

Nine challenges in modelling the emergence of novel pathogens

James O. Lloyd-Smith^{1,2,*}, Sebastian Funk³, Angela R. McLean⁴, Steven Riley^{2,5},
James L. N. Wood^{2,6}

1. Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA, USA.
2. Fogarty International Center, National Institutes of Health, Bethesda, MD, USA.
3. Center for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom
4. Department of Zoology, Oxford Martin School, University of Oxford, Oxford, United Kingdom
5. MRC Centre for Outbreak Analysis and Disease Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, United Kingdom
6. Disease Dynamics Unit, Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom

* Corresponding author; jlloydsmith@ucla.edu

Abstract

Studying the emergence of novel infectious agents involves many processes spanning host species, spatial scales, and scientific disciplines. Mathematical models play an essential role in combining insights from these investigations and drawing robust inferences from field and experimental data. We describe nine challenges in modelling the emergence of novel pathogens, emphasizing the interface between models and data.

Introduction

While humankind continues to battle ancient adversaries such as tuberculosis and malaria, there is constant concern about the emergence of new human pathogens from sources in non-human animals (1). At the very least, this concern is justified by devastating pandemic emergences of HIV-1, HIV-2, and Spanish influenza. We have also seen the near-establishment of SARS-Coronavirus, and a relentless series of zoonotic threats competing for our attention and public health resources. At the time of writing, influenza A H7N9 in China (2) and MERS-Coronavirus in the Saudi Arabian peninsula (3) are both causing substantial numbers of cases, and deaths, and health authorities are searching for effective responses.

This article focuses on challenges in modelling the emergence of pathogens that are novel to human hosts, such as MERS-CoV or zoonotic influenza strains. We consider problems at the interface of models and data that pertain to interpreting patterns in observed outbreaks, and contributing to rational and robust assessment of risks posed by putative emerging pathogens. We assume that candidate zoonotic pathogens are circulating in some non-human reservoir population or populations, from which they can spill over to infect humans. Humans infected

directly by animals are known as spillover or primary cases. If human-to-human transmission occurs, then subsequent cases infected by humans are termed non-primary.

In assessing pathogen emergence, it is useful to delineate what is known about a pathogen's ability to spread between humans. A crucial distinction exists between pathogens that are capable of sustained human-to-human transmission in some settings (i.e. $R_0 > 1$ in humans), and those that exhibit inefficient spread, with subcritical dynamics (i.e. $0 < R_0 < 1$). This latter group includes many pathogens viewed as significant future threats, such as influenza A H5N1, influenza A H7N9, MERS-CoV and monkeypox virus. Another group includes pathogens detected by 'prospecting' in various non-human animal populations, including many that are previously unknown to science (e.g. 4), the relevance of which is often unknown.

Challenge 1. Better capture the disease dynamics in proximal non-human species

One can imagine two extreme conceptual models for the dynamics of emergence from non-human hosts into humans. In 'static reservoir emergence', the dynamics of the pathogen in the reservoir do not change from their long-term pattern. Because of chance or some change in human behaviour, the pathogen spills over from this static reservoir system to cause human infection. In 'dynamic reservoir emergence', the ecology of the pathogen in its non-human hosts changes substantially prior to emergence in humans; changes could include transmission into domestic animals, or gains in transmissibility due to evolutionary changes in the pathogen. However, while the conceptual differentiation between static and dynamic reservoir emergence is attractive, key case studies point much more towards dynamic emergence. For example, Nipah virus caused outbreaks in pigs prior to infecting humans (5) and outbreaks of Sin Nombre virus infection (including the first identified outbreak) have been linked to elevated rodent population densities following periods of increased rainfall (6).

Current assessments of emergence risks from novel pathogens focus heavily on the frequency of particular pathogen genotypes (7) or predicted (static) distributions of reservoir species (8), and do not include dynamic factors in reservoir ecology. Therefore, an important, broad challenge is to use models in conjunction with available data to help detect and characterise potentially dangerous changes in the ecology of infectious diseases in key wildlife or livestock reservoirs.

Challenge 2. Expand models for cross-species spillover transmission from general principles to specific, mechanistic frameworks integrating all relevant data types.

Characterization of the spillover force of infection is crucial to emergence dynamics. Very general frameworks have been advanced, for instance to decompose the spillover force of infection into (9):

Spillover FOI = prevalence in reservoir \times reservoir-human contact rate \times P(infection | contact). We need a new generation of approaches that take advantage of broader developments in infectious disease dynamics and epidemiology. For instance, ecologic, economic or environmental factors giving rise to interactions among the three terms should be considered, and their dynamical consequences explored. Constructing more mechanistic models of spillover transmission will raise specific challenges, but may also present new solutions. For

instance, when human infection occurs via environmental reservoirs, or through food, it may be possible to integrate many complexities of reservoir ecology into their impact on environmental burden, then use dose-response relationships to understand risk to humans. Otherwise, there can be many challenges associated with finding a relevant characterization of prevalence in the reservoir, particularly when the system involves multiple host species and multiple pathogen strains, each possibly posing different risks to humans. Transmission dynamic models that incorporate data from sequence-based or niche modeling approaches may help to predict spillover risk more generally.

Epidemiology has well-developed frameworks for risk factor analysis, which can be applied to spillover because primary cases can be viewed as independent outcomes (at least approximately). Thus there are opportunities to integrate a biostatistical approach to primary cases with a stochastic model of subsequent transmission, creating a joint inference framework. For example, primary infection with Nipah virus in Bangladesh is associated with drinking date-palm sap, while on-going transmission is associated with close contacts among humans (10). A joint framework that links these co-factors using a mechanistic model may aid in distinguishing between primary and non-primary cases (see Challenge 4). Studies of age-based mixing patterns have shed light on transmission dynamics of endemic pathogens (11); there could be similar benefits to linking spillover risk factor information to data on mixing patterns in relevant human populations. Analogously, the coupling between spatial distribution of spillover risk and spatial factors influencing human-to-human transmission may govern the risk of a major outbreak (for instance, risk will be lower if spillover occurs chiefly in remote settlements than if it happens in crowded urban areas).

Challenge 3. Harness pathogen genetic data across the human-animal interface to map transmission and detect adaptation.

Pathogen sequence data could shed light on central questions in zoonotic emergence, by reconstructing transmission connections or looking for adaptation in a new host. However, numerous challenges persist. Because of historic interdisciplinary divisions, isolates from animals and humans have often been grown, detected or analyzed using different approaches, which effectively precludes useful inference. Pathogen isolates are often rare, particularly in difficult-to-culture genera, so isolates from linked cases are unusual. Animal sources of human spillover cases are often gone (dead, eaten, or moved away) by the time the human cases are detected and investigated, so that sequences come from other animal individuals that may not be closely linked. Any inferences about pathogen evolution must include uncertainty arising from (typically poorly-known) transmission and evolutionary processes in animal hosts. This situation is particularly challenging if multiple species are involved in circulation of the pathogen, as for avian influenza. Many current examples are based on coarse sampling and use appropriately coarse analyses, such as phylogenies, but higher-resolution methods (preferably not sensitive to missing samples) will be needed as isolate detection improves. These methods will need to explicitly link transmission mechanisms to sequence evolution.

Challenges also arise when trying to assess whether evolutionary adaptation played a role in a past emergence event, and when any adaptive mutations occurred (12). Often, there is no

baseline surveillance prior to emergence, so ancestral genotypes cannot be assessed, or available samples are separated by substantial gaps. There is typically poor information about pathogen diversity in animal hosts, let alone in individual animals.

Challenge 4. Improve methods to analyze stochastic dynamics after pathogen introduction, accounting for heterogeneities and imperfect observation

Substantial progress has been made on modeling the stochastic dynamics of early generations of transmission after a novel pathogen is introduced to a population, yet major challenges remain.

Particular challenges arise from heterogeneities (typically uncharacterized) in host contact patterns, host susceptibility and infectiousness, environmental factors, and possibly pathogen phenotypes. Which of these matter for a given outbreak, and how can this be determined? Further challenges arise from host population structure at scales from households to cities, and the resulting possibility that local pools of susceptibles will be depleted. These effects are often neglected for emerging pathogens, with the rationale that many hosts are available, but this assumption fails easily (13). Non-stationary dynamics are also challenging: any changes through time arising from control measures or behaviour change will be entangled with non-stationarities driven by contact network effects (whereby the most connected individuals are infected early) or superspreading effects (whereby transmission rates revert to the mean after outbreaks are kicked off by superspreaders).

These factors combine to shape the dynamics of early transmission chains, and the resulting data form the basis for inference about outbreaks of emerging infections. Outbreaks are often small, so data are commonly pooled from multiple introduction events at different times and locations, introducing further heterogeneities. Such 'mixed distributions' have been used to estimate reproductive numbers for post-elimination measles, for instance, though it is recognized that the largest outbreaks occur in populations with unusually low vaccination rates, where different parameters would apply (14). Furthermore, imperfect observation causes cases and outbreaks to be missing from data sets, likely in non-random ways. We need models that can account for these problems, and ideally correct for them, to enable robust inference of parameters of interest.

A particular set of challenges arises for subcritical pathogens, whose epidemiology is characterized by a mix of spillover events and self-limited chains of human-to-human transmission (so-called 'stuttering chains'). A primary and pervasive challenge is to disentangle contributions from these two sources. There has been recent progress in methods to estimate R_0 based on the distribution of chain lengths (15) or ratio of primary to non-primary cases (assuming these can be distinguished)(16). These approaches have complementary strengths and weaknesses, so hybrid or alternative approaches would be valuable. Also, little attention has been paid to the joint characterization of temporal variation in spillover hazard and human-to-human transmissibility. In some scenarios it may be possible to extract the contribution of subcritical human-to-human transmission, leaving a robust description of the spillover hazard

itself – or conversely to extract the contribution of spillover and see the human-to-human transmission more clearly (17).

Challenge 5. Improve data collection and analysis to learn from the frequency of singleton (sporadic) cases.

For many zoonotic infections, particularly those with low transmissibility among humans (or high variation in transmissibility), many primary cases fail to transmit and thus appear as singleton cases. Most attention is focused on larger disease outbreaks, when transmission in the human population raises concern. Sporadic singleton cases are viewed as low priorities for surveillance. Singletons are fundamentally more difficult to detect and data on singletons are even sometimes dropped from analyses (discussed in (18)).

Data on singletons are needed to achieve correct estimates of total spillover rates, as well as fatality rates, and also to study risk factors for primary infection. They comprise an important component of the distribution of outbreak sizes, which can be used to estimate R_0 (19). The frequency of singleton cases has surprising influence when R_0 and other parameters are estimated from chain size distributions (18). If R_0 can be estimated through other means, then measuring the frequency of singleton cases allows heterogeneity in transmission to be estimated (20). An important challenge for the field is to improve reliability of singleton case data, and hence to incorporate these data into epidemiological analyses – or else to develop robust methods for parameter estimation that account for missing or biased data on singletons.

Challenge 6. Develop theory and case studies for the role of intermediate hosts in pathogen emergence.

Many recent zoonotic emergence events feature an intermediate host species, acting as a transmission bridge between the ‘true reservoir’ where the pathogen is maintained, and the human ‘target’ population. Prominent examples include Nipah virus passing from flying foxes to pigs to humans in Malaysia, Hendra virus passing from flying foxes to horses to humans, and SARS-CoV passing from fruit bats to palm civets to humans. These intermediate hosts might contribute strictly via contact, by bridging between two host populations with no direct contact, or they might have a more biological role, e.g. as an ‘amplifying host’ that can generate high pathogen titers, or by facilitating pathogen evolution that increases transmission in the human host.

There is no general framework to define, compare, or contrast these various roles for intermediate hosts in emergence events, and modeling has been *ad hoc* and system-specific. We need to define general principles and defining characteristics of the different scenarios and then to apply them to case studies. Such a model framework could also guide decisions about allocation of surveillance or control effort.

Challenge 7. Expand models for emerging infections to account for host immunity.

Most models of emerging infections assume a completely susceptible host population, but this is not valid if parts of the population have been exposed to low doses of the pathogen or to less virulent ancestors or related pathogens. Those in frequent contact with animals may have

been exposed to zoonotic pathogens such as SARS-CoV or influenza, and more elderly sub-populations might have historic exposures.

These partially immune groups can cause profound dynamic effects. Having a population fraction immune, or partially immune, can facilitate disease persistence by reducing the chance of extinction in the post-epidemic trough (21). If those most at risk of exposure to a zoonotic pathogen are those with the highest levels of immunity (because of multiple previous exposures), they could form an effective barrier preventing an infection from spreading to the rest of the population; such a pattern is seen for influenza antibodies in numerous studies of swine industry workers (22). Data from this scenario might also lead to underestimation of the reproduction number of the pathogen if it spreads into a population that is truly naive. Models could help distinguish between pathogens that fail to spread because their transmissibility is low in all humans, versus those that fail because of low transmissibility in the human population in contact with the reservoir. We need to understand when these effects matter, and how to identify them.

Challenge 8. Devise approaches to measuring *infection* fatality rates

When a new infection emerges there is great concern to know its ‘case fatality rate’. This is formally defined as the number of deaths from infection divided by the number of cases – but this clear-sounding definition hides a deep source of confusion in the definition of a ‘case’. Often only severely ill cases are counted, leading to overestimation of case fatality rates. A basic tenet of infectious disease biology is that infection does not always lead to disease, and case fatality rates calculated with ‘cases of disease’ as the denominator can differ greatly from those calculated using ‘cases of infection’. Such differences have led to widely varying estimates of case fatality rates for new emerging infections, with some corresponding alarmism, e.g. for pandemic H1N1 swine flu when it first emerged in 2009.

Valuable progress has been made by focusing on case fatality rates in hospital-admitted cases or symptomatic cases only (23). However, to make sensible projections of the total expected mortality from novel infectious diseases, we need to move beyond hospital case fatality rates towards *infection* fatality rates that explicitly use the number of infections as the denominator. Estimating this quantity is challenging if many infections are asymptomatic or cause only mild illness. One possible solution is to use serology to estimate numbers infected; this was deployed in Hong Kong during the first wave of the 2009 H1N1 epidemic (24), though challenges could arise from unknown background seroprevalence and delayed availability of serological data. An alternative is to estimate total numbers of infections from analyses of pathogen sequence data, which is possible in principle during the exponential phase of an outbreak (25). Since pathogen sequences are often deposited in the public domain in real time during pandemics this might provide very early estimates of the numbers of infections, including those that are asymptomatic. Mathematical modelling could clarify what can be expected from such an approach. Given the numerous assumptions involved, would estimates of the numbers infected ever be accurate enough to make this approach useful? How many sequences are needed? Could we correct for sampling bias and heterogenous transmission? (See the challenges in phylodynamic inference, in this volume (26).) With the widespread

availability of viral sequence from the 2009 H1N1 pandemic and a 'gold standard' estimate of infection rates from published serology, that event could be a testing ground for new methods for calculating infection rates and thus infection fatality rates.

Challenge 9. Design robust and efficient approaches to empirical studies of novel pathogens aimed at risk assessment.

Massive effort is going into surveying for possible zoonotic pathogens in various wildlife and domestic animal populations. Wide application of sensitive technologies identifies many pathogens, but it is unclear how these results map onto public health risk, especially when based on detection of pathogen nucleic acid rather than pathogen isolation. Further work aims to assess risks from particular pathogens by focused laboratory or infection experiments.

In the distant future, one might be able to detect a pathogen, sequence it and then know enough to make a reasoned estimate of the risks associated with its emergence. Far too many gaps exist now for that to be feasible. However, modelling studies can help by forcing definitions of emergence in clear, quantitative frameworks, and keeping the focus on key processes. In particular, models can define what properties of novel pathogens need to be measured. Pathogen phenotypes are often assessed in animal models and modelling should be used to analyse such studies and shed greater light on their optimal design. For example, some influenza research has focused on infectiousness per unit time rather than the duration of infectiousness. Is this design sufficient to characterize risk? Another essential contribution would be developing modelling approaches to link data from experimental infections to data attainable in the field.

Models can also help to design surveillance programmes for emerging pathogens. Model-guided fieldwork (MGF) has been advanced recently as a useful tool for all of disease ecology (27), and could find useful application here. MGF approaches can help empirical scientists to focus on particular sample types (seroprevalence versus infection prevalence or incidence) and sub-populations, rather than following *ad hoc* or unfocussed data collection plans. Careful modelling can help define the best sentinel groups, how they should be surveyed, and with what sample sizes. MGF can also help to identify animal species (or groups of species) that act as disease reservoirs (a notoriously tough problem (28, 29)).

Summary

Research on emerging pathogens has highlighted the essential need to integrate insights from many disciplines, and to link processes acting at multiple scales (there are obvious connections to other articles in this issue (26, 28, 30, 31)). This drives home the need for focused modelling efforts, to link these disparate data types, explore case studies, and define priorities in data collection. There is a general pattern of serious challenges arising from missing information – and from not adequately addressing these data gaps. At the same time, expanded sampling efforts and new technologies are bringing a flood of data that must be analyzed with a focus on mechanistic principles and possible imbalances in sampling design. The long-term goal is to draw robust conclusions about past events and make appropriate assessments of risk (and uncertainty!) about pathogens that appear to be threatening.

Acknowledgements

The authors are grateful to the Isaac Newton Institute for Mathematical Sciences, for providing the setting for discussions leading to this paper. We acknowledge support from the Research and Policy for Infectious Disease Dynamics (RAPIDD) program of the Science and Technology Directory, Department of Homeland Security, and Fogarty International Center, National Institutes of Health. JLS was also supported by the National Science Foundation (EF-0928987, OCE-1335657) and the De Logi Chair in Biological Sciences. SF was supported by a UK Medical Research Council Career Development Award in Biostatistics. SR was supported by: Wellcome Trust Project Award 093488/Z/10/Z; R01 TW008246-01 from Fogarty International Centre; and The Medical Research Council (UK, Project Grant MR/J008761/1). JLNW was also supported by the Alborada Trust and the European Union FP7 project ANTIGONE (contract number 278976).

References

1. K. E. Jones *et al.*, Global trends in emerging infectious diseases. *Nature* **451**, 990 (Feb, 2008).
2. Centers for Disease Control & Prevention, Emergence of Avian Influenza A (H7N9) Virus Causing Severe Human Illness-China, February-April 2013. *MMWR. Morb. Mortal. Wkly. Rep.* **62**, 366 (2013).
3. P. Penttinen *et al.*, Taking stock of the first 133 MERS coronavirus cases globally—Is the epidemic changing. *Euro Surveill* **18**, 39 (2013).
4. S. J. Anthony *et al.*, A strategy to estimate unknown viral diversity in mammals. *MBio* **4**, e00598 (2013).
5. U. D. Parashar *et al.*, Case-control study of risk factors for human infection with a new zoonotic paramyxovirus, Nipah virus, during a 1998–1999 outbreak of severe encephalitis in Malaysia. *J. Infect. Dis.* **181**, 1755 (2000).
6. B. Hjelle, G. E. Glass, Outbreak of hantavirus infection in the Four Corners region of the United States in the wake of the 1997-1998 El Nino-southern oscillation. *J. Infect. Dis.* **181**, 1569 (May, 2000).
7. C. A. Russell *et al.*, The potential for respiratory droplet–transmissible A/H5N1 influenza virus to evolve in a mammalian host. *Science* **336**, 1541 (2012).
8. T. L. Fuller *et al.*, Predicting hotspots for influenza virus reassortment. *Emerg. Infect. Dis.* **19**, 581 (2013).
9. J. O. Lloyd-Smith *et al.*, Epidemic dynamics at the human-animal interface. *Science* **326**, 1362 (Dec 4, 2009).
10. E. S. Gurley *et al.*, Person-to-person transmission of Nipah virus in a Bangladeshi community. *Emerg. Infect. Dis.* **13**, 1031 (Jul, 2007).
11. J. Mossong *et al.*, Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine* **5**, e74 (2008).
12. K. M. Pepin, S. Lass, J. R. C. Pulliam, A. F. Read, J. O. Lloyd-Smith, Identifying genetic markers of adaptation for surveillance of viral host jumps. *Nature Reviews Microbiology* **8**, 802 (Nov, 2010).

13. P. C. Cross, P. L. F. Johnson, J. O. Lloyd-Smith, W. M. Getz, Utility of R_0 as a predictor of disease invasion in structured populations. *Journal of the Royal Society Interface* **4**, 315 (Apr, 2007).
14. A. King, P. Varughese, G. De Serres, G. A. Tipples, J. Waters, Measles elimination in Canada. *J. Infect. Dis.* **189**, S236 (May 1, 2004).
15. S. Blumberg, J. O. Lloyd-Smith, Inference of R_0 and transmission heterogeneity from the size distribution of stuttering chains. *PLoS Computational Biology* **9**, e1002993 (2013).
16. S. Cauchemez *et al.*, 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 Medicine* **10**, e1001399 (2013).
17. A. Kucharski *et al.*, Distinguishing Between Reservoir Exposure and Human-to-Human Transmission for Emerging Pathogens Using Case Onset Data. *PLOS Currents Outbreaks* **1**, (2014).
18. S. Blumberg, J. Lloyd-Smith, Comparing methods for estimating R_0 from the size distribution of subcritical transmission chains. *Epidemics* **5**, 131 (2013).
19. C. P. Farrington, M. N. Kanaan, N. J. Gay, Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics* **4**, 279 (Apr, 2003).
20. J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355 (2005).
21. J. R. Pulliam *et al.*, Agricultural intensification, priming for persistence and the emergence of Nipah virus: a lethal bat-borne zoonosis. *Journal of the Royal Society Interface* **9**, 89 (2012).
22. K. P. Myers *et al.*, Are swine workers in the United States at increased risk of infection with zoonotic influenza virus? *Clin. Infect. Dis.* **42**, 14 (2006).
23. H. Yu *et al.*, Human infection with avian influenza A H7N9 virus: an assessment of clinical severity. *The Lancet* **382**, 138 (2013).
24. S. Riley *et al.*, Epidemiological characteristics of 2009 (H1N1) pandemic influenza based on paired sera from a longitudinal community cohort study. *Plos Medicine* **8**, e1000442 (2011).
25. S. D. Frost, E. M. Volz, Viral phylodynamics and the search for an 'effective number of infections'. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**, 1879 (2010).
26. S. D. Frost *et al.*, Eight Challenges in Phylodynamic Inference. *Epidemics* **In press**, (2014).
27. O. Restif *et al.*, Model-guided fieldwork: practical guidelines for multidisciplinary research on wildlife ecological and epidemiological dynamics. *Ecol. Letters* **15**, 1083 (2012).
28. M. G. Buhnerkempe *et al.*, Eight challenges in modelling disease systems with multiple host species and multiple infectious agents. *Epidemics* **In press**, (2014).
29. M. Viana *et al.*, Assembling evidence for identifying reservoirs of infection. *Trends Ecol. Evol.* **29**, 270 (2014).
30. J. R. Gog *et al.*, Seven challenges in modelling pathogen dynamics within-host and across scales. *Epidemics* **In press**, (2014).

31. P. S. Wikramaratna *et al.*, Five challenges in modelling interacting strain dynamics. *Epidemics* **In press**, (2014).