Leveraging pathogen community distributions to understand outbreak and emergence potential

Tad A. Dallas[a,†], Colin J. Carlson[b,c,†] and Timothée Poisot[d]

[a] *Centre for Ecological Change, University of Helsinki, 00840 Helsinki, Finland*
[b] *National Socio-Environmental Synthesis Center, University of Maryland, Annapolis, Maryland 21401, USA.*
[c] *Department of Biology, Georgetown University, Washington, D.C. 20057, USA.*
[d] *Dépt de Sciences Biologiques, Univ. de Montréal, Montréal, Canada*
[†] *These authors contributed equally to this study.*

*Corresponding author: tad.a.dallas@gmail.com

Running title: Predicting pathogen emergence

Author contributions: TD, CJC, and TP conceived of the idea for the study. TD and TP designed the model. All authors contributed to the writing of the manuscript.

Data accessibility: `R` code is available on figshare at `https://doi.org/10.6084/m9.figshare.6364955`.

Ethics: This study used existing data on pathogen outbreak and emergence events in human populations.

Funding: This work was supported by the National Socio-Environmental Synthesis Center (SESYNC) under funding received from the National Science Foundation DBI-1639145.

Competing interests: The authors declare no competing interests.

Keywords: Emerging infectious disease, community ecology, community dissimilarity, disease forecasting, pathogen biogeography

Abstract

Understanding pathogen outbreak and emergence events has important implications to the management of infectious disease. Apart from preempting infectious disease events, there is considerable interest in determining why certain pathogens are consistently found in some regions, and why others spontaneously emerge or re-emerge over time. Here, we use a trait-free approach which leverages information on the global community of human infectious diseases to estimate the potential for pathogen outbreak, emergence, and re-emergence events over time. Our approach uses pairwise dissimilarities among pathogen distributions between countries and country-level pathogen composition to quantify pathogen outbreak, emergence, and re-emergence potential as a function of time (e.g., number of years between training and prediction), pathogen type (e.g., virus), and transmission mode (e.g., vector-borne). We find that while outbreak and re-emergence potential are well captured by our simple model, prediction of emergence events remains elusive, and sudden global emergences like an influenza pandemic seem beyond the predictive capacity of the model. While our approach allows for dynamic predictability of outbreak and re-emergence events, data deficiencies and the stochastic nature of emergence events may preclude accurate prediction; but our results make a compelling case for incorporating a community ecology perspective into existing disease forecasting efforts.

Introduction

The emergence of infectious diseases in humans and wildlife is a continuous and natural process that is nevertheless rapidly intensifying with global change (Jones *et al.*, 2008). Around the world, the diversity, and frequency, of infectious outbreaks is rising over time (Smith *et al.*, 2014; Jones *et al.*, 2008), and the vast majority of pathogens with zoonotic potential still have yet to emerge in human populations, with an estimated 600,000 minimum viruses with zoonotic potential (Carroll *et al.*, 2018). Intensifying pathways of contact between wildlife reservoirs and humans, and rapid spread of new pathogens among human populations around the globe, are considered major drivers in this accelerating process (Cleaveland *et al.*, 2007; Tatem *et al.*, 2006). Changes in climate and land-use, as well as food insecurity and geopolitical conflict, are expected to exacerbate feedbacks between socio-ecological change and emerging infectious diseases (EIDs). In the face of these threats, the anticipation of disease emergence events is a seminal but elusive challenge for public health research (Morse *et al.*, 2012).

One forecasting approach recognizes that the drivers of emergence events are distributed non-randomly in space and time, and follow predictable regional patterns that inherently predispose some areas to a higher burden of EIDs (Allen *et al.*, 2017). Different classes of emerging pathogens (e.g., new pathogens versus drug-resistant strains of familiar ones; vector-borne and/or zoonotically transmitted diseases) follow different spatial risk patterns at a global scale (Jones *et al.*, 2008). In part, this can be explained by the non-random distribution of host groups that disproportionately contribute to zoonotic emergence events, like bats and rodents (Johnson *et al.*, 2015a; Olival *et al.*, 2017), and are likely to continue to do so (Han *et al.*, 2016a,b, 2015). However, additional factors are strongly associated with the distribution of emerging infection risk; notably human population density, land cover, and land use change (Allen *et al.*, 2017). In addition to these

factors, deterministic emergence of disease is influenced by social, cultural, and economic factors (Bonds *et al.*, 2010; Farmer, 1996; McMichael, 2004; Murray & Schaller, 2010; Parkes *et al.*, 2005).

As a consequence of this heterogeneity in host distributions and other contributing factors, emerging pathogens may follow Tobler's First Law ("near things are more related than distant things"; Tobler (1970)), and fall into a handful of global biogeographic regions with similar pathogen communities (Murray *et al.*, 2015). However, with increasing global connectivity, both pathogens and the free-living organisms that host them are spreading around the world at an accelerating rate, and consequently the spatial structure of pathogen diversity is becoming less pronounced. One study examining a global pathogen-country network showed that modularity is decreasing while connectance is increasing over time: pathogen ranges are on average expanding, and over time, geographically-separate regions are facing more threats (Smith *et al.*, 2007; Poisot *et al.*, 2014). This process of biotic homogenization has critical implications for public health, as known diseases can become unfamiliar problems in novel locations, or can re-emerge in landscapes from which they were previously eradicated.

Leveraging disease ecology in global health settings requires models that consider disease emergence as a long-term process over space and time, extending beyond initial spillover events. Work that models the impact of human mobility networks has arisen out of the pandemic influenza literature (Balcan *et al.*, 2009; Russell *et al.*, 2008; Khan *et al.*, 2009), and has recently been successful in developing a multi-scale approach to anticipating emergence risk for hemorrhagic viruses in Africa (Pigott *et al.*, 2017). However, conceptually-similar work that models across global pathogen species is mostly unexplored. Murray *et al.* (2015) suggest that countries who share pathogens might be more likely targets during a given pathogen outbreak, but this approach does not leverage information on

the identity of the shared pathogens. Given the inherent need in estimating outbreak potential, and the current availability of data on outbreak events, the need to leverage existing data for the dynamics prediction of outbreak potential is a pressing research need.

Here, we examine the predictability of pathogen biogeography over time using a similarity-based approach that utilizes data on all pathogen outbreaks in all countries, but does not require information on pathogen traits or spatial structure. In the process of modeling outbreak predictability, we test a basic but important hypothesis: do recurring outbreaks have a more predictable signal than emergence events (and, implicitly, are emergence events predictable)? Within emergences, we further note the subtle difference between emergence and re-emergence, and hypothesize the factors driving these might be subtly different. While both may be driven by genetic shifts in pathogens or changing land use patterns enhancing transmission risk, re-emergence events are more likely to be related to weakened healthcare infrastructure, prematurely-terminated eradication campaigns (Chiappini *et al.*, 2013; Minor, 2004), or low detection long-term persistence of environmental pathogen reservoirs (e.g., anthrax spores in the soil; Carlson *et al.* (2018)).

Finally, we examine whether pathogens show any differences in predictability based on agent, class, or transmission mode. Diseases of zoonotic origin (i.e. with animal hosts) and with vector-borne transmission might be harder to predict due to hidden constraints on their distribution and more complicated outbreak dynamics than directly-transmitted pathogens have. On the other hand, commonalities between species that share vectors or reservoir hosts might lead to similarities in distributions (a common notion in pathogen biogeography, as in how dengue models were frequently used in the early days of the Zika pandemic, given the shared vector *Aedes aegypti*; Bogoch *et al.* (2016); Carlson *et al.* (2016)). In this case, community-based prediction could be more powerful for zoonotic and vector-borne

diseases. Differential frequency of zoonotic and vector-borne transmission might also make different pathogen classes (viruses, bacteria, fungi, and macroparasites) more or less predictable, as might different dispersal ability on a global scale, with respiratory viruses usually presumed to spread the fastest, and macroparasites generally treated as the most dispersal-limited. Understanding how the role of community structure changes for these different pathogens can help contextualize the method we use, and understand how it might be built upon to account for these differences.

Methods

Pathogen emergence data

Data from the Global Infectious Diseases and Epidemiology Network (GIDEON) contains pathogen outbreak information at the country level obtained from case reports, governmental agencies, and published literature records (Berger, 2005; Yu & Edberg, 2005). Records with multiple etiological agents (e.g., "*Aeromonas* and marine *Vibrio* infx.") and unresolved to agent level (e.g., "Respiratory viruses - miscellaneous") were excluded from the model. In a handful of cases, we kept divisions between clinical presentations from the same pathogens, like cutaneous versus visceral leishmaniasis. The data obtained were yearly records between 1990 and 2016, and consisted of pathogen outbreak and emergence events for 234 pathogens across 224 countries. While there are some data for pathogen events between 1980 and 1990, the number of pathogen events reported was fewer than from 1990 onward, suggesting some potential reporting or sampling bias in these earlier years. Therefore, we restrict our analyses to pathogen occurrences after 1990. Based on supplemental data from Smith *et al.* (2007) and updated with recent literature given several misclassifications, each was manually classified as a bacterial,

viral, fungal, protozoan, or macroparasitic disease, and as vector-borne and/or zoonotic or neither. In some rare cases, these were left as unknown; for example, Oropouche virus is vector-borne but its sylvatic cycle remains uncertain, while the environmental origin of Bas-Congo virus is altogether unknown.

While much can be gained by leveraging data on multiple pathogens to predict outbreak or emergence potential, there are some drawbacks. The most pronounced is that pandemic events may strongly influence model predictions, such that a pandemic of one pathogen will decrease model performance when attempting to predict outbreak or emergence potential of other pathogens. We explore this further in the supplement, where we see the inclusion of influenza and the corresponding 2009 flu pandemic noticeably affects our model performance. As such, we remove influenza from the main text analyses, and place analyses containing flu in the supplement for comparison.

We distinguish between three different types of pathogen events; outbreak, re-emergence, and emergence. Outbreaks are pathogen events are recurrent pathogen events, quantified as having occurred in a given country within three years of a given year. Re-emergence events are those that did not occur within three years, but have occurred at some time in a given country in the past (a cutoff we chose inspired by World Health Organization guidelines for certifying regional eradication of poliovirus or dracunculiasis). Lastly, emergence events were considered as the first record of a pathogen within a country.

Model structure

We developed a dissimilarity-based approach to forecast pathogen outbreak and emergence events that does not require country-level or pathogen traits data. Applying tools from community ecology, we calculated mean pairwise dissimilarity (Bray-Curtis index, $\overline{BC}$) values for countries (how dissimilar are the pathogen

communities between countries) and pathogens (how dissimilar are the geographic distributions of pathogens). For a given pair of countries a, b with P_a and P_b pathogens each, and S shared pathogens among those, the Bray-Curtis index is given as:

$$BC_{a,b} = 1 - \frac{2S}{P_a P_b} \tag{1}$$

This can be treated as a measure of dissimilarity between different countries' pathogen communities. We then considered the potential for a pathogen to be found in a country proportional to the product of these dissimilarity values. We also included year as a covariate, resulting in a set of four variables for model training.

Using these data, we applied a statistical approach previously used for species distribution modeling (Drake & Richards, 2017) and link prediction in ecological networks (Dallas *et al.*, 2017a) called `plug-and-play` (PNP). This approach utilizes information on pathogen occurrence events, and also on background interactions — country-pathogen pairs which did not have a recorded outbreak — to estimate the suitability of a country for pathogen emergence from a particular pathogen (Figure 1). These suitability values can then be used to quantify model performance on data not used to train the model. Model performance was quantified using Area Under the Curve (AUC), which captures the ability of the classifier to rank positive instances higher than negative instances.

Assessing model performance

We used the PNP modeling approach to address the possibility of predicting pathogen outbreak and emergence events, specifically examining three different potential scenarios. First, we examined how the inclusion of pathogen events from previous years influenced model accuracy. That is, we predicted pathogen events

of 2016 using data starting at 2015 and then including additional years until 1995. This was performed to determine the amount of data necessary to make accurate forecasts. Second, we examined how predictive accuracy was maintained as we attempted to predict both past (hindcast) and future (forecast) pathogen events. To do this, we trained models on a ten year period (either 2005-2015 for hindcasting, or 1990-2000 for forecasting), and used these models to predict pathogen events between 1990 and 2004 for hindcasting, and between 2001 and 2015 for forecasting. Lastly, we examined how the accuracy of predictions might have changed over time. Given increased surveillance in more recent years, predictive accuracy might be dependent on the time period at which models are trained and predictions made. To test this, we trained models along a rolling window of 4 years from 1990-2015, using these models to predict pathogen events in the year following the final year of model training (e.g., a model trained on 1990-1994 would be used to predict pathogen events in 1995).

Results

We find that our dissimilarity-based model can predict outbreak events accurately, re-emergence events slightly less accurately, and emergence events only slightly better than random. This makes intuitive sense, as outbreak events occur repeatedly, providing not only ample data for model training, but also a clear tendency of a pathogen to occur in a country. That is, if the model is allowed to see 5 years of data, and the country has an outbreak of a particular pathogen in 4 of the 5 years, a naive model would predict that an outbreak will likely occur with an 80% probability. Meanwhile, emergence events are determined by many unique drivers (Allen *et al.*, 2017), which may not be consistent across any two given emergence events, and which we evidently lack sufficient data to predict using our method.

While our model allows for dynamic predictability of outbreak and re-emergence events, data deficiencies and the stochastic nature of emergence events may thus preclude accurate prediction.

Our predictive model was sensitive to the number of training years (Figure 2), with accuracy plateauing around 5-10 years of training data; however, models also just trained on a single year (the temporally closest community matrix) seemed to perform disproportionately well, which would make sense if the community changes in a Markov-like process. We further examined the limits of predictability in terms of both hindcasting and forecasting pathogen outbreak and emergence events by training the model on a known period of 10 years, and then either forecasting or hindcasting t years into the past or future (Figure 3). Interestingly, our accuracy – measured as area under the receiver operating characteristic – did not decline at the same rate when hindcasting and forecasting. That is, model accuracy was higher when hindcasting relative to the accuracy of forecasts of the same duration of time away from the training data (Figure 3). This perhaps indicates that as the country-pathogen network becomes asymptotically more connected and stable (Poisot *et al.*, 2014), the network accumulates information content, reducing the time sensitivity of hindcasting performance.

Examining a rolling window of t years ($t = 4$ years) over the last two decades, we failed to detect evidence that the enhanced reporting and surveillance in more recent years influenced our model's predictive ability (Figure 4). This also suggests that even though there were annual variations in the sample size of both pathogens and countries, there was still consistency in the structure of the country-pathogen interaction matrix over time. We explore the sensitivity of this finding to the size of the rolling window in the Supplemental Materials.

Differences in PNP model accuracy among pathogen types existed when examining the effect of the amount of data used for model training (Figure 2), with

viruses having lower accuracy relative to bacteria, fungi, or other parasites. The simplest explanation for this is that accuracy is sensitive to the number of events. However, the average number of viral occurrences over time ($\bar{x} = 179$) was only slightly less than the average number of bacterial ($\bar{x} = 185$) occurrences, and far greater than the average number of fungal ($\bar{x} = 10$) or macroparasite ($\bar{x} = 17.7$) or protozoan parasite ($\bar{x} = 22.5$) occurrence events. The average number of pathogen occurrences over time is qualitatively proportional to the number of unique viruses ($n = 83$), bacteria ($n = 81$), fungi ($n = 14$), macroparasites ($n = 38$), and protozoans ($n = 15$) we examined. Interestingly, differences among pathogen types were not found when examining the ability of the modeling approach to hindcast/forecast (Figure 3) or when examining predictive accuracy along a rolling window (Figure 4).

For our 2016 explanatory PNP model, differentiating pathogens based on zoonotic and vector-borne transmission modes suggested that both classes of pathogens were more difficult to forecast (Figure 2). Though we suspected data imbalance might drive this pattern, this seems unlikely: the majority of pathogens (144 of 228) were zoonotic, and many (59 of 233) were vector-borne. A more compelling explanation is that this year was an anomalous result; transmission mode did not influence accuracy when hindcasting/forecasting (Figure S5) or when models were trained along a rolling window (Figure 4), though there was notable year-to-year variation in the latter.

Discussion

Community ecology and biogeography have a history as deeply linked fields, and both play an increasingly significant role in emerging infectious disease research. (Johnson *et al.*, 2015b; Murray *et al.*, 2015; Stephens *et al.*, 2016) However, research

connecting the two for global pathogen diversity is fairly limited so far. Our goal was to examine whether the intrinsic structure of pathogen biogeography, approached as a bipartite network, was predictable enough to enable forecasting of different outbreak types—even in the absence of any other mechanistic predictors, like transmission mode, phylogenetic data, or environmental covariates.

Despite obvious stochasticity and data limitations, the modeling approach performed well with as little as 7 to 10 years of training data, and when predicting country-pathogen network structure across large time windows. The model was able to capture pathogen outbreak and re-emergence potential well, suggesting that, at least at administrative levels, pathogen outbreak and re-emergence events are both recurrent and predictable (and that community assembly patterns are structured and predictive of outbreak potential). However, our model generally failed to forecast pathogen emergence events. This is maybe unsurprising, as predicting when and where the next major public health threat will emerge is an incredibly difficult task which remains unsolved despite having received decades of attention (Allen *et al.*, 2017; Jones *et al.*, 2008; Morse *et al.*, 2012). However, the failure of community information to help anticipate local emergences is still disappointing, especially given the proposal that biogeographic "co-zones" could be useful strategic tools for pandemic forecasting. (Murray & Schaller, 2010)

We found some indications of differences in the predictability of pathogen events as a function of pathogen type and transmission modes. In the 2016 model breakdown, bacteria were the most predictable while viruses were disproportionately unpredictable, as were zoonotic and vector-borne pathogens. Given how clearly unpredictable emergence events were, this might make intuitive sense: zoonotic pathogens make up the majority of emerging diseases (Jones *et al.*, 2008), and single-stranded RNA viruses (many vector-borne) have been responsible for many of the biggest recent emergence events (Johnson *et al.*, 2015a). However, this

pattern did not appear to hold up across all or even most years, and the factors that reduce model performance on a year-by-year basis are mostly unclear at the community level.

One contributor to interannual variation is large-scale events such as pandemics, which appeared to strongly influence prediction of the entire country-pathogen network. While pandemic spread may be predictable using detailed information on climate, human movement, and local environmental suitability (Morse *et al.*, 2012; Tizzoni *et al.*, 2012; Zhang *et al.*, 2017), our approach lacks these mechanistic predictors and is sensitive to these black swan events. This can be seen in reduced model performance during the 2009 flu pandemic, including for pathogens with no relationship to flu, although viruses and vector-borne pathogens are more severely affected (see Supplemental Materials). So while the model benefits from pathogen community data, rare and widespread events can strongly reduce model accuracy. Future work to differentially weight these stochastic events would probably improve model performance.

While this approach enhances estimation of outbreak and emergence potential for rare pathogens or poorly sampled countries, it is also worth nothing that our approach is *not* a valid standalone forecasting tool. This is in large part due to how time is used in the model: though year is a covariate, the model itself is not temporally explicit, meaning that the model can predict a certain link following on previous years, but it would be erroneous to interpret that as a forecast for a given point in time. However, the tool can be used to investigate pathogen outbreak and emergence potential under different pathogen range expansion scenarios. That is, researchers could construct artificial data which differs from empirical data slightly, and quantify the ability of the model to predict those novel events. Since the method is based on dissimilarity of countries and pathogen distributions at its core, it is possible to examine the expected outcome as pathogen distributions

become more (or less) homogeneous, or countries become more (or less) dissimilar in their pathogen communities.

Within infectious disease ecology, a disproportionate focus has emerged on the drivers and predictability of emergence events. (Allen *et al.*, 2017) Recent work offers a compelling case that community ecology might bring predictive tools to bear on that problem (Johnson *et al.*, 2015b), and modeling work suggests that community assembly data can be leveraged to better predict how pathogens spread (Murray *et al.*, 2015), the host range of emerging diseases (Dallas *et al.*, 2017b; Johnson *et al.*, 2015a), and the dynamics of diseases within an ecosystem (Parker *et al.*, 2015; Johnson *et al.*, 2013). Our results show how a simple model considering the entire pathogen community captures important global geographic variation in outbreak potential, but as a standalone tool, still struggles to predict when a pathogen will first arrive in a new region. Though this casts doubt on biogeographic tools like "co-zones" as standalone tools for surveillance or outbreak response, our study is a compelling indicator that community data could be very easily leveraged alongside other socioecological predictors to forecast disease emergence as an ecosystem process rather than a single-species one. With a Nipah virus outbreak in India and an Ebola virus outbreak in the Democratic Republic of the Congo alone both concurrent to the completion of this manuscript, the priority of prediction in emerging disease research only continues to grow.

Acknowledgements

We thank GIDEON (`https://www.gideononline.com/`) for their collection and curation of the data.

References

Allen, T., Murray, K.A., Zambrana-Torrelio, C., Morse, S.S., Rondinini, C., Di Marco, M., Breit, N., Olival, K.J. & Daszak, P. (2017). Global hotspots and correlates of emerging zoonotic diseases. *Nature Communications*, 8, 1124.

Balcan, D., Hu, H., Goncalves, B., Bajardi, P., Poletto, C., Ramasco, J.J., Paolotti, D., Perra, N., Tizzoni, M., Van den Broeck, W. *et al.* (2009). Seasonal transmission potential and activity peaks of the new influenza a (h1n1): a monte carlo likelihood analysis based on human mobility. *BMC Medicine*, 7, 45.

Berger, S.A. (2005). GIDEON: a comprehensive web-based resource for geographic medicine. *International journal of health geographics*, 4, 10.

Bogoch, I.I., Brady, O.J., Kraemer, M.U., German, M., Creatore, M.I., Kulkarni, M.A., Brownstein, J.S., Mekaru, S.R., Hay, S.I., Groot, E. *et al.* (2016). Anticipating the international spread of zika virus from brazil. *The Lancet*, 387, 335–336.

Bonds, M.H., Keenan, D.C., Rohani, P. & Sachs, J.D. (2010). Poverty trap formed by the ecology of infectious diseases. *Proceedings of the Royal Society of London B: Biological Sciences*, 277, 1185–1192.

Carlson, C.J., Dougherty, E.R. & Getz, W. (2016). An ecological assessment of the pandemic threat of zika virus. *PLoS Neglected Tropical Diseases*, 10, e0004968.

Carlson, C.J., Getz, W.M., Kausrud, K.L., Cizauskas, C.A., Blackburn, J.K., Bustos Carrillo, F.A., Colwell, R., Easterday, W.R., Ganz, H.H., Kamath, P.L., Økstad, O.A., Turner, W.C., Kolstø, A. & Stenseth, N.C. (2018). Spores and soil from six sides: interdisciplinarity and the environmental biology of anthrax (bacillus anthracis). *Biological Reviews*, 0, in press.

370 Carroll, D., Daszak, P., Wolfe, N.D., Gao, G.F., Morel, C.M., Morzaria, S., Pablos-
371 Méndez, A., Tomori, O. & Mazet, J.A. (2018). The global virome project.
372 *Science*, 359, 872–874.

373 Chiappini, E., Stival, A., Galli, L. & De Martino, M. (2013). Pertussis re-emergence
374 in the post-vaccination era. *BMC Infectious Diseases*, 13, 151.

375 Cleaveland, S., Haydon, D. & Taylor, L. (2007). Overviews of pathogen emergence:
376 which pathogens emerge, when and why? *Current Topics in Microbiology and*
377 *Immunology*, pp. 85–111.

378 Dallas, T., Huang, S., Nunn, C., Park, A.W. & Drake, J.M. (2017a). Estimating
379 parasite host range. *Proceedings of the Royal Society B: Biological Sciences*, 284,
380 20171250.

381 Dallas, T., Park, A.W. & Drake, J.M. (2017b). Predictability of helminth parasite
382 host range using information on geography, host traits and parasite community
383 structure. *Parasitology*, 144, 200–205.

384 Drake, J. & Richards, R. (2017). Estimating environmental suitability. *bioRxiv*,
385 p. 109041.

386 Farmer, P. (1996). Social inequalities and emerging infectious diseases. *Emerging*
387 *Infectious Diseases*, 2, 259.

388 Han, B.A., Kramer, A.M. & Drake, J.M. (2016a). Global patterns of zoonotic
389 disease in mammals. *Trends in Parasitology*, 32, 565–577.

390 Han, B.A., Schmidt, J.P., Alexander, L.W., Bowden, S.E., Hayman, D.T. & Drake,
391 J.M. (2016b). Undiscovered bat hosts of filoviruses. *PLoS Neglected Tropical*
392 *Diseases*, 10, e0004815.

393 Han, B.A., Schmidt, J.P., Bowden, S.E. & Drake, J.M. (2015). Rodent reservoirs
394 of future zoonotic diseases. *Proceedings of the National Academy of Sciences*,
395 112, 7039–7044.

396 Johnson, C.K., Hitchens, P.L., Evans, T.S., Goldstein, T., Thomas, K., Clements,
397 A., Joly, D.O., Wolfe, N.D., Daszak, P., Karesh, W.B. *et al.* (2015a). Spillover
398 and pandemic properties of zoonotic viruses with high host plasticity. *Scientific*
399 *Reports*, 5.

400 Johnson, P.T., De Roode, J.C. & Fenton, A. (2015b). Why infectious disease
401 research needs community ecology. *Science*, 349, 1259504.

402 Johnson, P.T., Preston, D.L., Hoverman, J.T. & LaFonte, B.E. (2013). Host and
403 parasite diversity jointly control disease risk in complex communities. *Proceed-*
404 *ings of the National Academy of Sciences*, 110, 16916–16921.

405 Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L. &
406 Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, 451,
407 990–993.

408 Khan, K., Arino, J., Hu, W., Raposo, P., Sears, J., Calderon, F., Heidebrecht, C.,
409 Macdonald, M., Liauw, J., Chan, A. *et al.* (2009). Spread of a novel influenza a
410 (h1n1) virus via global airline transportation. *New England Journal of Medicine*,
411 2009, 212–214.

412 McMichael, A.J. (2004). Environmental and social influences on emerging infec-
413 tious diseases: past, present and future. *Philosophical Transactions of the Royal*
414 *Society of London B: Biological Sciences*, 359, 1049–1058.

415 Minor, P.D. (2004). Polio eradication, cessation of vaccination and re-emergence
416 of disease. *Nature Reviews Microbiology*, 2, 473.

417 Morse, S.S., Mazet, J.A., Woolhouse, M., Parrish, C.R., Carroll, D., Karesh, W.B.,
418 Zambrana-Torrelio, C., Lipkin, W.I. & Daszak, P. (2012). Prediction and pre-
419 vention of the next pandemic zoonosis. *The Lancet*, 380, 1956–1965.

420 Murray, D.R. & Schaller, M. (2010). Historical prevalence of infectious diseases
421 within 230 geopolitical regions: A tool for investigating origins of culture. *Jour-*
422 *nal of Cross-Cultural Psychology*, 41, 99–108.

423 Murray, K.A., Preston, N., Allen, T., Zambrana-Torrelio, C., Hosseini, P.R. &
424 Daszak, P. (2015). Global biogeography of human infectious diseases. *Proceed-*
425 *ings of the National Academy of Sciences*, 112, 12746–12751.

426 Olival, K.J., Hosseini, P.R., Zambrana-Torrelio, C., Ross, N., Bogich, T.L. &
427 Daszak, P. (2017). Host and viral traits predict zoonotic spillover from mammals.
428 *Nature*, 546, 646–+.

429 Parker, I.M., Saunders, M., Bontrager, M., Weitz, A.P., Hendricks, R., Magarey,
430 R., Suiter, K. & Gilbert, G.S. (2015). Phylogenetic structure and host abundance
431 drive disease pressure in communities. *Nature*, 520, 542.

432 Parkes, M.W., Bienen, L., Breilh, J., Hsu, L.N., McDonald, M., Patz, J.A., Rosen-
433 thal, J.P., Sahani, M., Sleigh, A., Waltner-Toews, D. *et al.* (2005). All hands on
434 deck: transdisciplinary approaches to emerging infectious disease. *EcoHealth*, 2,
435 258–272.

436 Pigott, D.M., Deshpande, A., Letourneau, I., Morozoff, C., Reiner, R.C., Kraemer,
437 M.U., Brent, S.E., Bogoch, I.I., Khan, K., Biehl, M.H. *et al.* (2017). Local,
438 national, and regional viral haemorrhagic fever pandemic potential in africa: a
439 multistage analysis. *The Lancet*.

440 Poisot, T., Nunn, C. & Morand, S. (2014). Ongoing worldwide homogenization of
441 human pathogens. *bioRxiv*, p. 009977.

Russell, C.A., Jones, T.C., Barr, I.G., Cox, N.J., Garten, R.J., Gregory, V., Gust, I.D., Hampson, A.W., Hay, A.J., Hurt, A.C. *et al.* (2008). The global circulation of seasonal influenza a (h3n2) viruses. *Science*, 320, 340–346.

Smith, K.F., Goldberg, M., Rosenthal, S., Carlson, L., Chen, J., Chen, C. & Ramachandran, S. (2014). Global rise in human infectious disease outbreaks. *Journal of the Royal Society Interface*, 11, 20140950.

Smith, K.F., Sax, D.F., Gaines, S.D., Guernier, V. & Guégan, J.F. (2007). Globalization of human infectious disease. *Ecology*, 88, 1903–1910.

Stephens, P.R., Altizer, S., Smith, K.F., Alonso Aguirre, A., Brown, J.H., Budischak, S.A., Byers, J.E., Dallas, T.A., Jonathan Davies, T., Drake, J.M. *et al.* (2016). The macroecology of infectious diseases: a new perspective on global-scale drivers of pathogen distributions and impacts. *Ecology Letters*, 19, 1159–1171.

Tatem, A.J., Rogers, D.J. & Hay, S. (2006). Global transport networks and infectious disease spread. *Advances in Parasitology*, 62, 293–343.

Tizzoni, M., Bajardi, P., Poletto, C., Ramasco, J.J., Balcan, D., Gonçalves, B., Perra, N., Colizza, V. & Vespignani, A. (2012). Real-time numerical forecast of global epidemic spreading: case study of 2009 a/h1n1pdm. *BMC Medicine*, 10, 165.

Tobler, W.R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46, 234–240.

Yu, V.L. & Edberg, S.C. (2005). Global infectious diseases and epidemiology network (gideon): a world wide web-based program for diagnosis and informatics in infectious diseases. *Clinical Infectious Diseases*, 40, 123–126.

Zhang, Q., Sun, K., Chinazzi, M., y Piontti, A.P., Dean, N.E., Rojas, D.P., Merler, S., Mistry, D., Poletti, P., Rossi, L. *et al.* (2017). Spread of zika virus in the americas. *Proceedings of the National Academy of Sciences*, 114, E4334–E4343.

Figure captions

Figure 1: The dissimilarity-based model used takes mean dissimilarity values of pathogen distributions and between countries in a given year, and uses this information in addition to the product of these two values to train the PNP model. Pathogen occurrences among countries are present or absent (black dots in panel `a` indicate pathogen occurrences), and the density of dissimilarities where the pathogen occurred relative to the overall density of dissimilarities provides information on the suitability of pathogen occurrence in a given country (`b`), and forms the basis of the PNP model approach.

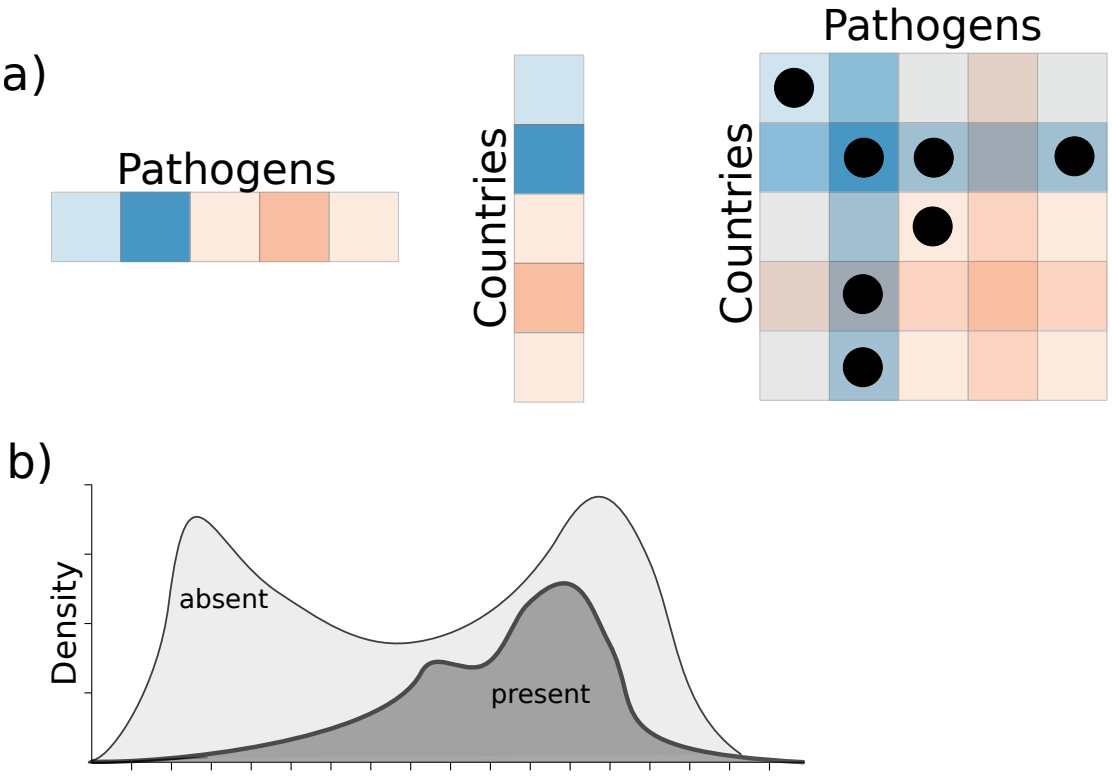
Figure 2: Pathogen events from previous years increased model predictive accuracy after an initial small decrease, suggesting that five years or more of data improves predictions, but accuracy could actually decrease in some data sparse situations where only two or three years of data were available.

Figure 3: Predictive accuracy decreased when attempting to forecast far into the past or future. Models were trained on either the period between 2005-2015 (for prediction into the past) or 1990-2000 (for prediction into the future).

Figure 4: Using a rolling window ($t = 4$ years), we found that predictive accuracy did not increase as a result of enhanced surveillance and data collection of more recent years.

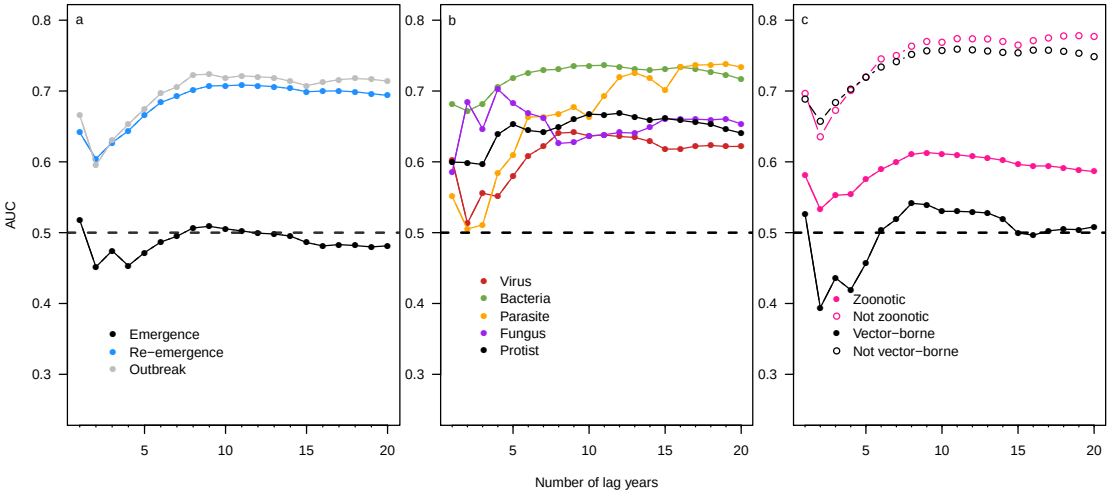
Figures

Figure 1



474 Figure 2

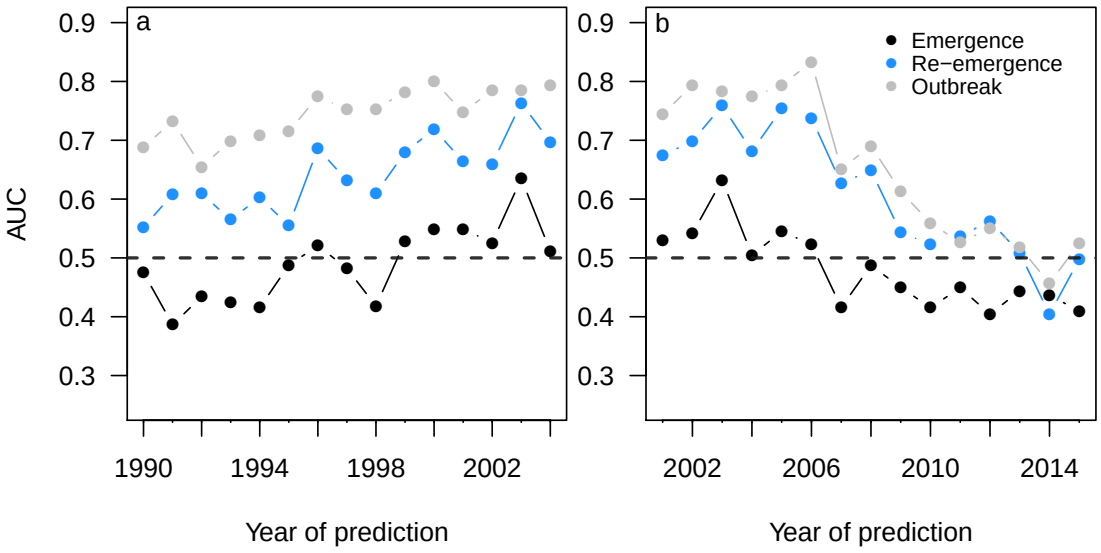
475



476

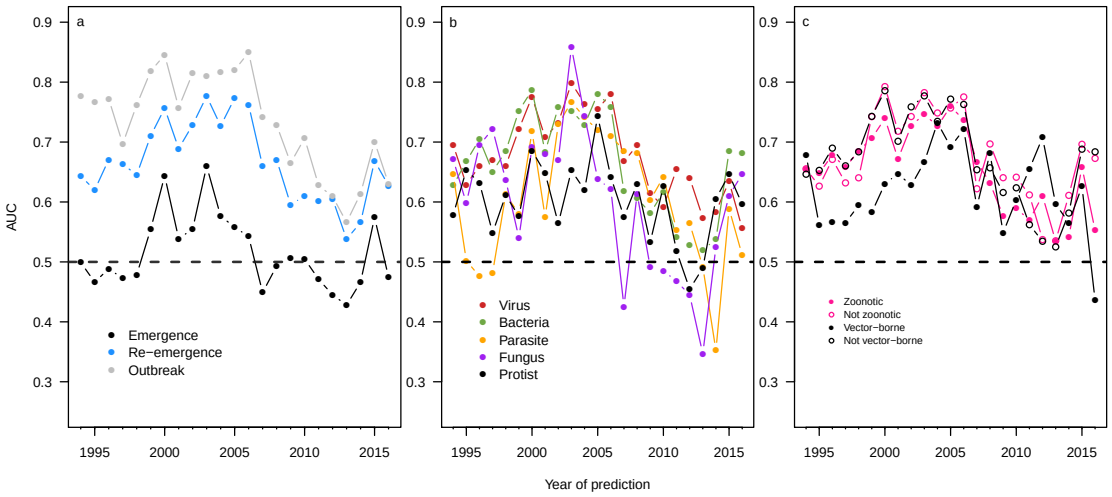
477 Figure 3

478



479

Figure 4



Supplemental materials

Effect of rolling window size

The size of the rolling window we used for model training prior to prediction could influence model performance. To examine this possibility, we used a rolling window of 7 years (compared to the 4 year window used in the main text), finding qualitatively similar results when flu was included (Figure S1) or excluded (Figure S2). We explored this further by examining rolling windows of 2, 4, and 6 years (Figure S3), with qualitatively similar findings. For this analysis, we excluded influenza, as we did in the main text.

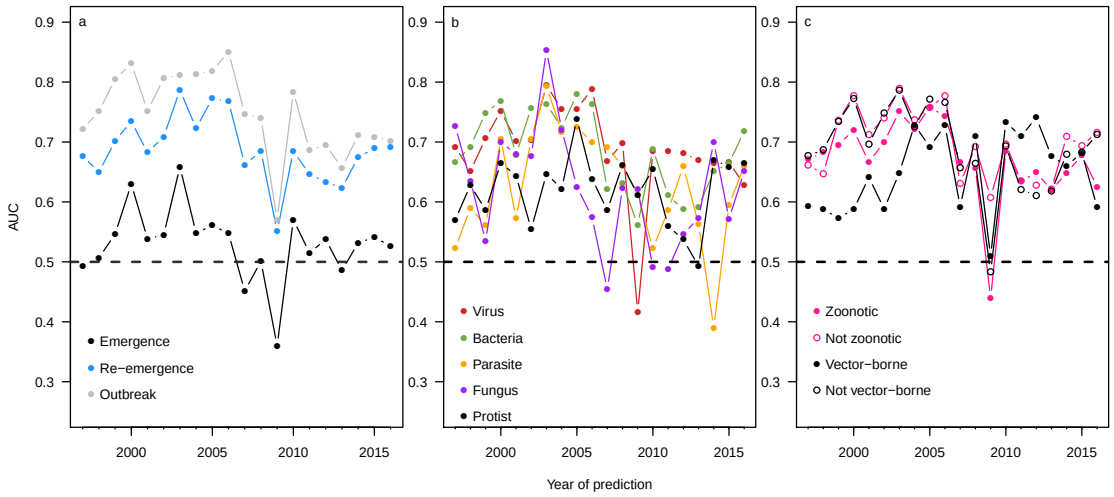


Figure S1: Rolling window size did not strongly influence model performance when considering next year prediction, as a window of 7 years produced qualitatively similar results to the window of 4 years we examine in the main text.

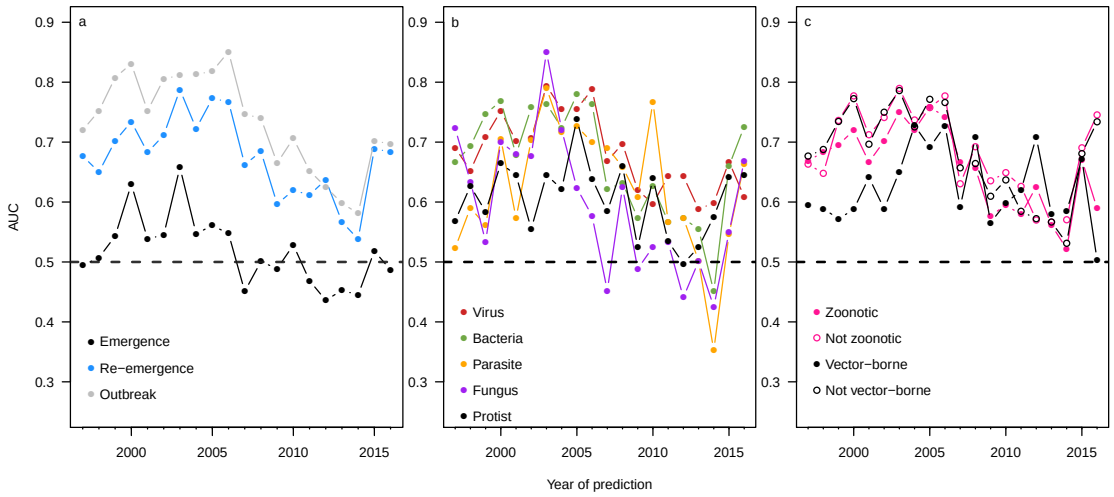


Figure S2: Rolling window size did not strongly influence model performance when considering next year prediction, as a window of 7 years produced qualitatively similar results to the window of 4 years we examine in the main text.

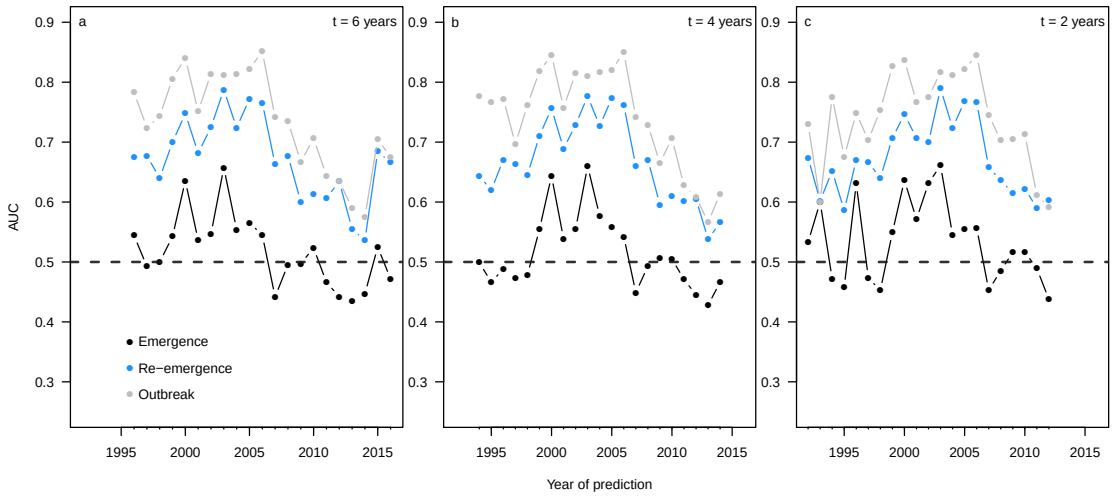


Figure S3: Examining rolling windows of 2, 4, and 6 years provides evidence that rolling window size did not strongly influence model performance.

Effect of pathogen traits on model hindcasting/forecasting ability

Here, we further explore the effect of pathogen type on model performance when hindcasting or forecasting pathogen outbreak or emergence event suitability. There was no predictable variation in model performance as a function of pathogen type (Figure S4) or whether the pathogen is classified as zoonotic or vector-borne (Figure S5).

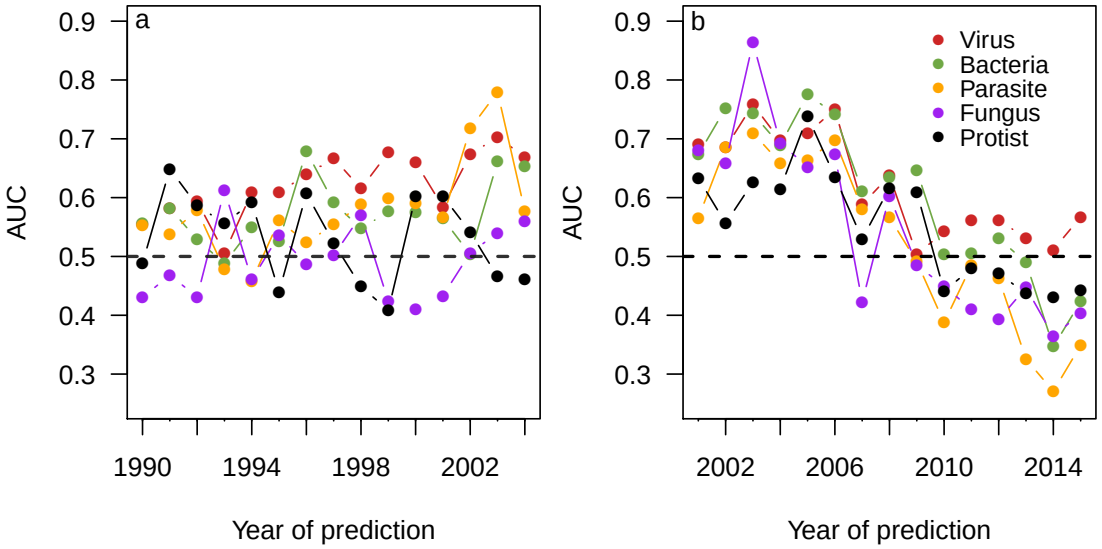


Figure S4: Predictive accuracy decreased when attempting to forecast far into the past or future, independent of pathogen type. Models were trained on either the period between 2005-2015 (for prediction into the past) or 1990-2000 (for prediction into the future).

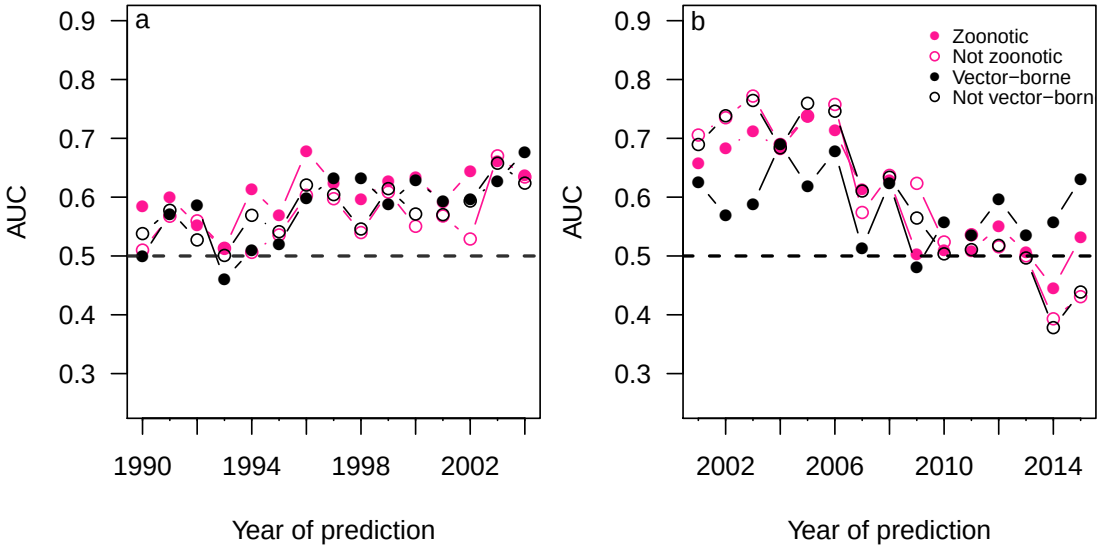


Figure S5: Predictive accuracy decreased when attempting to forecast far into the past or future. This was insensitive to whether the pathogen is considered zoonotic or vector-borne. Models were trained on either the period between 2005-2015 (for prediction into the past) or 1990-2000 (for prediction into the future).

The effect of including influenza

The 2009 influenza A pandemic fundamentally changed the network of countries and pathogens through the addition of many links to one pathogen (Figure S6). This may be an issue for approaches such as ours, which relies on extracting information from the similarity between pathogens in their distributions among countries, and similarity between countries in their pathogen composition. When the model wasn't expected to predict a pandemic event, the inclusion of influenza did not substantially influence model predictions when trained on differing numbers of years (Figure S7) or when forecasting or hindcasting to different time periods (Figure S8). However, the effect of the 2009 influenza pandemic can be seen in the substantial declines in model performance when attempting to forecast one year ahead after training on a rolling window of 4 years (Figure S9). Interestingly, the exclusion of influenza results in lower mean performance when the model doesn't have data on many years, likely because influenza is widespread and can influence the pathogen and country dissimilarity values used to train the model. However, once sufficient data is provided, model performance with and without influenza is nearly identical.

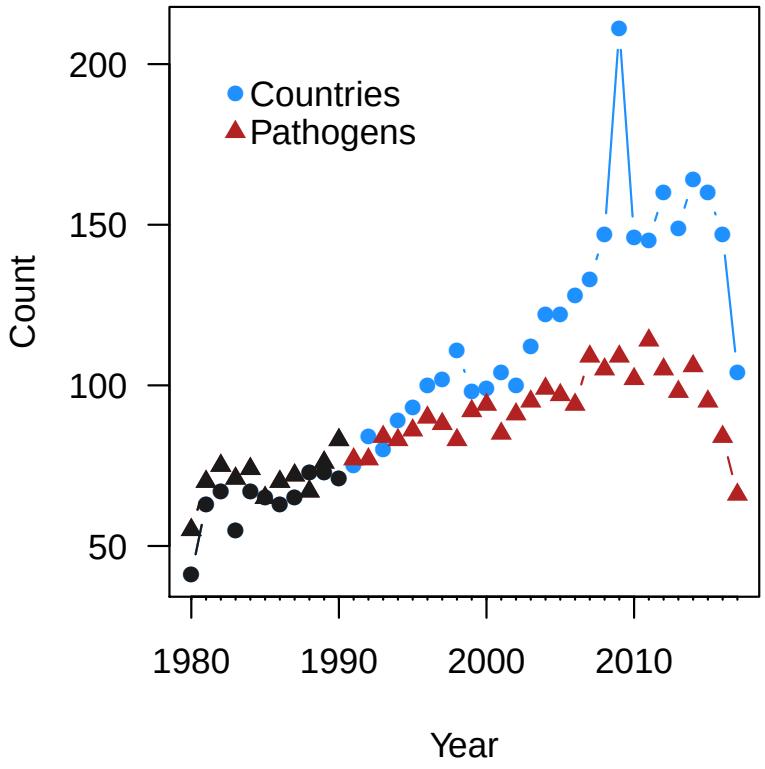
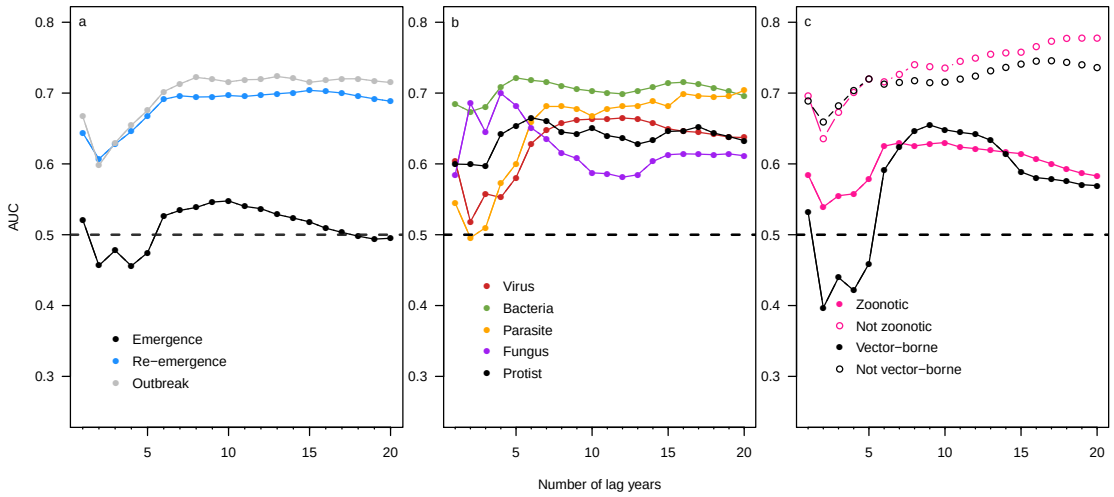


Figure S6: The number of countries with at least one outbreak event and the number of pathogens has increased over time, likely due to more vigilant sampling and description of emerging pathogens in a larger number of countries.



0

Figure S7: Pathogen events from previous years increased model predictive accuracy after an initial small decrease, suggesting that five years or more of data improves predictions, but accuracy could actually decrease in some data sparse situations where only two or three years of data were available.

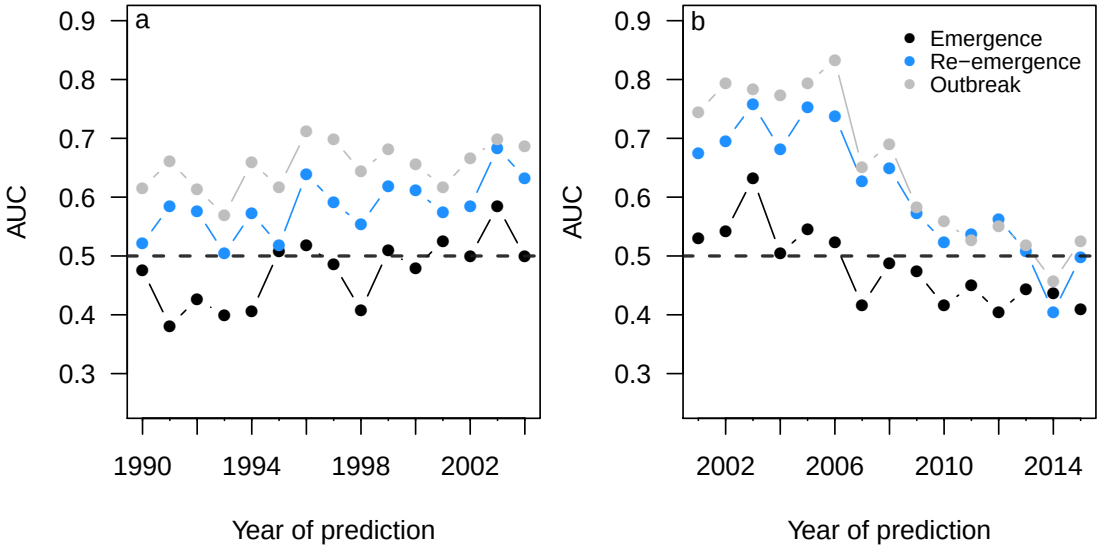


Figure S8: Predictive accuracy decreased when attempting to forecast far into the past or future. Models were trained on either the period between 2005-2015 (for prediction into the past) or 1990-2000 (for prediction into the future).

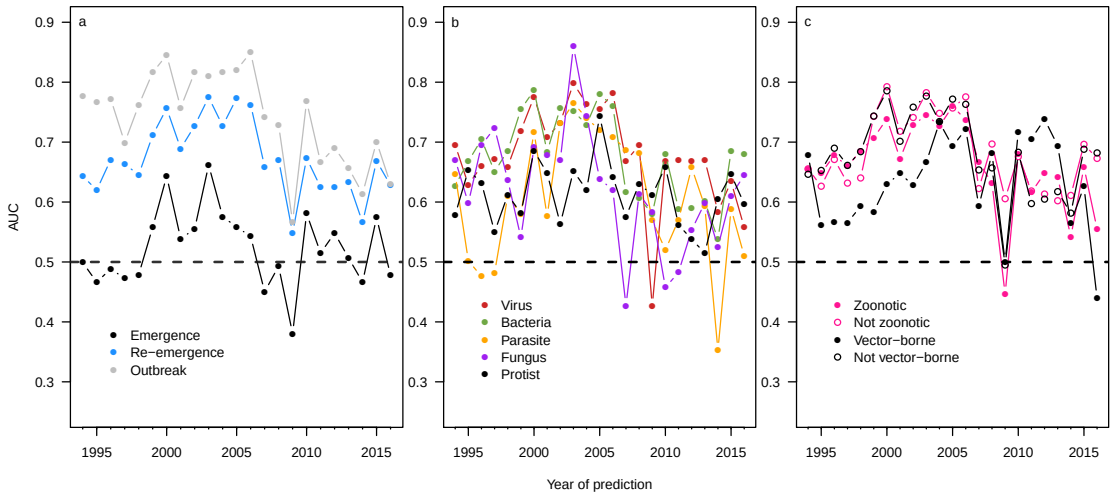


Figure S9: Using a rolling window ($t = 4$ years), we found that predictive accuracy did not increase as a result of enhanced surveillance and data collection of more recent years.