Epidemics 27 (2019) 59–65

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Confronting data sparsity to identify potential sources of Zika virus spillover infection among primates

Barbara A. Han[a,*], Subhabrata Majumdar[b,1,2], Flavio P. Calmon[c,2], Benjamin S. Glicksberg[d,2], Raya Horesh[e], Abhishek Kumar[2,3], Adam Perer[f,2], Elisa B. von Marschall[g], Dennis Wei[e], Aleksandra Mojsilović[e], Kush R. Varshney[e]

[a] Cary Institute of Ecosystem Studies, Box AB Millbrook, NY 12545, USA
[b] University of Florida Informatics Institute, 432 Newell Drive, CISE Bldg E251, Gainesville, FL 32611, USA
[c] Harvard University, 29 Oxford St, Cambridge, MA 02138, USA
[d] Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, 94158, USA
[e] IBM Research, 1101 Kitchawan Rd, Yorktown Heights, NY 10598, USA
[f] Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA
[g] IBM Watson Media & Weather, 550 Assembly St, Columbia, SC 29201, USA

ARTICLE INFO

Keywords:
Predictive analytics
Flavivirus
Arbovirus
Non-human primate
Machine learning
Bayesian multi-task learning
Imputation
Neotropical
Spillover
Spillback
Ecology
Surveillance

ABSTRACT

The recent Zika virus (ZIKV) epidemic in the Americas ranks among the largest outbreaks in modern times. Like other mosquito-borne flaviviruses, ZIKV circulates in sylvatic cycles among primates that can serve as reservoirs of spillover infection to humans. Identifying sylvatic reservoirs is critical to mitigating spillover risk, but relevant surveillance and biological data remain limited for this and most other zoonoses. We confronted this data sparsity by combining a machine learning method, Bayesian multi-label learning, with a multiple imputation method on primate traits. The resulting models distinguished flavivirus-positive primates with 82% accuracy and suggest that species posing the greatest spillover risk are also among the best adapted to human habitations. Given pervasive data sparsity describing animal hosts, and the virtual guarantee of data sparsity in scenarios involving novel or emerging zoonoses, we show that computational methods can be useful in extracting actionable inference from available data to support improved epidemiological response and prevention.

1. Introduction

In one of the largest zoonotic disease outbreaks in modern times, the American Zika virus epidemic spread from Brazil to more than 30 surrounding countries in South and Central America (hereafter, the Americas). Zika virus is one of several flaviviruses transmitted to humans by mosquito vectors, which cause several zoonotic diseases (yellow fever, Dengue, St. Louis encephalitis, Japanese encephalitis, and West Nile). The species in which these viruses are maintained in the wild (sylvatic reservoirs) are thought to be distinct from species that maintain flaviviruses transmitted primarily by tick vectors (e.g., leading to Kyasanur Forest disease, Powassan, and Omsk hemorrhagic fever among other zoonoses). Compared to the other mosquito-borne flaviviruses, American Zika virus was relatively understudied, and is thought to have infected over half a million people in the latest outbreak, causing neurologic disorders in over 3600 infants to date (Mitchell, 2016). While this Zika virus epidemic was de-categorized as a public health emergency of international concern in November 2016, the long-term public health and societal consequences of endemic Zika virus infection in the Americas could be substantial. In addition to teratogenic effects presenting as microcephaly and other serious brain anomalies, congenital Zika syndrome comprises a broad suite of abnormalities of vision, hearing, muscle tone, and joint movement (Baud et al., 2017). The immediate costs of outbreak response, via increased vector control and screening pregnant women, will be further extended by costs accumulating over the long-term to support a generation of

* Corresponding author.
E-mail address: hanb@caryinstitute.org (B.A. Han).
[1] Currently at AT&T Labs Research.
[2] Work done while at IBM Research.
[3] Currently at Google Brain.

https://doi.org/10.1016/j.epidem.2019.01.005
Received 7 October 2018; Received in revised form 17 January 2019; Accepted 31 January 2019
Available online 19 March 2019

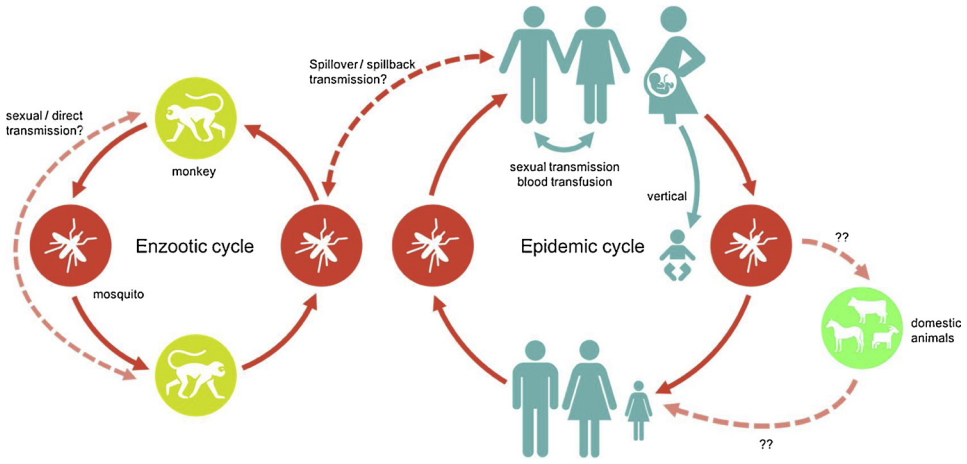

Fig. 1. A conceptual figure of Zika virus transmission routes. Zika virus is primarily acquired and transmitted in humans through mosquito vectors (red arrows). Once infected, humans can pass infection to other humans, either vertically, sexually, or through contaminated blood (teal arrows), or by providing an infected bloodmeal to competent mosquitoes (red arrow). Spillover transmission occurs when a competent mosquito acquires infection from a sylvatic source (reservoir) and transmits the infection to a human, whereas spillback transmission would occur if a competent mosquito acquires infection from a human and transmits it back to a competent wild host (red dashed arrows). Additional unconfirmed routes of transmission include sexual transmission among sylvatic hosts (pink dashed arrow, left), and vector-borne transmission to and from domesticated mammals (pink dashed arrows, right).

children with lifelong neurologic and developmental disorders, musculoskeletal deformities, and associated disabilities including blindness. These disorders are likely to strain a limited health care system, especially in countries where resources for the disabled are already limited.

Zika virus (ZIKV) was first discovered in a sentinel rhesus macaque (*Macaca mulatta*) in Uganda. Like other mosquito-borne flaviviruses (e.g., yellow fever virus, West Nile virus), ZIKV originated (and persists today) in a paleotropic sylvatic (wild) cycle involving non-human primate species (hereafter, *primates*) and forest-living *Aedes* mosquitoes (Bueno et al., 2016). These ancestral sylvatic strains spilled over into a transmission cycle maintained in humans and anthropophilic mosquitoes (e.g., dengue virus), thereby moving into global circulation. Primates acting as virus reservoirs or amplifying hosts can transmit Zika virus to human populations via competent mosquito vectors that feed on both human and non-human hosts (Fig. 1) (Althouse et al., 2018; Evans et al., 2017; Kuno et al., 2017). In addition to proximity with hosts of the sylvatic cycle, flavivirus spillover events are likely to be influenced by changing ecological patterns (e.g., climate warming) that support increasing vector population abundance, or increasing proximity of humans to sylvatic hosts (Kilpatrick and Randolph, 2012).

Mosquito-borne zoonoses repeated spillover to humans (e.g., West Nile and yellow fever) are best managed via multi-faceted control efforts that simultaneously address human susceptibility, vector abundance, and sylvatic sources of infection (Kilpatrick and Randolph, 2012). For example, vaccine therapy has proven remarkably effective for yellow fever, achieving close to 100% efficacy with a single vaccination that may confer lifelong immunity (Gotuzzo et al., 2013). Yet, outbreaks of yellow fever occur with regularity due to incomplete vaccine coverage in humans and repeated spillover infection from multiple sylvatic sources of infection, including non-human primates (Hanley et al., 2013). Thus, even if vaccine coverage and vector control were highly efficient, the successful mitigation of ZIKV and other mosquito-borne flaviviruses may still depend on effectively managing viral spillover from persistent sylvatic species that transmit infection to humans (Althouse et al., 2016; Bueno et al., 2016). Recent theoretical work exploring ZIKV risk in the Americas suggests a high probability of ZIKV being maintained by relatively small populations of susceptible primates – with as few as 6000 individuals and 10,000 mosquitoes, a sylvatic cycle would maintain a steady pool of infected hosts to facilitate repeated spillover transmission to humans (Althouse et al., 2016).

A critical question is whether and in which species ZIKV will establish a sylvatic cycle in its new range in the Americas. Identifying which species are most likely to serve as reservoirs is paramount for targeting surveillance, especially in regions where biodiversity is exceptionally high. Fundamentally, the ability for primate hosts to support and sustain a sylvatic virus cycle is at least partially dependent on biologically encoded interactions between hosts and pathogens. In hosts, intrinsic traits - for example, basal metabolic rates, traits indicating a 'slow' vs. 'fast' pace of life (e.g., litter sizes, reproductive rates), behavior, and biogeography – are readily observable features that distinguish one host species from another, but also recapitulate an unobservable evolutionary history underlying the capacity of certain species to perpetuate zoonotic pathogens (Han et al., 2016a). This general pattern, that organismal traits can serve as useful indicators of zoonotic capacity, has been supported by various studies; For example, Cable et al. linked host metabolic rates to rates of pathogenesis across multiple hosts and multiple zoonoses (Cable et al., 2007); Han et al. (a) showed that traits associated with a fast life history strategy were more common in rodents known to be reservoirs of zoonotic disease (Han et al., 2015); and (b) utilized organismal trait information to predict novel bat hosts of filoviruses like Ebola (Han et al., 2016b), predictions which were subsequently validated by independent field surveillance (Goldstein et al., 2018; Yang et al., 2017).

Like many zoonotic disease systems, the Zika virus system suffers from data sparsity. Among global primates, only two species have been confirmed positive for Zika virus, and an incomplete understanding of primate biology and ecology, even for relatively common species, limits our ability to make data-driven decisions about surveillance and spillover prevention. Here, we leverage intrinsic biological features of the world's primates together with data on which primates have tested positive for six mosquito-borne flaviviruses. To overcome data sparsity in both primate traits and ZIKV positivity, we apply a multiple imputation approach and a Bayesian multi-label machine learning approach. Applied in tandem, these methods preserve biological dependency patterns among species traits and leverage information on primate hosts known to be positive for mosquito-borne flaviviruses to identify particular species whose intrinsic trait profiles suggest a high probability of serving as hosts for ZIKV and other mosquito-borne flaviviruses.

2. Methods

We compiled data on intrinsic traits (a.k.a. *features*) describing all primate species (sample size $n = 376$) from PanTHERIA, of which 18 species were confirmed (in peer-reviewed publications available at the time this study was undertaken) as reservoirs of one or more of six mosquito-borne flaviviruses (FLAV) with non-human mammalian reservoirs according to the GIDEON database (Berger, 2005; Jones et al., 2009). FLAV-positivity for all positive primate species is summarized in Table 1. At the time of analysis, mosquito-borne flaviviruses include the

Table 1
Primate species that are positive for various mosquito-borne flaviviruses (SLEV = St. Louis encephalitis virus; YFV = Yellow Fever virus; ZIKV = Zika virus; WNV = West Nile virus; DENV = Dengue virus; JEV = Japanese encephalitis virus).

Species	FLAV positivity
Alouatta caraya	SLEV, YFV
Alouatta seniculus	YFV
Callithrix jacchus	ZIKV
Cebus apella	YFV
Cebus libidinosus	ZIKV
Cercocebus atys	WNV
Chlorocebus aethiops	ZIKV
Erythrocebus patas	ZIKV
Lemur catta	WNV
Macaca fascicularis	DENV, JEV
Macaca fuscata	JEV
Macaca leonina	DENV, JEV
Macaca mulatta	WNV
Macaca nemestrina	DENV, JEV, WNV
Macaca sinica	DENV
Macaca sylvanus	WNV
Mandrillus sphinx	WNV, YFV
Pithecia pithecia	YFV
Saguinus midas	YFV
Trachypithecus cristatus	DENV

dengue viruses (all serotypes; DENV), yellow fever virus (YFV), St. Louis encephalitis (SLEV), Japanese encephalitis (JEV), and West Nile virus (WNV). In this database, a *reservoir* is defined as a species for which there is public health consensus that some populations maintain continuous infection and may therefore serve as the source of spillover infection to human populations. This definition excludes species that suffer acute disease-induced mortality, sentinel species, and laboratory-infected animals that are not known to be infected in the wild (Holzmann et al., 2010). However, species that have seroconverted and those testing positive for antibodies against mosquito-borne flaviviruses are included, as are species that are sampled from captive colonies or species that are most likely only incidentally infected (e.g., WNV). We label all such species as positive for mosquito-borne flaviviruses ("FLAV +", or "flavivirus positivity") in this analysis, though clearly the wide variation in diagnostics means that our label combines species that are susceptible or competent for infection together with species that are true reservoirs. Among these 6 mosquito-borne viruses, only three are currently recognized as having primate reservoirs contributing to a sylvatic transmission cycle (YFV, DENV, ZIKV). We expanded this to the more conservative list of six flaviviruses to borrow information across diseases, thus augmenting the otherwise very sparsely labeled dataset of FLAV + primates available for applying our machine learning algorithm. Moreover, the public health consequences of this approach, namely, labeling a FLAV- species as FLAV+ (type I error), was preferred to the alternative (inflating type II error, i.e., failing to predict FLAV + primate species). Among the 18 primate species that were FLAV + in this dataset, two species are confirmed reservoirs of ZIKV (*Chlorocebus aethiops* and *Erythrocebus patas*). Both of these species are Old World (paleotropical) primates found in the Sahel and Sudan-Guinea grassland biomes across northern Africa.

Our initial data set on primate characteristics contained 50 features for 376 species. These features describe various aspects of host species including their ecology (e.g., social group size, terrestriality, age to sexual maturity), physiology (e.g., metabolic rate, neonate size), biogeography (e.g., geographic range area, home range size), etc (Supplementary Tables S1 for variable names and definitions). However, not all features were known for all species, even though primates are among the best-studied mammal groups. Upon filtering this dataset to remove sparse features (variables with data missing for more than 80% of species), as well as extremely understudied species (data present for fewer than two features) and *Homo sapiens*, our final dataset contained 33 features for 364 primate species. In this matrix, approximately one third of species (34.9%) was missing data for one or more features. We imputed these missing trait data using the Multiply Imputed Chained Equations (MICE) method (Penone et al., 2014; Raghunathan et al., 2001). This method retains biological relationships while reducing bias compared to the common practice of dropping those species that are missing data from the analysis, even when there is underlying structure in missing data (i.e., data are not missing completely at random) (Penone et al., 2014). Given a set of initial values for a given species, the MICE method predicts missing entries of a particular feature by iteratively leveraging the information available across the other variables, thus preserving biological dependency patterns among features. This process is repeated until the entries across a number of imputed datasets reach a stable distribution, indicated by the value of the Gelman-Rubin $\hat{R}$ statistic being less than 1.1 (Gelman and Rubin, 1992).

We obtained ten such imputed datasets from the original data matrix and applied a recently developed supervised learning method, Bayesian Multi-label Learning via Positive Labels (BMLPL), to build predictive models (Rai et al., 2015). This method achieves superior performance in settings where the goal is to assign a subset of labels to samples in a highly sparse matrix with correlated labels. In our study, seropositivity of primate species to any of 6 mosquito-borne flaviviruses comprises species-level vectors of binary labels that are highly correlated (phylogenetically), where the labels (seropositive status) are highly sparse, especially for Zika virus. Using BMLPL, we leveraged correlations between labels by simultaneously modeling as binary responses the host status for all 6 major mosquito-borne flaviviruses across all primate species, and considering species-level features as predictors.

We applied BMLPL to each of the ten imputed data matrices, and calculated probability of each primate species being a reservoir of ZIKV by taking the mean of predicted probabilities across these ten models. We assessed the classification accuracy of this model using the area under the receiver operating curve (AUC) plotted using these probabilities (Fig. 2, top left panel). Furthermore, we used components of the trained BMLPL model objects to construct a variable importance measure and averaged them across the imputed datasets to identify features that are particularly influential for correctly assigning ZIKV positivity by our model.

To evaluate out-of-sample performance of the model, we used stratified ten-fold cross-validation. For this, we first randomly divided the set of known FLAV + primates and the set of primates with unknown FLAV status each into ten folds. Then we combined pairs of these two types of folds to generate ten folds out of the full set of primates that have roughly the same proportion of FLAV + and undetected primates. Finally, we generated probability scores for primates inside each fold using a model trained with the data on all primates outside that fold. Fig. 2 (top right panel) plots the receiver operating curve from these probability scores.

See Supplementary Methods S3 for additional details on the imputation technique, BMLPL models, variable importance and validations.

3. Data availability

The datasets generated during and/or analysed during the current study are available in the Figshare repository, https://doi.org/10.6084/m9.figshare.5459176.v1. Code to reproduce these analyses are available at https://github.com/shubhobm/Predict-zika-res.

4. Results

Our model leveraged information on primate hosts known to be positive for mosquito-borne flaviviruses to identify which additional primate species are the most likely to be positive for ZIKV. The

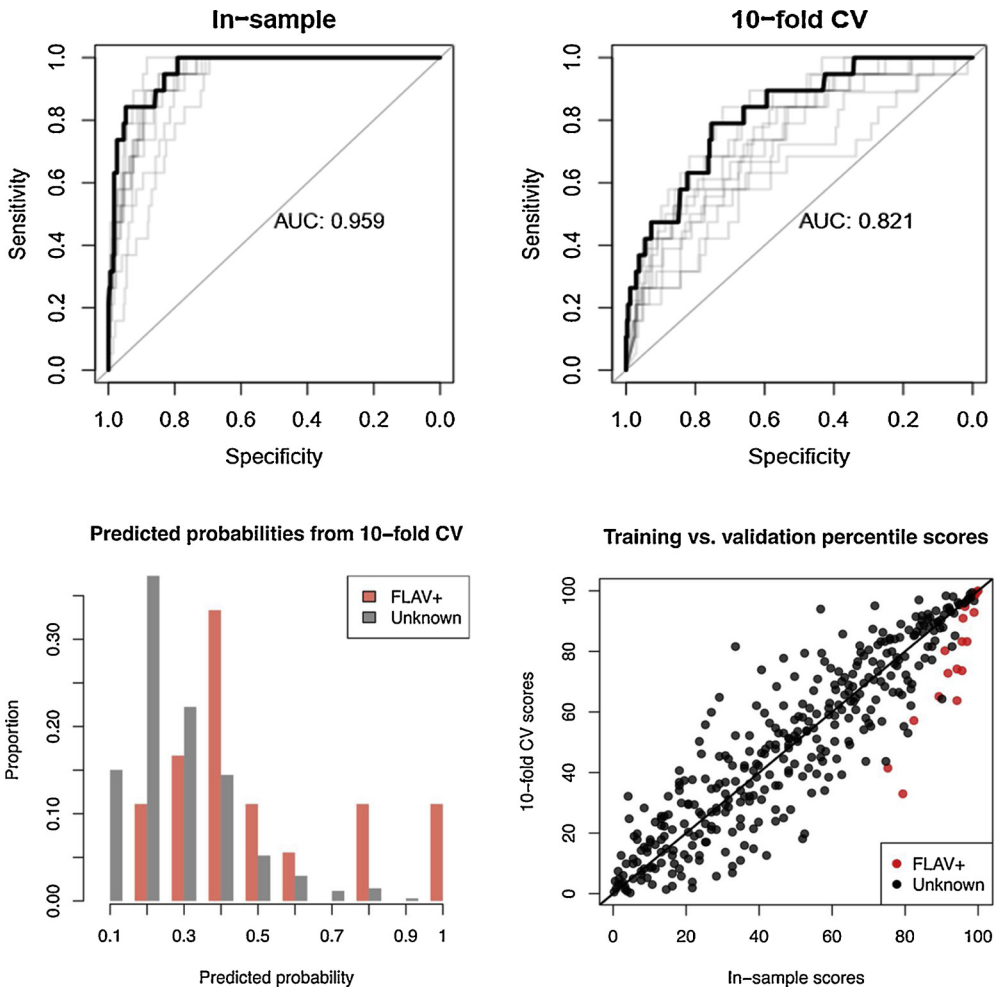

Fig. 2. The performance accuracy of the predictive model is illustrated by two ROC curves obtained from the model trained on the full data (in-sample, top left panel) and using 10-fold stratified cross-validation (top right panel). Grey lines in these two panels indicate curves obtained from analyzing individual imputed datasets, while the black lines are the curves obtained from mean probabilities taken over all 10 imputed datasets. A histogram (bottom left panel) depicts the distribution of predicted probabilities for both FLAV + and FLAV- species, and the scatterplot (bottom right panel) gives risk scores of carrying mosquito-borne flaviviruses for all primate species in our analysis, with red bars/points representing primates that are confirmed positive (using any diagnostic method) for any of the 6 mosquito-borne flaviviral diseases. Among 18 confirmed primate species, those with the lowest model-predicted risk scores include *Lemur catta* (33·0), *Pithecia pithecia* (41·5), *Saguinus midas* (57·1), *Alouatta caraya* (63·7), *Chlorocebus aethiops* (65·1).

classification task was therefore to identify those species with and without an intrinsic capacity to carry ZIKV infection based on trait similarities. An AUC = 1.0 for any classification model indicates that the model is able to distinguish perfectly between classes across all samples. Our model achieved an AUC of 0.96 when trained on the full data for all 364 primates included in our study (Fig. 2, top left). The cross-validation process gives a better idea about how the model performs when predicting the risk of an undetected primate being a ZIKV reservoir. To do this, the BMLPL method was tasked with predicting reservoir probabilities for primates that are in one of the 10 cross-validation folds using a model trained on the other nine folds. This was repeated for samples in all folds, and then averaged over the 10 imputed datasets, achieving prediction accuracy (measured by AUC) of 0.82 (Fig. 2, top right).

For each primate species, we report a ZIKV reservoir risk score (i.e., the probability of testing positive for ZIKV (ZIKV+)) and the corresponding percentile of this risk compared to all other primates (Supplementary Table S2). The model assigned high risk scores to the majority (13 of 18) of primate species that are known reservoirs of other mosquito-borne flaviviruses (Fig. 2, bottom panel). However, for five of these 18 species (Fig. 2, bottom panel, red points), the cross-validated risk scores were low, indicating that in the absence of *a priori* confirmations as flavivirus-positives, deficiencies in basic biological data about these species would have precluded our capacity to identify them as posing a risk of carrying mosquito-borne flaviviruses.

In the Americas, we identify six species that were at or above the 90th percentile probability of being ZIKV+ (Table 2, Supplementary Table S2). None of these species have yet tested positive for ZIKV. Moreover, three species have yet to test positively for any mosquito-borne flavivirus (*Cebus albifrons*, the white-fronted capuchin; *Saimiri boliviensis*, the black-capped squirrel monkey; *Cebus capucinus*, the white-faced capuchin). The model identified 29 species spanning sub-Saharan Africa, India, and Southeast Asia that were at or above the 90th percentile probability of ZIKV positivity (Fig. 3). While we discuss species comprising the 90th percentile probability in more detail below, this probability cutoff was chosen arbitrarily and additional species fall just below this cutoff (e.g., six neotropical species within the 85th percentile (mean score > 0.228): *Cebus olivaceus, Alouatta palliata, Saimiri sciureus, Callicebus torquatus, Ateles paniscus, and Chiropotes albinasus*). The full list of primate species and their associated scores for FLAV positivity can be found in Table S2. In addition, the geographic ranges of all confirmed and predicted Zika-positive primate species are plotted in Fig. 3, which is also available as an interactive map at

Table 2
New world primate species whose risk scores for testing positive for Zika virus ($ZIKV^+$) were above the 90th percentile, and other mosquito-borne flaviviruses for which each species has tested positively (YFV = yellow fever virus; SLEV = St. Louis encephalitis virus; Undetected = the primate species is currently unknown to be positive for any mosquito-borne flaviviruses).

Species	Status	Percentile risk $ZIKV^+$
Cebus apella	YFV +	99.7
Cebus albifrons	Undetected	97.3
Alouatta seniculus	YFV +	95.9
Alouatta caraya	YFV +, SLEV +	94.2
Saimiri boliviensis	Undetected	92.9
Cebus capucinus	Undetected	90.7

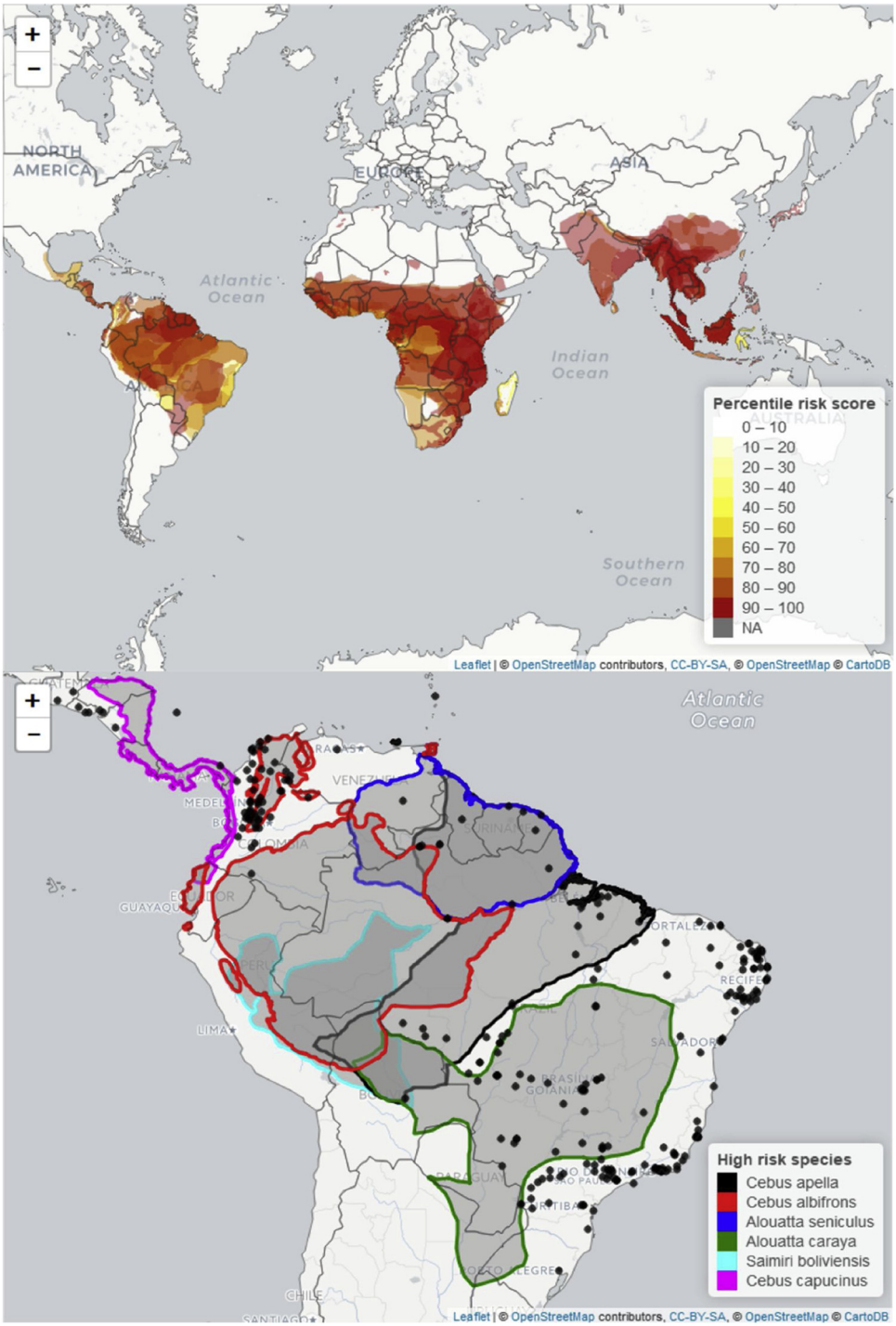

Fig. 3. Maps depicting the overlapping geographical ranges of 29 primate species at or above the 90th percentile probability for Zika virus positivity, globally (top), and in Brazil and surrounding countries in South America overlaid with locations of human Zika virus cases in black points (bottom) (Messina et al., 2016).

https://smajumdar.shinyapps.io/Zika_risk_map/.

The five most influential features for predicting ZIKV positivity (in descending order of relative importance) are maximum latitude, body mass, interbirth interval, age at first birth, and maximum longitude. Relative importance for all the features can be found in Fig. 7 (c) of the Supplementary Methods S3.

5. Discussion

In the Americas, Zika virus efforts have shifted from a state of emergency response to a longer-term goal of ZIKV control and eradication. The likelihood of achieving these goals depends fundamentally on whether ZIKV has spilled back from the predominantly human-centered transmission cycle to establish a persistent sylvatic cycle (Althouse et al., 2016). Our model leveraged information describing biological, ecological, and life history characteristics of primates to predict the zoonotic capacity in some species to serve as suitable reservoirs for ZIKV. On the basis of high-level features alone, our model was able to distinguish primate species that were positive for mosquito-borne flaviviruses with 82% accuracy and identified particular species with high risk of ZIKV positivity. While there are relatively few species that have been found FLAV+, and even fewer species that

have been found ZIKV+, these species are broadly distributed geographically and overlap in some places with dense human population centers.

In the Americas, there were five species with risk scores in the 90th percentile. The proximity of these species to human settlements and opportunities for human contacts at high frequencies suggests prioritizing them for ZIKV surveillance in the Americas. The tufted capuchin (*Cebus apella*), the Venezuelan red howler (*Alouatta seniculus*), and the white faced capuchin (*Cebus capucinus*) are exceptionally well-adapted to co-existing in dense human settlements, with some populations of the white faced capuchin considered to be commensal with humans (McKinney, 2011). In the neotropics, the Venezuelan red howler monkey (*Alouatta seniculus*) and the spider monkey (*Saimiri boliviensis*) are hunted for bushmeat in parts of their range, while the tufted capuchin and the white fronted capuchin (*Cebus albifrons*) are commonly kept as pets and are hunted for live trade export (Peres, 2000; Peres and Dolman, 2000). Although the hunting pressure on black howlers (*Alouatta caraya*) is somewhat lower compared to the capuchins, this species has frequent contact with livestock arising from a combination of diminishing natural habitat and the encroachment of agricultural land into patchily deforested areas (Crockett, 1998).

Neotropical primates with the highest ZIKV + risk scores are all relatively common species that are predominantly arboreal. This habitat stratification between humans (terrestrial) and non-human primates (arboreal) may suggest that spatial segregation will reduce spillover risk to humans. However, ZIKV was first isolated from a forest-dwelling mosquito vector (*Aedes africanus*) caught in the forest canopy as well as on the ground (Haddow et al., 1964). Moreover, Evans et al. assigned high likelihoods of ZIKV competence to a number of forest-dwelling primatophilic mosquito based on shared traits with known flavivirus vectors (Evans et al., 2017). ZIKV was recently isolated from one of these species, *Culex quinquefasciatus*, which easily adapts to both forested and human altered landscapes (Song et al., 2017). Thus, in addition to controlling *Aedes aegypti* vectors that are predominantly responsible for human-to-human transmission, surveillance should expand to consider other possible vector species that may perpetuate a sylvatic cycle and play a role in spillover transmission in the long-term (Evans et al., 2017).

We observed patterns of geographic overlap between predicted ZIKV + primates and human cases in Central and South America that suggest particular species that may be further prioritized for ZIKV surveillance to assess possible spill-back from humans (Fig. 3; https://smajumdar.shinyapps.io/Zika_risk_map/) (Messina et al., 2016). The geographic ranges of two monkey species (*Alouatta palliata*, ZIKV + risk score = 86·8; *Ateles geoffroyi*, ZIKV + risk score = 71·3) overlap with all recorded human ZIKV cases (as of January 2016) in Mexico, Guatemala, El Salvador, Honduras, Nicaragua, and Panama (Messina et al., 2016). Similar geographic overlaps are apparent in the paleotropics. For example, in addition to one known ZIKV reservoir (*Erythrocebus patas*) in Uganda, there are four primate species with high ZIKV + risk scores whose ranges overlap with human cases of ZIKV (*Papio anubis*, risk score = 97·4; *Colobus guereza*, risk score = 94·8; *Chlorocebus pygerythrus*, risk score = 92·4; *Galago senegalensis*, risk score = 83) (Supplementary Table S2; ZIKV + NHPs in https://smajumdar.shinyapps.io/Zika_risk_map/). Surveillance efforts could additionally be prioritized by species' proximity to centers of human population density, or to zones of high activity/traffic (border crossings, where there can be high, but transient, human density). In light of ongoing transmission in southeast Asia, the current outbreak of Zika virus in India ("WHO | Zika virus infection: India," 2018), and recent evidence that the American strain of Zika virus is more efficiently transmitted by *Aedes aegypti* than the Asian strain (Pompon et al., 2017), species-specific predictions of primate reservoirs for Zika virus may also aid in prioritizing ongoing ZIKV surveillance in Asia for both old and new virus strains. Similarly, improving ZIKV surveillance within its native range (Africa) will validate and improve model predictions for ZIKV as well as the other mosquito-borne flaviviruses. Multiple data streams that include, for example, locations of known and potential human cases, and data-driven predictions of both mosquito vectors and ZIKV + primates may combine to offer more immediately actionable insight to ZIKV management, which may be especially important given the long periods of latent ZIKV infection observed in primate hosts (Hirsch et al., 2017) and the potential for underreported latency in humans (Cardona-Ospina et al., 2018).

Unsurprisingly, the variables that were most important for model accuracy included those delimiting known geographic range limits of primate species, with ZIKV + probability increasing for species with ranges limits in the paleotropics (i.e., Africa, where ZIKV + primate species occur) (partial dependence plots in Supplementary Fig. S4). Primates who are ZIKV + tend to have a slightly larger body size as adults compared to ZIKV- species. We also find that ZIKV + species tend to give birth earlier in life, with shorter intervals between births, compared to ZIKV- primate species. There are many reasons these traits may be related to ZIKV positivity. For example, small species may have greater ZIKV competence (viral amplification leading to higher titers of circulating virus) due to a relative lack of adaptive immunity, which is expensive to maintain and therefore represents an energetic trade-off with reproductive investments. On the other hand, larger species may invest more heavily in adaptive immune responses, leading to greater tolerance to flavivirus infections (persistent infections without detrimental effects), a strategy that may also be advantageous for contending with a greater diversity of parasites encountered over longer lifespans. Tests of such hypotheses require more specific data collection, for example, on comparative immunology in ZIKV + species that differ in size and rate of fitness output and species' investments that affect near vs. long-term fitness (Zuk and Stoehr, 2002; Lochmiller and Derenberg, 2000).

While our modeling approaches successfully assigned high risk scores to most of the species for which there was prior evidence of FLAV positivity (Fig. 2), they could not compensate for extreme data-deficiency. For instance, when we employed the ten-fold cross-validation process by re-labeling FLAV + species in each holdout fold as FLAV- to observe model-assigned risk scores, we found that some confirmed FLAV + species were still assigned relatively low risk scores (Fig. 2, bottom two panels), likely due to the large percentage of trait data that were imputed even for very common species such as *Lemur catta* (ring-tailed lemur), *Sanguinus midas* (golden-handed tamarin), and *Alouatta caraya* (black howler monkey) (Fig. 2, red dots). Likewise, with FLAV positivity, predictions from the BMLPL model depend on a number of latent variables that are fixed in the beginning of the modeling process, and we found little variation in predictions across models that varied in the number of latent variables (considering from one to six latent variables). This suggests that the primate species identified by our Zika-specific model should also be considered as having a high probability of being positive for one or more of the other five mosquito-borne flaviviruses included in our analysis. The sparsity of positive labels meant that a true external validation, where the model is built on a (larger) subset of the samples and then validated on the remaining samples, would not be very informative. In such scenarios, 10-fold cross validation provides a more accurate estimate of out-of-sample predictive capability of BMLPL compared to in-sample validation. While this validation approach is more appropriate for dealing with label sparsity, when comparing predictive accuracy across models it is important to note that models validated using in-sample cross-validation will generate higher AUCs compared to those validated using typical out-of-sample procedures.

The limits imposed by a pervasive lack of basic biological information, even for very common species that frequently co-occur with humans, underlie the predictive capacity of more complex phenomena, such as disease competence and spillover risk to humans (Han and Drake, 2016). Here, we show that even in the presence of limited and irregular species-level data, methodological and computational

advances can be leveraged to build predictive models that capitalize on existing species-level data while simultaneously supporting research to improve empirical baselines about sylvatic cycles of ZIKV and other zoonoses. Iterative feedbacks between modeling and empirical data collection will lead to long-term efficiency gains in the mitigation and prevention of zoonotic spillover infection to humans.

Author contributions

Generated idea (BH, KV, AM), designed the study (BH, KV, AM), collated data (BH), developed analytical methods (SM, FC, DW, KV), validated results (BH, FC, BG, AP, DW, KV), visualized data and results (SM, BG, AP), wrote the manuscript (BH, SM, KV), coordinated research activities (DW, KV, AM), and acquired funding (KV, AM, BH). All authors were involved in improving study design following interpretation of analyses, and in reviewing the final manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgements

The authors would like to thank P. Feinberg and V. Ramesh for supporting data collation; C. Hough for generating the conceptual figure; Drs. K. Hanley, S. LaDeau, E. Han; C. Carlson and two anonymous reviewers for valuable manuscript comments; and M. Gillespie for editorial assistance. This work was conducted under the auspices of the IBM Science for Social Good initiative. Funding for B. Han was provided by NSF EEID grant (DEB-1717282).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.epidem.2019.01.005.

References

Althouse, B.M., Vasilakis, N., Sall, A.A., Diallo, M., Weaver, S.C., Hanley, K.A., 2016. Potential for Zika virus to establish a sylvatic transmission cycle in the Americas. PLoS Negl. Trop. Dis. 10, e0005055. https://doi.org/10.1371/journal.pntd.0005055.

Althouse, B.M., Guerbois, M., Cummings, D.A.T., Diop, O.M., Faye, O., Faye, A., Diallo, D., Sadio, B.D., Sow, A., Faye, O., Sall, A.A., Diallo, M., Benefit, B., Simons, E., Watts, D.M., Weaver, S.C., Hanley, K.A., 2018. Role of monkeys in the sylvatic cycle of chikungunya virus in Senegal. Nat. Commun. 9, 1046. https://doi.org/10.1038/s41467-018-03332-7.

Baud, D., Gubler, D.J., Schaub, B., Lanteri, M.C., Musso, D., 2017. An update on Zika virus infection. Lancet. https://doi.org/10.1016/S0140-6736(17)31450-2.

Berger, S.A., 2005. GIDEON: a comprehensive Web-based resource for geographic medicine. Int. J. Health Geogr. 4, 10. https://doi.org/10.1186/1476-072X-4-10.

Bueno, M.G., Martinez, N., Abdalla, L., Duarte Dos Santos, C.N., Chame, M., 2016. Animals in the Zika virus life cycle: what to expect from megadiverse Latin American Countries. PLoS Negl. Trop. Dis. 10, e0005073. https://doi.org/10.1371/journal.pntd.0005073.

Cable, J.M., Enquist, B.J., Moses, M.E., 2007. The allometry of host-pathogen interactions. PLoS One 2, e1130. https://doi.org/10.1371/journal.pone.0001130.

Cardona-Ospina, J.A., Alvarado-Arnez, L.E., Escalera-Antezana, J.P., Bandeira, A.C., Musso, D., Rodríguez-Morales, A.J., 2018. Sexual transmission of arboviruses: more to explore? Int. J. Infect. Dis. 76, 126–127. https://doi.org/10.1016/j.ijid.2018.08.022.

Crockett, C.M., 1998. Conservation biology of the genus Alouatta. Int. J. Primatol. 19, 549–578.

Evans, M.V., Dallas, T.A., Han, B.A., Murdock, C.C., Drake, J.M., 2017. Data-driven identification of potential Zika virus vectors. Elife 6, 077966. https://doi.org/10.7554/eLife.22053.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7, 457–472.

Goldstein, T., Anthony, S.J., Gbakima, A., Bird, B.H., Bangura, J., Tremeau-Bravard, A., Belaganahalli, M.N., Wells, H.L., Dhanota, J.K., Liang, E., Grodus, M., Jangra, R.K., DeJesus, V.A., Lasso, G., Smith, B.R., Jambai, A., Kamara, B.O., Kamara, S., Bangura, W., Monagin, C., Shapira, S., Johnson, C.K., Saylors, K., Rubin, E.M., Chandran, K., Lipkin, W.I., Mazet, J.A.K., 2018. The discovery of Bombali virus adds further support for bats as hosts of ebolaviruses. Nat. Microbiol. 3, 1084–1089. https://doi.org/10.1038/s41564-018-0227-2.

Gotuzzo, E., Yactayo, S., Córdova, E., 2013. Efficacy and duration of immunity after yellow fever vaccination: systematic review on the need for a booster every 10 years. Am. J. Trop. Med. Hyg. 89, 434–444. https://doi.org/10.4269/ajtmh.13-0264.

Haddow, A.J., Williams, M.C., Woodall, J.P., Simpson, D.I., Goma, L.K., 1964. Twelve isolations of Zika virus from *Aedes (stegomyia) africanus (theobald)* taken in and above a Uganda forest. Bull. World Health Organ. 31, 57–69.

Han, B.A., Drake, J.M., 2016. Future directions in analytics for infectious disease intelligence: toward an integrated warning system for emerging pathogens. EMBO Rep. 17, 785–789. https://doi.org/10.15252/embr.201642534.

Han, B.A., Schmidt, J.P., Bowden, S.E., Drake, J.M., 2015. Rodent reservoirs of future zoonotic diseases. Proc. Natl. Acad. Sci. U. S. A. 112, 7039–7044. https://doi.org/10.1073/pnas.1501598112.

Han, B.A., Kramer, A.M., Drake, J.M., 2016a. Global patterns of zoonotic disease in mammals. Trends Parasitol. 32, 565–577. https://doi.org/10.1016/j.pt.2016.04.007.

Han, B.A., Schmidt, J.P., Alexander, L.W., Bowden, S.E., Hayman, D.T.S., Drake, J.M., 2016b. Undiscovered bat hosts of filoviruses. PLoS Negl. Trop. Dis. 10, e0004815. https://doi.org/10.1371/journal.pntd.0004815.

Hanley, K.A., Monath, T.P., Weaver, S.C., Rossi, S.L., Richman, R.L., Vasilakis, N., 2013. Fever versus fever: the role of host and vector susceptibility and interspecific competition in shaping the current and future distributions of the sylvatic cycles of dengue virus and yellow fever virus. Infect. Genet. Evol. 19, 292–311. https://doi.org/10.1016/j.meegid.2013.03.008.

Hirsch, A.J., Smith, J.L., Haese, N.N., Broeckel, R.M., Parkins, C.J., Kreklywich, C., DeFilippis, V.R., Denton, M., Smith, P.P., Messer, W.B., Colgin, L.M.A., Ducore, R.M., Grigsby, P.L., Hennebold, J.D., Swanson, T., Legasse, A.W., Axthelm, M.K., MacAllister, R., Wiley, C.A., Nelson, J.A., Streblow, D.N., 2017. Zika Virus infection of rhesus macaques leads to viral persistence in multiple tissues. PLoS Pathog. 13, e1006219. https://doi.org/10.1371/journal.ppat.1006219.

Holzmann, I., Agostini, I., Areta, J.I., Ferreyra, H., Beldomenico, P., Di Bitetti, M.S., 2010. Impact of yellow fever outbreaks on two howler monkey species (Alouatta guariba clamitans and A. caraya) in Misiones, Argentina. Am. J. Primatol. 72, 475–480. https://doi.org/10.1002/ajp.20796.

Jones, K.E., Bielby, J., Cardillo, M., Fritz, S.A., O'Dell, J., Orme, C.D.L., Safi, K., Sechrest, W., Boakes, E.H., Carbone, C., Connolly, C., Cutts, M.J., Foster, J.K., Grenyer, R., Habib, M., Plaster, C.A., Price, S.A., Rigby, E.A., Rist, J., Teacher, A., Bininda-Emonds, O.R.P., Gittleman, J.L., Mace, G.M., Purvis, A., 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology 90https://doi.org/10.1890/08-1494.1. 2648–2648.

Kilpatrick, A.M., Randolph, S.E., 2012. Drivers, dynamics, and control of emerging vector-borne zoonotic diseases. Lancet 380, 1946–1955. https://doi.org/10.1016/S0140-6736(12)61151-9.

Kuno, G., Mackenzie, J.S., Junglen, S., Hubálek, Z., Plyusnin, A., Gubler, D.J., 2017. Vertebrate reservoirs of arboviruses: myth, synonym of amplifier, or reality? Viruses 9. https://doi.org/10.3390/v9070185.

Lochmiller, R.L., Deerenberg, C., 2000. Trade-offs in evolutionary immunology: just what is the cost of immunity? Oikos 88, 87–98.

McKinney, T., 2011. The effects of provisioning and crop-raiding on the diet and foraging activities of human-commensal white-faced capuchins (Cebus capucinus). Am. J. Primatol. 73, 439–448. https://doi.org/10.1002/ajp.20919.

Messina, J.P., Kraemer, M.U., Brady, O.J., Pigott, D.M., Shearer, F.M., Weiss, D.J., Golding, N., Ruktanonchai, C.W., Gething, P.W., Cohn, E., Brownstein, J.S., Khan, K., Tatem, A.J., Jaenisch, T., Murray, C.J., Marinho, F., Scott, T.W., Hay, S.I., 2016. Mapping global environmental suitability for Zika virus. Elife 5, e15272. https://doi.org/10.7554/eLife.15272.

Mitchell, C., 2016. PAHO WHO | Zika Cumulative Cases [WWW Document]. Pan American Health Organization / World Health Organization. URL http://www.paho.org/hq/index.php?option=com_content&view=article&id=12390:zika-cumulative-cases&catid=8424:contents&Itemid=42090&lang=en (Accessed 15 May 2017).

Penone, C., Davidson, A.D., Shoemaker, K.T., Di Marco, M., Rondinini, C., Brooks, T.M., Young, B.E., Graham, C.H., Costa, G.C., 2014. Imputation of missing data in life-history trait datasets: which approach performs the best? Methods Ecol. Evol. 5, 961–970. https://doi.org/10.1111/2041-210X.12232.

Peres, C.A., 2000. Effects of subsistence hunting on vertebrate community structure in amazonian forests. Conserv. Biol. 14, 240–253. https://doi.org/10.1046/j.1523-1739.2000.98485.x.

Peres, C.A., Dolman, P.M., 2000. Density compensation in neotropical primate communities: evidence from 56 hunted and nonhunted Amazonian forests of varying productivity. Oecologia 122, 175–189. https://doi.org/10.1007/PL00008845.

Pompon, J., Morales-Vargas, R., Manuel, M., Huat Tan, C., Vial, T., Hao Tan, J., Sessions, O.M., da Vasconcelos, P.C., Ng, L.C., Missé, D., 2017. A Zika virus from America is more efficiently transmitted than an Asian virus by Aedes aegypti mosquitoes from Asia. Sci. Rep. 7, 1215. https://doi.org/10.1038/s41598-017-01282-6.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P., 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv. Methodol. 27, 85–96.

Rai, P., Hu, C., Henao, R., Carin, L., 2015. Large-scale bayesian multi-label learning via topic-based label embeddings. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural InFormation Processing Systems 28. Curran Associates, Inc., pp. 3222–3230.

Song, S., Li, Y., Fu, S., Liu, H., Li, X., Gao, X., Xu, Z., Liu, G., Wang, D., Tian, Z., Zhou, J., He, Y., Lei, W., Wang, H., Wang, B., Lu, X., Liang, G., 2017. Could Zika virus emerge in Mainland China? Virus isolation from nature in Culex quinquefasciatus, 2016. Emerg. Microbes Infect. 6, e93. https://doi.org/10.1038/emi.2017.80.

WHO, 2018. Zika Virus Infection: India.

Yang, Xing-Lou, Zhang, Yun-Zhi, Jiang, Ren-Di, Guo, Hua, Zhang, Wei, Li, Bei, Wang, Ning, Wang, Li, Waruhiu, Cecilia, Zhou, Ji-Hua, Li, Shi-Yue, Daszak, Peter, Wang, Lin-Fa, Shi, Zheng-Li, 2017. Genetically diverse filoviruses in rousettus and eonycteris spp. bats, China, 2009 and 2015. Emerging Infect. Dis. 23, 482. https://doi.org/10.3201/eid2303.161119.

Zuk, M., Stoehr, A.M., 2002. Immune defense and host life history. Am. Nat. 160 (Suppl. 4), S9–S22. https://doi.org/10.1086/342131.