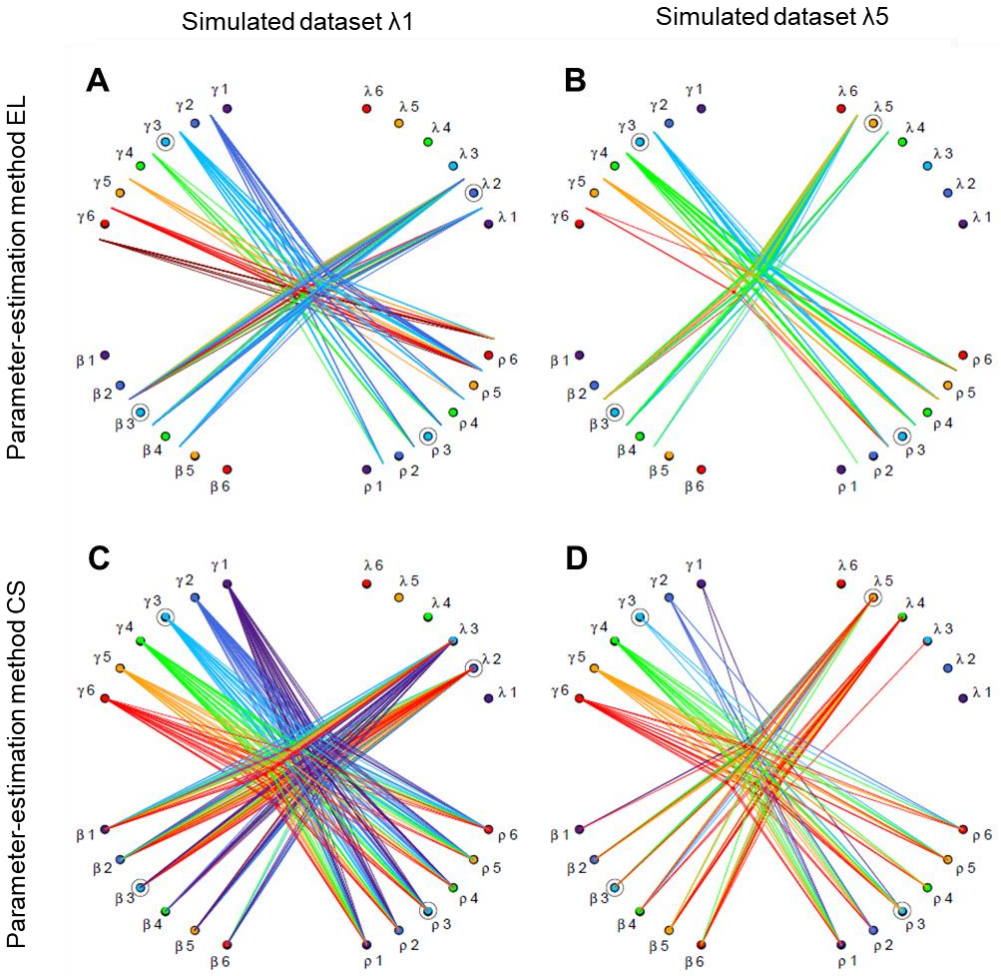


Figure S3.3. Tanglegrams of parameter sets used in intervention simulations; shown for two datasets simulated using different spillover rates (columns) and for two of the parameter-estimate generating methods (rows). True parameter values are circled. In each tanglegram, there are four parameter groups, one for the three spillover parameters (λ_c, λ_j, and λ_n), one for the two probabilities a pig was infected at a show (γ_c and γ_j), one for the two probabilities a pig started an outbreak at a show (β_c and β_j), and one parameter describing transmission on home farms (ρ). The values used in each parameter group are given in Table S3.3. Each estimated parameter set is represented in the tanglegram by four lines radiating out from a centroid point. Each line points to one of the six parameter sets from each parameter group. Lines are colored according to the γ value. Tanglegrams in the first row (**A-B**) show parameter sets estimated using the EL parameter-estimation method (Table S3.1). Because parameter values within each parameter group are estimated individually, the endpoints of lines were calculated as the weighted average of each parameter in that group. So the location of the line in the λ group was determined by (λ_c* n_c + λ_j* n_j + λ_n* n_n) / (n_c + n_j + n_n) where n_i is the number of shows of type i. Tanglegrams in the second row (**C-D**) show parameter sets estimated using the CS parameter-estimation method (Table S3.1). Here, parameters within a parameter group were not estimated independently (a

grid search over the same 6^4 parameter sets represented in the tanglegrams was used to generate parameter estimates) so lines point directly to one of the six values within each parameter group.

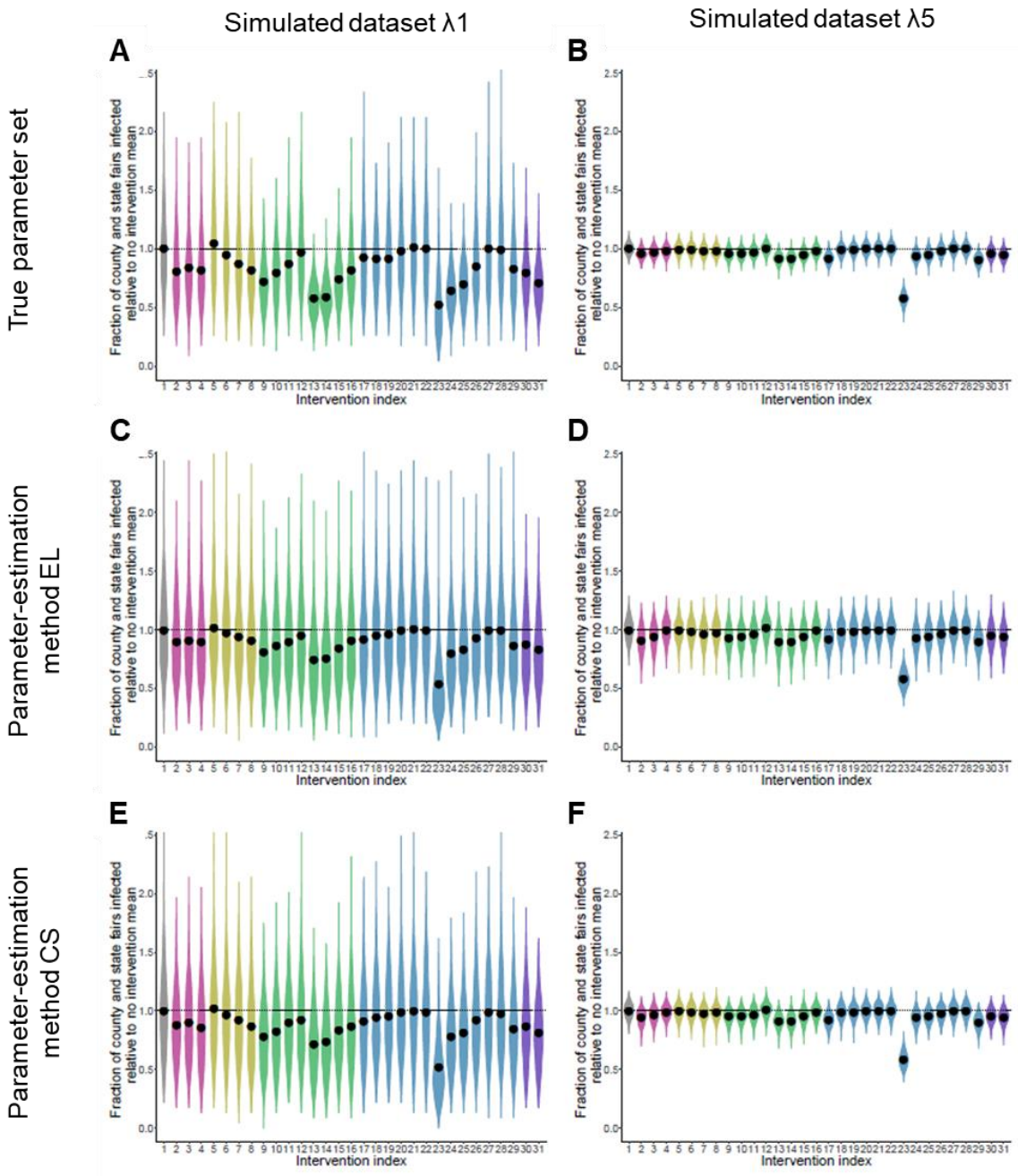


Figure S3.4. Violin plots show the results of the 1000 intervention simulations (Table S3.2) run for two test datasets (columns) using true parameter values as well as parameter values obtained using two estimation methods (rows). The first row (**A**-**B**) shows the fraction of IAV-positive county and state shows relative to the no-intervention scenario, with intervention simulations run using the true parameter values. These values can be compared with the results obtained using parameter sets estimated with the EL method (row 2, **C**-**D**) and the CS method (row 3, **E**-**F**) (Table S3.1). The mean value for each intervention is indicated by a black dot. The first column shows the results for a dataset simulated with low spillover rates (λ1) and the second column

shows the results for a dataset simulated with high spillover rates (λ5). The results for all other parameter sets are shown in Table S3.7.

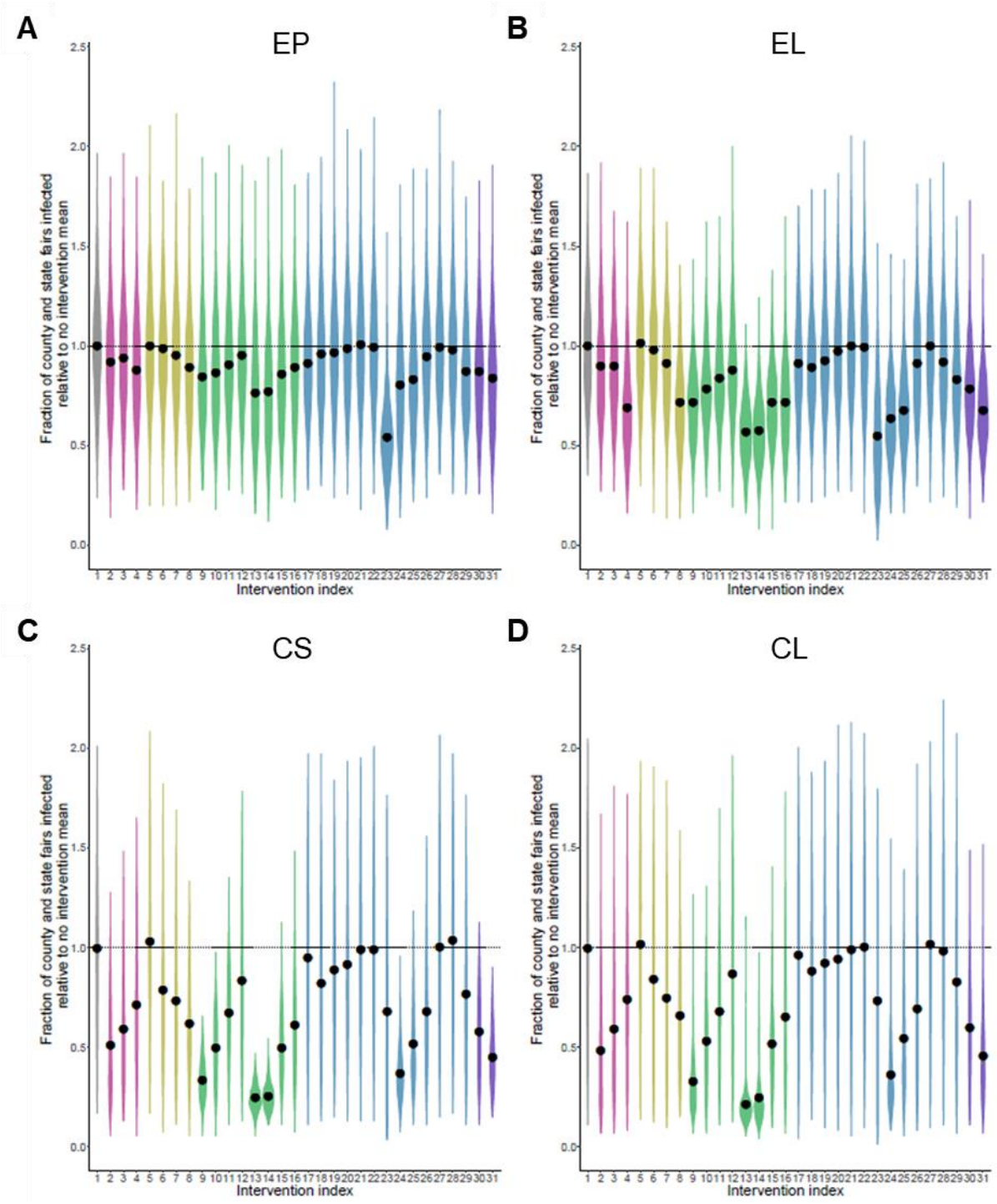


Figure S3.5. Fraction of county and state shows IAV positive under different intervention scenarios (Table S3.2) relative to the mean fraction positive with no intervention for the H3-2010 HA lineage. Violin plots show the results of 1000 simulations for each intervention scenario. The mean value is indicated by a black dot. Each subplot shows the results from a different method of generating parameter estimates: **A.** EP (Epidemiological data & Posterior distribution), **B.** EL

(Epidemiological data & Likelihood), **C.** CS (Clade-size distribution data & Summary statistics), and **D.** CL (Clade-size distribution data & Likelihood) (Table S3.1).

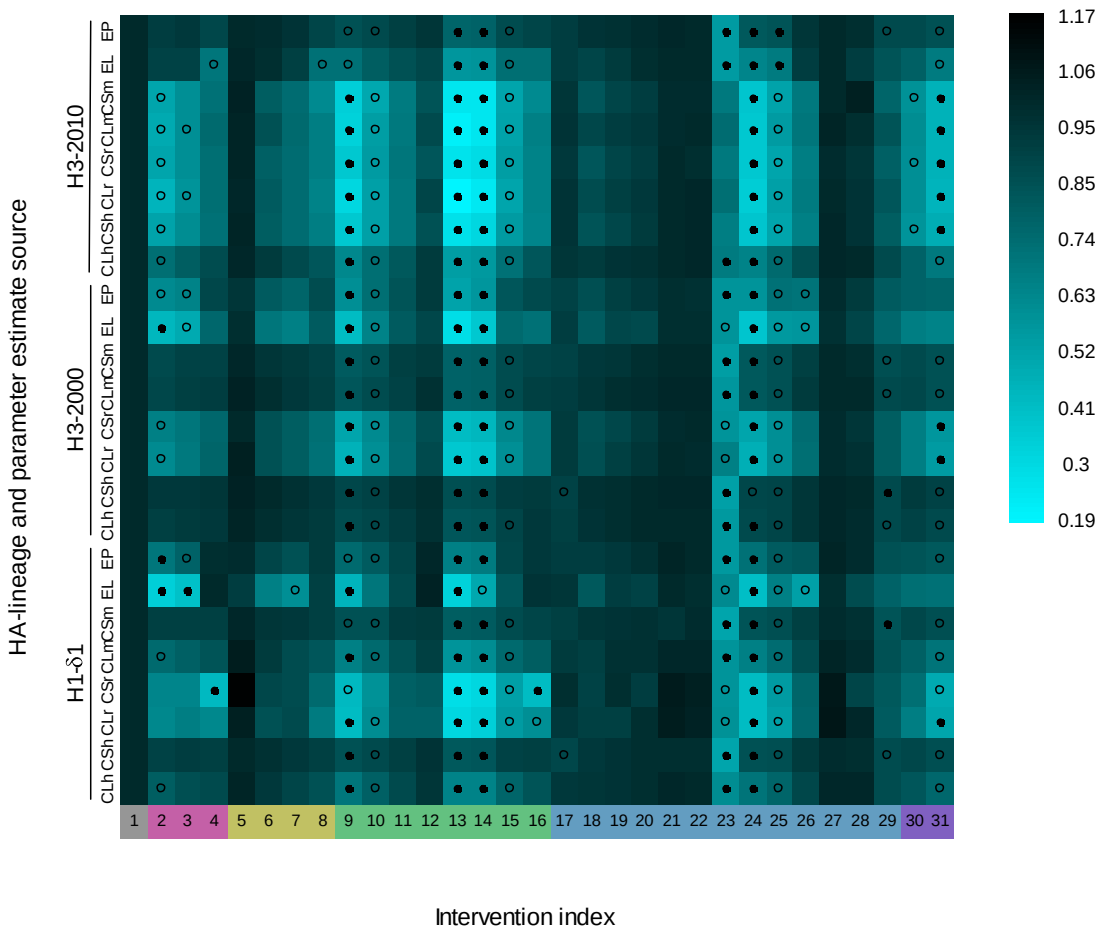


Figure S3.6. The expected fraction of IAV-positive county and state shows under different intervention scenarios relative to no-intervention is indicated by the color of a grid cell where lighter values indicate a more effective intervention. Each row corresponds to a HA-lineage / parameter-estimate-source pair. For each HA-lineage, the results from four parameter-estimation methods are shown (Table S3.1). For the clade-size distribution methods, results from all three tip-sampling assumptions are shown: the moderate-assortativity tip-sampling assumption (CSm and CLm), the completely-random tip-sampling assumption (CSr and CLr), and the highly-assortative tip-sampling assumption (CSh and CLh). Each column corresponds to a tested intervention. The top five interventions for each row are indicated by filled black circles and the next five best interventions are indicated by open circles.

Table S3.1. Four different approaches were taken to obtain the 1000 parameter sets used in the intervention simulations. The approaches differ in the data source and methods used to estimate the parameter sets that best match the data. For each HA lineage in the real-world dataset and for each test-simulated dataset, four groups of 1000 parameter set estimates were calculated. Each of these groups was then used to run simulations under different intervention scenarios.

Method name	**Data source**	**Approach used to evaluate parameter sets**	**Selection of 1000 parameter sets for use in intervention simulations**
EP (Epidemiological data & Posterior distribution)	Epidemiological data: IAV positive/negative status of each sampled show (row 1 of Figure 3.1)	The likelihood of a parameter set was calculated based on the network model of IAV transmission; MCMC used to search parameter space	Draw 1000 parameter sets from posterior (draw from MCMC chain with replacement)
EL (Epidemiological data & Likelihood)			Take the 100 parameter sets associated with the highest likelihoods from MCMC chain; repeat each value 10 times to get the 1000 sets for simulations
CS (Clade-size distribution data & Summary statistics)	Clade-size distribution data: the sizes of IAV-only clades in the observed phylogenetic tree (row 3 of Figure 3.1)	The network model of IAV transmission and the IAV phylogenetic tree observation model were used to generated 100 simulated clade-size distributions for a grid of 6^4 parameter sets	Use the parameter sets that satisfy the summary statistics (within the 90% CI for all eight statistics); repeat these parameter sets as necessary until have 1000 sets for simulations
CL (Clade-size distribution data & Likelihood)			Take 1000 draws (with replacement) from the 6^4 parameter sets, with the probability of a parameter set being drawn proportional to the likelihood calculated using the number of clades and the clade size distribution

Table S3.2. The names, models, and parameter values associated with each of the thirty-one interventions tested.

Intervention number	Intervention name	Model used	Parameter values used
1	No intervention	Original network model	Unmodified values from Ω
2	No jackpot	Removes all jackpot shows from the original network model	Unmodified values from Ω
3	Half jackpot	Removes half of the jackpot shows from the original network model (removed shows are randomly selected)	Unmodified values from Ω
4	No national	Removes all national shows from the original network model	Unmodified values from Ω
5	Jackpots one week earlier	Shifts all jackpot shows one week earlier in the season	Unmodified values from Ω
6	Jackpots two weeks earlier	Shifts all jackpot shows two weeks earlier in the season	Unmodified values from Ω
7	Jackpots four weeks earlier	Shifts all jackpot shows four weeks earlier in the season	Unmodified values from Ω
8	National shows four weeks earlier	Shifts all national shows four weeks earlier in the season	Unmodified values from Ω
9	One week downtime before all shows	Farms must allow at least one week to pass between the end date of one show and the start date of the next show they attend	Unmodified values from Ω
10	One week downtime before county/state	Farms must not attend any shows during the week before they attend a county or state fair	Unmodified values from Ω
11	One week downtime before 2/3 of county/state	A randomly selected 2/3 of county fairs enforce the rule that farms must not attend any shows during the week before they attend the county or state fair	Unmodified values from Ω
12	One week downtime after national shows	After attending a national show, farms must allow at least one week to pass before attending another show	Unmodified values from Ω
13	Two weeks downtime before	Farms must allow at least two weeks to pass between the end date of one show and the start date of the next	Unmodified values from Ω

	all shows	show they attend	
14	Two weeks downtime before county/state	Farms must not attend any shows during the two weeks before they attend a county or state fair	Unmodified values from Ω
15	Two weeks downtime before 2/3 of county/state	A randomly selected 2/3 of county fairs enforce the rule that farms must not attend any shows during the two weeks before they attend the county or state fair	Unmodified values from Ω
16	Two weeks downtime after national shows	After attending a national show, farms must allow at least two weeks to pass before attending another show	Unmodified values from Ω
17	Reduce spillover to 90%	Original network model	Replace λ_c with 0.9* λ_c Replace λ_j with 0.9* λ_j Replace λ_n with 0.9* λ_n
18	Reduce transmission at show to 90%	Original network model	Replace γ_c, γ_j, β_c, and β_j with 0.9*γ_c, 0.9*γ_j, 0.9*β_c, and 0.9*β_j, respectively
19	Reduce transmission at county/state shows to 90%	Original network model	Replace γ_c and β_c with 0.9*γ_c and 0.9*β_c *only at county and state fairs*
20	Reduce transmission at jackpot shows to 90%	Original network model	Replace γ_j and β_j with 0.9*γ_j and 0.9*β_j
21	Reduce transmission at national shows to 90%	Original network model	Replace γ_c and β_c with 0.9*γ_c and 0.9*β_c *only at national shows*
22	Reduce transmission on all farms to 90%	Original network model	Replace ρ with 0.9* ρ
23	Reduce spillover to 50%	Original network model	Replace λ_c with 0.5* λ_c Replace λ_j with 0.5* λ_j Replace λ_n with 0.5* λ_n
24	Reduce	Original network model	Replace γ_c, γ_j, β_c, and β_j

	transmission at all shows to 50%		with 0.5*γ_c, 0.5*γ_j, 0.5*β_c, and 0.5*β_j, respectively
25	Reduce transmission at county/state shows to 50%	Original network model	Replace γ_c and β_c with 0.5*γ_c and 0.5*β_c *only at county and state fairs*
26	Reduce transmission at jackpot shows to 50%	Original network model	Replace γ_j and β_j with 0.5*γ_j and 0.5*β_j
27	Reduce transmission at national shows to 50%	Original network model	Replace γ_c and β_c with 0.5*γ_c and 0.5*β_c *only at national shows*
28	Reduce transmission on all farms to 50%	Original network model	Replace ρ with 0.5* ρ
29	Reduce show transmission and spillover to 90%	Original network model	Replace γ_c, γ_j, β_c, β_j, λ_c, and λ_j with 0.9*γ_c, 0.9*γ_j, 0.9*β_c, 0.9*β_j, 0.9*λ_c, and 0.9*λ_j, respectively
30	One week downtime before 2/3 of county/state and reduce transmission at shows to 90%	A randomly selected 2/3 of county and state fairs enforce the rule that farms must not attend any shows during the week before they attend the county or state fair	Replace γ_c, γ_j, β_c, and β_j, with 0.9*γ_c, 0.9*γ_j, 0.9*β_c, and 0.9*β_j, respectively
31	Two weeks downtime before 2/3 of county/state and reduce transmission at shows to 90%	A randomly selected 2/3 of county and state fairs enforce the rule that farms must not attend any shows during the two weeks before they attend the county or state fair	Replace γ_c, γ_j, β_c, and β_j, with 0.9*γ_c, 0.9*γ_j, 0.9*β_c, and 0.9*β_j, respectively

Table S3.3. Values used for each parameter in a parameter group (parameters are described in Table 3.1). For the twenty-one simulated test datasets, the default values for parameter groups λ, γ, β, and ρ are $\lambda 2$, $\gamma 3$, $\beta 3$, and $\rho 3$, respectively (indicated in bold in the table). The first simulated test dataset used all default values and the subsequent twenty datasets used default values for all but one parameter group. Simulated datasets are named according to the parameter group that does not take default values. For example, in test dataset $\lambda 1$, default values were used for γ, β, and ρ, and $\lambda 1$ was used for λ.

Parameter group	**Values – 1**	**Values – 2**	**Values – 3**	**Values – 4**	**Values – 5**	**Values – 6**
λ (λ_c, λ_j, λ_n)	0.001, 0.001, 0.1	**0.05, 0.05, 0.3**	0.15, 0.15, 0.6	0.3, 0.3, 0.9	0.6, 0.6, 0.92	0.9, 0.9, 0.95
γ (γ_c, γ_j)	0.01, 0.007	0.1, 0.07	**0.3, 0.21**	0.5, 0.35	0.7, 0.49	0.9, 0.63
β (β_c, β_j)	0.01, 0.007	0.1, 0.07	**0.3, 0.21**	0.5, 0.35	0.7, 0.49	0.9, 0.63
ρ	0.01	0.1	**0.3**	0.5	0.7	0.9

Table S3.4. Absolute error between true parameters and estimates (averaged across the twenty-one tested true parameter sets and across the 1000 parameter sets generated for each true parameter set), reported for eight different methods of generating parameter estimates.

Parameter-estimation method	λ_c	λ_j	λ_n	γ_c	γ_j	β_c	β_j	ρ	**Mean**
EP	0.05	0.07	0.22	0.27	0.24	0.24	0.24	0.30	0.21
EL	0.03	0.05	0.21	0.25	0.26	0.28	0.23	0.29	0.21
CS	0.02	0.02	0.03	0.27	0.19	0.25	0.17	0.30	0.18
CL	0.05	0.05	0.06	0.28	0.20	0.26	0.19	0.30	0.19
CSr	0.07	0.07	0.07	0.30	0.21	0.29	0.20	0.29	0.19
CLr	0.07	0.07	0.07	0.30	0.21	0.28	0.20	0.30	0.20
CSh	0.03	0.03	0.08	0.27	0.19	0.27	0.19	0.31	0.21
CLh	0.07	0.07	0.12	0.28	0.20	0.26	0.18	0.30	0.21

Table S3.5. Absolute error between true parameters and estimates (averaged across the four parameter-estimation methods and across the 1000 parameter sets generated for each true parameter set), reported for each of twenty-one tested true parameter sets. Simulated datasets are named according to the parameter group that does not take default values. For example, in test dataset $\lambda 1$, default values were used for γ, β, and ρ, and $\lambda 1$ was used for λ.

Parameter set	λ_c	λ_j	λ_n	γ_c	γ_j	β_c	β_j	ρ	**Mean**
Default	0.03	0.04	0.18	0.26	0.19	0.20	0.25	0.31	0.19
$\lambda 1$	0.00	0.01	0.17	0.30	0.26	0.27	0.24	0.27	0.21
$\lambda 3$	0.09	0.07	0.20	0.25	0.19	0.33	0.20	0.26	0.22
$\lambda 4$	0.05	0.10	0.08	0.21	0.20	0.29	0.15	0.31	0.19
$\lambda 5$	0.10	0.17	0.10	0.25	0.24	0.22	0.26	0.26	0.21
$\lambda 6$	0.15	0.15	0.12	0.34	0.19	0.34	0.28	0.29	0.25
$\gamma 1$	0.03	0.02	0.08	0.30	0.26	0.24	0.20	0.30	0.20
$\gamma 2$	0.01	0.02	0.07	0.19	0.30	0.24	0.17	0.26	0.17
$\gamma 4$	0.03	0.03	0.14	0.28	0.23	0.25	0.25	0.29	0.20
$\gamma 5$	0.01	0.04	0.09	0.41	0.15	0.28	0.19	0.21	0.21
$\gamma 6$	0.02	0.04	0.10	0.37	0.24	0.22	0.25	0.26	0.21
$\beta 1$	0.02	0.03	0.12	0.26	0.25	0.28	0.16	0.27	0.19
$\beta 2$	0.02	0.02	0.09	0.23	0.19	0.18	0.18	0.26	0.16
$\beta 4$	0.02	0.01	0.11	0.23	0.27	0.21	0.15	0.24	0.17
$\beta 5$	0.02	0.07	0.25	0.25	0.18	0.21	0.22	0.23	0.19
$\beta 6$	0.02	0.02	0.15	0.31	0.21	0.33	0.21	0.24	0.21
$\rho 1$	0.01	0.02	0.07	0.26	0.23	0.29	0.31	0.44	0.22
$\rho 2$	0.05	0.08	0.22	0.23	0.17	0.24	0.18	0.36	0.20
$\rho 4$	0.03	0.04	0.27	0.23	0.27	0.23	0.18	0.35	0.21
$\rho 5$	0.01	0.02	0.06	0.26	0.25	0.32	0.17	0.35	0.20
$\rho 6$	0.02	0.03	0.08	0.24	0.20	0.25	0.18	0.49	0.20

Table S3.6. Relative error between true parameters and estimates (averaged across four methods for generating parameter estimates and across the 1000 parameter sets generated for each true parameter set), reported for each of twenty-one tested true parameter sets.

Parameter set	λ_c	λ_j	λ_n	γ_c	γ_j	β_c	β_j	ρ	**Mean**
Default	0.70	0.71	0.60	0.85	0.91	0.68	1.17	1.05	0.70
$\lambda 1$	2.87	7.18	1.66	1.01	1.23	0.89	1.12	0.90	2.87
$\lambda 3$	0.61	0.48	0.33	0.82	0.88	1.10	0.95	0.88	0.61
$\lambda 4$	0.18	0.33	0.09	0.71	0.94	0.97	0.71	1.04	0.18
$\lambda 5$	0.16	0.28	0.11	0.84	1.16	0.73	1.22	0.88	0.16
$\lambda 6$	0.17	0.17	0.13	1.13	0.89	1.14	1.32	0.97	0.17
$\gamma 1$	0.52	0.49	0.26	30.38	36.63	0.81	0.95	1.00	0.52
$\gamma 2$	0.25	0.31	0.23	1.88	4.27	0.80	0.80	0.88	0.25
$\gamma 4$	0.68	0.66	0.46	0.55	0.67	0.84	1.19	0.96	0.68
$\gamma 5$	0.30	0.71	0.29	0.59	0.30	0.93	0.92	0.70	0.30
$\gamma 6$	0.38	0.82	0.34	0.41	0.39	0.75	1.17	0.86	0.38
$\beta 1$	0.37	0.51	0.41	0.85	1.18	27.55	23.37	0.91	0.37
$\beta 2$	0.36	0.32	0.29	0.78	0.90	1.75	2.57	0.87	0.36
$\beta 4$	0.46	0.30	0.36	0.77	1.29	0.43	0.42	0.82	0.46
$\beta 5$	0.41	1.30	0.82	0.85	0.87	0.30	0.44	0.78	0.41
$\beta 6$	0.44	0.36	0.49	1.02	1.00	0.36	0.34	0.80	0.44
$\rho 1$	0.26	0.37	0.22	0.88	1.11	0.95	1.46	44.32	0.26
$\rho 2$	1.07	1.51	0.72	0.75	0.83	0.80	0.84	3.57	1.07
$\rho 4$	0.70	0.88	0.92	0.76	1.30	0.78	0.84	0.69	0.70
$\rho 5$	0.27	0.38	0.21	0.87	1.17	1.05	0.81	0.51	0.27
$\rho 6$	0.36	0.67	0.26	0.78	0.97	0.83	0.86	0.54	0.36

Table S3.7. Comparison of intervention simulation results using true and parameter sets estimated using four methods. Rows are the parameter set used to generate a simulated dataset. Values are the error in mean fraction of IAV-positive county/state shows relative to no-intervention scenario.

Parameter set	**EP** (Epidemiological data & Posterior distribution)	**EL** (Epidemiological data & Likelihood)	**CS** (Clade-size distribution data & Summary statistics)	**CL** (Clade-size distribution data & Likelihood)	Average across parameter-estimation methods
Default	0.06	0.05	0.06	0.02	0.05
$\lambda 1$	0.15	0.18	0.14	0.15	0.16
$\lambda 3$	0.05	0.04	0.04	0.05	0.04
$\lambda 4$	0.01	0.07	0.01	0.01	0.02
$\lambda 5$	0.01	0.00	0.05	0.02	0.02
$\lambda 6$	0.01	0.01	0.00	0.10	0.03
$\gamma 1$	0.33	0.36	0.04	0.07	0.20
$\gamma 2$	0.06	0.09	0.02	0.07	0.06
$\gamma 4$	0.12	0.04	0.12	0.20	0.12
$\gamma 5$	0.09	0.11	0.03	0.03	0.06
$\gamma 6$	0.12	0.07	0.01	0.02	0.06
$\beta 1$	0.09	0.08	0.03	0.03	0.06
$\beta 2$	0.07	0.16	0.04	0.03	0.07
$\beta 4$	0.08	0.05	0.05	0.05	0.06
$\beta 5$	0.14	0.06	0.04	0.03	0.07
$\beta 6$	0.10	0.04	0.03	0.03	0.05
$\rho 1$	0.03	0.19	0.06	0.02	0.07
$\rho 2$	0.09	0.06	0.06	0.06	0.07
$\rho 4$	0.07	0.14	0.04	0.06	0.08
$\rho 5$	0.08	0.20	0.10	0.03	0.10
$\rho 6$	0.10	0.09	0.03	0.03	0.06
Average across all parameter sets	0.09	0.10	0.05	0.05	0.07

Table S3.8. Comparison of intervention simulation results obtained with true parameter sets and results obtained with parameter sets estimated using misspecified tip-sampling assumptions. Error in mean fraction of county/state shows infected relative to no-intervention.

Parameter set	**CSr** (completely-random tip-sampling assumption)	**CLr** (completely-random tip-sampling assumption)	**CSh** (high-assortativity tip-sampling assumption)	**CLh** (high-assortativity tip-sampling assumption)
Default	0.09	0.13	0.07	0.06
$\lambda 1$	0.14	0.18	0.17	0.15
$\lambda 3$	0.22	0.21	0.05	0.03
$\lambda 4$	0.14	0.16	0.01	0.02
$\lambda 5$	0.18	0.19	0.01	0.01
$\lambda 6$	0.22	0.24	0.00	0.03
$\gamma 1$	0.07	0.06	0.02	0.08
$\gamma 2$	0.09	0.15	0.02	0.04
$\gamma 4$	0.11	0.10	0.12	0.21
$\gamma 5$	0.04	0.04	0.14	0.05
$\gamma 6$	0.02	0.03	0.10	0.10
$\beta 1$	0.06	0.04	0.05	0.03
$\beta 2$	0.03	0.02	0.04	0.02
$\beta 4$	0.07	0.07	0.10	0.04
$\beta 5$	0.03	0.04	0.09	0.06
$\beta 6$	0.03	0.03	0.11	0.04
$\rho 1$	0.05	0.02	0.06	0.02
$\rho 2$	0.11	0.14	0.10	0.09
$\rho 4$	0.12	0.14	0.06	0.05
$\rho 5$	0.03	0.07	0.09	0.06
$\rho 6$	0.07	0.10	0.06	0.08
Average across all parameter sets	0.09	0.10	0.07	0.06

Table S3.9. Mean (and 95% CI) of the 1000 parameter estimates used from each method of generating parameter sets and for each of the HA lineages.

HA lineage	**Parameter-estimation method**	λ_c	λ_j	λ_n	γ_c	γ_j	β_c	β_j	ρ
H3-2010	**EP** (Epidemiological data & Posterior distribution)	0.15 (0.05-0.25)	0.14 (0.01-0.32)	0.79 (0.36-0.99)	0.39 (0.06-0.93)	0.16 (0-0.64)	0.27 (0.03-0.8)	0.39 (0.01-0.95)	0.43 (0.06-0.97)
H3-2010	**EL** (Epidemiological data & Likelihood)	0.07 (0.04-0.11)	0.06 (0.02-0.16)	0.86 (0.65-0.99)	0.3 (0.13-0.56)	0.08 (0.03-0.18)	0.44 (0.16-0.62)	0.55 (0.13-0.93)	0.29 (0.09-0.78)
H3-2010	**CS** (Clade-size distribution data & Summary statistics; moderate-assortativity tip-sampling assumption)	0.05 (0.05-0.05)	0.05 (0.05-0.05)	0.3 (0.3-0.3)	0.53 (0.3-0.9)	0.37 (0.21-0.63)	0.59 (0.3-0.7)	0.41 (0.21-0.49)	0.6 (0.01-0.9)
H3-2010	**CL** (Clade-size distribution data & Likelihood; moderate-assortativity tip-sampling assumption)	0.06 (0.05-0.15)	0.06 (0.05-0.15)	0.32 (0.3-0.6)	0.63 (0.3-0.9)	0.44 (0.21-0.63)	0.64 (0.3-0.9)	0.45 (0.21-0.63)	0.45 (0.01-0.9)
H3-2010	**CS** (Clade-size distribution data & Summary statistics; completely-random tip-sampling assumption)	0.05 (0.05-0.05)	0.05 (0.05-0.05)	0.3 (0.3-0.3)	0.57 (0.3-0.9)	0.4 (0.21-0.63)	0.57 (0.3-0.9)	0.4 (0.21-0.63)	0.39 (0.01-0.9)
H3-2010	**CL** (Clade-size distribution data & Likelihood; completely-random tip-sampling assumption)	0.05 (0.05-0.05)	0.05 (0.05-0.05)	0.3 (0.3-0.3)	0.64 (0.3-0.9)	0.45 (0.21-0.63)	0.67 (0.3-0.9)	0.47 (0.21-0.63)	0.45 (0.01-0.9)

H3-2010	**CS** (Clade-size distribution data & Summary statistics; high-assortativity tip-sampling assumption)	0.05 (0.05-0.05)	0.05 (0.05-0.05)	0.3 (0.3-0.3)	0.59 (0.3-0.9)	0.41 (0.21-0.63)	0.56 (0.3-0.9)	0.39 (0.21-0.63)	0.42 (0.01-0.9)
H3-2010	**CL** (Clade-size distribution data & Likelihood; high-assortativity tip-sampling assumption)	0.16 (0.05-0.31)	0.16 (0.05-0.31)	0.56 (0.3-0.9)	0.51 (0.01-0.9)	0.35 (0.01-0.63)	0.49 (0.01-0.9)	0.34 (0.01-0.63)	0.42 (0.01-0.9)
H3-2000	**EP** (Epidemiological data & Posterior distribution)	0.03 (0-0.08)	0.1 (0.01-0.24)	0.63 (0.14-0.94)	0.36 (0.01-0.95)	0.49 (0.12-0.97)	0.09 (0.01-0.3)	0.57 (0.19-0.97)	0.32 (0.02-0.94)
H3-2000	**EL** (Epidemiological data & Likelihood)	0.01 (0-0.03)	0.03 (0-0.06)	0.49 (0.14-0.78)	0.48 (0.09-0.85)	0.5 (0.21-0.78)	0.07 (0.02-0.17)	0.77 (0.49-0.98)	0.26 (0.03-0.88)
H3-2000	**CS** (Clade-size distribution data & Summary statistics; moderate-assortativity tip-sampling assumption)	0.1 (0.05-0.15)	0.1 (0.05-0.15)	0.44 (0.3-0.6)	0.34 (0.01-0.9)	0.24 (0.01-0.63)	0.32 (0.01-0.9)	0.22 (0.01-0.63)	0.4 (0.01-0.9)
H3-2000	**CL** (Clade-size distribution data & Likelihood; moderate-assortativity tip-sampling assumption)	0.11 (0.05-0.3)	0.11 (0.05-0.3)	0.47 (0.3-0.9)	0.34 (0.01-0.9)	0.24 (0.01-0.63)	0.33 (0.01-0.9)	0.23 (0.01-0.63)	0.39 (0.01-0.9)
H3-2000	**CS** (Clade-size distribution data & Summary statistics; completely-random tip-sampling assumption)	0.05 (0.05-0.05)	0.05 (0.05-0.05)	0.3 (0.3-0.3)	0.5 (0.1-0.9)	0.35 (0.07-0.63)	0.47 (0.1-0.9)	0.33 (0.07-0.63)	0.41 (0.01-0.9)
H3-	**CL** (Clade-size	0.05	0.05	0.3 (0.3-	0.49 (0.1-	0.34	0.49 (0.1-	0.34	0.45

2000	distribution data & Likelihood; completely-random tip-sampling assumption)	(0.05-0.05)	(0.05-0.05)	0.3)	0.9)	(0.07-0.63)	0.9)	(0.07-0.63)	(0.01-0.9)
H3-2000	**CS** (Clade-size distribution data & Summary statistics; high-assortativity tip-sampling assumption)	0.1 (0.05-0.15)	0.1 (0.05-0.15)	0.46 (0.3-0.6)	0.31 (0.01-0.9)	0.21 (0.01-0.63)	0.3 (0.01-0.9)	0.21 (0.01-0.63)	0.4 (0.01-0.9)
H3-2000	**CL** (Clade-size distribution data & Likelihood; high-assortativity tip-sampling assumption)	0.13 (0.05-0.3)	0.13 (0.05-0.3)	0.54 (0.3-0.9)	0.33 (0.01-0.9)	0.23 (0.01-0.63)	0.3 (0.01-0.9)	0.21 (0.01-0.63)	0.39 (0.01-0.9)
H1-δ1	**EP** (Epidemiological data & Posterior distribution)	0.04 (0.01-0.09)	0.15 (0.03-0.3)	0.29 (0.02-0.7)	0.21 (0-0.84)	0.38 (0.01-0.95)	0.21 (0.01-0.82)	0.3 (0.02-0.9)	0.41 (0.01-0.96)
H1-δ1	**EL** (Epidemiological data & Likelihood)	0.02 (0-0.04)	0.05 (0.01-0.1)	0.27 (0.05-0.46)	0.08 (0-0.21)	0.64 (0.28-0.97)	0.33 (0.02-0.8)	0.41 (0.12-0.89)	0.19 (0.02-0.6)
H1-δ1	**CS** (Clade-size distribution data & Summary statistics; moderate-assortativity tip-sampling assumption)	0.05 (0.05-0.05)	0.05 (0.05-0.05)	0.3 (0.3-0.3)	0.29 (0.01-0.9)	0.2 (0.01-0.63)	0.31 (0.01-0.9)	0.22 (0.01-0.63)	0.41 (0.01-0.9)
H1-δ1	**CL** (Clade-size distribution data & Likelihood; moderate-assortativity tip-sampling assumption)	0.04 (0-0.05)	0.04 (0-0.05)	0.26 (0.1-0.3)	0.36 (0.01-0.9)	0.25 (0.01-0.63)	0.36 (0.01-0.9)	0.25 (0.01-0.63)	0.42 (0.01-0.9)
H1-δ1	**CS** (Clade-size distribution data &	0.01 (0-	0.01 (0-	0.16 (0.1-	0.61 (0.1-	0.43 (0.07-	0.54 (0.1-	0.38 (0.07-	0.49 (0.1-

	Summary statistics; completely-random tip-sampling assumption)	0.05)	0.05)	0.3)	0.9)	0.63)	0.9)	0.63)	0.9)
H1-δ1	**CL** (Clade-size distribution data & Likelihood; completely-random tip-sampling assumption)	0.01 (0-0.05)	0.01 (0-0.05)	0.14 (0.1-0.3)	0.46 (0.1-0.9)	0.32 (0.07-0.63)	0.44 (0.1-0.9)	0.31 (0.07-0.63)	0.45 (0.01-0.9)
H1-δ1	**CS** (Clade-size distribution data & Summary statistics; high-assortativity tip-sampling assumption)	0.05 (0.05-0.05)	0.05 (0.05-0.05)	0.3 (0.3-0.3)	0.29 (0.01-0.9)	0.2 (0.01-0.63)	0.3 (0.01-0.9)	0.21 (0.01-0.63)	0.39 (0.01-0.9)
H1-δ1	**CL** (Clade-size distribution data & Likelihood; high-assortativity tip-sampling assumption)	0.06 (0-0.15)	0.06 (0-0.15)	0.31 (0.1-0.6)	0.34 (0.01-0.9)	0.24 (0.01-0.63)	0.34 (0.01-0.9)	0.24 (0.01-0.63)	0.42 (0.01-0.9)

3.8 References

1. Smith G, Vijaykrishna, Bahl, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature [Internet]. 2009;459(7250):1122. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=f5h&AN=42428767&site=ehost-live

2. Nfon CK, Berhane Y, Hisanaga T, Zhang S, Handel K, Kehler H, et al. Characterization of H1N1 swine influenza viruses circulating in Canadian pigs in 2009. J Virol [Internet]. 2011 Jun 21; Available from: http://jvi.asm.org/content/early/2011/06/22/JVI.00801-11.abstract

3. Strelioff CC, Vijaykrishna D, Riley S, Guan Y, Peiris JSM, Lloyd-Smith JO. Inferring patterns of influenza transmission in swine from multiple streams of surveillance data. Proc R Soc B Biol Sci [Internet]. 2013 Jul 7;280(1762). Available from: http://rspb.royalsocietypublishing.org/content/280/1762/20130872.abstract

4. Nelson MI, Stratton J, Killian ML, Janas-Martindale A, Vincent AL. Continual Reintroduction of Human Pandemic H1N1 Influenza A Viruses into Swine in the United States, 2009 to 2014. Sandri-Goldin RM, editor. J Virol [Internet]. 2015 Jun 15;89(12):6218–26. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4474294/

5. Vincent A, Awada L, Brown I, Chen H, Claes F, Dauphin G, et al. Review of Influenza A Virus in Swine Worldwide: A Call for Increased Surveillance and Research. Zoonoses Public Health [Internet]. 2013 Apr 5;61(1):4–17. Available from: https://doi.org/10.1111/zph.12049

6. Lewis NS, Russell CA, Langat P, Anderson TK, Berger K, Bielejec F, et al. The global antigenic diversity of swine influenza A viruses. Jit M, editor. Elife [Internet]. 2016 Apr 22;5:e12217. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4846380/

7. Nelson M, Culhane MR, Rovira A, Torremorell M, Guerrero P, Norambuena J. Novel Human-like Influenza A Viruses Circulate in Swine in Mexico and Chile. PLoS Curr [Internet]. 2015 Aug 13;7:ecurrents.outbreaks.c8b3207c9bad98474eca3013fa933c. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4551470/

8. Nelson MI, Viboud C, Vincent AL, Culhane MR, Detmer SE, Wentworth DE, et al. Global migration of influenza A viruses in swine. Nat Commun [Internet]. 2015 Mar 27;6:6696. Available from: http://dx.doi.org/10.1038/ncomms7696

9. Nelson MI, Worobey M. Origins of the 1918 Pandemic: Revisiting the Swine “Mixing

Vessel." Am J Epidemiol [Internet]. 2018 Jul 26;kwy150-kwy150. Available from: http://dx.doi.org/10.1093/aje/kwy150

10. Animal Production and Health. Agriculture and Consumer Protection Department. Pigs and... [Internet]. Food and Agriculture Organization of the United Nations. 2016. Available from: http://www.fao.org/ag/AGAInfo/themes/en/pigs/home.html

11. Anderson TK, Campbell BA, Nelson MI, Lewis NS, Janas-Martindale A, Killian ML, et al. Characterization of co-circulating swine influenza A viruses in North America and the identification of a novel H1 genetic clade with antigenic significance. Virus Res [Internet]. 2015;201:24–31. Available from: http://www.sciencedirect.com/science/article/pii/S0168170215000799

12. Nelli RK, Kuchipudi S V, White GA, Perez BB, Dunham SP, Chang K-C. Comparative distribution of human and avian type sialic acid influenza receptors in the pig. BMC Vet Res [Internet]. 2010 Jan 27;6:4. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2832630/

13. Van Poucke SGM, Nicholls JM, Nauwynck HJ, Van Reeth K. Replication of avian, human and swine influenza viruses in porcine respiratory explants and association with sialic acid distribution. Virol J [Internet]. 2010;7(1):38. Available from: https://doi.org/10.1186/1743-422X-7-38

14. Trebbien R, Larsen LE, Viuff BM. Distribution of sialic acid receptors and influenza A virus of avian and swine origin in experimentally infected pigs. Virol J [Internet]. 2011 Sep 8;8:434. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3177912/

15. Nelson MI, Gramer MR, Vincent AL, Holmes EC. Global transmission of influenza viruses from humans to swine. J Gen Virol [Internet]. 2012 Oct 6;93(Pt 10):2195–203. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3541789/

16. Nelson MI, Wentworth DE, Culhane MR, Vincent AL, Viboud C, LaPointe MP, et al. Introductions and Evolution of Human-Origin Seasonal Influenza A Viruses in Multinational Swine Populations. García-Sastre A, editor. J Virol [Internet]. 2014 Sep 16;88(17):10110–9. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4136342/

17. Nelson MI, Vincent AL. Reverse zoonosis of influenza to swine: new perspectives on the human-animal interface. Trends Microbiol [Internet]. 2015 Mar 4;23(3):142–53. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4348213/

18. Centers for Disease Control and Prevention. Case Count: Detected US Human Infections with H3N2v by State since August 2011 [Internet]. 2017. Available from: https://www.cdc.gov/flu/swineflu/h3n2v-case-count.htm

19. Prescott K, Kroona S, et al. Limited Human-to-Human Transmission of Novel Influenza A (H3N2) Virus - Iowa, November 2011. Morb Mortal Wkly Rep. 2011;60(47):1615–7.

20. Cauchemez S, Epperson S, Biggerstaff M, Swerdlow D, Finelli L, Ferguson NM. Using Routine Surveillance Data to Estimate the Epidemic Potential of Emerging Zoonoses: Application to the Emergence of US Swine Origin Influenza A H3N2v Virus. PLoS Med [Internet]. 2013 Mar 5;10(3):e1001399. Available from: http://dx.doi.org/10.1371%2Fjournal.pmed.1001399

21. Nelson MI, Wentworth DE, Das SR, Sreevatsan S, Killian ML, Nolting JM, et al. Evolutionary Dynamics of Influenza A Viruses in US Exhibition Swine. J Infect Dis [Internet]. 2016 Jan 15;213(2):173–82. Available from: http://dx.doi.org/10.1093/infdis/jiv399

22. National Assembly of State Animal Health Officials (NASAHO) and National Association of State Public Health Veterinarians (NASPHV). Measures to minimize influenza transmission at swine exhibitions, 2018. 2018; Available from: http://nasphv.org/documentsCompendiaZoonoticInfluenza.html

23. Bliss N, Nelson SW, Nolting JM, Bowman AS. Prevalence of Influenza A Virus in Exhibition Swine during Arrival at Agricultural Fairs. Zoonoses Public Health [Internet]. 2016 Sep 1;63(6):477–85. Available from: http://dx.doi.org/10.1111/zph.12252

24. Bowman AS, Nolting JM, Nelson SW, Slemons RD. Subclinical Influenza Virus A Infections in Pigs Exhibited at Agricultural Fairs, Ohio, USA, 2009–2011. Emerg Infect Dis [Internet]. 2012;18(12):1945–50. Available from: https://dx.doi.org/10.3201/eid1812.121116

25. Bowman AS, Workman JD, Nolting JM, Nelson SW, Slemons RD. Exploration of risk factors contributing to the presence of influenza A virus in swine at agricultural fairs. Emerg Microbes Infect [Internet]. 2014 Jan 22;3(1):e5. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3913824/

26. Bowman A. Evaluation of influenza dynamics in exhibition swine at jackpot shows [Internet]. 2017. Available from: https://www.pork.org/research/evaluation-influenza-dynamics-exhibition-swine-jackpot-shows/

27. Nelson MI, Stucker KM, Schobel SA, Trovão NS, Das SR, Dugan VG, et al. Introduction, Evolution, and Dissemination of Influenza A Viruses in Exhibition Swine in the United States during 2009 to 2013. Schultz-Cherry S, editor. J Virol [Internet]. 2016 Dec 1;90(23):10963–71. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5110178/

28. Lauterbach SE, Zentkovich MM, Nelson SW, Nolting JM, Bowman AS. Environmental

surfaces used in entry-day corralling likely contribute to the spread of influenza A virus in swine at agricultural fairs. Emerg Microbes Infect [Internet]. 2017 Feb 22;6(2):e10. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5322325/

29. Bliss N, Stull J, Moeller S, Rajala-Schultz P, Bowman A. Movement patterns of exhibition swine and associations of influenza A virus infection with swine management practices. J Am Vet Med Assoc [Internet]. 2017;251(6):706–13. Available from: https://doi.org/10.2460/javma.251.6.706

30. Ferguson NM, Donnelly CA, Anderson RM. The Foot-and-Mouth Epidemic in Great Britain: Pattern of Spread and Impact of Interventions. Science (80-) [Internet]. 2001;292(5519):1155–60. Available from: http://www.sciencemag.org/content/292/5519/1155.abstract

31. Kiss IZ, Green DM, Kao RR. The network of sheep movements within Great Britain: network properties and their implications for infectious disease spread. J R Soc Interface [Internet]. 2006 Oct 22;3(10):669 LP-677. Available from: http://rsif.royalsocietypublishing.org/content/3/10/669.abstract

32. Buhnerkempe MG, Grear DA, Portacci K, Miller RS, Lombard JE, Webb CT. A national-scale picture of U.S. cattle movements obtained from Interstate Certificate of Veterinary Inspection data. Prev Vet Med [Internet]. 2013;112(3):318–29. Available from: http://www.sciencedirect.com/science/article/pii/S0167587713002493

33. Buhnerkempe MG, Tildesley MJ, Lindström T, Grear DA, Portacci K, Miller RS, et al. The Impact of Movements and Animal Density on Continental Scale Cattle Disease Outbreaks in the United States. PLoS One [Internet]. 2014 Mar 26;9(3):e91724. Available from: https://doi.org/10.1371/journal.pone.0091724

34. Brooks-Pollock E, de Jong MCM, Keeling MJ, Klinkenberg D, Wood JLN. Eight challenges in modelling infectious livestock diseases. Epidemics [Internet]. 2015;10:1–5. Available from: http://www.sciencedirect.com/science/article/pii/S1755436514000401

35. Tildesley MJ, Ryan SJ. Disease Prevention versus Data Privacy: Using Landcover Maps to Inform Spatial Epidemic Models. PLOS Comput Biol [Internet]. 2012 Nov 1;8(11):e1002723. Available from: https://doi.org/10.1371/journal.pcbi.1002723

36. Carslake D, Grant W, Green LE, Cave J, Greaves J, Keeling M, et al. Endemic cattle diseases: comparative epidemiology and governance. Philos Trans R Soc B Biol Sci [Internet]. 2011 Jul 12;366(1573):1975 LP-1986. Available from: http://rstb.royalsocietypublishing.org/content/366/1573/1975.abstract

37. Brooks-Pollock E, Conlan AJK, Mitchell AP, Blackwell R, McKinley TJ, Wood JLN. Age-dependent patterns of bovine tuberculosis in cattle. Vet Res [Internet]. 2013;44(1):97.

Available from: https://doi.org/10.1186/1297-9716-44-97

38. Keeling MJ, Woolhouse MEJ, Shaw DJ, Matthews L, Chase-Topping M, Haydon DT, et al. Dynamics of the 2001 UK Foot and Mouth Epidemic: Stochastic Dispersal in a Heterogeneous Landscape. Science (80-) [Internet]. 2001;294(5543):813–7. Available from: http://www.sciencemag.org/content/294/5543/813.abstract

39. Woolhouse MEJ, Shaw DJ, Matthews L, Liu W-C, Mellor DJ, Thomas MR. Epidemiological implications of the contact network structure for cattle farms and the 20–80 rule. Biol Lett [Internet]. 2005 Sep 22;1(3):350 LP-352. Available from: http://rsbl.royalsocietypublishing.org/content/1/3/350.abstract

40. Kao RR, Danon L, Green DM, Kiss IZ. Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. Proc R Soc B Biol Sci [Internet]. 2006 Aug 22;273(1597):1999 LP-2007. Available from: http://rspb.royalsocietypublishing.org/content/273/1597/1999.abstract

41. Boender GJ, Hagenaars TJ, Bouma A, Nodelijk G, Elbers ARW, de Jong MCM, et al. Risk Maps for the Spread of Highly Pathogenic Avian Influenza in Poultry. PLOS Comput Biol [Internet]. 2007 Apr 20;3(4):e71. Available from: https://doi.org/10.1371/journal.pcbi.0030071

42. Backer JA, Hagenaars TJ, van Roermund HJW, de Jong MCM. Modelling the effectiveness and risks of vaccination strategies to control classical swine fever epidemics. J R Soc Interface [Internet]. 2008 Jan 1; Available from: http://rsif.royalsocietypublishing.org/content/early/2009/02/09/rsif.2008.0408.abstract

43. Volkov I, Pepin KM, Lloyd-Smith JO, Banavar JR, Grenfell BT. Synthesizing within-host and population-level selective pressures on viral populations: the impact of adaptive immunity on viral immune escape. J R Soc Interface. 2010;7(50):1311–8.

44. Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc R Soc B Biol Sci [Internet]. 2012;279(1728):444–50. Available from: http://rspb.royalsocietypublishing.org/content/279/1728/444.abstract

45. Heinen PP, de Boer-Luijtze EA, Bianchi ATJ. Respiratory and systemic humoral and cellular immune responses of pigs to a heterosubtypic influenza A virus infection. Vol. 82, Journal of General Virology. 2001. 2697-2707 p.

46. Reeth K Van, Brown I, Essen S, Pensaert M. Genetic relationships, serological cross-reaction and cross-protection between H1N2 and other influenza A virus subtypes endemic in European pigs. Virus Res [Internet]. 2004;103(1):115–24. Available from: http://www.sciencedirect.com/science/article/pii/S0168170204001224

47. Van Reeth K, Braeckmans D, Cox E, Van Borm S, van den Berg T, Goddeeris B, et al. Prior infection with an H1N1 swine influenza virus partially protects pigs against a low pathogenic H5N1 avian influenza virus. Vaccine [Internet]. 2009;27(45):6330–9. Available from: http://www.sciencedirect.com/science/article/pii/S0264410X09004046

48. Busquets N, Segalés J, Córdoba L, Mussá T, Crisci E, Martín-Valls GE, et al. Experimental infection with H1N1 European swine influenza virus protects pigs from an infection with the 2009 pandemic H1N1 human influenza virus. Vet Res [Internet]. 2010 Sep;41(5). Available from: https://doi.org/10.1051/vetres/2010046

49. Qiu Y, De hert K, Van Reeth K. Cross-protection against European swine influenza viruses in the context of infection immunity against the 2009 pandemic H1N1 virus: studies in the pig model of influenza. Vet Res [Internet]. 2015;46(1):105. Available from: https://doi.org/10.1186/s13567-015-0236-6

50. Bowman AS, Walia RR, Nolting JM, Vincent AL, Killian ML, Zentkovich MM, et al. Influenza A(H3N2) Virus in Swine at Agricultural Fairs and Transmission to Humans, Michigan and Ohio, USA, 2016. Emerg Infect Dis J [Internet]. 2017;23(9):1551. Available from: http://wwwnc.cdc.gov/eid/article/23/9/17-0847

51. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, et al. The Influenza Virus Resource at the National Center for Biotechnology Information . J Virol [Internet]. 2008 Jan 17;82(2):596–601. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2224563/

52. Sents N. World Pork Expo 2018 Highlights. Successful Farming [Internet]. 2018; Available from: https://www.agriculture.com/news/world-pork-expo-2018-highlights

53. Romagosa A, Allerson M, Gramer M, Joo HS, Deen J, Detmer S, et al. Vaccination of influenza a virus decreases transmission rates in pigs. Vet Res [Internet]. 2011 Dec 20;42(1):120. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3258204/

54. Brookes SM, Núñez A, Choudhury B, Matrosovich M, Essen SC, Clifford D, et al. Replication, Pathogenesis and Transmission of Pandemic (H1N1) 2009 Virus in Non-Immune Pigs. PLoS One [Internet]. 2010 Feb 5;5(2):e9068. Available from: https://doi.org/10.1371/journal.pone.0009068

55. Skowronski DM, Moser FS, Janjua NZ, Davoudi B, English KM, Purych D, et al. H3N2v and Other Influenza Epidemic Risk Based on Age-Specific Estimates of Sero-Protection and Contact Network Interactions. McVernon J, editor. PLoS One [Internet]. 2013 Jan 11;8(1):e54015. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3543419/

56. Dorjee S, Poljak Z, Revie CW, Bridgland J, McNab B, Leger E, et al. A Review of Simulation Modelling Approaches Used for the Spread of Zoonotic Influenza Viruses in

Animal and Human Populations. Zoonoses Public Health [Internet]. 2013 Sep 1;60(6):383–411. Available from: http://dx.doi.org/10.1111/zph.12010

57. Reynolds JJH, Torremorell M, Craft ME. Mathematical Modeling of Influenza A Virus Dynamics within Swine Farms and the Effects of Vaccination. PLoS One [Internet]. 2014 Aug 27;9(8):e106177. Available from: https://doi.org/10.1371/journal.pone.0106177

58. Pitzer VE, Aguas R, Riley S, Loeffen WLA, Wood JLN, Grenfell BT. High turnover drives prolonged persistence of influenza in managed pig herds. J R Soc Interface [Internet]. 2016 Jun 29;13(119). Available from: http://rsif.royalsocietypublishing.org/content/13/119/20160138.abstract

59. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2018.