Mapping the global distribution of zoonoses of public health importance

A thesis submitted for the degree of *Doctor of Philosophy*

David Michael Pigott

Magdalen College, University of Oxford

Trinity 2015

Mapping the global distribution of zoonoses of public health importance

David Michael Pigott

Magdalen College

University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity 2015

Medical cartography can provide valuable insights into the epidemiology and ecology of infectious diseases, providing a quantitative representation of the distribution of these pathogens. Such methods therefore provide a key step in informing public health policy decisions ranging from prioritising sites for further investigation to identifying targets for interventions. By increasing the resolution at which risk is defined, policymakers are provided with an increasingly informed approach for considering next steps as well as evaluating past progress. In spite of their benefits however, global maps of infectious disease are lacking in both quality and comprehensiveness.

This thesis sets out to investigate the next steps for medical cartography and details the use of species distribution models in evaluating global distributions of a variety of zoonotic diseases of public health importance.

Chapter 2 defines a methodology by which global targets for infectious disease mapping can be quantitatively assessed by comparing the global burden of each disease with the demand from national policymakers, non-governmental organisations and academic communities for global assessments of disease distribution. Chapter 3 introduces the use of boosted regression trees for mapping the distribution of a group of vector-borne diseases identified as being a high priority target, the leishmaniases. Chapter 4 adapts these approaches to consider Ebola virus disease. This technique shows that the West African outbreak was ecologically consistent with past infections and suggests a much wider area of risk than previously considered. Chapter 5 investigates Marburg virus disease and considers the variety of different factors relating to all aspects of the transmission cycle that must be considered in these analyses. Chapters 6 and 7 complete the mapping of the suite of viral haemorrhagic fevers by assessing the distribution of Crimean-Congo haemorrhagic fever and Lassa fever. Finally, Chapter 8 considers the risk that these viral haemorrhagic fevers present to the wider African continent, quantifying potential risk of spillover infections, local outbreaks and more widespread infection.

This thesis addresses important information gaps in global knowledge of a number of pathogens of public health importance. In doing so, this work provides a template for considering the global distribution of a number of other zoonotic diseases.

Statement of Contribution and Associated Publications

The chapters presented in this thesis represent articles that have either been published or are in preparation for submission to peer-review journals. Due to the broad nature of these articles, they represent collaborative efforts, involving many members of the Spatial Ecology and Epidemiology group (SEEG) in which I am based, but also individuals from various other institutions across the world. While full author contributions are outlined at the end of many of the articles, I summarise below my own contributions to each chapter.

Chapter 2 - Identifying targets for infectious disease mapping.

This chapter has been accepted for publication and is included in its final form. I was joint first author with REH. I was responsible for overseeing all aspects of the final publication and contributed to the significant revisions of the original analyses and manuscript prepared by REH. The study was conceived through a number of discussions with colleagues at the Bill & Melinda Gates Foundation (SD, THF, AJG, AMK, LKK, CHS, SJB) and members of SEEG involved with the Hay *et al.* (2013) "Global Mapping of Infectious Disease" paper (DMP, REH, KEB, PWG, CLM, SIH). I devised the cluster grouping (with REH) and was responsible for assigning disability adjusted life year (DALY) values (with REH, HHK, TV, CJLM). Policy interest was calculated by REH and AW. I performed the final analysis (with NG) and wrote the final draft of the manuscript, which was subsequently edited and approved by all authors.

Pigott, D. M.*, Howes, R. E.*, Wiebe, A., Battle, K. E., Golding, N., Gething P. W., Dowell, S., Farag, T. H., Garcia, A. J., Kimball, A. M., Krause, L. K., Smith, C. H., Brooker, S. J., Kyu, H. H., Vos, T. Murray, C. J. L., Moyes, C. L. and Hay, S. I. (2015). Prioritising infectious disease mapping. *PLoS Neglected Tropical Diseases*. 9 (6):e0003756 doi:10.371/jounal.pntd.0003756

Chapter 3 - A framework for mapping vector-borne diseases: the leishmaniases.

This chapter has been published and is included in its final form. I was responsible for overseeing all aspects of this work. I conceived the study (with DBG and SIH), generated the evidence consensus layer (with OJB), geopositioned the occurrence dataset (with KAD, KEB, JPM, MFM), liaised with international collaborators (YB, PB, FP, JSB, CCF, SRM) and curated their databases to complement the literature-based survey, implemented the modelling tasks (with SB, NG and PWG), analysed the data (with NG, RR, SIH) and wrote the first draft of the manuscript, which was subsequently edited and approved by all authors.

Pigott, D. M., Bhatt, S., Golding, N., Duda, K. A., Battle, K. E., Brady, O. J., Messina, J. P., Balard, Y., Bastien, P. Pratlong, F., Brownstein, J. S., Freifeld, C. C., Mekaru, S. R., Gething, P. W., George, D. B., Myers, M. F., Reithinger, R. and Hay, S. I. (2014). Global distribution maps of the leishmaniases. *eLife*.3:e02851 doi: 10.7554/eLife.02851.001

Chapter 4 - A framework for mapping zoonotic disease: Ebola virus disease.

This chapter has been published and is included in its final form. I was joint first author with NG. I collected and assembled the outbreak and animal infection data (with AM, ZH, AJH, OJB), analysed the final models (with NG, OJB, MUGK, SIH) and wrote the first draft of the manuscript which was approved and edited by all authors.

Pigott, D. M.*, Golding, N.*, Mylne, A., Huang, Z., Henry, A. J., Weiss, D. J., Brady, O. J., Kraemer, M. U. G., Smith, D. L., Moyes, C. L., Bhatt, S., Gething, P. W., Horby, P. W., Bogoch, I. I., Brownstein, J. S., Mekaru, S. R., Tatem, A. J., Khan, K., Hay, S. I. (2014) Mapping the zoonotic niche of Ebola virus disease in Africa. *eLife*. 3:e04395 doi:10.7554/eLife.04395

Chapter 5 - Understanding risk maps: Marburg virus disease.

This chapter has been published and is included in its final form. I was responsible for overseeing all aspects of this work. I collected and geopositioned the outbreak and animal

infection data (with AM, ZH, OJB), ran the models, analysed the data (with NG, OJB, MUGK, SIH) and wrote the first draft of the manuscript which was approved and edited by all authors.

Pigott, D. M., Golding, N., Mylne, A., Huang, Z., Weiss, D. J., Brady, O. J., Kraemer, M. U. G. and Hay, S. I. (2015). Mapping the zoonotic niche of Marburg virus disease in Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. 109:366-78 doi: 10.1093/trstmh/trv024

Chapter 6 - Mapping a tick-borne zoonotic disease: Crimean-Congo haemorrhagic fever.

This chapter consists of a paper accepted for publication and included in its final form. I was joint first author with JPM. I generated the evidence consensus (with KAD), collated and geopositioned the occurrence database (with JPM, KAD, JSB, MFM) and analysed the model outputs (with JPM, NG, TPR, MG, GRWW, PAN, SIH).

Messina, J. P.*, Pigott, D. M.*, Golding, N., Duda, K. A., Brownstein, J. S., Weiss, D. J., Gibson, H., Robinson, T. P., Gilbert, M., Wint, G. R. W., Nuttall, P. A., Gething, P. W., Myers, M. F., George, D. B. and Hay, S. I. (2015) The global distribution of Crimean-Congo hemorrhagic fever. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. 109:503-513 doi:10.1093/trstmh/trv050

Chapter 7 - Mapping a rodent-borne zoonotic disease: Lassa fever.

This chapter consists of a paper accepted for publication and included in its final form. I was joint first author with AM. I generated the evidence consensus, advised on the occurrence database collection and curation (with AM), analysed the models (with AM and NG) and co-wrote the first draft of the manuscript (with AM) which was approved and edited by all authors.

Mylne, A.*, Pigott, D. M.*, Longbottom, J., Shearer, F., Duda, K. A., Messina, J. P., Weiss, D. J., Moyes, C. L., Golding, N. and Hay, S.I. (2015) Mapping the zoonotic niche of Lassa fever in

Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. 109:483-492 doi:10.1093/trstmh/trv047

Chapter 8 - Translating maps into policy

This chapter has been prepared for submission to a peer-review journal. I was responsible for all aspects of this work. I conceived the study (with SIH), performed the analyses and wrote the first draft of the manuscript. The provisional authorship is as follows:

Pigott, D. M., Golding, N., Messina, J. P., Mylne, A., Shearer, F., Brady, O. J., Kraemer, M. U. G., Bhatt, S. J., Gething, P. W., Weiss, D. J., Moyes, C. L., Tatem, A. J. and Hay, S. I. (2015) The co-distribution of contagious viral haemorrhagic fevers in Africa. in prep

* * *

In addition to these articles, a number of other papers, complementary to this work, were completed. They have been added as appendices as reference materials.

This article describes a survey of the methods available for mapping infectious disease at a global scale and reviews existing disease maps for 174 clinical conditions.

Hay, S. I., Battle, K. E., Pigott, D. M., Smith, D. L., Moyes, C. L., Bhatt, S., Brownstein, J. S., Collier, N., Myers, M. F., George, D. B., Gething, P. W. (2013) Global mapping of infectious disease risk. *Philosophical Transactions of the Royal Society B.* 368:20120250 doi:10.1098/Rstb.2012.0250.

This article provides more detail on the occurrence database creation for the leishmaniasis project.

Pigott, D. M., Golding, N., Messina, J. P., Battle, K. E., Duda, K. A., Balard, Y., Bastien, P., Pratlong, P., Brownstein, J. S., Freifeld, C. C., Mekaru, S. R., Madoff, L. C., George, D. B.,

Myers, M. F. and Hay, S. I. (2014) Global database of leishmaniasis occurrence locations, 1960-2012. *Scientific Data*. 1:140036 doi:10.1038/sdata.2014.36.

This article provides more detail on the occurrence database creation for the Ebola virus disease project and provides additional information on the geographic spread of past outbreaks.

Mylne, A.*, Brady, O. J.*, Huang, Z., Pigott, D. M., Golding, N., Kraemer, M. U. G. and Hay, S. I. (2014) A comprehensive database of the geographic spread of past human Ebola outbreaks. *Scientific Data*. 1:140042. doi:10.1038/sdata.2014.42.

This article provides more detail on the occurrence database creation for the Crimean-Congo haemorrhagic fever project.

Messina, J. P., Pigott, D. M., Duda, K. A., Brownstein, J. S., Myers, M. F., George, D. B. and Hay, S. I. (2015) A global compendium of human Crimean-Congo haemorrhagic fever virus occurrence. *Scientific Data*. 2:150016. doi:10.1038/sdata.2015.16

* * *

The following articles incorporate the outputs from this thesis:

Pigott, D. M. and Kraemer, M. U. G. (2014) Enhancing infectious disease mapping with open access resources. *Eurosurveillance*. 19:e20989.

Golding, N., Moyes, C. L., Wilson, A. L., Cano, J., Pigott, D. M., Velayudhan, R., Brooker, S. J., Smith, D. L., Hay, S. I. and Lindsay, S. W. (2015) Integrating vector control across diseases. *BMC Medicine*. In press.

* * *

The following papers were produced in parallel to this thesis, but are not directly related to it:

Brady, O. J., Johansson, M. A., Guerra, C. A., Bhatt, S., Golding, N., Pigott, D. M., Delatte, H., Grech, M. G., Leisnham, P. T., Maciel-de-Freitas, R., Styer, L. M., Smith, D. L., Scott, T. W., Gething, P. W. and Hay, S. I. (2013) Modelling adult *Aedes aegypti* and *Aedes albopictus* survival at different temperatures in laboratory and field settings. *Parasites and Vectors*. 6:351 doi:10.1186/1756-3305-6-351.

Brady, O. J., Golding, N., Pigott, D. M., Kraemer, M. U. G., Messina, J. P., Reiner, R. C., Scott, T. W., Smith, D. L., Gething, P. W. and Hay, S. I. (2014) Global temperature constraints on *Aedes aegypti* and *Ae. albopictus* persistence and competence for dengue virus transmission. *Parasites and Vectors*. 7:388 doi:10.1186/1756-3305-7-338.

Messina, J. P., Brady, O. J., Scott, T. W., Zou, C. T., Pigott, D. M., Duda, K. A., Bhatt, S., Katzelnick, L., Howes, R. E., Battle, K. E., Simmons, C. P. and Hay, S. I. (2014) Global spread of dengue virus types: mapping the 70 year history. *Trends in Microbiology*. 22 (3):138-146 doi:10.1016/j.tim.2013.12.011.

Messina, J. P., Brady, O. J., Pigott, D. M., Brownstein, J. S., Hoen, A. G. and Hay, S. I. (2014) A global compendium of human dengue virus occurrence. *Scientific Data*. 1:140004 doi:10.1038/sdata.2014.4.

Messina, J. P., Brady, O. J., Pigott, D. M., Golding, N., Kraemer, M. U. G., Scott, T. W., Wint, G. R. W., Smith, D. L. and Hay, S. I. (2015) The many projected futures of dengue. *Nature Reviews Microbiology*. 13 (4):230-239 doi:10.1038/nrmicro3430.

Reiner, R. C., Perkins, T. A., Barker, C. M., Niu, T, Chaves, L. F., Ellis, A. M., George, D. B., Le Menach, A., Pulliam, J., Bisanzio, D., Buckee, C., Chiyaka, C., Cummings, D. A. T., Garcia, A. J., Gatton, M. L., Gething, P. W., Hartley, D. M., Johnston, G., Klein, E. Y., Michael, E., Linsday, S. W., Lloyd, A. L., Pigott, D. M., Reisen, W. K., Ruktanonchai, N., Singh, B., Tatem,

A. J., Kitron, U., Hay, S. I., Scott, T. W. and Smith, D. L. (2013) A systematic review of mathematical models of mosquito-borne pathogen transmission: 1970-2010. *Journal of the Royal Society Interface*. 10:20120921 doi:10.1098/Rsif.2012.0921.

Smith, D. L., Perkins, T. A., Reiner, R. C., Barker, C. M., Niu, T. C., Chaves, L. F., Ellis, A. M., George, D. B., Le Menach, A., Pulliam, J. R. C., Bisanzio, D., Buckee, C., Chiyaka, C., Cummings D. A. T., Garcia, A. J., Gatton, M. L., Gething, P. W., Hartely, D. M., Johnston, G., Klein, E. Y., Michael, E., Lloyd, A. L., Pigott, D. M., Reisen, W. K., Ruktanonchai, N., Singh, B. K., Stoller, J., Tatem, A. J., Kitron, U., Godfray, H. C. J., Cohen, J. M., Hay, S. I. and Scott, T. W. (2014) Recasting the theory of mosquito-borne pathogen transmission dynamics and control. *Transactions on the Royal Society of Tropical Medicine and Hygiene*. 108 (4):185-197 doi:10.1093/trstmh/tru026.

* * *

As his supervisor, I certify that the statements of contribution listed here are a fair representation of David Pigott's work.

Prof Simon Hay

10th June 2015

Table of Contents

Appendix

Acknowledgements

I would like to thank my supervisor and committee for their support over the last three years, Simon Hay, Peter Gething, Samir Bhatt, Nick Golding and Angela McLean. They have all provided encouragement and insight over this period and have helped with all aspects of my work, from statistical analyses to tips on writing styles.

In addition, I would like to thank my colleagues, and friends, who have put up with my presence in the office – from Kate, Fred and Ros who were there from the research assistant days to Adrian, Bonnie, Dan, Donal, Ewan, Freya, Janey, Moritz, Oli and Ursula who were there towards the end.

I have Desmond Morris to thank for inadvertently starting this all off through a brief chat with my mother when he came into the bank to cash in some cheques. I must also thank Angela McLean, Charles Godfray and the Christensen Fund for providing financial support through the Sir Richard Southwood Graduate Scholarship.

Finally I have to thank my Mum and Dad for always being there for me over the last twenty-five years. I would not have been able to reach this point without their unconditional love and support.

Abbreviations

AFRO – African Region (WHO Regional Office)

AMRO – Region of the Americas (WHO Regional Office)

APOC – African Programme for Onchocerciasis Control

BRT – Boosted regression trees

CAR – Central African Republic

CCHF – Crimean-Congo haemorrhagic fever

CCHFV – Crimean-Congo haemorrhagic fever virus

CL – Cutaneous leishmaniasis

DALY – Disability adjusted life year

DRC – Democratic Republic of Congo

EBOV – Ebola virus

EMRO – Eastern Mediterranean Region (WHO Regional Office)

EURO – European Region (WHO Regional Office)

EVD – Ebola virus disease

GAHI – Global atlas of helminth infections

GAT – Global Atlas of Trachoma

GBD – Global Burden of Disease

GBIF – Global Biodiversity Information Facility

GIDEON – Global Infectious Diseases and Epidemiology Online Network

HIV – Human Immunodeficiency Virus

INFORM – Information for risk management

IUCN – International Union for Conservation of Nature

LASV – Lassa virus

MAP – Malaria Atlas Project

MARV – Marburg virus

MBG – Model-based geostatistics

MVD – Marburg virus disease

NW – New World

OW – Old World

PAR – Population at risk

SDM – Species distribution models

SEARO – South-East Asian Region (WHO Regional Office)

SEEG – Spatial Ecology and Epidemiology Group

STH – Soil-transmitted helminth

VBD – Vector-borne disease

VHF – Viral haemorrhagic fever

VL – Visceral leishmaniasis

WHO – World Health Organization

WPRO – Western Pacific Region (WHO Regional Office)

YLD – Years lived with disability

YLL – Years of life lost

Chapter 1

Introduction

1.1. Diseases and Mapping

Using maps to aid epidemiological investigations is not a new phenomenon *(Koch, 2011; Koch, 2014)*. The first reported disease maps date from 1796 where they were used to investigate an outbreak of Yellow Fever in New York City *(Koch, 2011)*. In 1854 however, the real utility of this approach was demonstrated; John Snow famously mapped the cases of cholera in Soho, London, and indicated that they were clustered around a water pump on Broad Street *(Snow, 1855)*. In doing so, he demonstrated the first clear example of medical cartography informing public health intervention as access to the pump was subsequently denied and the cases declined. Even global maps were published during this time period, charting the spread of cholera across the Old World *(Koch, 2014; Lancet editorial, 1831)*. Since these days, the role of the disease map, and importantly, the information that they can convey has evolved well beyond mere visualisation to being a key component in downstream analyses and in forming an evidence-base for informing public health policy decisions *(Hay et al., 2013b)*.

Advances in statistical techniques and computing power, coupled with increasingly detailed information, at ever higher spatial and temporal resolutions, have meant that maps, rather than acting as a means for investigating trends, can themselves provide answers and solutions. We can now incorporate information on vector life history traits in order to spatially delimit transmission of vector-borne diseases *(Brady et al., 2014; Gething et al., 2011b)*; we can use prevalence and incidence data in order to provide continuous spatial surfaces of transmission risk and identify areas where incidence and mortality rates are expected to be highest *(Gething et al., 2012; Gething et al., 2011a; Hay et al., 2010)*; and we can use information from existing cases to infer areas of potential risk based upon shared environmental similarities *(Bhatt et al.,*

2013; Gilbert et al., 2014). Maps can be used to understand a number of aspects of disease biology ranging from determinants of disease macroecology *(Guernier et al., 2004)*, to disease emergence *(Jones et al., 2008)* and potential antimicrobial resistance *(Van Boeckel et al., 2015)*. As their use becomes more widespread, some outputs have now become integrated into important global health policy documents *(CDC, 2013; WHO, 2014)*.

1.2. Methods in disease cartography

A variety of methodological approaches have now been developed to evaluate the spatial and temporal distributions of diseases. At their most basic, maps can simply display data. This can include data from specific geographic areas, such as location of reported cases or prevalence in a defined population as well as data aggregated to a geographic or political unit *(Alvar et al., 2012)*. Such maps, whilst useful as a display of initial information, are of limited use. The aim should be to use and interpret these data to identify what factors influence a given distribution or whether data gaps represent true absence of cases, or absence of reporting, often through modelling approaches. In some circumstances however, this is the only option available given particular data constraints *(Hay et al., 2013a)*.

More advanced statistical techniques now allow for inference from existing data into areas where reported cases or surveys are not present. A vast array of approaches are available to extrapolate and interpolate from existing data *(Pfeiffer et al., 2008)*, using such approaches to produce smoothed surfaces of variation in data, or more sophisticated means of relating spatial variation to underlying causes, correlates and covariates *(Bhatt et al., 2013; Gething et al., 2011a)*. Advances in computing power combined with much richer streams of data have meant that increasingly computationally intensive statistical approaches can now be implemented *(Hay et al., 2013a; Hay et al., 2013b)*. Bearing this in mind, two groups of models are becoming more widely used in mapping diseases across larger spatial scales.

1.3. Species distribution models

One such sophisticated analytical technique is species distribution modelling (SDM, also referred to as ecological niche modelling) which has been co-opted from its initial use in understanding the distribution of plants and animals for use in disease mapping. A variety of different approaches have been developed but they are essentially variations on a similar theme. SDMs compare reported occurrences of the species of interest to environmental covariates at these locations to define an environmental profile that describes areas of reported presence. This ruleset can then be applied to locations for which data is available on environmental correlates, but not presence or absence of the species, and the location's environment is compared to that describing where reports have occurred *(Elith and Leathwick, 2009; Franklin, 2009)*. The model therefore attempts to spatially define the niche of a given species *(Peterson et al., 2011)*.

These various models can be divided into four main approaches; presence-only, presence-absence, presence-background and presence-pseudoabsence *(Peterson et al., 2011)*. Whilst retaining the same core concept, these approaches vary based upon the input data provided. All the models require input occurrence data, or presence data, which are records of the species of interest at a particular location. Presence-only approaches (such as BIOCLIM *(Busby, 1991)*) use this data to fit "environmental envelopes" enclosing known occurrences, independent of other information *(Peterson et al., 2011)*.

The other methods contrast this information with alternative datasets (background, pseudoabsence and absence data) in order to produce a more refined estimate of a species' distribution. Absence records refer to data on the species of interest where its absence has been noted during collection survey. Background data, in contrast, is randomly sampled from across the entire geographic space that is to be modelled. Pseudoabsence records provide a more rational way of generating background data – here sampling is biased towards locations that have already been determined as being unsuitable for the species present. Given the similarities

in the way the data is utilised, some models can be adapted to use all three data types but care must be taken in interpreting the results *(Fithian et al., 2015; Guillera-Arroita et al., 2014).* Over the last decade a number of regression-related, machine learning approaches have been developed and they have been shown to outperform older techniques *(Elith et al., 2006)* with boosted regression trees (BRT) *(Elith et al., 2008),* MaxEnt *(Elith et al., 2011; Phillips et al., 2006),* generalised dissimilarity models *(Ferrier et al., 2002a; Ferrier et al., 2002b)* and multivariate adaptive regression splines *(Leathwick et al., 2005)* being the four methods shown to have the greatest discriminatory power. The relative merit of each of these approaches is heavily dependent on the modelling scenario and data content *(Elith and Graham, 2009).*

Integrating diseases into species distribution modelling approaches is non-trivial. Extending niche theories to pathogens becomes complex, particularly when multiple species are involved in the transmission cycle *(Peterson et al., 2011).* Visual aids such as BAM diagrams have been used to help characterise this problem (Figure 1.1) *(Peterson, 2008)* – here the presence of a species can be considered as the intersection of three overlapping circles representing suitable space: one defining areas with suitable biotic interactions (B); one defining suitable abiotic conditions (A); and the other defined by the species dispersal ability (M). Once additional species are incorporated, the problem becomes significantly more complex and difficult to represent diagrammatically. When interpreting SDM-based risk maps it is therefore important to consider how suitable the use of SDMs was, particularly given the BAM determinants of any given distribution and where the model may have failed to account for one aspect. Often compromises and assumptions must be made during modelling and these may subtly change what the resulting prediction map conveys.

Recently BRTs have been increasingly used to map diseases *(Gilbert et al., 2014; Martin et al., 2011),* particularly with the inclusion of randomly generated pseudo-data *(Bhatt et al., 2013).* The model is consistently regarded as one of the best performers in a variety of scenarios *(Elith et al., 2006)* and has been shown to be robust when dealing with complex epidemiological

systems *(Bhatt et al., 2013; Leathwick et al., 2008)*. Boosted regression trees combine regression trees, where decision rules on environmental predictors are based upon data apportioning resulting from binary splits *(De'ath, 2007)*, with boosting, which selects sets of trees that maximise fit to the data *(Elith et al., 2008)*. In doing so, a statistically robust characterisation of the environmental space occupied by the input data can be achieved.

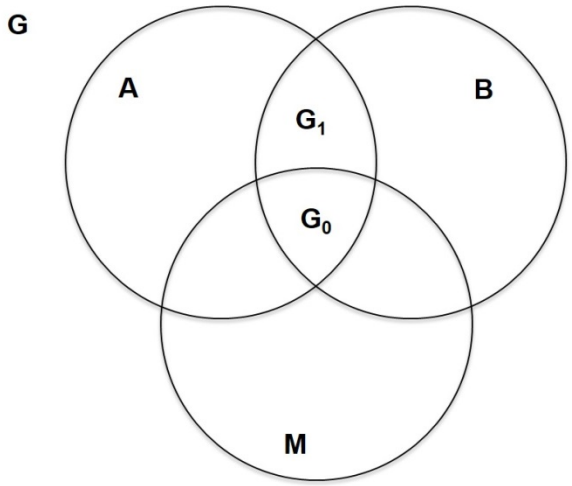


Figure 1.1. Visualising species distributions using BAM diagrams. Here the three circles represent the range of possible values for a given species. A indicates areas abiotically suitable, B defines regions where biotic interactions allow for the species to be present and M outlines regions where the organism can disperse or move to. G represents all potential geographic space. G_1 denotes the region where any two circles overlap, whilst G_0 represents the intersection of all three. G_0 therefore denotes the geographic space where the species will be distributed (picture credit: Freya Shearer).

Since robust information on the absence of diseases is generally unavailable (particularly at a global scale) applying BRT and similar approaches to disease mapping requires the use of pseudo-data. Bias in underlying datasets is a key problem for all SDMs *(Phillips et al., 2009)*, however methods have been developed to minimise their impact within pseudo-data generation. The incorporation of independent measures of disease status *(Brady et al., 2012)* as well as

ensembling a variety of different model parameter permutations can allow for more robust predictions *(Bhatt et al., 2013)*.

The use of SDMs for disease mapping is a relatively new field and as such is still developing in terms of defining best practice for modelling, interpretation and visualisation. An awareness of the limitations of these models is therefore crucial when interpreting results from any given output. There is definite utility however in providing a first attempt at spatial assessments of disease risk as it acts as a means for identifying information gaps and areas for future research. Continued refinement of these surfaces with new and diverse data sources will only help aid our knowledge of disease epidemiology *(Plowright et al., 2015)*.

1.4. Model-based geostatistics

In spite of new developments in SDMs, when large, high quality detailed datasets are available, techniques involving the use of model-based geostatistics (MBG) are preferred. Disease distributions are rarely at equilibrium, particularly when considering infectious diseases that are under a variety of both environmental and human induced pressures. Rather than disease presence remaining consistent over long periods of time, we have seen significant shifts in distributions in response to both natural and human changes *(Gething et al., 2010)*. However, SDMs require the assumption that diseases are in equilibrium (or additional information that can compensate for this *(Brady et al., 2012)*). In addition, the link between SDM suitability outputs and key clinical information such as prevalence and incidence is still uncertain *(Bhatt et al., 2013)*. For many diseases the availability of population representative datasets is the biggest limiting factor. However for a select few, significant global datasets are available *(Global Atlas of Helminth Infections, 2014; Moyes et al., 2013)*, and these facilitate the use of MBG approaches. These methods use more data intensive statistical techniques which can accommodate additional spatial autocorrelation that influence the spatial and temporal distribution of disease *(Gething et al., 2011a)*.

Model-based geostatistics is used as a catch-all term for these more detailed modelling approaches *(Diggle and Ribeiro, 2010)*, the most advanced of which invoke Bayesian inference to inform disease cartography. These models combine not only information on climatic and human covariates, but also spatial relationships to interpret data and make informed inferences to areas of unknown status. Bayesian approaches have the additional benefit of being able to use informative priors to guide the modelling process (although their inclusion is not essential) *(Basanez et al., 2004; Patil et al., 2011)*.

This approach can evaluate a number of different epidemiological values such as disease prevalence rates *(Gething et al., 2012; Gething et al., 2011a)* and disease intensity *(Clements et al., 2009)*. Importantly, given the flexibility of the model, information on dynamic processes such as public health intervention measures *(Bhatt et al., 2015)* can be incorporated, allowing for non-equilibria states to be considered alongside long-term, more stable parameters such as temperature *(Gething et al., 2011b)*. This ability to combine various data types, quantify epidemiological parameters and propagate uncertainty throughout makes MBG techniques very powerful tools. However, the inclusion of a variety of disparate sources of variation, combined with the detailed information required to inform the model mean that data acquisition is often the limiting factor. This trade-off between modelling complexity and data requirements represents a significant challenge to the disease mapping community.

1.5. Reviewing global mapping of infectious disease

Given data availability is a critical determinant of the suitability of these different mapping approaches, there is a need to provide a comprehensive guide to assessing the appropriateness of these different techniques for a given dataset. A recent survey conducted by myself and colleagues defined a data-driven means for categorising infectious diseases *(Hay et al., 2013a)*. Five distinct categories were outlined (Figure 1.2).

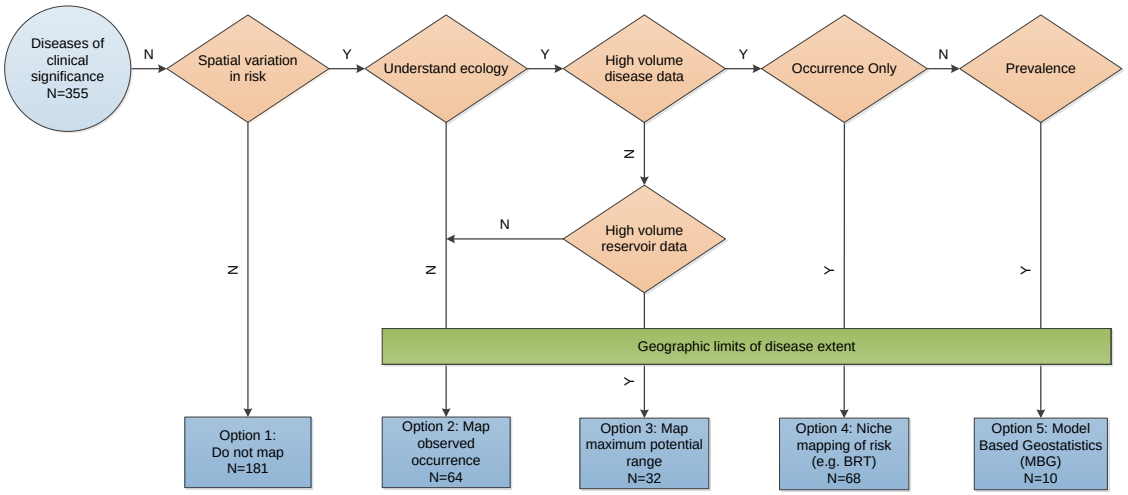


Figure 1.2. Classifying infectious diseases by mapping approach. 355 infectious diseases can be divided into five groupings based upon this decision tree. Their suggested mapping outputs are outlined (from Hay *et al.* (2013a)).

Option 1 describes those diseases that show no spatial variation in occurrence globally – essentially their distribution is tied to that of the human populations they infect *e.g.* Epstein Barr virus *(Macsween and Crawford, 2003)*. Whilst some of these diseases do indeed show spatial trends in prevalence (for instance meningococcal meningitis *(Trotter and Greenwood, 2007)*) for many there is insufficient data present to reliably categorise this at a global level (contrast this with HIV/AIDS for instance *(UNAIDS, 2014)*). Options 2 to 5 therefore represent infectious diseases that do show spatial variation in their distribution.

Availability and access to appropriate data is key for determining the relative categorisation for each disease. Hay *et al.* (2013a) used the number of PubMed search results as a proxy for the amount of information available for creating maps for these diseases. Using the Global Infectious Disease and Epidemiology Online Network (GIDEON) as a guide for the number of countries that have reported cases *(Edberg, 2005)*, we estimated the number of potential occurrences suitable for mapping per country by dividing the total PubMed hits by the total number of endemic countries. If this was below 25 occurrences per country, Hay *et al.* recommended plotting these reported cases as locations on a map or summarise number of case

reports by administrative unit status (*i.e.* state or province-level records). This was classified as Option 2. In instances where we have sufficient information on the transmission cycle to implicate a reservoir host or vector species, it is possible to determine a maximum extent by defining the range of the reservoir or vector. Pathogens that could be described in this way where classed as Option 3 diseases. Quantitative estimates of the distributions of many arthropod vectors are increasingly available *(Kraemer et al., 2015; Sinka et al., 2012)* and knowledge of the ranges of mammalian species has been the focus of large international conservation projects resulting in a much richer information base than that available for many diseases *(Schipper et al., 2008).* As a result we can begin to define the sub-national variation in these conditions by first looking at their vector and reservoir species.

Where sufficient occurrence data was available, we could use SDMs to better understand sub-national variation, Hay *et al.* classified these pathogens as Option 4 diseases. Here we can use global environmental covariate sources *(Hay et al., 2006; Hijmans et al., 2005; Weiss et al., 2014)* to infer from existing cases the likely presence of a disease in other locations due to shared environmental conditions, both climatic and relating to human factors. For some diseases, widespread and rich sources of information incorporating measures of disease prevalence and incidence are available *(Hay and Snow, 2006),* which allows for the use of MBG. These conditions, which include soil-transmitted helminths, tuberculosis, lymphatic filariasis, HIV and malaria, were categorised as Option 5 diseases.

We can therefore consider all diseases that show spatial variation within this framework of five categories. As novel data sources continue to be exploited to obtain disease occurrence data *(Brownstein et al., 2009)* and routine disease reporting becomes more comprehensive, more diseases will continue to be added to the Option 5 category.

1.6. Contemporary status of global diseases maps

Having categorised all infectious diseases, an important next step was to evaluate the standard of existing maps. As part of the review process Hay *et al.* (2013a) assessed the global mapping status of 176 diseases of clinical significance. Initially, existing disease maps were collated. Each map was evaluated on the quality of the data used (measured by contemporariness of the data, diagnostic accuracy and geographical precision), relative population assessed (population living in areas covered by the map divided by those living in countries considered by GIDEON as being at risk) as well as the methodology used (evaluated within the context of the diseases' Option classification). In total only 7% of infectious diseases had global maps that were defined as "sufficient". These were dengue *(Rogers et al., 2006)*, Lassa fever *(Fichet-Calvet and Rogers, 2009)*, mayaro *(Powers et al., 2006)*, monkeypox *(Levine et al., 2007)*, *Plasmodium falciparum* malaria *(Gething et al., 2011a)* and *P. vivax* malaria *(Gething et al., 2012)*. Despite the utility of such outputs, the paucity of high quality global maps is both surprising and disappointing. Figure 1.3 highlights just how much opportunity there is for improvements to disease mapping with the plot dominated by blank space. A large number of unmapped conditions fall within the Option 4 classification – using SDMs to map infectious diseases therefore could have a significant impact on global understanding of what drives pathogen distributions.

1.7. Aims and structure of the thesis

Using this review as a call to action, this thesis investigates a number of techniques that can be used not only to produce disease maps but also to inform future global health and policy discussions. Chapter 2 outlines a methodology for prioritising infectious diseases for mapping. Unlike existing platforms *(Krause and Prioritisation Working Group, 2008)*, this approach integrates quantitative information rather than qualitative or expert opinion based relative categorisation. Diseases were grouped together into clusters based upon shared taxonomic and epidemiological similarities. This was done since mapping projects would likely be able to share outputs or covariates based upon these similarities. Leveraging these shared resources will

hopefully speed up the mapping process for related conditions. Constituent diseases within these clusters were evaluated based upon their burden on public health (measured as disability-adjusted life years (DALYs) *(GBD 2013 Disease and Injury Incidence and Prevalence Collaborators, 2015; GBD 2013 Mortality and Causes of Death Collaborators, 2015)*) and measures of perceived demand within the public health sector, determined using disease-specific h-indices, disease notification statuses and inclusion in non-governmental organisation mission statements. The resulting relative ranking not only recognises the importance of HIV, Tuberculosis and Malaria as key infectious diseases for mapping, but also identifies a number of important groupings, particularly within the neglected tropical diseases *(Hotez, 2013; WHO, 2009, 2010, 2012)*, reinforcing calls for their more mainstream inclusion in policy settings.

Chapter 3 introduces SDMs to mapping infectious diseases. The Hay *et al.* (2013a) review and Chapter 2 identified the significant but often neglected impact that zoonotic diseases can have on human health. The leishmaniases, identified as one of the top 10 clusters to be prioritised, represent an intriguing test case – not only are large amounts of data available *(Pigott et al., 2014)*, but a large number of reservoir, vector and *Leishmania* species are involved in the transmission cycle. The flexibility of BRTs, combined with the use of evidence consensus at a sub-national level, can account for much of this complexity. The outputs presented are the first evidence-based maps of the distributions of these conditions presented at a global scale and provide an important first step in further refining our understanding of these pathogens. A variety of different information streams exist *(Foley et al., 2012; Schipper et al., 2008)* that can be incorporated in future iterations to strengthen their predictive value.

Figure 1.3. Radial plot indicating completeness of contemporary disease mapping efforts. Existing maps for each disease were evaluated and their score plotted radially in black. The diseases with the most comprehensive maps will have near complete black segments; mostly background blank space in a segment indicates that existing coverage is poor (adapted from Hay *et al.* (2013a). Picture credit: Nick Golding).

The outbreak of Ebola virus disease in West Africa was a wake-up call *(Chan, 2014; Farrar and Piot, 2014)*. As Peter Piot described, West Africa represented a "perfect storm" of dysfunctional governance in countries not previously thought to be at risk of filoviral disease *(Piot, 2014)*. Chapter 4 therefore sets out to understand the spatial processes that may influence

areas at risk of Ebola virus disease spillover events - viral transmission from animal reservoirs to humans. As we have seen from genetic evidence *(Dudas and Rambaut, 2014; Gire et al., 2014)*, the West African outbreak came from just one such spillover event with subsequent cases driven solely by human-to-human transmission. Characterising the areas where similar outbreaks could originate is therefore of importance for limiting the potential public health impact of future haemorrhagic fever outbreaks. We incorporated occurrence data on three bat species implicated as reservoir hosts *(Olival and Hayman, 2014)* into a BRT approach in order to identify regions at risk of zoonotic spillover events. This was the first study to show that the Guinean index case was in a region environmentally suitable for Ebola virus persistence in animal reservoirs and also identified a much wider area than first thought as being at similar risk of viral transmission. As the current outbreak reaches an endpoint, these outputs can be used to guide surveillance, particularly in areas of predicted risk where no infection has previously been reported.

Ebola virus disease represents just one contagious viral haemorrhagic fever in Africa. Marburg virus disease (caused by a related filovirus) shares similar characteristics to its sister virus in having bat reservoir hosts and its capacity to cause outbreaks by secondary human-to-human transmission *(Towner et al., 2006)* as well as long distance infections in returning travellers *(Fujita et al., 2009; Timen et al., 2009)*. Chapter 5 uses SDMs to predict the geographical extent of the zoonotic niche of Marburg virus disease, again suggesting large swathes of Africa are environmentally suitable for disease transmission. The number of occurrence records available for mapping this virus is limited due to the rarity of the condition. As a result this chapter also provides an important discussion in identifying important information gaps in how the virus spreads in reservoir populations and how humans interact with these species. The differences between models that include animal data and those that do not suggests that mapping zoonotic disease needs to account not only for where the pathogen is distributed within the reservoir hosts, but also where humans interact with these hosts.

In addition to Ebola and Marburg viruses, Crimean-Congo haemorrhagic fever (CCHF) and Lassa fever share life history patterns and pose similar risks to both local and international communities *(Bannister, 2010)*. Chapters 6 and 7 describe how BRTs were used to predict their disease distributions, re-introducing evidence consensus techniques, as well as refining covariate selection to include land cover types (reflecting the requirements for CCHF tick vectors) and rodent reservoir hosts (specifically the Natal multimammate mouse, *Mastomys natalensis*, which carries Lassa virus [LASV]).

Bridging the gap between primary research and policymakers is a critical step in the scientific process, but is often underappreciated. Chapter 8 analyses the risk that these African viral haemorrhagic fevers pose to public health by pairing the potential zoonotic niche maps for each disease with information assessing the healthcare infrastructure and vulnerability to crises of at-risk countries. This provides a rationale for targeting specific countries based upon the inherent risk of an outbreak starting and the ability of a given nation to stop further transmission – information which can play a key role in future preparedness planning.

This thesis consists of six papers of which five have either been published, accepted for publication or are in press at peer-reviewed scientific journals and one chapter which is prepared for submission.

1.8. References

Alvar J, Velez ID, Bern C, Herrero M, Desjeux P, et al. 2012. Leishmaniasis worldwide and global estimates of its incidence. *PLoS One* **7**: e35671. doi:10.1371/journal.pone.0035671.

Bannister B. 2010. Viral haemorrhagic fevers imported into non-endemic countries: risk assessment and management. *Brit Med Bull* **95**: 193-225. doi:10.1093/Bmb/Ldq022.

Basanez MG, Marshall C, Carabin N, Gyorkos T, Joseph L. 2004. Bayesian statistics for parasitologists. *Trends Parasitol* **20**: 85-91. doi:10.1016/J.Pt.2003.11.008.

Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, et al. 2013. The global distribution and burden of dengue. *Nature* **496**: 504-507. doi:10.1038/Nature12060.

Bhatt S, Weiss DJ, Mappin B, Dalrymple U, Cameron E, et al. 2015. Insecticide-treated nets (ITNs) in Africa 2000-2017: coverage, system efficiency and future needs to achieve international targets. *eLife*: in submission.

Brady OJ, Gething PW, Bhatt S, Messina JP, Brownstein JS, et al. 2012. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLoS Negl Trop Dis* **6**: e1760. doi:10.1371/Journal.Pntd.0001760.

Brady OJ, Golding N, Pigott DM, Kraemer MU, Messina JP, et al. 2014. Global temperature constraints on *Aedes aegypti* and *Ae. albopictus* persistence and competence for dengue virus transmission. *Parasit Vectors* **7**: 338. doi:10.1186/1756-3305-7-338.

Brownstein JS, Freifeld CC, Madoff LC. 2009. Digital disease detection - harnessing the web for public health surveillance. *N Engl J Med* **360**: 2153-2157. doi:10.1056/Nejmp0900702.

Busby JR. 1991. BIOCLIM - a bioclimate analysis and prediction system. *Plant Prot Q* **6**: 8-9.

CDC. 2013. CDC health information for international travel 2014. New York: Oxford University Press. 688 p.

Chan M. 2014. Ebola virus disease in West Africa - no early end to the outbreak. *N Engl J Med* **371**: 1183-1185. doi:10.156/NEJMp1409859.

Clements ACA, Firth S, Dembele R, Garba A, Toure S, et al. 2009. Use of Bayesian geostatistical prediction to estimate local variations in *Schistosoma haematobium* infection in western Africa. *Bull World Health Organ* **87**: 921-929. doi:10.2471/Blt.08.058933.

De'ath G. 2007. Boosted trees for ecological modeling and prediction. *Ecology* **88**: 243-251. doi:10.1890/0012-9658(2007)88[243:Btfema]2.0.Co;2.

Diggle PJ, Ribeiro PJ. 2010. Model-based geostatistics: Springer. 246 p.

Dudas G, Rambaut A. 2014. Phylogenetic analysis of Guinea 2014 EBOV ebolavirus outbreak. *PLoS Curr* **6**: ecurrents.outbreaks.84eefe85ce43ec89dc80bf0670f0677b0678b0417d. doi:10.1371/currents.outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d.

Edberg SC. 2005. Global infectious diseases and epidemiology network (GIDEON): a world wide web-based program for diagnosis and informatics in infectious diseases. *Clin Infect Dis* **40**: 123-126. doi:10.1086/426549.

Elith J, Graham CH. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography* **32**: 66-77. doi:10.1111/j.1600-0587.2008.05505.x.

Elith J, Graham CH, Anderson RP, Dudik M, Ferrier S, et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**: 129-151. doi:10.1111/j.2006.0906-7590.04596.x.

Elith J, Leathwick JR. 2009. Species distribution models: ecological explanation and prediction across space and time. *Annu Rev Ecol Evol Syst* **40**: 677-697. doi:10.1146/annurev.ecolsys.110308.120159.

Elith J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. *J Anim Ecol* **77**: 802-813. doi:10.1111/j.1365-2656.2008.01390.x.

Elith J, Phillips SJ, Hastie T, Dudik M, Chee YE, et al. 2011. A statistical explanation of MaxEnt for ecologists. *Divers Distrib* **17**: 43-57. doi:10.1111/j.1472-4642.2010.00725.x.

Farrar JJ, Piot P. 2014. The Ebola emergency - immediate action, ongoing strategy. *N Engl J Med* **371**: 1545-1546. doi:10.1056/Nejme1411471.

Ferrier S, Drielsma M, Manion G, Watson G. 2002a. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodivers Conserv* **11**: 2309-2338. doi:10.1023/A:1021374009951.

Ferrier S, Watson G, Pearce J, Drielsma M. 2002b. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodivers Conserv* **11**: 2275-2307. doi:10.1023/A:1021302930424.

Fichet-Calvet E, Rogers DJ. 2009. Risk maps of Lassa fever in West Africa. *PLoS Negl Trop Dis* **3**: e388. doi:10.1371/journal.pntd.0000388.

Fithian W, Elith J, Hastie T, Keith DA. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods Ecol Evol* **6**: 424-438. doi:10.1111/2041-210x.12242.

Foley DH, Wilkerson RC, Dornak LL, Pecor DB, Nyari AS, et al. 2012. SandflyMap: leveraging spatial data on sand fly vector distribution for disease risk assessments. *Geospat Health* **6**: S25-S30.

Franklin J. 2009. Mapping species distributions. Cambridge: Cambridge University Press. 320 p.

Fujita N, Miller A, Miller G, Gershman K, Gallagher N, et al. 2009. Imported case of Marburg hemorrhagic fever - Colorado, 2008. *MMWR Morb Mortal Wkly Rep* **58**: 1377-1381. doi:mm5849a2.

GBD 2013 Disease and Injury Incidence and Prevalence Collaborators. 2015. Global, regional, and national incidence, prevalence and YLDs for 301 acute and chronic diseases and injuries for 188 countries, 1990-2013: a systemtatic analysis for the Global Burden of Disease Study 2013. *Lancet*: in press.

GBD 2013 Mortality and Causes of Death Collaborators. 2015. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematics analysis for the Global Burden of Disease Study 2013. *Lancet* **385**: 117-171.

Gething PW, Elyazar IRF, Moyes CL, Smith DL, Battle KE, et al. 2012. A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. *PLoS Negl Trop Dis* **6**: e1814. doi:10.1371/Journal.Pntd.0001814.

Gething PW, Patil AP, Smith DL, Guerra CA, Elyazar IRF, et al. 2011a. A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar J* **10**: 378. doi:10.1186/1475-2875-10-378.

Gething PW, Smith DL, Patil AP, Tatem AJ, Snow RW, et al. 2010. Climate change and the global malaria recession. *Nature* **465**: 342-U394. doi:10.1038/Nature09098.

Gething PW, Van Boeckel TP, Smith DL, Guerra CA, Patil AP, et al. 2011b. Modelling the global constraints of temperature on transmission of *Plasmodium falciparum* and *P. vivax*. *Parasit Vectors* **4**: 92. doi:10.1186/1756-3305-4-92.

Gilbert M, Golding N, Zhou H, Wint GR, Robinson TP, et al. 2014. Predicting the risk of avian influenza A H7N9 infection in live-poultry markets across Asia. *Nat Commun* **5**: 4116. doi:10.1038/ncomms5116.

Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, et al. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**: 1369-1372. doi:10.1126/science.1259657.

Global Atlas of Helminth Infections. GAHI: global atlas of helminth infections. Available: http://www.thiswormyworld.org/. Accessed: July 2014

Guernier V, Hochberg ME, Guegan JFO. 2004. Ecology drives the worldwide distribution of human diseases. *PLoS Biol* **2**: 740-746. doi:10.1371/journal/pbio.0020141.

Guillera-Arroita G, Lahoz-Monfort JJ, Elith J. 2014. Maxent is not a presence-absence method: a comment on Thibaud *et al. Methods Ecol Evol* **5**: 1192-1197. doi:10.1111/2041-210x.12252.

Hay SI, Battle KE, Pigott DM, Smith DL, Moyes CL, et al. 2013a. Global mapping of infectious disease. *Philos Trans R Soc Lond B Biol Sci* **368**: 20120250. doi:10.1098/Rstb.2012.0250.

Hay SI, George DB, Moyes CL, Brownstein JS. 2013b. Big data opportunities for global infectious disease surveillance. *PLoS Med* **10**: e1001413. doi:10.1371/journal.pmed.1001413.

Hay SI, Okiro EA, Gething PW, Patil AP, Tatem AJ, et al. 2010. Estimating the global clinical burden of *Plasmodium falciparum* malaria in 2007. *PLoS Med* **7**: e1000290. doi:10.1371/journal.pmed.1000290.

Hay SI, Snow RW. 2006. The Malaria Atlas Project: developing global maps of malaria risk. *PLoS Med* **3**: 2204-2208. doi:10.1371/journal.pmed.0030473.

Hay SI, Tatem AJ, Graham AJ, Goetz SJ, Rogers DJ. 2006. Global environmental data for mapping infectious disease distribution. *Adv Parasit* **62**: 37-77. doi:10.1016/S0065-308x(05)62002-7.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol* **25**: 1965-1978. doi:10.1002/Joc.1276.

Hotez PJ. 2013. Forgotten people, forgotten diseases: the neglected tropical diseases and their impact on global health and development. Washington DC: ASM Press. 275 p.

Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, et al. 2008. Global trends in emerging infectious diseases. *Nature* **451**: 990-994. doi:10.1038/Nature06536.

Koch T. 2011. Disease maps: epidemics on the ground. Chicago: University of Chicago Press. 368 p.

Koch T. 2014. 1831: the map that launched the idea of global health. *Int J Epidemiol* **43**: 1014-1020. doi:10.1093/Ije/Dyu099.

Kraemer MUG, Sinka ME, Duda KA, Mylne A, Shearer F, et al. 2015. The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. *eLife*: accepted manuscript.

Krause G, Prioritisation Working Group. 2008. How can infectious diseases be prioritized in public health? A standardized prioritization scheme for discussion. *EMBO Rep* **9**: S22-S27. doi:10.1038/embor.2008.76.

Lancet editorial. 1831. History of the rise, progress, ravages, etc. of the blue cholera of India. *Lancet* **17**: 241-284.

Leathwick JR, Elith J, Chadderton WL, Rowe D, Hastie T. 2008. Dispersal, disturbance and the contrasting biogeographies of New Zealand's diadromous and non-diadromous fish species. *J Biogeogr* **35**: 1481-1497. doi:10.1111/j.1365-2699.2008.01887.x.

Leathwick JR, Rowe D, Richardson J, Elith J, Hastie T. 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biol* **50**: 2034-2052. doi:10.1111/j.1365-2427.2005.01448.x.

Levine RS, Peterson AT, Yorita KL, Carroll D, Damon IK, et al. 2007. Ecological niche and geographic distribution of human monkeypox in Africa. *PLoS One* **2**: e176. 10.1371/journal.pone.0000176.

Macsween KF, Crawford DH. 2003. Epstein-Barr virus - recent advances. *Lancet Infect Dis* **3**: 131-140. doi:10.1016/S1473-3099(03)00543-7.

Martin V, Pfeiffer DU, Zhou XY, Xiao XM, Prosser DJ, et al. 2011. Spatial distribution and risk factors of highly pathogenic avian influenza (HPAI) H5N1 in China. *PLoS Pathog* **7**: e1001308. doi:10.1371/journal.ppat.1001308.

Moyes CL, Temperley WH, Henry AJ, Burgert CR, Hay SI. 2013. Providing open access data online to advance malaria research and control. *Malar J* **12**: e161. doi:10.1186/1475-2875-12-161.

Olival KJ, Hayman DTS. 2014. Filoviruses in bats: current knowledge and future directions. *Viruses* **6**: 1759-1788. doi:10.3390/V6041759.

Patil AP, Gething PW, Piel FB, Hay SI. 2011. Bayesian geostatistics in health cartography: the perspective of malaria. *Trends Parasitol* **27**: 245-252. doi:10.1016/J.Pt.2011.01.003.
Peterson AT. 2008. Biogeography of diseases: a framework for analysis. *Naturwissenschaften* **95**: 483-491. doi:10.1007/s00114-008-0352-5.
Peterson AT, Soberon J, Pearson RG, Anderson RP, Martinez-Meyer E, et al. 2011. Ecological niches and geographic distributions. Princeton: Princeton University Press. 314 p.
Pfeiffer DU, Robinson TP, Stevenson M, Stevens KB, Rogers DJ, et al. 2008. Spatial analysis in epidemiology. Oxford: Oxford University Press. 142 p.
Phillips SJ, Anderson RP, Schapire RE. 2006. Maximum entropy modeling of species geographic distributions. *Ecol Model* **190**: 231-259. doi:10.1016/j.ecolmodel.2005.03.026.
Phillips SJ, Dudik M, Elith J, Graham CH, Lehmann A, et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol Appl* **19**: 181-197. doi:10.1890/07-2153.1.
Pigott DM, Golding N, Messina JP, Battle KE, Duda KA, et al. 2014. Global database of leishmaniasis occurrence locations, 1960-2012. *Sci Data* **1**: 140036. doi:10.1038/sdata.2014.36.
Piot P. 2014. Ebola's perfect storm. *Science* **345**: 1221. doi:10.1126/science.1260695.
Plowright RK, Eby P, Hudson PJ, Smith IL, Westcott D, et al. 2015. Ecological dynamics of emerging bat virus spillover. *Proc R Soc Lond B Biol Sci* **282**: e20142124. doi:10.1098/Rspb.2014.2124.
Powers AM, Aguilar PV, Chandler LJ, Brault AC, Meakins TA, et al. 2006. Genetic relationships among Mayaro and Una viruses suggest distinct patterns of transmission. *Am J Trop Med Hyg* **75**: 461-469.
Rogers DJ, Wilson AJ, Hay SI, Graham AJ. 2006. The global distribution of yellow fever and dengue. *Adv Parasit* **62**: 181-220. doi:10.1016/S0065-308x(05)62006-4.
Schipper J, Chanson JS, Chiozza F, Cox NA, Hoffmann M, et al. 2008. The status of the world's land and marine mammals: diversity, threat, and knowledge. *Science* **322**: 225-230. doi:10.1126/science.1165115.
Sinka ME, Bangs MJ, Manguin S, Rubio-Palis Y, Chareonviriyaphap T, et al. 2012. A global map of dominant malaria vectors. *Parasit Vectors* **5**: 69. doi:10.1186/1756-3305-5-69.
Snow J. 1855. On the mode of communication of cholera. London: John Churchill.
Timen A, Koopmans MPG, Vossen ACTM, van Doornum GJJ, Gunther S, et al. 2009. Response to imported case of Marburg hemorrhagic fever, the Netherlands. *Emerg Infect Dis* **15**: 1171-1175. doi:10.3201/eid1508.090051.
Towner JS, Khristova ML, Sealy TK, Vincent MJ, Erickson BR, et al. 2006. *Marburgvirus* genomics and association with a large hemorrhagic fever outbreak in Angola. *J Virol* **80**: 6497-6516. doi:10.1128/JVI.00069-06.
Trotter CL, Greenwood BM. 2007. Meningococcal carriage in the African meningitis belt. *Lancet Infect Dis* **7**: 797-803. doi:10.1016/S1473-3099(07)70288-8.
UNAIDS. UNAIDS. Available: http://www.unaids.org/en/. Accessed: July 2014
Van Boeckel TP, Brower C, Gilbert M, Grenfell BT, Levin S, et al. 2015. Global trends in antimicrobial use in food animals. *Proc Natl Acad Sci USA* **112**: 5649-5654. doi:10.1073/pnas.1503141112.
Weiss DJ, Atkinson PM, Bhatt S, Mappin B, Hay SI, et al. 2014. An effective approach for gap-filling continental scale remotely sensed time-series. *ISPRS J Photogramm Remote Sens* **98**: 106-118. doi:10.1016/j.isprsjprs.2014.10.001.
WHO. 2009. Neglected tropical diseases, hidden successes, emerging opportunities. Geneva: World Health Organization. 59 p.
WHO. 2010. Working to overcome the global impact of neglected tropical diseases. First WHO report on neglected tropical diseases. Geneva: World Health Organization. 172 p.
WHO. 2012. Accelerating work to overcome the global impact of neglected tropical diseases - a roadmap for implementation. Geneva: World Health Organization. 16 p.
WHO. 2014. World malaria report 2014. Geneva: World Health Organization. 242 p.

Chapter 2

Identifying targets for infectious disease mapping.

Using the Hay *et al.* (2013) analysis as a starting point, this chapter illustrates a quantitative method that can be used to identify the next targets for disease mapping. In an environment where funding can often constrain opportunity for research, and where increasing methodological complexity and difficulty in data abstraction and interpretation mean that the mapping process becomes increasingly time-consuming, a rational way for selecting next priorities is key. This work has been published in *PLoS Neglected Tropical Diseases* and is included here in its final form. The additional information referred to in this chapter is included in the Appendix of this thesis.

PLOS | NEGLECTED TROPICAL DISEASES

RESEARCH ARTICLE

Prioritising Infectious Disease Mapping

David M. Pigott[1‡], Rosalind E. Howes[1‡], Antoinette Wiebe[1], Katherine E. Battle[1], Nick Golding[2], Peter W. Gething[1], Scott F. Dowell[3], Tamer H. Farag[3], Andres J. Garcia[3], Ann M. Kimball[3], L. Kendall Krause[3], Craig H. Smith[3], Simon J. Brooker[4], Hmwe H. Kyu[5], Theo Vos[5], Christopher J. L. Murray[5], Catherine L. Moyes[2], Simon I. Hay[2,5,6]*

1 Spatial Ecology & Epidemiology Group, Department of Zoology, University of Oxford, Oxford, United Kingdom, **2** Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, **3** Bill & Melinda Gates Foundation, Seattle, Washington, United States of America, **4** London School of Hygiene & Tropical Medicine, London, United Kingdom, **5** Institute for Health Metrics and Evaluation, University of Washington, Seattle, Washington, United States of America, **6** Fogarty International Center, National Institutes of Health, Bethesda, Maryland, United States of America

‡ These authors share first authorship on this work.
* simon.hay@well.ox.ac.uk

Citation: Pigott DM, Howes RE, Wiebe A, Battle KE, Golding N, Gething PW, et al. (2015) Prioritising Infectious Disease Mapping. PLoS Negl Trop Dis 9 (6): e0003756. doi:10.1371/journal.pntd.0003756

Editor: Xiao-Nong Zhou, National Institute of Parasitic Diseases China CDC, CHINA

Received: October 5, 2014

Accepted: April 13, 2015

Published: June 10, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The initial work to develop a platform that will map a range of infectious diseases is funded by the Bill & Melinda Gates Foundation (Global Health Grant Number OPP1093011), and the authors are grateful to the Surveillance and Vaccine Development team for their contribution to the prioritisation work described in this paper. DMP is funded by a Sir Richard Southwood Graduate Scholarship from the Department of Zoology at the University of Oxford; REH is financially supported by a Wellcome Trust Senior Research Fellowship to SIH (#095066) which also supports AW and KEB; NG is funded by a grant

Abstract

Background

Increasing volumes of data and computational capacity afford unprecedented opportunities to scale up infectious disease (ID) mapping for public health uses. Whilst a large number of IDs show global spatial variation, comprehensive knowledge of these geographic patterns is poor. Here we use an objective method to prioritise mapping efforts to begin to address the large deficit in global disease maps currently available.

Methodology/Principal Findings

Automation of ID mapping requires bespoke methodological adjustments tailored to the epidemiological characteristics of different types of diseases. Diseases were therefore grouped into 33 clusters based upon taxonomic divisions and shared epidemiological characteristics. Disability-adjusted life years, derived from the Global Burden of Disease 2013 study, were used as a globally consistent metric of disease burden. A review of global health stakeholders, existing literature and national health priorities was undertaken to assess relative interest in the diseases. The clusters were ranked by combining both metrics, which identified 44 diseases of main concern within 15 principle clusters. Whilst malaria, HIV and tuberculosis were the highest priority due to their considerable burden, the high priority clusters were dominated by neglected tropical diseases and vector-borne parasites.

Conclusions/Significance

A quantitative, easily-updated and flexible framework for prioritising diseases is presented here. The study identifies a possible future strategy for those diseases where significant knowledge gaps remain, as well as recognising those where global mapping programs have already made significant progress. For many conditions, potential shared epidemiological information has yet to be exploited.

from the Bill & Melinda Gates Foundation (OPP1053338). SJB is supported by a Wellcome Trust Senior Research Fellowship (#092765) and acknowledges the support of the Bill & Melinda Gates Foundation (OPP1033751). CLM is funded by a grant from the Bill & Melinda Gates Foundation (OPP1093011). The funders had no role in study design, data colleciton and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: SJB is a deputy editor of PLoS Neglected Tropical Diseases.

Author Summary

Maps have long been used to not only visualise, but also to inform infectious disease control efforts, identify and predict areas of greatest risk of specific diseases, and better understand the epidemiology of disease over various spatial scales. In spite of the utilities of such outputs, globally comprehensive maps have been produced for only a handful of infectious diseases. Due to limited resources, it is necessary to define a framework to prioritise which diseases to consider mapping globally. This paper outlines a framework which compares each disease's global burden with its associated interest from the policy community in a data-driven manner which can be used to determine the relative priority of each condition. Malaria, HIV and TB are, unsurprisingly, ranked highest due to their considerable health burden, while the other priority diseases are dominated by neglected tropical diseases and vector-borne diseases. For some conditions, global mapping efforts are already in place, however, for many neglected conditions there still remains a need for high resolution spatial surveys.

Introduction

Maps provide an essential evidence-base to support progress towards global health commitments [1]. For example, they provide important baseline estimates of disease limits [2–7], transmission [8–10] and clinical burden [11–14]; underpin surveillance systems and outbreak tracking [15,16]; help target resource allocation from the macro- [17,18] through the meso- [19–22] to the micro-scale [23]; and inform international travel guidelines [24–26]. Significant developments in mapping techniques have occurred over the last few decades, particularly through the use of species distribution models and model-based geostatistics [1,27]. Similarly, disease data has become more widespread and easier to share [28]. Despite these advances however, a recent review of 355 clinically-significant infectious diseases (IDs) indicated that of the 174 IDs for which an opportunity for mapping was identified, only 4% had been comprehensively mapped [1]. For many of these conditions, there is a significant shortfall between existing maps and what can be achieved with contemporary methods and datasets.

Traditional mapmaking has focussed on a vertical, single-species approach, requiring highly labour intensive, and therefore expensive, manual data identification and assembly [13,21,29–31]. The present era of open-access big data, high computational capacity, and rapid software development offers new opportunities for scaling-up the spatial mapping of IDs, primarily through the automation of data gathering and geopositioning but ultimately also to mapping. The Atlas of Baseline Risk Assessment for Infectious Disease (abbreviated ABRAID, as in, "to awake") is a developing software platform designed to exploit this opportunity and has the ambition to produce continuously updated maps for 174 IDs globally [28]. Realising automation of data retrieval and positioning at this scale is a practically non-trivial but conceptually simple, logistic scaling exercise. In order to automate mapping for each ID so that it is continuously updated and improved as new information becomes available, the spatial inference methods used need to be tailored to each unique ID epidemiology [28]. In some cases this will require disease-specific methodological developments. This requires substantial investment, so an objective and systematic approach is required to determine the order in which IDs are to be mapped.

The first stage in this process is to organise all IDs using a schema based upon shared biological and epidemiological traits; for example, "the mosquito-borne arboviruses". Such groups will likely have similar mapping requirements, enabling synergies in data collation, covariate selection, increased efficiency (*i.e.* in software development), and more robust validation of outputs [32,33]. We refer to these disease groups as "mapping clusters" and they form the basic architecture of the prioritisation process.

To rationally prioritise mapping of these conditions, the diseases within each mapping cluster were evaluated based upon their global burden (both morbidity and mortality), as well as the disease's importance amongst public health stakeholders. Data inputs are quantitative in nature and reliant on either independently derived data or data sourced from entire communities rather than selected expert individuals. Therefore, this proposed framework is unaffected by much of the subjectivity associated with other prioritisation studies, and also provides a platform for rapidly incorporating changes to existing diseases, as well as emerging novel public health threats. The prioritisation exercise helps to guide the order in which diseases are mapped to best support public health priorities; we argue that all relevant diseases can and should eventually be mapped.

A comprehensive atlas of IDs is of central importance in providing geographical context to the understanding of tropical disease and global health [34–36]. Moreover, as the atlas becomes more complete the overlay of maps will provide opportunities for investigating patterns of global disease diversity [37,38] and the process of disease emergence [39].

Methods

Method summary

In order to generate disease prioritisation standards, diseases with shared taxonomy and transmission characteristics were grouped together to create clusters. Diseases within each cluster were evaluated based upon two factors reflecting their importance from a public health perspective: (a) the global burden of the disease and (b) the current public health focus on the condition. Both metrics were assessed simultaneously in order to rank the clusters, and specific diseases were then identified for prioritisation.

Selection of diseases for mapping

This study aimed to be comprehensive in its scope of IDs. All diseases identified in a previous review as meriting mapping were included [1]. This earlier study categorised 355 diseases into five classes: Option 1, indicating that the disease was unsuitable for occurrence based mapping methods; Option 2, mapping the observed occurrence of the disease; Option 3, mapping the maximum potential range of the disease using knowledge of vector, intermediate host and reservoir species; Option 4, using niche mapping methods such as boosted regression trees; and Option 5, where sufficient data exist to allow for global maps of variation in prevalence of infection and/or disease. Option 1 diseases included those that showed no sustained spatial variation in occurrence (*i.e.* had a cosmopolitan distribution) and had insufficient evidence to allow for the global mapping of variation in prevalence using advanced statistical methods such as model-based geostatistics. In cases where such information does exist, these diseases were promoted to Option 5 status. Revisions to the *Hay et al. (2013)* paper have led to the inclusion of tuberculosis, ascariasis, trichuriasis and trachoma—all previously listed as Option 1—as Option 5 diseases. Further revisions included the exclusion of New and Old World Spotted Fever Rickettsiosis and New and Old World Phlebovirus because their constituent diseases were included. In addition, *Plasmodium knowlesi* was included due to the increasing appreciation of its significance to human health in Southeast Asia [40,41]. The new revised total of diseases

that warrant mapping was therefore 176. Those diseases not considered for mapping due to Option 1 classification are outlined and justified in the Supporting Information.

Creating disease clusters

Diseases were grouped into clusters based on characteristics relevant to spatial epidemiology. Diseases were placed in the same cluster if they had the potential to mutually reinforce each other in terms of data assembly, mapping requirements and cross-validation of data by comparison of outputs. Clustering classifications were therefore based on the key factors influencing the approach taken for mapping.

At the coarsest level, pathogens were grouped by agent type (virus/bacteria/fungus/other) and the larger agent groupings were split into specific phyla (*e.g.* Nematoda and Platyhelminths) [42]. These relatively coarse groups reflect fundamental differences in life histories and epidemiology as well as the most basic taxonomic divisions. Within these broad groupings, the mode of transmission was used to create the final disease clusters. This is an important factor when mapping IDs, as the mode of transmission has a large influence on which abiotic correlates are relevant to the mapping process. For instance, the transmission limits of vector-borne diseases are restricted in part by the environmental suitability for the vector species in question, thus diseases spread by similar vectors will share covariates [43]. Similarly, sexually-transmitted diseases are likely to share mapping methods linked to human distribution and behaviour, whilst pathogens spread by water contact would share common traits linked to the environment; these groupings can therefore be logically considered together within a mapping framework. The mode of transmission classifications are defined in the previous publication [1].

Assessing disease burden

The burden of each disease was assessed using the disability-adjusted life year (DALY) estimates from the 2013 Global Burden of Disease Study (GBD 2013) [44,45]. DALYs quantify both morbidity and mortality attributed to each disease and therefore better capture the total impact of a disease than do clinical cases or mortality alone [44,45]. The GBD's systematic approach across a wide spectrum of diseases provides an extremely valuable resource from which to compare the relative impact of diseases on human health.

Wherever possible, direct links were made between the GBD estimates and diseases in the mapping list. The GBD disease categories, which are based upon the International Classification of Diseases and Related Health Problems (ICD-10) [46], do not always specify particular infectious agents, but rather focus on the clinical symptoms of infection, or non-specified disease groups. These aggregated DALY estimates had to be split across the relevant causative diseases in the mapping list, therefore the *Hay et al. (2013)* study was reconciled with the ICD-10 codes and then GBD categories in order to disaggregate the broader classifications such as "other diarrheal diseases" and "other neglected tropical diseases". The full process is outlined in the associated Supporting Information, S1 Text.

Overall, 11 of the 176 mapping diseases could not be reconciled to the GBD categories. Some were not considered due to having unknown pathogenic agents (*e.g.* tropical sprue) and others were very rare and fell into ICD-10 categories that were assigned over various groupings (*e.g.* pentastomiasis). These diseases were allocated a nominal DALY of 100; this value, while arbitrary, is low enough to avoid skewing the analysis. For each cluster, the total DALYs for all diseases was calculated and contributed to the final analysis.

Assessing global health community interest

Of equal importance is the need to produce maps for those diseases where there is the greatest demand, whether from international organisations or from local public health authorities. Measuring this factor was achieved by surveying a representative subset of potential end-users, to identify which diseases have been prioritised by major public health stakeholders: state-funded public health agencies, private companies (*e.g.* vaccine developers), political bodies, non-governmental advocates and practitioners, as well as the scientific research community. For each disease, the final policy score was the sum of three component scores: public health, stakeholder interest, status as a notifiable disease, and h-index.

Cases from the different categories of public health stakeholders were included to capture the spectrum of interest groups (see S1 Text for full listing). Each organisation's mission statement and project pages were reviewed to identify the diseases contained in their public health portfolio. Depending on the type of stakeholder, this would indicate that the organisation would, for example, dedicate funding and effort towards control of that disease, advocate for the disease to governments or public health agencies, or dedicate research funding to the disease. Each disease was allocated one point per stakeholder reporting an interest in it. An inclusive approach was followed, whereby diseases were considered to be of interest to a stakeholder, irrespective of any hierarchy within the agency's prioritisation system.

Another point was allocated to diseases which were notifiable to national disease control agencies. In order to mitigate spatial bias in the notifiable disease listed by different agencies, a search for countries which had readily-accessible and clearly defined domestic policy relating to named pathogens was performed, and one country from each of the main GBD defined regions was selected: USA (High Income), Brazil (Latin America and the Caribbean), Zambia (sub-Saharan Africa), United Arab Emirates (North Africa and the Middle East), India (South Asia), Malaysia (South East Asia, East Asia and Oceania) and Croatia (Central and Eastern Europe and Central Asia). Interest in these diseases at a domestic level suggests that there will be interest in maps of these diseases, as demonstrated by the presence of subscription-only online databases of maps including GIDEON [47] and the rapid expansion of real-time maps to which physicians are encouraged to contribute [15,28].

Academic output, a proxy of funder agency awards, but also of high-quality data availability [1], was quantified based on the *h*-index of each disease [48], as reported by Scopus [49]. More commonly used to assess a scientist's productivity and impact, the *h*-index is used here to quantify the level of active interest across the academic community in each disease [50]. The *h*-index is the number of published papers (referring to a particular disease) that have been cited by at least as many other papers. In other words, an *h*-index of 7 signifies that 7 published papers including that disease name have been cited at least 7 times. For each disease in turn, Scopus citation numbers were generated for all publications referring to the disease (document search for "Disease Name" in "Article Title, Abstract, Keywords"). This Scopus search generates a Citation Tracker file showing the number of citations to each publication referring to the "Disease Name". Diseases were then categorised according to their *h*-index. Those for which there was evidence of very high scientific output scored 2 (*h*-index >100), those with intermediate *h*-index (>50–100) scored 1.5, while diseases with *h*-index of <50 scored 1.

The diseases classified as Option 4 (use niche modelling methods) and Option 5 (model prevalence or incidence) have the most epidemiological data available and have the greatest potential to benefit from a dedicated mapping exercise, but also require the most resources. Option 2 and 3 diseases are data-poor and both require mapping of occurrence data only [1], and therefore are significantly less time-intensive to map, limited to more simplistic analyses, than those diseases categorised as Option 4 and 5. Option 3 disease mapping relates potential

transmission limits to aspects of vector biology. In cases where Option 4 and 5 diseases also have the same vector, the Option 3 disease will be considered as part of mapping these complementary diseases; where this is not the case, a disease's transmission limits can be assessed through a mixture of literature surveys and occurrence data overlap. Option 4 and 5 diseases within the disease clusters were therefore prioritised and for each cluster, the average policy score for the Option 4 and 5 diseases was calculated and contributed to the final analysis. These diseases should be the primary focus of future cartographic efforts as these require the most attention and bespoke inputs to be generated.

Mapping prioritisation ranking of diseases

The final step in the process was to combine these assessments to produce a ranking of disease clusters and therefore recommend diseases to prioritise for mapping. Each cluster was plotted on a graph based on its total DALYs and the average policy priority of its Option 4 and 5 diseases. Option 2 and 3 diseases were included in the cluster DALY scores in order to reflect the relative importance that each cluster represented in terms of burden of disease. One cluster may consist of a large number of minor diseases which, as a collective grouping, represent a significant problem—by retaining the DALY score, this burden is reflected, With the policy priority score however, the opposite is the case; inclusion of multiple low scoring diseases would down-weight the cluster as a whole. In scenarios where clusters consist of a diverse grouping of pathogens, averaging policy score across all conditions misrepresents those with a high policy priority and therefore masks these diseases in comparison to clusters that only consist of those diseases with high policy priority scores.

Each cluster was then evaluated based upon its distance from a hypothetical cluster which had the highest DALYs (*i.e.* that of HIV) and the highest policy score (*i.e.* that of Malaria) relative to a line drawn from this cluster to the origin; those closer to this hypothetical cluster, along this axis, were prioritised higher. As a result, the relative influence of burden and policy priority could be considered both simultaneously and independently. Within each cluster, the diseases to be prioritised (*i.e.* Option 4 or 5) were then reported (Table 1). The code to replicate this methodology is freely available from: https://github.com/SEEG-Oxford/prioritisation.

Results

Organization of the mapping clusters

The 176 diseases identified as having a rationale for mapping were organised into 33 clusters, based upon the biological and taxonomic classifications of the causative pathogen, modes of transmission and the mapping method recommended in a previous review [1] (Fig 1). Seven of these clusters included only a single disease due to their unique transmission within their broader taxonomic grouping (HIV, poliomyelitis, avian influenza, pythiosis, South American bartonellosis, tuberculosis and babesiosis). Conversely, the mosquito-borne arbovirus cluster was the largest cluster, consisting of 26 diseases, many of which have the potential to benefit from modelled maps.

Prioritising mapping diseases

Fig 2 brings together the two indices selected to prioritise diseases for mapping—disability adjusted life-year (DALY) burden and relative stakeholder interest. These plots demonstrate that the HIV, malaria and tuberculosis clusters are exceptional in representing an overwhelming share of DALY burden [51] and being of highest priority to the global health community with their placement in the top right quadrant of the graph. These three clusters contain five

Table 1. Clusters indicated as mapping priorities with their constituent diseases recommended for distribution modelling and current global mapping projects identified.

Cluster (main diseases to map / total diseases in cluster)	Diseases within cluster, to map	Total cluster DALYs	Average policy score	Current global mapping projects
1. Malaria (n = 3/5)	*Plasmodium falciparum*	65,493,135	11.8	MAP [13,29,40]; WHO [81]
	P. knowlesi			
	P. vivax			
2. HIV (n = 1/1)	HIV	69,480,661	11	GBD [51]; UNAIDS [82]
3. Tuberculosis (n = 1/1)	Tuberculosis	49,816,215	11	GBD [51]
4. Food/Water-borne (Bacteria) (n = 1/4)	Cholera	9,962,003	8	
5. Water-borne (Platyhelminth) (n = 3/7)	*Schistosoma haematobium*	3,062,843	7.7	GAHI [83]; Global NTD database [31]
	S. japonicum			
	S. mansoni			
6. Trypanosomiasis (n = 2/2)	African trypanosomiasis	728,564	7.5	WHO [68,84]
	American trypanosomiasis			
7. Filariasis (n = 3/3)	Bancroftian filariasis	2,022,099	6.2	GAHI [85]
	Brugia malayi			
	B. timori			
8. Soil Transmitted Helminths (n = 3/3)	Ascariasis	4,029,403	5.3	GAHI [5,14]
	Hookworm			
	Trichuriasis			
9. Leishmaniasis (n = 3/3)	Cutaneous leishmaniasis (Old World)	4,283,139	5.2	SEEG [8]
	Cutaneous leishmaniasis (New World)			
	Visceral leishmaniasis			
10. Unknown agent (n = 1/4)	Tropical sprue	3,609,400	4	
11. Picornaviridae (n = 1/1)	Polio	116,065	6	The Global Polio Eradication Initiative [86]
12. Food/Water-borne (Nematode) (n = 1/13)	Dracunculiasis	422,476	2.5	
13. Fly-borne (Nematode) (n = 2/5)	Loiasis	711,246	4.3	WHO and APOC [87] [88]
	Onchocerciasis			
14. Direct contact (Bacteria) (n = 4/6)	Anthrax	1,030,777	4	
	Brazilian purpuric fever			
	Leprosy			
	Trachoma			GAT [10,89]
15. Mosquito-borne (Virus) (n = 15/26)	Barmah Forest disease	4,219,569	2.6	
	California serogroup viruses			
	Chikungunya			
	Dengue			SEEG [2,12]
	Japanese encephalitis			
	Murray Valley encephalitis			
	Rift Valley fever			
	Rocio			
	Ross River virus			
	Sindbis			
	St. Louis encephalitis			
	Venezuelan equine encephalitis			
	Western equine encephalitis			

(*Continued*)

Table 1. (*Continued*)

Cluster (main diseases to map / total diseases in cluster)	Diseases within cluster, to map	Total cluster DALYs	Average policy score	Current global mapping projects
	West Nile fever			
	Yellow fever			

* Indicates default null value.

MAP—Malaria Atlas Project; WHO—World Health Organization; GBD—Global Burden of Disease; GAHI—Global Atlas of Helminth Infections; SEEG—Spatial Ecology and Epidemiology Group; APOC—African Programme for Onchocerciasis Control; GAT—Global Atlas of Trachoma

doi:10.1371/journal.pntd.0003756.t001

individual diseases that are a mapping priority, malaria (*Plasmodium falciparum*, *P. vivax*, and *P. knowlesi*), HIV and tuberculosis. Table 1 shows the top 15 disease clusters (*i.e.* those in the top right of Fig 2), representing 44 individual diseases, with their associated scores. Fig 3 demonstrates that there exists a group of approximately 45 diseases that are the collective focus of public health agencies. The 44 diseases prioritised by this study include all those diseases that represent a significant cartographic challenge (*i.e.* those diseases requiring either species distribution modelling approaches to produce occurrence maps or model-based geostatistics to produce prevalence maps, n = 33) identified by these public health agencies, save rabies and avian influenza. The clusters are ranked in order, whilst the diseases within each cluster are alphabetical and should be considered equal on the basis of this prioritisation.

The top ten priority clusters account for over 92% of all DALYs for those IDs which require mapping (*i.e.* the 176 IDs identified); if this is expanded to the top 15 clusters containing 44 diseases to map, this value increases to 95% (Fig 4). Within these 44 diseases, 19 of the 29 neglected tropical diseases (NTD) highlighted by the WHO are represented. Within the top ten prioritised clusters, 14 individual diseases relate to these same NTDs [52,53]. The top 15 prioritised clusters include some diseases, such as the picornaviridae (polio), that have a low DALY burden but a high public health ranking because they are high on the eradication agenda.

Disease burden

It was possible to establish a direct correspondence with GBD estimates for 34 of the 176 diseases with a strong rationale for mapping as listed by *Hay et al. (2013)* [44,45]. DALY estimates were allocated to a further 132 diseases by linking diseases with ICD-10 codes [46] and their respective GBD category definitions. Whilst these burden values are not accurate absolute values, and should not be interpreted as such, this DALY allocation does allow relative burdens to be determined. The remaining 11 diseases were given the baseline DALY allocation of 100, a value not intended to represent an estimate of the "true" DALYs associated with these diseases, but rather to distinguish them from diseases which were considered to cause a major burden in the GBD analysis. It is safe to assume that if such diseases were not assigned a specific GBD classification, their global impact on mortality and morbidity is relatively small.

In total, the 176 diseases with a strong rationale for mapping [1] represent over 230 million DALYs, approximately 10% of the global DALY burden and 47% of the global ID DALY burden. At the cluster level, HIV, malaria and tuberculosis represent 80% of the overall mapping-disease DALY burden (Fig 5A). Apart from these three conditions, the only other IDs in the top 50 highest DALYs globally are not currently recommended for mapping because they do not show spatial variation in their occurrence and have insufficient data to map variation in disease prevalence with model-based geostatistical analyses. The high-burden diseases not currently considered for mapping include respiratory diseases, meningitis, and many diarrhoeal

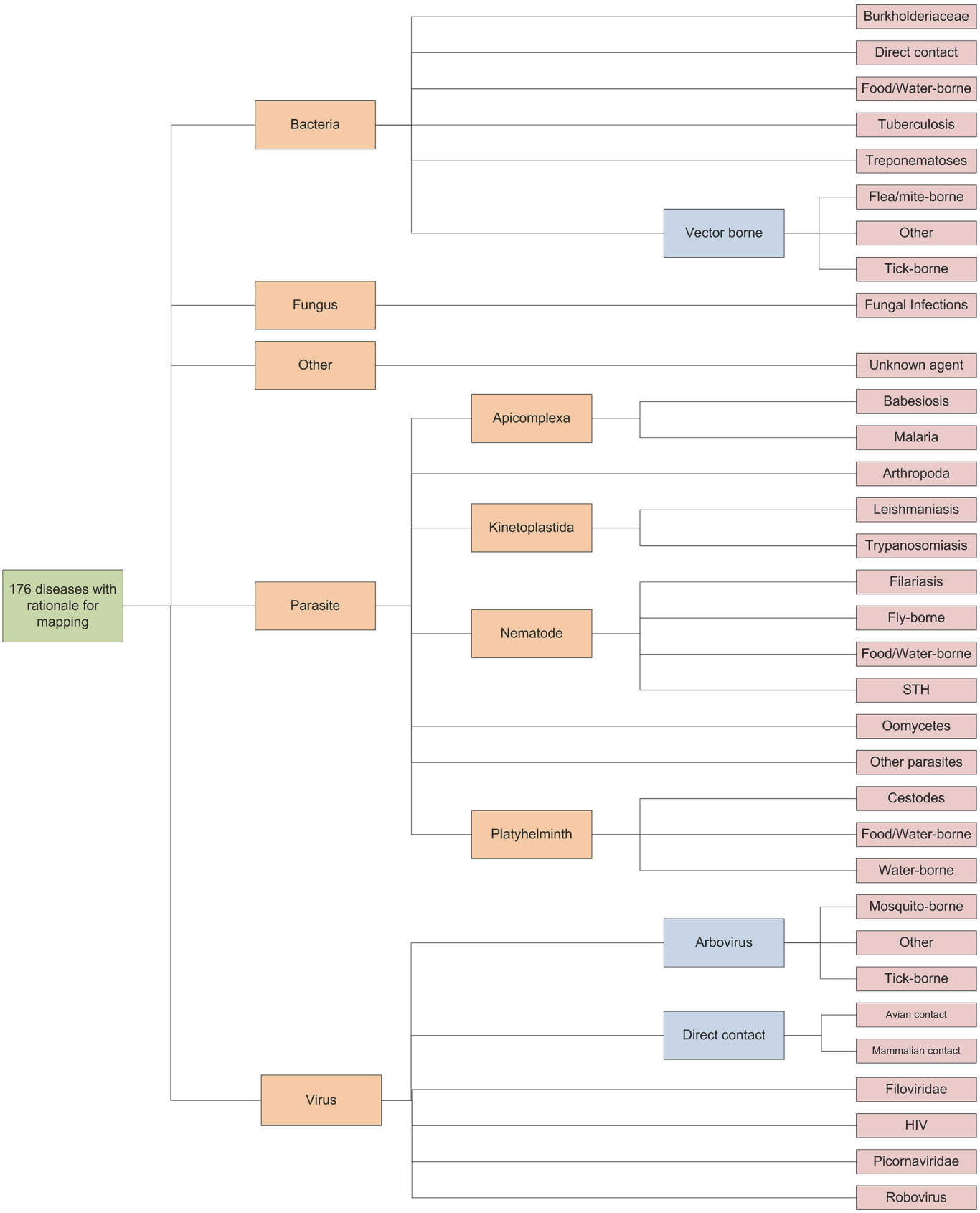

176 diseases with rationale for mapping
Bacteria
Burkholderiaceae
Direct contact
Food/Water-borne
Tuberculosis
Treponematoses
Vector borne
Flea/mite-borne
Other
Tick-borne
Fungus
Fungal Infections
Other
Unknown agent
Parasite
Apicomplexa
Babesiosis
Malaria
Arthropoda
Kinetoplastida
Leishmaniasis
Trypanosomiasis
Nematode
Filariasis
Fly-borne
Food/Water-borne
STH
Oomycetes
Other parasites
Platyhelminth
Cestodes
Food/Water-borne
Water-borne
Virus
Arbovirus
Mosquito-borne
Other
Tick-borne
Direct contact
Avian contact
Mammalian contact
Filoviridae
HIV
Picornaviridae
Robovirus

Fig 1. Hierarchical organisation of the 33 clusters. The 176 diseases with strong rationale for mapping were first sorted by taxonomy of pathogenic agent (in orange) and then structured by common epidemiological and transmission characteristics into sub-groupings (in blue) and finally clusters (in red). STH = soil transmitted helminth, VBD = vector borne disease.

doi:10.1371/journal.pntd.0003756.g001

infections. Alternative approaches to mapping broader symptom groupings (severe pneumonia, severe diarrhoea and severe febrile illnesses) and then differentiating constituent disease

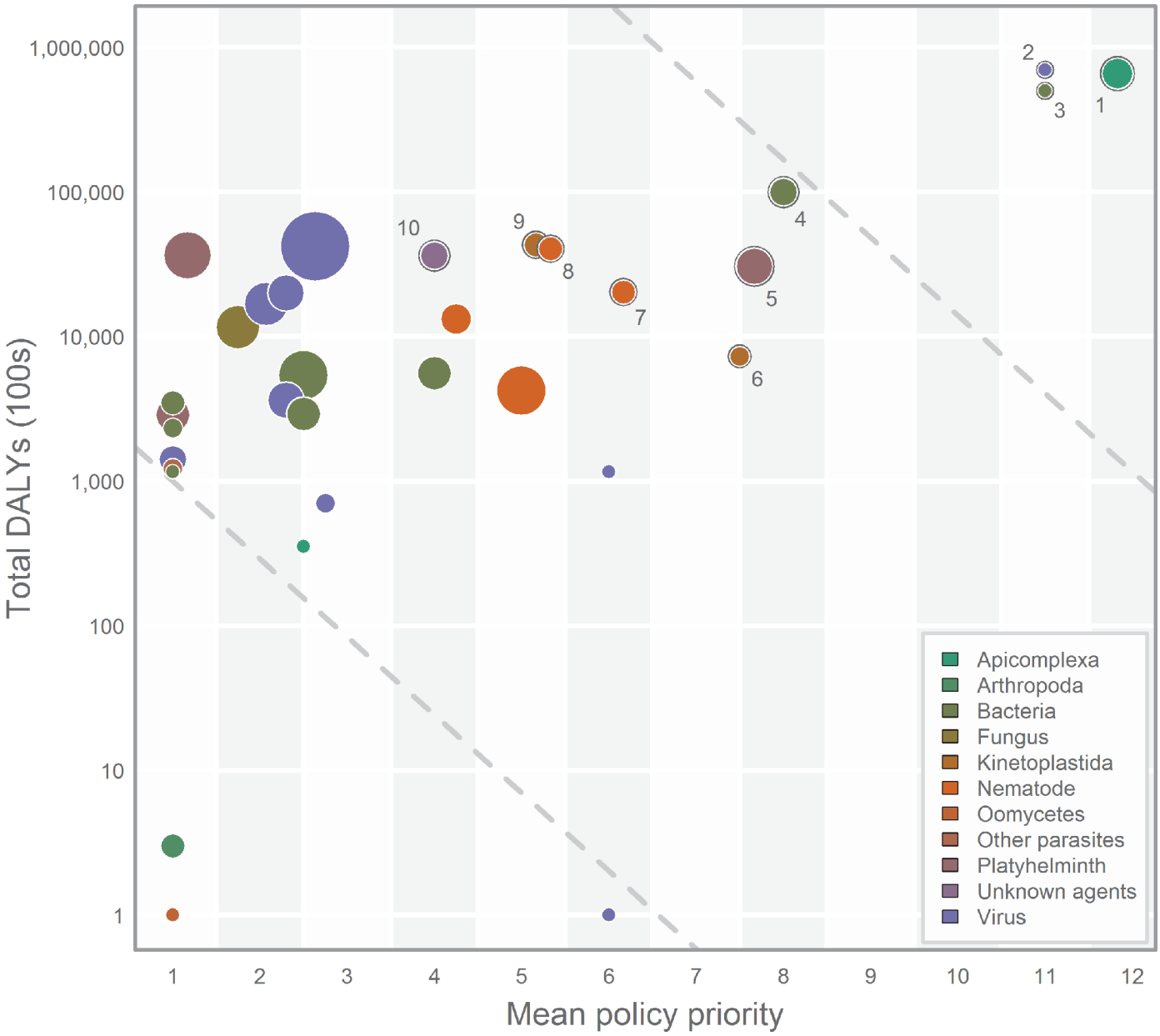


Fig 2. Disease prioritisation. Plot showing the 33 clusters of diseases as ranked by burden of disease DALYs (y-axis—logarithmic scale) and mean policy priority score of occurrence mapping and prevalence mapping diseases (x-axis—linear scale). The top ten clusters circled and numbered as identified in Table 1. The size of the circle is determined by the total number of diseases contained and colour is based upon taxonomy (as outlined by Fig 1; the web appendix contains the full disease listing for each cluster). The dashed guidelines are perpendicular to the axis along which prioritisation order for the clusters was determined; those closer to the top right, along this axis, were prioritised higher.

doi:10.1371/journal.pntd.0003756.g002

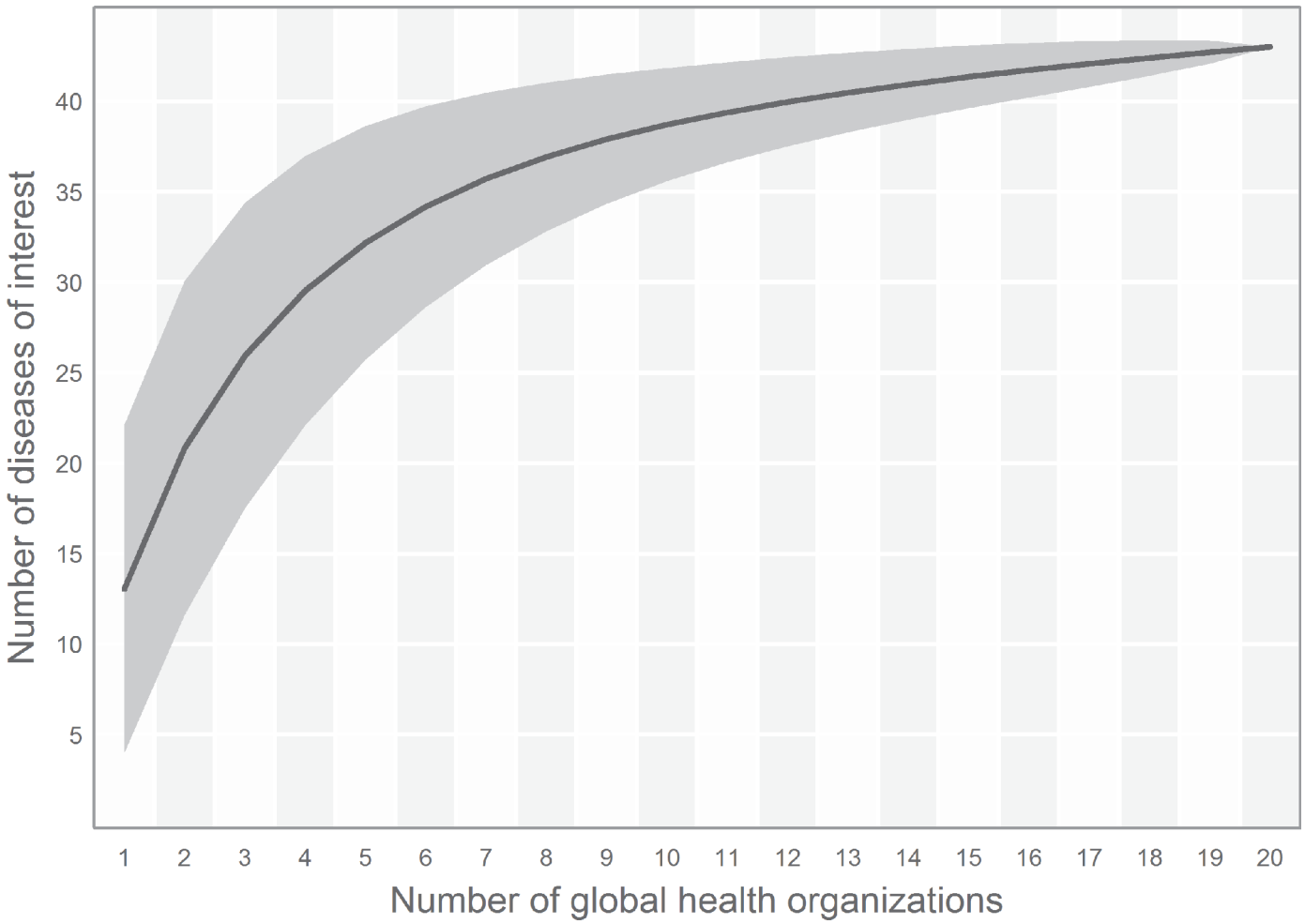


Fig 3. A "species accumulation" curve showing the cumulative number of diseases of interest sampled by increasing numbers of public health stakeholders examined. The diseases of interest of twenty global health stakeholders was indexed and plotted (see Methods). As additional organisations are sampled beyond the fifteen used in this study, the number of unique diseases identified plateaus at around 42. Thus not all public health stakeholders need to be sampled to capture the global diversity of diseases of public health interest.

doi:10.1371/journal.pntd.0003756.g003

components, are being developed. Together, this would map 80% of all DALYs caused by communicable diseases.

A higher resolution focus on the clusters excluding HIV, malaria and tuberculosis (Fig 5B) shows that over 60% of DALYs associated with the 176 IDs are accounted for by the other top ten prioritised clusters; approximately three quarters of the remaining DALYs are accounted for when the remaining prioritised clusters of diseases are included.

Global health community interest

The treemap in Fig 5C displays the repartition of interest from the global health community across the clusters. Interest was scored in terms of: 1) the stated priorities of a survey of assorted public health stakeholders who are expected to be end-users of the maps, 2) status as a notifiable disease, and 3) prominence in the academic literature.

A total of 20 diverse stakeholders were surveyed. This was found to be a sufficiently large number to sample based on an analysis similar to a species accumulation curve that demonstrates the diminishing returns from increasing sampling effort [54]. The number of new

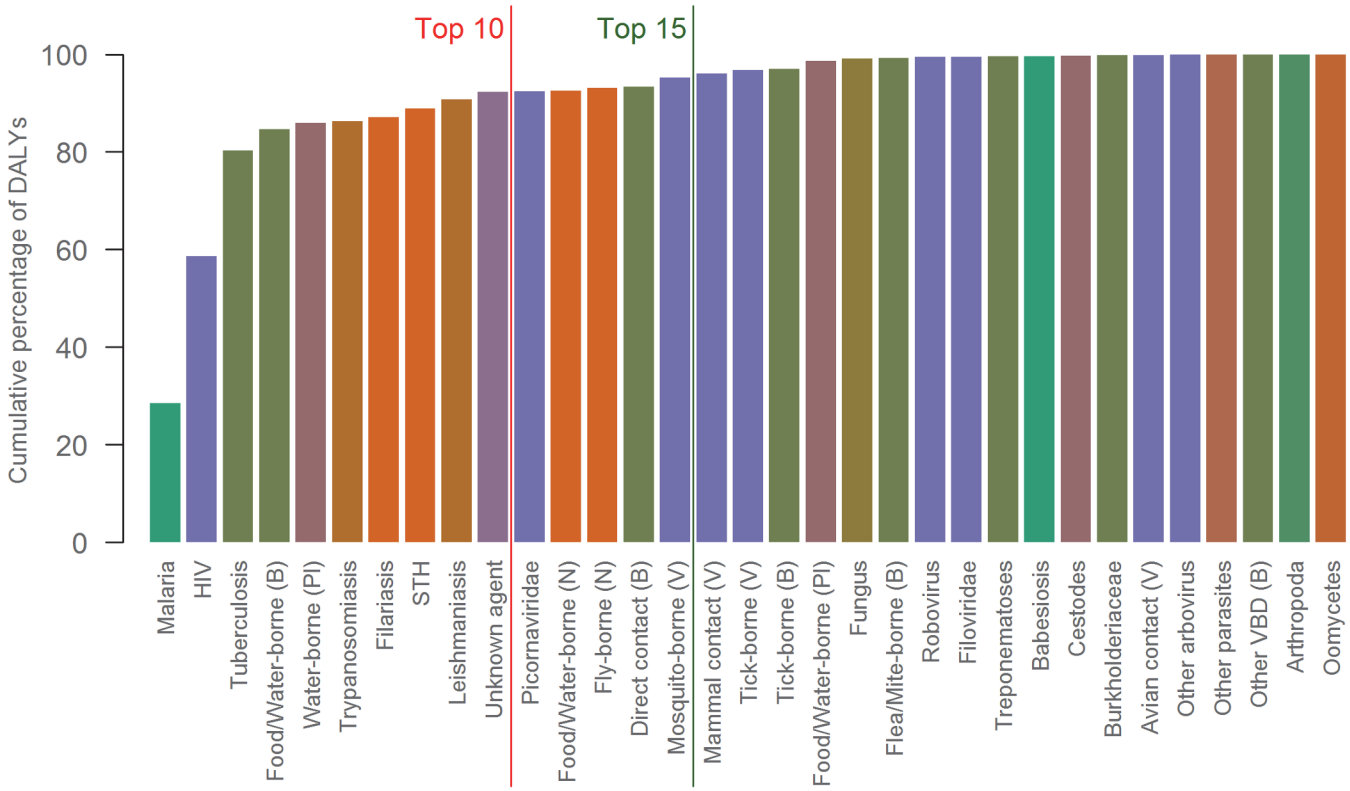


Fig 4. Cumulative percentage barplot indicating the cumulative percentage of DALYs accounted for by each cluster. The colouring is based upon taxonomy, as in Fig 2. The red line indicates the top ten clusters, the dark green indicates the top 15.

doi:10.1371/journal.pntd.0003756.g004

diseases reported levelled off at around 15 organisations sampled (Fig 3) and so the 20 organisations used for this analysis was sufficient to capture the diseases of public health priority. Of the 176 diseases recommended for mapping [1], 24% were prioritised by at least one public health agency, and 55% were notifiable to at least one of the national disease control agencies.

Of those diseases that represent the greatest cartographic challenge, all were prioritised by at least one public health agency and two thirds were notifiable diseases. Of the 176 diseases, thirty diseases (17%) had an *h*-index [48] above 100 (with HIV having the highest *h*-index of 461), while 64% of the diseases had an *h*-index of 50 or less. Of the occurrence mapping and prevalence mapping diseases, 30% had an *h*-index above 100 and only 37% had an *h*-index of 50 or less.

Unlike the DALY burden, which was allocated at the disease level (S1 Text), the stated priority diseases were often grouped to the cluster level by the surveyed stakeholders. For instance, rather than specifying "*Plasmodium vivax*" or "visceral leishmaniasis" as a focus, "malaria" and "leishmaniasis" would be more commonly stated targets. Each component disease of these clusters would therefore be allocated a point, meaning that the number of component diseases in each cluster strongly inflated the overall interest score allocated at the cluster aggregate. Interest scores were calibrated in the final prioritisation assessment to the number of diseases classified as occurrence or prevalence mapping within each cluster (*i.e.* those requiring the more advanced geostatistical techniques, see Methods for more details), so as to avoid being unduly skewed by the size of the cluster.

Overall, malaria, HIV and tuberculosis were the leading clusters of interest, with scores of 11.8, 11 and 11, respectively. A further seven clusters received repeated interest, including

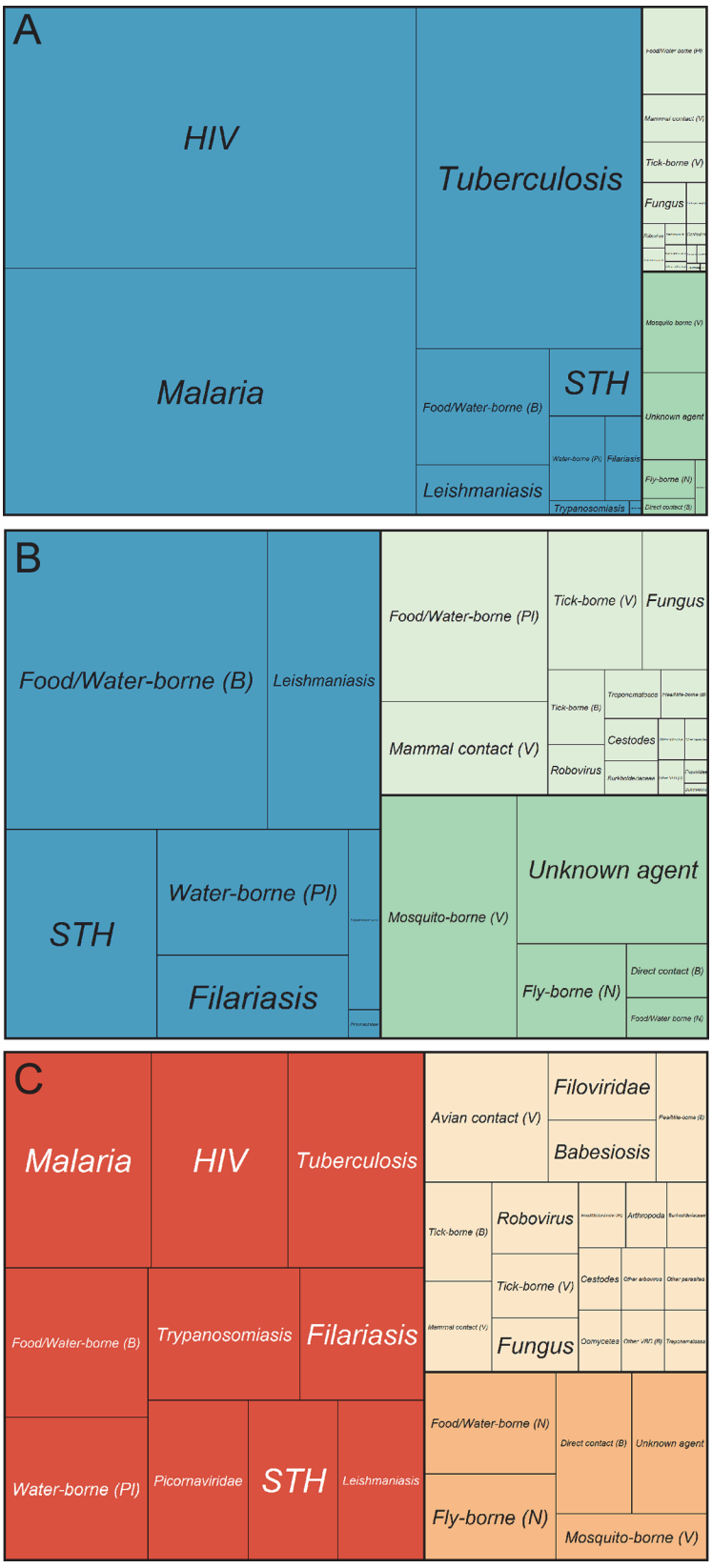

A
HIV
Tuberculosis
Malaria
Food/Water-borne (B)
STH
Leishmaniasis
Water-borne (PI)
Filariasis
Trypanosomiasis
Mosquito-borne (V)
Unknown agent
Fly-borne (N)
Direct contact (B)
Tick-borne (V)
Fungus
Mammal contact (V)
B
Food/Water-borne (B)
Leishmaniasis
Food/Water-borne (PI)
Tick-borne (V)
Fungus
Tick-borne (B)
Cestodes
Robovirus
Mammal contact (V)
STH
Water-borne (PI)
Filariasis
Mosquito-borne (V)
Unknown agent
Fly-borne (N)
Direct contact (B)
C
Malaria
HIV
Tuberculosis
Avian contact (V)
Filoviridae
Babesiosis
Food/Water-borne (B)
Trypanosomiasis
Filariasis
Tick-borne (B)
Robovirus
Arthropoda
Cestodes
Tick-borne (V)
Fungus
Water-borne (PI)
Picornaviridae
STH
Leishmaniasis
Food/Water-borne (N)
Direct contact (B)
Unknown agent
Fly-borne (N)
Mosquito-borne (V)

Fig 5. Plots indicating the relative importance of each mapping cluster. (A) Area of each section is determined by the total DALY contribution of each of the 33 clusters. Blue indicates a cluster contributing to the top ten clusters to be prioritised, green indicates top 44 diseases (n = 5 clusters) and light green represents the remaining disease clusters (n = 18). (B) Area of each section is determined by the total DALY contribution of 30 clusters, with HIV, tuberculosis and malaria excluded. Blue indicates a cluster contributing to the top ten clusters to be prioritised (n = 7), green indicates top 44 diseases (n = 5 clusters) and light green represents the remaining disease clusters (n = 18). STH = soil-transmitted helminth, (B)—bacteria, (N)—nematode, (Pl)—platyhelminth, (V)—virus. (C) Area of each section is determined by the total policy interest score of each of the 33 clusters. Red indicates a cluster within the top ten to be prioritised, orange indicates one of top 44 diseases (n = 5) and light pink represents the remaining disease clusters (n = 18). STH = soil-transmitted helminth, (B)—bacteria, (N)—nematode, (Pl)—platyhelminth, (V)—virus.

doi:10.1371/journal.pntd.0003756.g005

food-borne/water-borne bacteria (score = 8) and water-borne trematodes (7.7), trypanosomiasis (7.5), filariasis (6.2), picornaviridae (6), avian contact viruses (6), soil-transmitted helminths (5.3) and leishmaniasis (5.2) all scoring highly, indicating their importance to the public health community. These scores are relative and intended to reveal general trends across the clusters rather than quantitatively reflect the weighting that any one institution places on a particular disease.

Discussion

A review of all clinically significant IDs identified 176 with a strong rationale for mapping, of which only 4% have been adequately mapped [1]. The current study was undertaken to define a ruleset for determining which diseases, from a cartographic and public health perspective, should be prioritised when sequentially addressing this shortfall. Diseases were clustered together based upon shared characteristics (such as basic taxonomic division and mode of transmission) in order to consider together those diseases that would synergise operationally in terms of data collection, covariate selection and methodology used. Given the large number of diseases identified, prioritisation is necessary; we addressed this by evaluating both within the context of disease burden as well as considering the diseases' influence within public health organisations and the wider academic community. It is important to stress that the study was focussed on priorities for mapping, and was not a general prioritisation of IDs; this is particularly important to emphasise given that a number of high-burden diseases, including meningitis, pneumonia and some diarrhoeal diseases, were not included in the list of 176 diseases [1,44,45].

Malaria is the infectious disease for which the most detailed and robust global risk maps exist [13,29]. The work of the Malaria Atlas Project [33,55] along with a proliferation of national and local-scale studies [56] has established a mature and sophisticated methodological approach centred on the use of model-based geostatistics to generate continuous surfaces of risk. This has been possible, in part, due to the long history of population-based malaria infection prevalence surveys where researchers and control programmes have used microscopy or rapid diagnostic tests to establish the proportion of randomly sampled individuals testing positive for malaria parasitaemia [30,57]. Crucial for geospatial mapping, such data are increasingly georeferenced with a latitude and longitude for each observation established *via* gazetteer methods (recorded location names linked to digital atlases) or directly using Global Positioning System (GPS) technology at the time of survey [58,59].

The high prioritisation of HIV and tuberculosis shown in the current study brings into sharp focus the need for similar mapping activities to be established for HIV and tuberculosis. All three diseases have an established history of routine and survey-based data collection that, in comparison to many other diseases, is of relatively high quality and consistency, laying the foundation for similar statistical mapping approaches to those used for malaria to be applied.

A cornerstone of HIV surveillance over the last several decades has been routine blood testing for HIV infection in mothers attending sentinel antenatal clinics. Such data provide rich longitudinal observations of prevalence in this demographic group and the potential exists to combine these with cross-sectional data from nationally representative household surveys [60] to generate optimal space-time models of the changing geographical pattern of infection across individual countries. Unlike HIV and malaria, population-based tuberculosis prevalence testing is not currently included as part of the major international survey programmes [58,61]. However, such surveys (reporting on the prevalence of bacteriologically-confirmed pulmonary tuberculosis) have been undertaken in a number of high-burden countries in recent years, with many more planned in the near future [62]. In a similar way to HIV, the prospect exists of a mapping methodology that could combine survey-based data with the rich health-system based data on new case notifications and other metrics, leveraging the respective strengths of community- and facility-based data. A longer-term goal must be the development of a data assimilation and modelling architecture for all three of these major global diseases to support robust and regularly updated global maps detailing their joint distribution and its evolution though time which can be used to assess the impact of control and international financing efforts [18].

The current analysis identifies a number of different NTDs as priority diseases for mapping, a finding which is consistent with the emphasis given to mapping by the global NTD community in order to geographically target NTDs interventions [63,64]. Specifically, for those NTDs where morbidity control is the goal, including soil-transmitted helminths (STH) and schistosomiasis, interventions are most cost-effective when they are targeted to areas of highest transmission [21]. For those NTDs which are identified for elimination, such as onchocerciasis and lymphatic filariasis, it is essential to know where transmission occurs and when it has been successfully halted following control measures. As a consequence of these operational requirements, large-scale mapping initiatives are underway for each of the main NTDs (Table 1). A challenge for mapping the NTDs, and indeed for mapping many IDs, is the need to continually update maps in order to help track the progress in control. As interventions reduce transmission levels and therefore distributions become more focalised, the need for mapping will only increase.

Unsurprisingly, the top 44 diseases for prioritisation are dominated by those with the highest global burden. However, certain clusters stand out as having high public health attention without a high burden, particularly the picornaviridae cluster and its constituent disease, polio. Although cases are now restricted to a few hundred each year, polio has been identified as an eradication target and is a high priority for many public health stakeholders despite recent obstacles in the eradication schedule [65,66]. In these eradication and elimination scenarios, the role of mapping changes subtly to both identifying areas where cases continue to occur, and in highlighting potential future risks and improving surveillance [67]. Following a similar logic, diseases such as dracunculiasis, African trypanosomiasis and onchocerciasis, in spite of relatively low burdens, remain high policy priorities due to elimination efforts in various parts of the globe [68,69]. These examples demonstrate the utility of the approach used in this study of using assessments of the public health burden as well as metrics of public health attention.

The disease prioritisation methodology used here differs from existing approaches, such as the "Delphi panel method", in that it does not include a panel of experts scoring various criteria associated with the diseases being considered [70–74]. In contrast, this study uses a simplified methodology, placing importance in reproducibility and flexibility, using clearly defined rules to assess available evidence and remove potentially subjective expert-opinion. The methods employed are reliant on independent, third party information, and are assessed in a consistent manner, which can easily respond to changes either in burden or public health focus. The

relative importance of these diseases will most likely change over time, so an approach that can easily accommodate this is preferable. Burden estimation using the GBD is crucial, since it is the leading globally consistent measure by which to compare these various diseases and the effects of their many different clinical manifestations. Any global assessment of 301 causes of mortality and morbidity, and associated sequelae, will be subject to the limitations of data availability and epidemiological understanding as well as model assumptions and implementation [53,75,76], and will require frequent updates in a rapidly changing world. The technique presented here has the advantage of being rapidly updateable, and we will reproduce these numbers with each new iteration of the GBD project. As a consequence, public health authorities can also easily create bespoke prioritisation lists based upon a selection of disease inclusion criteria (such as those endemic to their particular country or region). This can more easily be achieved with the availability of sub-national estimates of disease burden from the GBD study. Country specific estimates of the interest scores can also be generated with greater specificity, and can therefore avoid some of the potential biases resulting from the use of other countries as representatives of each GBD region used in this study.

Additional factors that may influence the disease priority, such as potential economic impact [77–79], were not used in this analysis because insufficient information was available to include these metrics. The methodology outlined above benefits from two metrics that can be applied globally to quantify DALYs and public health priority. As and when measures of additional disease impacts become available, they can and should be incorporated into assessments such as this.

The study also identifies some high DALY groupings that do not have high-level policy interest. Three groupings (Tick-borne (Bacterial), Tick-borne (Viral) and Mammal contact (Viral)) have a cumulative high DALY burden, but relatively low policy rankings and therefore are just outside the top 15 cluster listing. This may reflect the large number of diverse pathogens that make up these groupings, many of which are relatively restricted in distribution and hence would not commonly be prioritised by globally focussed organisations. That said, the high DALY value indicates that these diseases are of international interest, particularly when secondary human-to-human transmission is a possibility such as with Lassa fever and Crimean Congo Haemorrhagic Fever [80]. These conditions further advocate the utility of regional and national level priority estimates.

The exclusion of diseases not suited for occurrence based mapping, and therefore omitted from the prioritisation process (so called Option 1 diseases [1]), is entirely based on cartographic considerations. Some of these diseases are inherently linked to human-to-human interactions, others are endogenous in origin, with the pathogen essentially ubiquitous amongst humans and only occasionally causing opportunist infections in certain scenarios, whilst some have the potential to cause infection anywhere across the globe due to the cosmopolitan distribution of their sources of infection, whether they be environmental or human based. Many of these diseases can vary spatially, as evidenced by the African meningitis belt, although such variation, when considered relative to the rest of the world, is due to differences in prevalence or intensity, not presence or absence. Occurrence based mapping methods, such as boosted regression trees, rely on binary presence/absence data. For conditions such as the common cold, diphtheria or respiratory syncytial virus, which have the potential to occur across the globe, these mapping techniques are ineffective. It is only through using more advanced methods, such as model-based geostatistics, that maps analysing the variation in intensity of these diseases can be produced. The limitation of this methodology is the amount of prevalence survey data required, which for many diseases is not comprehensive or detailed enough to allow for global analyses. Basic human related covariates, such as population density, urban extent profiles and national vaccination statistics can be used to explain a degree of the global variation in

these diseases, but fall short of the wealth of information that can be derived from comprehensive global prevalence datasets, such as those available for malaria. As we continue to explore additional data avenues, there will be an increasing number of diseases where such data become available.

The disease prioritisation outlined in this study offers a logical framework for proceeding with disease mapping, which reinforces the necessity of existing programmes and identifies those diseases to focus on next (Table 1). Diseases which will form the initial focus of future study comprise both those with the highest-burden and those of greatest concern to the global health community. The initial top-priority diseases include a range of disease agents and transmission routes, and therefore present a variety of challenges for mapping. The prioritisation and clustering of these diseases presents a clear plan of action designed to maximise the effectiveness and value of future cartographic efforts.

Supporting Information

S1 Text. Supporting Information providing more specific details on the rationale for mapping, linking the Global Burden of Disease and those identified as mapping targets, feeds for the identifying diseases of interest to public health stakeholders and a full cluster ranking.
(DOCX)

Acknowledgments

We thank Maria Devine, Kirsten Duda and Moritz Kraemer for proofreading.

Author Contributions

Conceived and designed the experiments: SIH. Performed the experiments: REH DMP AW. Analyzed the data: REH DMP KEB AW NG. Contributed reagents/materials/analysis tools: SJB AW DMP REH HHK TV CJLM. Wrote the paper: DMP REH AW KEB NG PWG THF AJG AMK CHS SJB CLM SIH SFD LKK HHK TV CJLM.

References

1. Hay SI, Battle KE, Pigott DM, Smith DL, Moyes CL, et al. (2013) Global mapping of infectious disease. Philos Trans R Soc Lond B Biol Sci 368: 20120250. doi: 10.1098/rstb.2012.0250 PMID: 23382431
2. Brady OJ, Gething PW, Bhatt S, Messina JP, Brownstein JS, et al. (2012) Refining the global spatial limits of dengue virus transmission by evidence-based consensus. PLoS Negl Trop Dis 6: e1760. doi: 10.1371/journal.pntd.0001760 PMID: 22880140
3. Guerra CA, Gikandi PW, Tatem AJ, Noor AM, Smith DL, et al. (2008) The limits and intensity of *Plasmodium falciparum* transmission: Implications for malaria control and elimination worldwide. PLoS Med 5: 300–311.
4. Guerra CA, Howes RE, Patil AP, Gething PW, Van Boeckel TP, et al. (2010) The international limits and population at risk of *Plasmodium vivax* transmission in 2009. PLoS Negl Trop Dis 4: e774. doi: 10.1371/journal.pntd.0000774 PMID: 20689816
5. Pullan RL, Brooker SJ (2012) The global limits and population at risk of soil-transmitted helminth infections in 2010. Parasit Vectors 5: 81. doi: 10.1186/1756-3305-5-81 PMID: 22537799
6. Ellis CK, Carroll DS, Lash RR, Peterson AT, Damon IK, et al. (2012) Ecology and geography of human monkeypox case occurrences across Africa. J Wildl Dis 48: 335–347. PMID: 22493109
7. Cano J, Rebollo MP, Golding N, Pullan RL, Crellen T, et al. (2014) The global distribution and transmission limits of lymphatic filariasis: past and present. Parasit Vectors 7: 466. PMID: 25303991
8. Pigott DM, Bhatt S, Golding N, Duda KA, Battle KE, et al. (2014) Global distribution maps of the leishmaniases. eLife: e02851.

9. Fichet-Calvet E, Rogers DJ (2009) Risk maps of Lassa fever in West Africa. PLoS Negl Trop Dis 3: e388. doi: 10.1371/journal.pntd.0000388 PMID: 19255625

10. Smith JL, Flueckiger RM, Hooper PJ, Polack S, Cromwell EA, et al. (2013) The geographical distribution and burden of trachoma in Africa. PLoS Negl Trop Dis 7: e2359. doi: 10.1371/journal.pntd.0002359 PMID: 23951378

11. Hay SI, Okiro EA, Gething PW, Patil AP, Tatem AJ, et al. (2010) Estimating the global clinical burden of *Plasmodium falciparum* malaria in 2007. PLoS Med 7: e1000290. doi: 10.1371/journal.pmed.1000290 PMID: 20563310

12. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, et al. (2013) The global distribution and burden of dengue. Nature 496: 504–507. doi: 10.1038/nature12060 PMID: 23563266

13. Gething PW, Elyazar IRF, Moyes CL, Smith DL, Battle KE, et al. (2012) A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. PLoS Negl Trop Dis 6: e1814. doi: 10.1371/journal.pntd.0001814 PMID: 22970336

14. Pullan RL, Smith JL, Jasrasaria R, Brooker SJ (2014) Global numbers of infection and disease burden of soil transmitted helminth infections in 2010. Parasit Vectors 7: 37. doi: 10.1186/1756-3305-7-37 PMID: 24447578

15. Brownstein JS, Freifeld CC, Madoff LC (2009) Digital disease detection—harnessing the web for public health surveillance. N Engl J Med 360: 2153–2157. doi: 10.1056/NEJMp0900702 PMID: 19423867

16. Freifeld CC, Mandl KD, Ras BY, Bronwnstein JS (2008) HealthMap: global infectious disease monitoring through automated classification and visualization of internet media reports. J Am Med Inform Assn 15: 150–157. PMID: 18096908

17. Tatem AJ, Smith DL, Gething PW, Kabaria CW, Snow RW, et al. (2010) Ranking of elimination feasibility between malaria-endemic countries. Lancet 376: 1579–1591. doi: 10.1016/S0140-6736(10)61301-3 PMID: 21035838

18. Pigott DM, Atun R, Moyes CL, Hay SI, Gething PW (2012) Funding for malaria control 2006–2010: a comprehensive global assessment. Malar J 11: 246. doi: 10.1186/1475-2875-11-246 PMID: 22839432

19. Gyapong JO, Kyelem D, Kleinschmidt I, Agbo K, Ahouandogbo F, et al. (2002) The use of spatial analysis in mapping the distribution of bancroftian filariasis in four West African countries. Ann Trop Med Parasitol 96: 695–705. PMID: 12537631

20. Zoure HGM, Wanji S, Noma M, Amazigo UV, Diggle PJ, et al. (2011) The geographic distribution of *Loa loa* in Africa: results of large-scale implementation of the Rapid Assessment Procedure for Loiasis (RAPLOA). PLoS Negl Trop Dis 5: e1210. doi: 10.1371/journal.pntd.0001210 PMID: 21738809

21. Brooker S, Kabatereine NB, Gyapong JO, Stothard JR, Utzinger J (2009) Rapid mapping of schistosomiasis and other neglected tropical diseases in the context of integrated control programmes in Africa. Parasitology 136: 1707–1718. doi: 10.1017/S0031182009005940 PMID: 19450373

22. Sturrock HJW, Picon D, Sabasio A, Oguttu D, Robinson E, et al. (2009) Integrated mapping of neglected tropical diseases: epidemiological findings and control implications for northern Bahr-el-Ghazal state, southern Sudan. PLoS Negl Trop Dis 3: e537. doi: 10.1371/journal.pntd.0000537 PMID: 19859537

23. Sturrock HJW, Hsiang MS, Cohen JM, Smith DL, Greenhouse B, et al. (2013) Targeting asymptomatic malaria infections: active surveillance in control and elimination. PLoS Med 10: e1001467. doi: 10.1371/journal.pmed.1001467 PMID: 23853551

24. WHO (2014) International travel and health: 2014 updates. Geneva: World Health Organization. 248 p.

25. Field VK, Ford L, Hill DR, editors (2010) Health information for overseas travel. London: National Travel Health Network and Centre. 398 p.

26. CDC (2013) CDC health information for international travel 2014. New York: Oxford University Press. 688 p.

27. Diggle PJ, Ribeiro PJ (2010) Model-based geostatistics: Springer. 246 p.

28. Hay SI, George DB, Moyes CL, Brownstein JS (2013) Big data opportunities for global infectious disease surveillance. PLoS Med 10: e1001413. doi: 10.1371/journal.pmed.1001413 PMID: 23565065

29. Gething PW, Patil AP, Smith DL, Guerra CA, Elyazar IRF, et al. (2011) A new world malaria map: *Plasmodium falciparum* endemicity in 2010. Malar J 10: 378. doi: 10.1186/1475-2875-10-378 PMID: 22185615

30. Guerra CA, Hay SI, Lucioparedes LS, Gikandi PW, Tatem AJ, et al. (2007) Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project. Malar J 6: 17. PMID: 17306022

31. Hurlimann E, Schur N, Boutsika K, Stensgaard AS, Laserna de Himpsl M, et al. (2011) Toward an open-access global database for mapping, control, and surveillance of neglected tropical diseases. PLoS Negl Trop Dis 5: e1404. doi: 10.1371/journal.pntd.0001404 PMID: 22180793

32. Brooker S, Rowlands M, Haller L, Savioli L, Bundy DAP (2000) Towards an atlas of human helminth infection in sub-Saharan Africa: the use of geographical information systems (GIS). Parasitol Today 16: 303–307. PMID: 10858650

33. Hay SI, Snow RW (2006) The Malaria Atlas Project: developing global maps of malaria risk. PLoS Med 3: 2204–2208.

34. Wertheim HFL, Horby P, Woodall JP (2012) Atlas of human infectious diseases, https://infectionatlas.org/. Oxford: Wiley-Blackwell. 280 p.

35. Magill AJ, Hill DR, Solomon T, Ryan ET, editors (2013) Hunter's tropical medicine and emerging infectious disease. London: Elsevier. 1190 p.

36. Farrar JJ, Hotez PJ, Junghanss T, Kang G, Lalloo D, et al., editors (2014) Manson's tropical diseases. London: Elsevier. 1337 p.

37. Guernier V, Hochberg ME, Guegan JFO (2004) Ecology drives the worldwide distribution of human diseases. PLoS Biol 2: 740–746.

38. Dunn RR, Davies TJ, Harris NC, Gavin MC (2010) Global drivers of human pathogen richness and prevalence. Proc R Soc Lond B Biol Sci 277: 2587–2595.

39. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, et al. (2008) Global trends in emerging infectious diseases. Nature 451: 990–994. doi: 10.1038/nature06536 PMID: 18288193

40. Moyes CL, Henry AJ, Golding N, Huang Z, Singh B, et al. (2014) Defining the geographical range of the *Plasmodium knowlesi* reservoir. PLoS Negl Trop Dis 8: e2780. doi: 10.1371/journal.pntd.0002780 PMID: 24676231

41. Barber BE, William T, Grigg MJ, Menon J, Auburn S, et al. (2013) A prospective comparative study of knowlesi, falciparum, and vivax malaria in Sabah, Malaysia: high proportion with severe disease from *Plasmodium knowlesi* and *Plasmodium vivax* but no mortality with early referral and artesunate therapy. Clin Infect Dis 56: 383–397. doi: 10.1093/cid/cis902 PMID: 23087389

42. Berman JJ (2012) Taxonomic guide to infectious diseases: understanding the biologic classes of pathogenic organisms. London: Elsevier. 355 p.

43. Brady OJ, Golding N, Pigott DM, Kraemer MU, Messina JP, et al. (2014) Global temperature constraints on *Aedes aegypti* and *Ae. albopictus* persistence and competence for dengue virus transmission. Parasit Vectors 7: 338. doi: 10.1186/1756-3305-7-338 PMID: 25052008

44. GBD 2013 Disease and Injury Incidence and Prevalence Collaborators (2015) Global, regional, and national incidence, prevalence and YLDs for 301 acute and chronic diseases and injuries for 188 countries, 1990–2013: a systemtatic analysis for the Global Burden of Disease Study 2013. Under submission.

45. GBD 2013 Mortality and Causes of Death Collaborators (2015) Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematics analysis for the Global Burden of Disease Study 2013. Lancet 385: 117–171. doi: 10.1016/S0140-6736(14)61682-2 PMID: 25530442

46. WHO International statistical classification of diseases and related health problems 10th revision. http://apps.who.int/classifications/icd10/browse/2010/en. Accessed: June 2014

47. Edberg SC (2005) Global infectious diseases and epidemiology network (GIDEON): a world wide web-based program for diagnosis and informatics in infectious diseases. Clin Infect Dis 40: 123–126. PMID: 15614701

48. Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA 102: 16569–16572. PMID: 16275915

49. Scopus Homepage. www.scopus.com. Accessed: July 2014

50. McIntyre KM, Hawkes I, Waret-Szkuta A, Morand S, Baylis M (2011) The *h*-Index as a quantitative indicator of the relative impact of human diseases. PLoS One 6: e19558. doi: 10.1371/journal.pone.0019558 PMID: 21625581

51. Murray CJ, Ortblad KF, Guinovart C, Lim SS, Wolock TM, et al. (2014) Global, regional, and national incidence and mortality for HIV, tuberculosis, and malaria during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet.

52. WHO (2012) Accelerating work to overcome the global impact of neglected tropical diseases—a roadmap for implementation. Geneva: World Health Organization. 16 p.

53. Hotez PJ, Alvarado M, Basanez MG, Bolliger I, Bourne R, et al. (2014) The Global Burden of Disease study 2010: interpretation and implications for the Neglected Tropical Diseases. PLoS Negl Trop Dis 8: e2865. doi: 10.1371/journal.pntd.0002865 PMID: 25058013

54. Southwood TRE, Henderson PA (2000) Ecological methods. Oxford: Wiley-Blackwell. 592 p.

55. Malaria Atlas Project Homepage. http://www.map.ox.ac.uk/. Accessed: July 2014

56. Clements ACA, Reid HL, Kelly GC, Hay SI (2013) Further shrinking the malaria map: how can geospatial science help to achieve malaria elimination? Lancet Infect Dis 13: 709–718. doi: 10.1016/S1473-3099(13)70140-3 PMID: 23886334

57. Hay SI, Smith DL, Snow RW (2008) Measuring malaria endemicity from intense to interrupted transmission. Lancet Infect Dis 8: 369–378. doi: 10.1016/S1473-3099(08)70069-0 PMID: 18387849

58. The DHS Program Homepage. http://dhsprogram.com/. Accessed: July 2014

59. Moyes CL, Temperley WH, Henry AJ, Burgert CR, Hay SI (2013) Providing open access data online to advance malaria research and control. Malar J 12: e161.

60. The DHS Program HIV/AIDS survey indicators database. http://hivdata.dhsprogram.com/. Accessed: July 2014

61. UNICEF Multiple indicator cluster survey (MICS). http://www.unicef.org/statistics/index_24302.html. Accessed: July 2014

62. WHO (2013) Global tuberculosis report 2013. Geneva: World Health Organization. 289 p.

63. Pullan RL, Gething PW, Smith JL, Mwandawiro CS, Sturrock HJW, et al. (2011) Spatial modelling of soil-transmitted helminth infections in Kenya: a disease control planning tool. PLoS Negl Trop Dis 5: e958. doi: 10.1371/journal.pntd.0000958 PMID: 21347451

64. WHO (2010) Working to overcome the global impact of neglected tropical diseases. First WHO report on neglected tropical diseases. Geneva: World Health Organization. 172 p.

65. GPEI (2013) Eighth report—October 2013. London: Global Polio Eradication Initiative. 59 p.

66. Upfill-Brown AM, Lyons HM, Pate MA, Shuaib F, Baig S, et al. (2014) Predictive spatial risk model of poliovirus to aid prioritization and hasten eradication in Nigeria. BMC Med 12: 92. doi: 10.1186/1741-7015-12-92 PMID: 24894345

67. Andre M (2013) Assessing the risks for Poliovirus outbreaks in polio-free countries—Africa 2012–2013. MMWR Morb Mortal Wkly Rep 62: 768–772. PMID: 24048153

68. WHO (2013) Control and surveillance of human African trypanosomiasis: report of a WHO expert committee. Geneva: World Health Organization. 237 p.

69. Eberhard M (2013) Progress toward elimination of onchocerciasis in the Americas-1993-2012. MMWR Morb Mortal Wkly Rep 62: 405–408. PMID: 23698606

70. Balabanova Y, Gilsdorf A, Buda S, Burger R, Eckmanns T, et al. (2011) Communicable diseases prioritized for surveillance and epidemiological research: results of a standardized prioritization procedure in Germany, 2011. PLoS One 6: e25691. doi: 10.1371/journal.pone.0025691 PMID: 21991334

71. DISCONTOOLS Project (2012) Approaches to the prioritisation of diseases to focus and prioritise research in animal health: a worldwide review of existing methodologies. Brussels: IFAH-Europe. 18 p.

72. Gilsdorf A, Krause G (2011) Prioritisation of infectious diseases in public health: feedback on the prioritisation methodology, 15 July 2008 to 15 January 2009. Euro Surveill 16: 15–21.

73. Krause G, Prioritisation Working Group (2008) How can infectious diseases be prioritized in public health? A standardized prioritization scheme for discussion. EMBO Rep 9: S22–S27. doi: 10.1038/embor.2008.76 PMID: 18578019

74. WHO (2006) Setting priorities in communicable disease surveillance. Geneva: World Health Organization. 29 p.

75. Byass P, de Courten M, Graham WJ, Laflamme L, McCaw-Binns A, et al. (2013) Reflections on the Global Burden of Disease 2010 estimates. PLoS Med 10: e1001477. doi: 10.1371/journal.pmed.1001477 PMID: 23843748

76. de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, et al. (2012) Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. Lancet Oncol 13: 607–615. doi: 10.1016/S1470-2045(12)70137-7 PMID: 22575588

77. Bonds MH, Keenan DC, Rohani P, Sachs JD (2010) Poverty trap formed by the ecology of infectious diseases. Proc R Soc Lond B Biol Sci 277: 1185–1192.

78. Hotez PJ (2013) Forgotten people, forgotten diseases: the neglected tropical diseases and their impact on global health and development. Washington DC: ASM Press. 275 p.

79. Molyneux DH (2014) Neglected tropical diseases: now more than just 'other diseases'—the post-2015 agenda. Int Health 6: 172–180. doi: 10.1093/inthealth/ihu037 PMID: 24969646

80. Bannister B (2010) Viral haemorrhagic fevers imported into non-endemic countries: risk assessment and management. Brit Med Bull 95: 193–225. doi: 10.1093/bmb/ldq022 PMID: 20682627

81. WHO (2013) World malaria report 2013. Geneva: World Health Organization. 255 p.

82. UNAIDS UNAIDS. http://www.unaids.org/en/. Accessed: July 2014

83. Global Atlas of Helminth Infections GAHI: global atlas of helminth infections. http://www.thiswormyworld.org/. Accessed: July 2014

84. Simarro PP, Cecchi G, Paone M, Franco JR, Diarra A, et al. (2010) The Atlas of human African trypanosomiasis: a contribution to global mapping of neglected tropical diseases. Int J Health Geogr 9: e57.

85. Cano J, Rebollo MP, Golding N, Pullan RL, Crellen T, et al. (2014) The global distribution and transmission limits of lymphatic filariasis: past and present. PLoS Negl Trop Dis under submission.

86. GPEI Data and monitoring. http://www.polioeradication.org/Dataandmonitoring.aspx. Accessed: July 2014

87. African Programme for Onchocerciasis Control Country profiles. http://www.who.int/apoc/countries/en/. Accessed: July 2014

88. Zoure HGM, Noma M, Tekle AH, Amazigo UV, Diggle PJ, et al. (2014) The geographic distribution of onchocerciasis in the 20 participating countries of the African Programme for Onchocerciasis control: (2) pre-control endemicity levels and estimated number infected. Parasit Vectors 7: 326. doi: 10.1186/1756-3305-7-326 PMID: 25053392

89. Trachoma Atlas Global atlas of trachoma. http://www.trachomaatlas.org/. Accessed: July 2014

Chapter 3

A framework for mapping vector-borne diseases: the leishmaniases.

As identified in Chapter 2, the leishmaniases have the greatest disability adjusted life years amongst the neglected tropical diseases and are one of the top 10 prioritised clusters. It is surprising therefore that no global, evidence-based approach to mapping the distribution of these various conditions had been implemented; indeed only one map, consisting of expert opinion ranges, had been published, as part of a World Health Organization report. This paper uses a comprehensive database of reported cases, coupled with assessments on the reliability of information on the presence or absence of the disease in each district of the world (termed evidence consensus), to model the environmental suitability for leishmaniasis transmission across the globe, using boosted regression trees. This work has been published in *eLife* and is included here in its final form. The additional information referred to in this chapter is included in the Appendix of this thesis. Also included in the Appendix is a more detailed paper discussing the occurrence data generation and curation protocols.

eLIFE elifesciences.org

RESEARCH ARTICLE

Global distribution maps of the leishmaniases

David M Pigott[1]*, Samir Bhatt[1], Nick Golding[1], Kirsten A Duda[1], Katherine E Battle[1], Oliver J Brady[1], Jane P Messina[1], Yves Balard[2], Patrick Bastien[2,3], Francine Pratlong[2,3], John S Brownstein[4,5], Clark C Freifeld[5,6], Sumiko R Mekaru[5], Peter W Gething[1], Dylan B George[7], Monica F Myers[1], Richard Reithinger[8], Simon I Hay[1,7]

[1]Spatial Ecology and Epidemiology Group, Department of Zoology, University of Oxford, Oxford, United Kingdom; [2]Laboratoire de Parasitologie–Mycologie, UFR Médecine, Université Montpellier 1 and UMR 'MiVEGEC', CNRS 5290/IRD 224, Montpellier, France; [3]Departement de Parasitologie–Mycologie, CHRU de Montpellier, Centre National de Référence des Leishmanioses, Montpellier, France; [4]Department of Pediatrics, Harvard Medical School, Boston, United States; [5]Children's Hospital Informatics Program, Boston Children's Hospital, Boston, United States; [6]Department of Biomedical Engineering, Boston University, Boston, United States; [7]Fogarty International Center, National Institutes of Health, Bethesda, United States; [8]Global Health Group, RTI International, Washington DC, United States

Abstract The leishmaniases are vector-borne diseases that have a broad global distribution throughout much of the Americas, Africa, and Asia. Despite representing a significant public health burden, our understanding of the global distribution of the leishmaniases remains vague, reliant upon expert opinion and limited to poor spatial resolution. A global assessment of the consensus of evidence for leishmaniasis was performed at a sub-national level by aggregating information from a variety of sources. A database of records of cutaneous and visceral leishmaniasis occurrence was compiled from published literature, online reports, strain archives, and GenBank accessions. These, with a suite of biologically relevant environmental covariates, were used in a boosted regression tree modelling framework to generate global environmental risk maps for the leishmaniases. These high-resolution evidence-based maps can help direct future surveillance activities, identify areas to target for disease control and inform future burden estimation efforts.
DOI: 10.7554/eLife.02851.001

***For correspondence:** david.pigott@zoo.ox.ac.uk

Competing interests: The authors declare that no competing interests exist.

Funding: See page 16

Received: 23 March 2014
Accepted: 26 June 2014
Published: 27 June 2014

Reviewing editor: Stephen Tollman, Wits University, South Africa

Introduction

The leishmaniases are a group of protozoan diseases transmitted to humans and other mammals by phlebotomine sandflies (***Murray et al., 2005***; ***WHO, 2010***). Considered as one of the neglected tropical diseases (NTD) (***WHO, 2009***), the leishmaniases can be caused by around 20 *Leishmania* species and include a complex life cycle involving multiple arthropod vectors and mammalian reservoir species (***Ashford, 1996***; ***Ready, 2013***). Sandflies belonging to either *Phlebotomus* spp. (Old World) or *Lutzomyia* spp. (New World) are the primary vectors; domestic dogs, rodents, sloths, and opossums are amongst a long list of mammals that are either incriminated or suspected reservoir hosts. Non-vector transmission (e.g., by accidental laboratory infection, blood transfusion, or organ transplantation) is possible, but rare (***Cardo, 2006***). Transmission of the leishmaniases can be either anthroponotic or zoonotic. The leishmaniases rank as the leading NTD in terms of mortality and morbidity with an estimated 50,000 deaths in 2010 (***Lozano et al., 2012***) and 3.3 million disability adjusted life years (***Murray et al., 2012***).

eLife digest Each year 1–2 million people are diagnosed with a tropical disease called leishmaniasis, which is caused by single-celled parasites. People are infected when they are bitten by sandflies carrying the parasite, and often develop skin lesions around the bite site. Though mild cases may recover on their own or with treatment, sometimes the parasites multiply and spread elsewhere causing further skin lesions and facial disfigurement. Furthermore, the parasites can also infect internal organs such as the spleen and the liver, which without treatment can be fatal.

The parasites that cause leishmaniasis are found in 88 countries around the world, mainly in South and Central America, Africa, Asia, and southern Europe. However, over 90% of potentially fatal infections occur in just six countries: Brazil, Ethiopia, Sudan, South Sudan, India, and Bangladesh. Although a few studies have looked at the distribution of leishmaniasis within different countries, we still do not have a complete picture of the distribution of the disease on a global scale.

To address this, Pigott et al. set out to create detailed maps of the distribution of leishmaniasis and the factors that promote its spread. Similar techniques had been previously used to map dengue fever, another tropical disease. Computer modelling was used to generate the maps based on data about the environment at the locations of known cases of leishmaniasis. This information was then used to infer the likelihood of leishmaniasis being present at other locations across the globe.

Based on their maps, Pigott et al. estimate that about 1.7 billion people, or one quarter of the world's population, live in areas where they are at potential risk of leishmaniasis. People living in built-up areas outside of cities are at the greatest risk, likely because some sandfly species prefer to live near dwellings, but other social and economic factors also contribute to the risk of catching this disease.

The factors driving the transmission of leishmaniasis differed in the Old World (Europe, Africa and Asia) and the New World (the Americas): built-up areas were more likely to be at risk in the Old World, while temperature and rainfall were bigger factors affecting risk in the New World. It is hoped that the maps created by Pigott et al. will help inform future estimates of the burden of leishmaniasis and target surveillance and disease control efforts more effectively to combat this tropical disease.

DOI: 10.7554/eLife.02851.002

Symptoms of *Leishmania* infection can take many different and diverse forms (***Banuls et al., 2011***), the two main outcomes being cutaneous leishmaniasis (CL) and visceral leishmaniasis (VL). Cutaneous leishmaniasis typically presents as cutaneous nodules or lesions at the site of the sandfly bite (localised cutaneous leishmaniasis). In some cases, parasites disseminate through the skin and present as multiple non-ulcerative nodules (diffuse cutaneous leishmaniasis, DCL) or propagate through the lymphatic system resulting in nasobronchial and buccal mucosal tissue destruction (mucosal leishmaniasis, ML) (***Reithinger et al., 2007***; ***Dedet and Pratlong, 2009***). Localised CL may resolve spontaneously and usually responds well to treatment; management of DCL and ML cases is more difficult and cases may take considerably longer to resolve, if at all. Visceral leishmaniasis generally affects the spleen, liver, or other lymphoid tissues, and, if left untreated, is fatal; a fraction of successfully treated VL cases may result in maculopapular or nodular rashes (post-kala-azar dermal leishmaniasis) (***Murray et al., 2005***; ***Dedet and Pratlong, 2009***). While the *Leishmania* species determines which of the main two forms of the leishmaniases will result from infection, establishment, progression, and severity of infection as well as treatment regimen and outcome is dependent on a range of other factors, including parasite strain, characteristics of sandfly saliva, parasite infection with *Leishmania* RNA virus, host genetics, and immunosuppression, particularly due to HIV co-infection (***Reithinger et al., 2007***; ***Ives et al., 2011***; ***Novais et al., 2013***).

Species distribution models provide a robust means of mapping these diseases at a global level. These models define a set of conditions, from a selection of environmental covariates, which best categorise known occurrences. Through this categorisation, areas of unknown pathogen presence can be identified and thus a global evaluation of environmental suitability for presence can be made.

A variety of factors can influence the distribution of an organism, including an array of environmental and other abiotic characteristics as well as biotic factors (***Peterson, 2008***). Whilst many areas may be environmentally suitable for a given species, other factors may prevent the species from being present in all of these locations. This distinction is often referred to as the difference between the fundamental and the realised niche of the species, the former describing a potential distribution based upon specific features of the environment whilst the latter indicates the distribution we observe. Such a framework can be applied just as successfully in the context of pathogens and their vectors as with macroorganisms (***Peterson et al., 2011***) and has already been applied to the mapping of malaria vectors (***Sinka et al., 2010***, ***2010***, ***2011***) and dengue (***Bhatt et al., 2013***). The relationships between the leishmaniases and environmental and socioeconomic factors known to influence their distribution at a global scale has not previously been considered in a comprehensive and quantitative manner (***Hay et al., 2013***). This study uses these modelling techniques in order to define the first evidence-based global environmental risk maps of the leishmaniases.

Results

Evidence of leishmaniasis

For each province or state across the globe (classed as Admin 1 by the Food and Agriculture Organization's Global Administrative Unit Layers (***FAO, 2008***), totalling some 3450) evidence was collected regarding CL and VL presence or absence. An assessment of the consensus of this evidence ranging from comprehensive agreement on disease presence (+100%) to consensus of disease absence (−100%) was made. ***Figures 1A–4A*** present these evidence consensus maps, with full reasoning for each administrative unit's score outlined in the associated data set (Dryad data set doi: 10.5061/dryad.05f5h). For Brazil, it was possible to perform this analysis at the district level (classed as Admin 2) totalling some 5510 units. In total, 950 Admin 1 units from 84 countries reported a consensus on CL presence greater than indeterminate (a score of 0), with 310 Admin 1 units from 42 countries reporting a complete consensus on the presence of CL. In Brazil, 2469 Admin 2 regions recorded CL cases over the period of investigation. Consensus on the presence of VL (score greater than 0) was reported in 793 Admin 1 units from 77 countries, with 88 Admin 1 units from 32 countries reporting complete consensus on VL. In Brazil, 1320 Admin 2 units recorded VL cases.

Of the 10 countries (Afghanistan, Colombia, Brazil, Algeria, Peru, Costa Rica, Iran, Syria, Ethiopia, and Sudan) that contribute 75% of the global estimated CL incidence (***Alvar et al., 2012***), only Algeria did not have regions of complete evidence consensus on presence due to incomplete and non-contemporary case data. Similarly, of the six countries (Brazil, Ethiopia, Sudan, South Sudan, India, and Bangladesh) that report 90% of all VL cases (***Alvar et al., 2012***), all six had regions of complete consensus on VL.

Figures 1A–4A also show the spatial distribution of occurrence data, defined as one or more reports of leishmaniasis in a given calendar year, collated from a variety of sources. Overall, there is a relatively broad geographic spread and good correspondence with the evidence consensus maps for each disease. Tunisia, Morocco and Brazil report the highest number of unique CL occurrences in any given year, whilst India reported the largest proportion of the VL occurrence data.

Table 1 reports the sources and types of data within the occurrence database. Whilst the majority of occurrence records contain accurate point data (62%), the remainder were recorded at a provincial or district level. Occurrence records for the two diseases were relatively similar in number with a total of 6426 records for CL and 6137 for VL.

Modelled distribution of the leishmaniases

Figures 1B–4B show the global predicted environmental risk maps for CL and VL. ***Table 2*** identifies the top five predictor variables in each of the four modelled regions (since CL and VL were modelled separately in the Old World and New World) as measured by average contribution to the boosted regression trees (BRT) submodels. Peri-urban and urban land cover is an important predictor of the distribution of CL in the Old World and of VL globally. Abiotic factors such as land surface temperature (LST) were better predictors of CL than of VL. In total, LST variables (annual minimum, maximum and mean) explain 21.99% of CL distribution in the Old World and 43.65% of CL distribution in the New World (with maximum LST having the highest relative contribution). Abiotic factors combined (including LST, normalised difference vegetation index (NDVI) and precipitation) accounted for 29.02% and

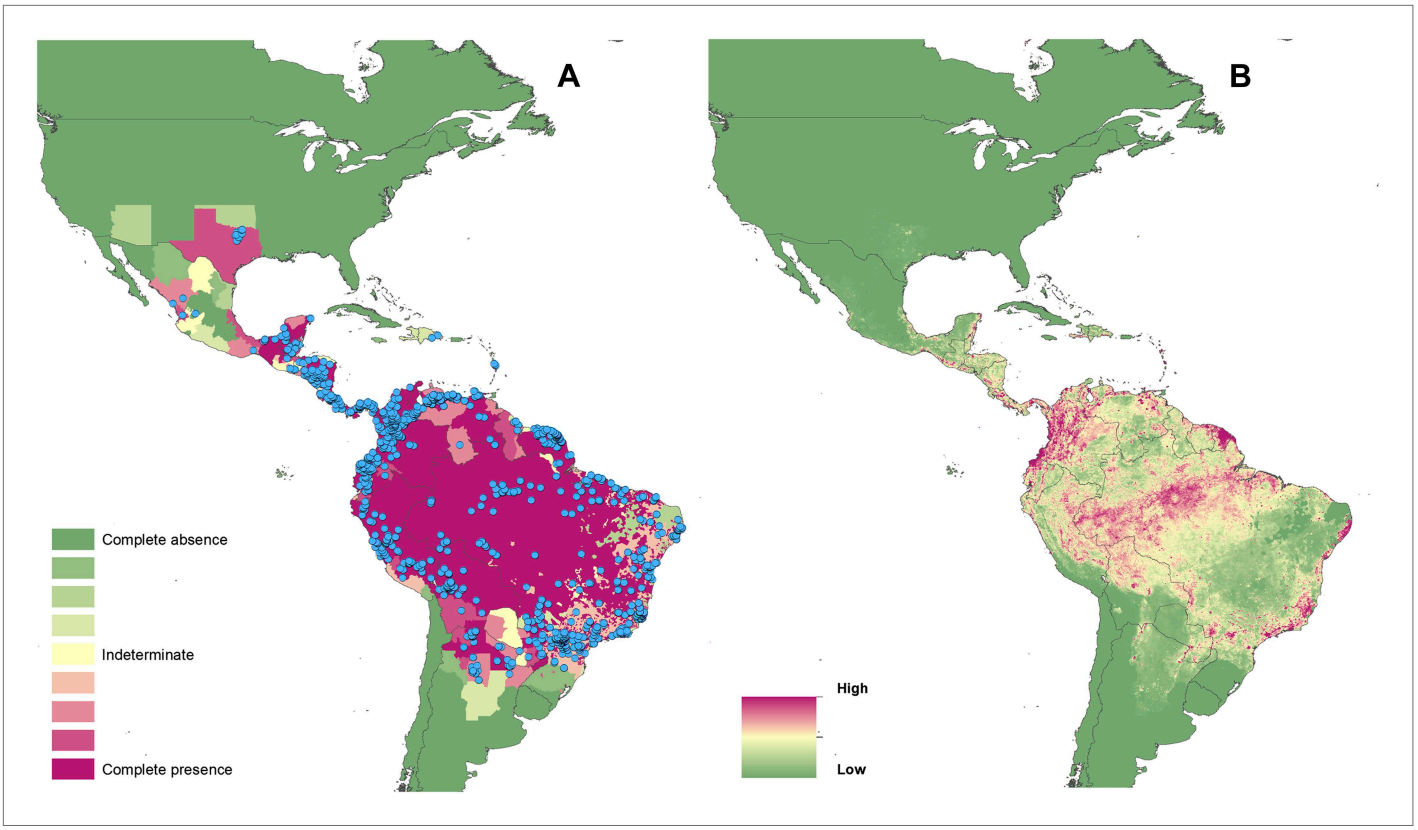


Figure 1. Reported and predicted distribution of cutaneous leishmaniasis in the New World. (**A**) Evidence consensus for presence of the disease ranging from green (complete consensus on the absence: −100%) to purple (complete consensus on the presence of disease: +100%). The blue spots indicate occurrence points or centroids of occurrences within small polygons. (**B**) Predicted risk of cutaneous leishmaniasis from green (low probability of presence) to purple (high probability of presence).

DOI: 10.7554/eLife.02851.003

The following figure supplements are available for figure 1:

Figure supplement 1. Uncertainty associated with predictions in ***Figure 1B***.

DOI: 10.7554/eLife.02851.004

48.55% of VL distribution in the Old World and New World, respectively. Validation statistics for all models were high with a mean area under the receiver operator curve (AUC) above 0.97 and mean correlations above 0.85 for all models.

In the New World, CL is predicted to occur primarily within the Amazon basin and other areas of rainforest. By contrast, VL is predicted to occur mainly along the coastline of Brazil, with sporadic foci across the rest of Southern and Central America. Outside of their main foci, both diseases are strongly associated with urban and peri-urban areas, resulting in a focal distribution throughout much of the New World.

In the Old World, both CL and VL are predicted to be present from the Mediterranean Basin across the Near East to Northwest India, with a few foci in Central China as well as in a thin band of predicted risk across West Africa and in the Horn of Africa. The predicted distribution of VL also extends into Northeast India and China with a large predicted focus in the northwest.

The populations living in areas predicted to be subject to environmental risk of CL and VL are estimated to be 1.71 billion and 1.69 billion, respectively, approximately a quarter of the world's population. ***Figure 4—figure supplement 4*** compares these national estimates to the annual case incidence data from all countries for which at least one case *per annum* was estimated by ***Alvar et al. (2012)***. There is a strong positive association between the two measures of disease occurrence. We provide estimates of the populations at risk in 90 countries for which no human cases of CL or VL were regularly reported (***Alvar et al., 2012***). A full table of this information is presented in the associated Dryad data set (doi: 10.5061/dryad.05f5h). For many of these countries, ***Alvar et al. (2012)*** reported a handful of

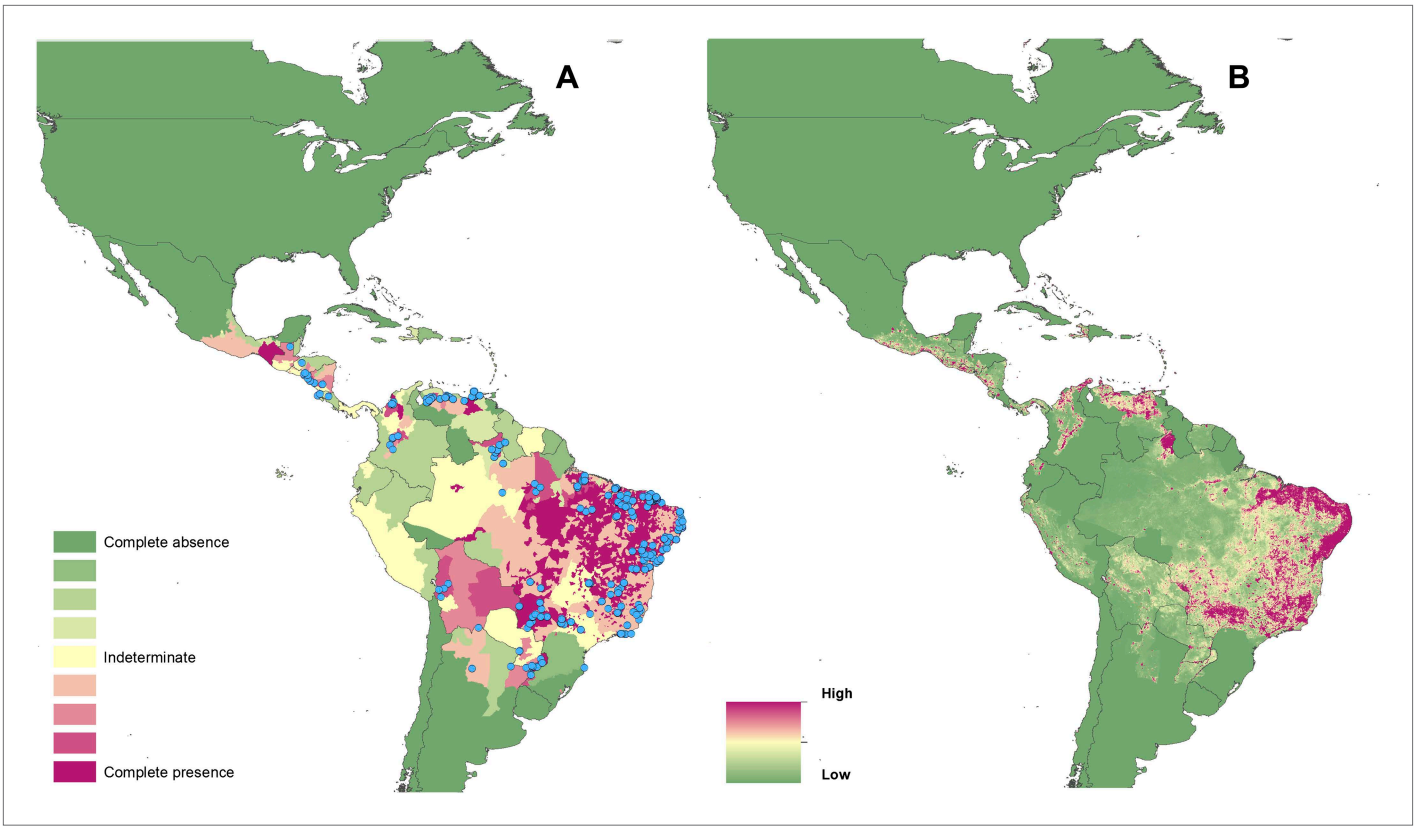


Figure 2. Reported and predicted distribution of visceral leishmaniasis in the New World. (**A**) Evidence consensus for presence of the disease ranging from green (complete consensus on the absence: −100%) to purple (complete consensus on the presence of disease: +100%). The blue spots indicate occurrence points or centroids of occurrences within small polygons. (**B**) Predicted risk of visceral leishmaniasis from green (low probability of presence) to purple (high probability of presence).

DOI: 10.7554/eLife.02851.005

The following figure supplements are available for figure 2:

Figure supplement 1. Uncertainty associated with predictions in ***Figure 2B***.

DOI: 10.7554/eLife.02851.006

sporadic cases over the years indicating very rare occurrence of infection, whilst the remainder were countries with inconclusive evidence of disease presence or absence. It is important to note that the relationship between environmental risk and true incidence of disease remains to be elucidated; however the association between populations living in areas of environmental risk and national level estimates of incidence suggests that the modelled occurrence–incidence relationship approach used by ***Bhatt et al. (2013)*** for dengue could be applied if the necessary longitudinal cohort study data were available.

Discussion

This work has compiled a large body of qualitative and quantitative information on the global distribution of the leishmaniases and employed a statistical modelling framework to generate the first published high-resolution global distribution maps of these diseases.

The evidence consensus maps provide a useful assessment of both global and regional knowledge of these diseases. Whilst in many countries consensus on presence or absence of the leishmaniases exists, in other areas, including large parts of Africa and many states in India, these assessments reveal significant uncertainty in assessing disease presence or absence using currently available evidence. It is in these data-poor countries that increased surveillance efforts should be concentrated to improve our knowledge of the global distribution of the leishmaniases. In some locations, cases have been reported as locally transmitted without the presence of proven vector species, which could indicate a false positive. However, the overall consensus score will reflect any uncertainty associated with the validity

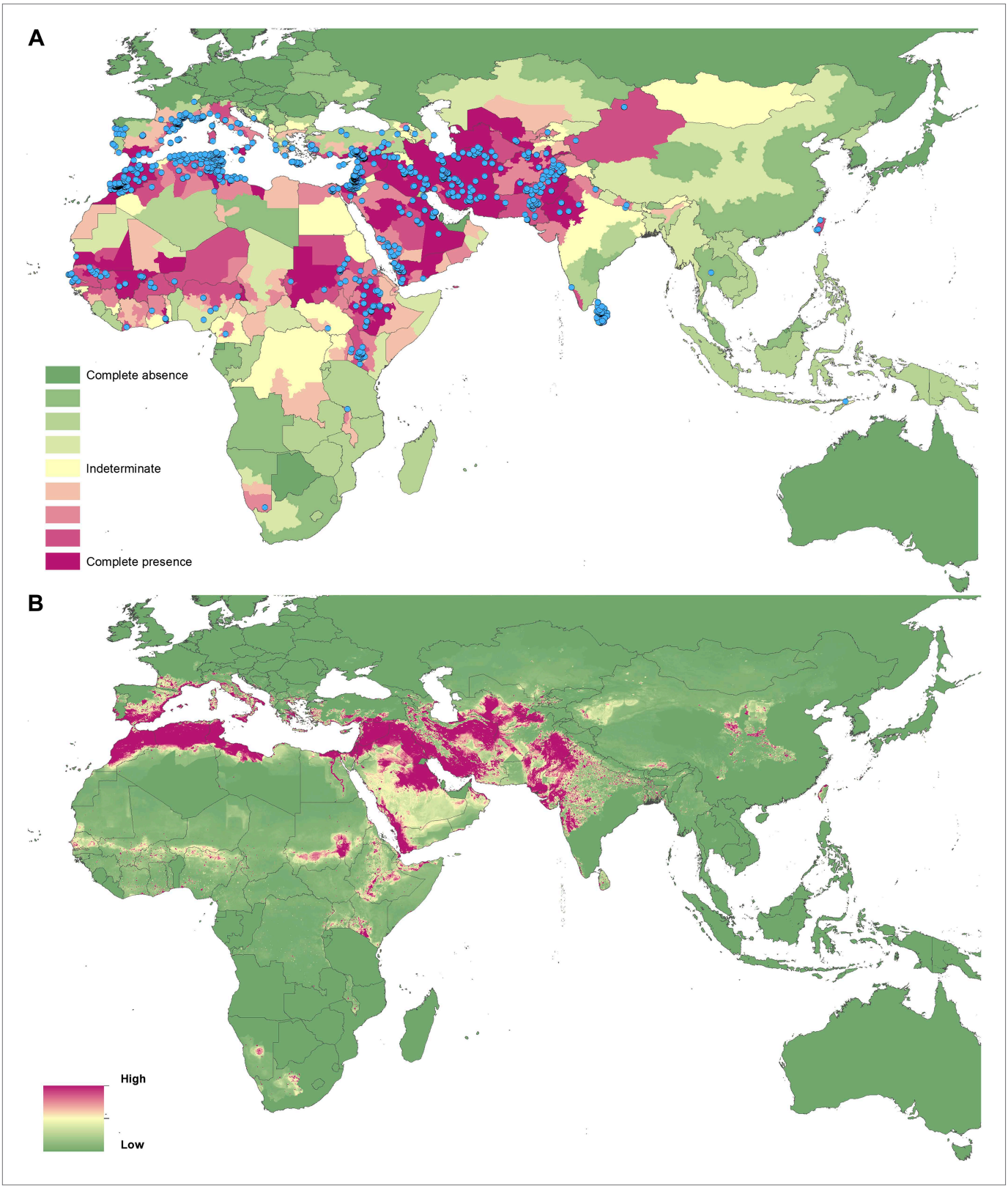


Figure 3. Reported and predicted distribution of cutaneous leishmaniasis in the Old World. (**A**) Evidence consensus for presence of the disease ranging from green (complete consensus on the absence: −100%) to purple (complete consensus on the presence of disease: +100%). The blue spots indicate

Figure 3. Continued on next page

Figure 3. Continued

occurrence points or centroids of occurrences within small polygons. (**B**) Predicted risk of cutaneous leishmaniasis from green (low probability of presence) to purple (high probability of presence).

DOI: 10.7554/eLife.02851.007

The following figure supplements are available for figure 3:

Figure supplement 1. Uncertainty associated with predictions in ***Figure 3B***.

DOI: 10.7554/eLife.02851.008

Figure supplement 2. Reported and predicted distribution of cutaneous leishmaniasis in northeast Africa.

DOI: 10.7554/eLife.02851.009

Figure supplement 3. Reported and predicted distribution of cutaneous leishmaniasis across the Near East, including Syria, Iran and Afghanistan.

DOI: 10.7554/eLife.02851.010

of these reports; if multiple independent sources report autochthonous cases, this increased certainty will be reflected in a higher consensus score. Similarly, whilst the occurrence database contains data from across the globe, this data set is inevitably subject to spatial bias in reporting, with more data reported from more economically developed countries where we already have a good knowledge of the disease (e.g., Spain, France, and Italy).

The complexity and diversity of transmission cycles involving not just humans, but also a multitude of vectors and reservoirs, necessitated a modelling approach which can account for highly non-linear effects of covariates on probability of disease presence. The BRT modelling approach employed is able to do this and has previously been shown to produce highly accurate predictions across a wide range of species (***Elith et al., 2006***, ***2008***). This ecological niche modelling approach is therefore able to deal with not only the variation in parasites causing infection, but also the various life-histories and habitat preferences associated with the different vector species.

A restriction of the BRT approach (in common with other species distribution modelling approaches) is the need for absence data in addition to occurrence data. Since reliable absence data were not available at this spatial scale, the incorporation of pseudo-data into the modelling framework was necessary. The methodology employed in this study attempted to minimise the problems this can cause, by using a probabilistic approach to generate the pseudo-data which incorporates the evidence consensus and distance from existing occurrence points. Similarly, reporting bias within the occurrence database is an issue with all presence-only species distribution models (***Peterson et al., 2011***). If bias is unaccounted for, there is the potential that the model merely reflects factors that correlate with the probability of reporting disease occurrence rather than the disease itself, such as healthcare expenditure (***Phillips et al., 2009***; ***Syfert et al., 2013***). The pseudo-data selection procedures (which included information from both the occurrence data set and the less-biased evidence consensus map) coupled with the model ensembling approach aimed to minimise this potential source of bias.

The differences in the most important predictors of disease presence between the two forms of the disease and between the Old and New Worlds highlight the complex and spatially variable epidemiology of the leishmaniases. Similar to a recent study of the spatial predictors of dengue occurrence (***Bhatt et al., 2013***), environmental and socioeconomic factors were found to be important contributors to the distribution of both CL and VL. For VL, both Old World and New World distributions are driven by peri-urban (and to a lesser degree urban) land cover. This reflects recent trends observed, for instance, in Brazil and Bihar state in India, where areas of highest risk have been found in peridomestic settings (***Bern et al., 2010***; ***Harhay et al., 2011***). This risk factor may well be linked back to aspects of vector bionomics, with many vectors in these regions associating with or near households in general (***Singh et al., 2008***; ***Poche et al., 2011***; ***Uranw et al., 2013***). Furthermore, whilst significant anthroponotic transmission of *L. donovani* occurs across parts of the Old World, zoonotic cycles of VL, primarily tied to canine hosts, dominate *L. infantum* transmission (***Chamaille et al., 2010***; ***Ready, 2013***), with infection in dogs shown to be closely associated with human population density.

Important predictors of CL distribution differed markedly between the Old and New World. Whilst peri-urban land cover was the most important predictor of the disease in the Old World, in the New World temperature was the highest predictor, with abiotic factors predicting 74.18% of CL distribution. This difference in the relative importance of climatic drivers reflects the fact that in the Old World

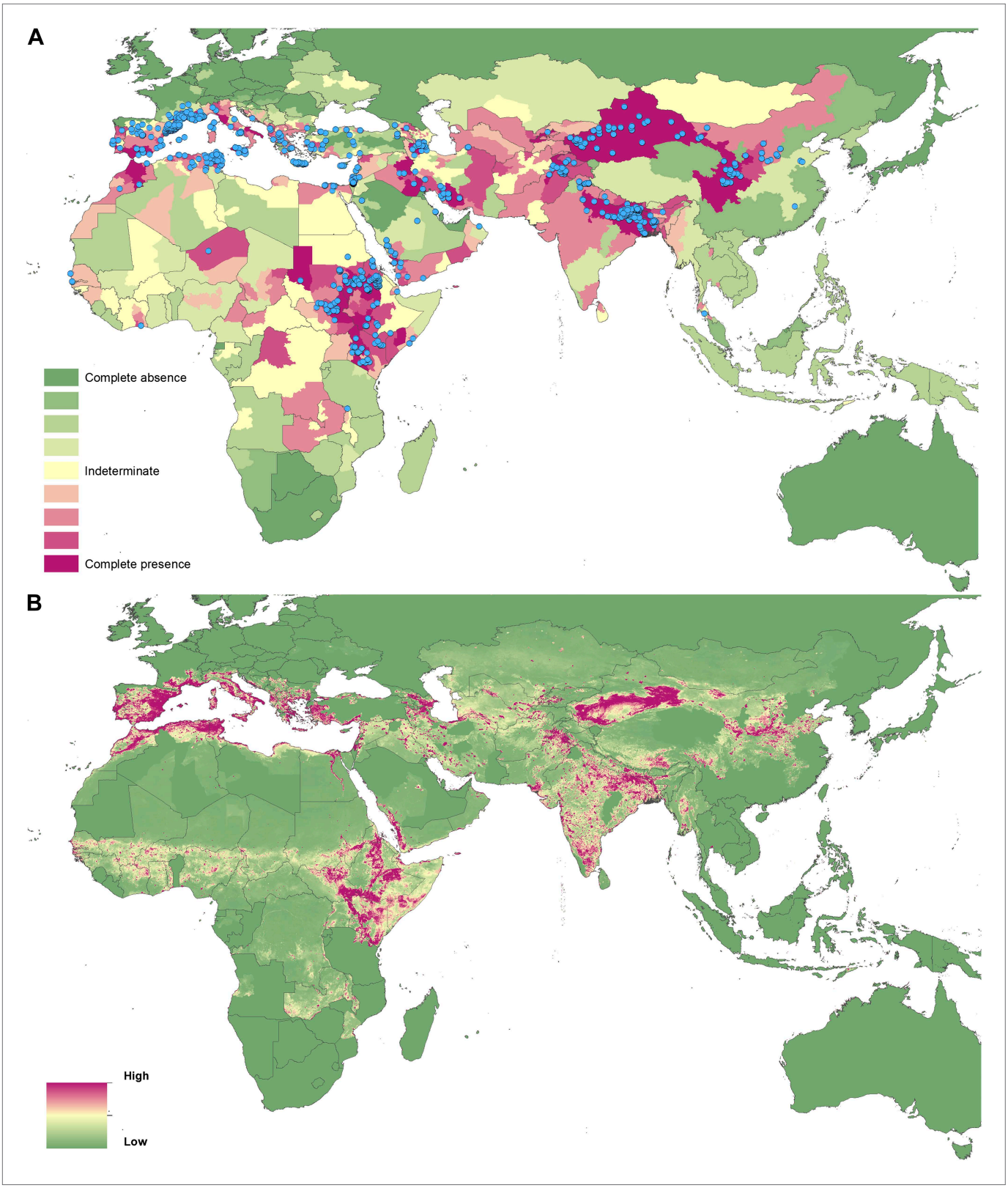


Figure 4. Reported and predicted distribution of visceral leishmaniasis in the Old World. (**A**) Evidence consensus for presence of the disease ranging from green (complete consensus on the absence: −100%) to purple (complete consensus on the presence of disease: +100%). The blue spots indicate occurrence points or centroids of occurrences within small polygons. (**B**) Predicted risk of visceral leishmaniasis from green (low probability of presence) to purple (high probability of presence).

Figure 4. Continued on next page

Figure 4. Continued

DOI: 10.7554/eLife.02851.011

The following figure supplements are available for figure 4:

Figure supplement 1. Uncertainty associated with predictions in ***Figure 4B***.

DOI: 10.7554/eLife.02851.012

Figure supplement 2. Reported and predicted distribution of visceral leishmaniasis in northeast Africa.

DOI: 10.7554/eLife.02851.013

Figure supplement 3. Reported and predicted distribution of visceral leishmaniasis in the Indian subcontinent.

DOI: 10.7554/eLife.02851.014

Figure supplement 4. Population at risk estimates for leishmaniasis.

DOI: 10.7554/eLife.02851.015

the main endemic CL areas are due to both anthroponotically transmitted *L. tropica* and zoonotic cycles of *L. major*, whereas in the New World the disease is primarily associated with sylvatic and zoonotic cycles with a variety of different *Leishmania* spp. and wild reservoir hosts implicated (***Ashford, 1996***; ***Reithinger et al., 2007***; ***WHO, 2010***; ***Lima et al., 2013***; ***Ready, 2013***).

The distribution maps represent a spatially refined assessment of the global environmental risk of leishmaniasis and provide a starting point for various public health activities including targeting areas for control and assessing disease burden. The maps compare favourably to the WHO Expert Committee on the Control of Leishmaniases outputs (***WHO, 2010***), have high model validation statistics and improve upon the existing body of work by providing a finer resolution of risk at a subnational level. Similarly, the countries indicated by ***Alvar et al. (2012)*** as having 90% of all VL and 75% of all CL cases, were all predicted by our maps to have risk for VL and CL, respectively.

There are a number of regions in which our maps do not correspond as closely to these previous findings. Regions such as Northwest China are predicted to have high risk for VL, though the low population densities in this area are likely to lead to very few cases and, given its remoteness, even fewer reported cases. Other regions, such as the Mediterranean coastline of Europe, are predicted to be highly suitable for leishmaniasis, but we see few human cases. This is because the maps presented predict the probability of disease presence in an area, rather than directly infer measures of incidence or burden, which can be influenced by a variety of other factors (e.g., in the Mediterranean coastline of Europe, VL has been associated with immunosuppression). The evidence consensus layer, used to mask out regions with high consensus on leishmaniasis absence, acts as a rough filter on the environmental risk

Table 1. Origin and spatial resolution of leishmaniasis occurrence data

Origin and resolution of occurrence data

	Point data	Province level data	District level data	Total
Cutaneous leishmaniasis				
Literature	3680	879	1220	5779
CNR-L	531	47	31	609
HealthMap	31	–	–	31
GenBank	6	–	1	7
Total	4248	926	1252	6426
Visceral leishmaniasis				
Literature	3050	1500	1068	5618
CNR-L	429	24	29	482
HealthMap	32	1	–	33
GenBank	3	–	1	4
Total	3514	1525	1098	6137

Each cell gives the number of occurrence records added to the data set by considering each additional datasource after removing duplicate records. Occurrence records are separated by spatial resolution—whether they are recorded as points (typically representing settlements) or as province level (admin 1) or district level (admin 2) data.

DOI: 10.7554/eLife.02851.016

Table 2. Mean relative contribution of predictor variables to the ensemble BRT models of CL and VL in both the Old and New World

Top predictors of CL	Relative contribution	Top predictors of VL	Relative contribution
Old world			
Peri-urban extents	47.34	Peri-urban extents	51.50
Minimum LST	18.36	Urban extents	17.38
Urban extents	9.01	Maximum NDVI	7.87
G-Econ	7.33	Minimum LST	5.87
Minimum Precipitation	4.95	Maximum Precipitation	4.00
New World			
Maximum LST	36.91	Peri-urban extents	25.90
Peri-urban extents	18.61	Urban extents	21.24
Maximum precipitation	12.06	Mean LST	9.18
Minimum precipitation	6.21	Mean NDVI	7.83
Minimum LST	4.39	Maximum LST	6.40

LST = Land Surface Temperature, G-Econ = Geographically based Economic data, NDVI = Normalised Difference Vegetation Index.

DOI: 10.7554/eLife.02851.017

maps. However, in order to model the true relationship between environmental risk and disease incidence, a global data set of geopositioned disease incidence data would be required; at present this is unavailable.

Estimates of the populations living in areas of environmental risk are therefore supplied as a proxy for the true burden of disease. However, they cannot be directly compared with other global estimates of the leishmaniases' disease burden, such as the WHO estimates of clinical burden of around 350 million (***WHO, 2010***). ***Figure 4—figure supplement 4*** shows a strong, positive relationship between population at risk estimates and estimated annual incidence from ***Alvar et al. (2012)***. The exceptions to this relationship (e.g., Egypt, Nigeria, and Côte d'Ivoire) are all countries with indeterminate evidence consensus scores, indicating a genuine lack of knowledge regarding both the distribution and incidence of disease.

Previous estimates of the leishmaniases' global burden have been complicated by poor knowledge of the global distribution of the diseases (***Bern et al., 2008***; ***Reithinger, 2008***). It is hoped that the maps presented here will help to increase the accuracy of future estimates. Ideally, future improvements to the global distribution maps presented here would distinguish between the different *Leishmania* species and sandfly vectors. Species-specific models at the same level of detail as those presented here are not currently possible due to a lack of suitable data. Developments in the use of 'big data' approaches to disease mapping (such as the incorporation of informal internet resources) may enable the construction of data sets which could be used in these analyses (***Hay et al., 2013***). A further complication with burden estimation is the epidemic nature of the disease, as evidenced by the national case time series in ***Alvar et al. (2012)***, leading to significant interannual variation in burden. Therefore, any burden estimation would have to account for this and the temporal spread of data would therefore be critical.

It should be noted that non-environmental drivers of transmission and morbidity, such as HIV immunosuppression and risk of infection via blood transfusions and intravenous drug usage, are not incorporated into our present models. The maps presented here can help inform the wider discussion of these factors and their impact on leishmaniasis (e.g., by identifying regions with greater risk for HIV and leishmaniasis co-infection) (***Desjeux and Alvar, 2003***). Similarly, the niche based models used here could enable a decoupling of environmental from social factors to assess the importance of the latter on leishmaniasis transmission in particular areas. It may indeed be the case that in some specific localities it is these non-environmental risk factors that are the main determinants of disease distribution.

Conclusions

These maps represent evidence-based estimates of the current global distribution of the leishmaniases incorporating a comprehensive occurrence database and a rigorous statistical modelling framework

with associated uncertainty statistics. We estimate that 1.71 billion and 1.69 billion individuals live in areas that are suitable for CL and VL transmission, respectively. These figures highlight the need for much greater awareness of this disease at a global scale. These maps provide an important baseline assessment and a strong foundation on which to base future burden estimates, target regions for control efforts and inform public health decisions.

Materials and methods

A boosted regression tree (BRT) modelling framework was used to generate global predicted environmental risk maps for CL and VL. This framework required four key information components: (i) a map of the consensus of evidence for the global extents of the leishmaniases; (ii) a comprehensive data set of geopositioned CL and VL occurrence records; (iii) a suite of global, gridded data sets on environmental correlates of the leishmaniases; and (iv) pseudo-data to augment the occurrence records. In order to better capture the realised niche of these diseases, prediction by the model is restricted to those areas of known disease transmission, or where transmission is uncertain, as defined by the evidence consensus layer (i). The full procedures used to generate these components and the resulting risk and prevalence maps are outlined below.

Evidence consensus

The methodology used for generating the definitive extents for the leishmaniases was adapted from work on dengue (***Brady et al., 2012***). Four primary evidence categories were used to determine a consensus on the presence or absence of the leishmaniases: (i) health reporting organisations; (ii) peer-reviewed evidence of local autochthonous transmission; (iii) case data; (iv) supplementary information. Cutaneous and visceral leishmaniasis were the two symptomatologies investigated: other forms of the disease were subset within these two – whilst VL contained cases of post-kala-azar dermal leishmaniasis, CL included diffuse, disseminated, and mucosal forms of the disease. Although limited amounts of data were available for some of these forms, their epidemiology is similar, and consequently this categorisation was seen as appropriate. Information was collected at provincial level (termed Admin 1 units by the Food and Agriculture Organization's (FAO) Global Administrative Unit Layers (GAUL) coding (***FAO, 2008***)) to better capture the focal nature of these diseases.

Health Reporting Organisation Evidence (scores between −3 and +3)

Two health reporting organisations were referenced, the Global Infectious Diseases and Epidemiology Online Network (GIDEON) (***Edberg, 2005***) and the World Health Organization (WHO) (***WHO, 2010***). The status of disease was recorded for each Admin 1 unit as either present, absent or unspecified. If both reported the disease as present, +3 was scored, if both reported absence, −3 was scored, with +2/−2 scored if one reporting body did not specify the presence or absence of the disease. If the two disagreed, or both were non-specific, 0 was scored reflecting the lack of a consensus on the status of that region.

Peer-reviewed evidence (scores between +2 and +6)

A review of reported leishmaniases' cases was performed. Using PubMed and Web of Knowledge with '[admin1 province] leish*' as the search parameters, articles from January 1960 until September 2012 were abstracted. Each abstract was imported into Endnote X4 and assessed for relevance. Papers that included reported cases on either CL or VL were then obtained. Cases were included if there was sufficient evidence to suggest that local autochthonous transmission had occurred. Where individuals from a non-endemic country had travelled to an endemic country (e.g., tourists and military personnel) and returned with an infection, this was included (as evidence for leishmaniasis in the foreign destination) since these typically represent immunologically naive individuals who have undergone more rigorous diagnostics in their home country, and thus represent a potentially more informed data source. Each paper was assessed for contemporariness and diagnostic accuracy. Contemporariness was graded in 3 bands: 2005–2012 = 3, 1997–2004 = 2 and 1997 and earlier = 1, as was diagnostic accuracy where 1 was scored for data that reported 'confirmed' cases without detailing methodologies implemented; 2 was scored where evidence of microscopy, serology, or the Montenegro skin test had been used; 3 was awarded to those studies that had used PCR or other molecular techniques (***Reithinger and Dujardin, 2007***). Contemporariness bins were based upon the potentially lengthy intrinsic incubation periods present with some *Leishmania* spp. as well as to accommodate the potential for epidemic cycles, where cases may only be detected in peak years and missed in the intervening

baseline periods. The most contemporary and diagnostically accurate papers were then subset to maximise the consensus score for any given area.

Case data (scores between −6 and +6)

Case data were derived from reports on the leishmaniases provided by national health officials (***Alvar et al., 2012***). A threshold value of 12 CL cases and 7 VL cases in a given province in a given year was deemed suitable by the authors to distinguish significant disease events from sporadic cases within that region. If cases were reported at or above the threshold and were dated no later than 2005, +6 was scored. If data existed below this threshold, indicating sporadic cases, or data indicated a history of reported cases in the region but with no evidence of time period, scores were assigned stratified by total annual healthcare expenditure (HE) per capita at average US$ exchange rates (***WHO, 2011***). This was used as a proxy to determine genuine sporadic reporting from inadequate surveillance. Three categories were defined—HE Low (<$100), HE Medium ($100 ≤ HE < $500), and HE High (≥$500). If sporadic cases were reported in an HE Low country, +4 was scored, whilst in an HE Medium country, +2 was scored, and in an HE High country, 0 was scored. If there were no reported case data available, HE Low countries scored +2, HE Medium countries scored −2 and HE High countries scored −6 (***Brady et al., 2012***).

Supplementary evidence

Supplementary evidence was provided in cases where a consensus on presence or absence could not be reached using the aforementioned evidence types, typically with areas where the consensus value was close to 0%. For these regions, additional literature searches were undertaken to determine whether known vector species or infected reservoir hosts were reported in the region. The justification for each provincial scenario is outlined in the associated online databases (Dryad data set doi: 10.5061/dryad.05f5h). In total, this assessment was required in 24 countries.

An overall consensus score for each administrative region was calculated by the sum of the scores in each category, divided by the maximum possible score, then expressed as a percentage. Consensus was defined as either complete (±75% to ±100%), good (±50% to ±74%), moderate (±25% to ±49%), poor (±1% to ±24%), or indeterminate (0%). Such a classification is intended more as a guide to the quality of evidence for the leishmaniases in an area, rather than as a strict classification of certainty. The full scores for each country are laid out in the associated online data sets (Dryad data set doi: 10.5061/dryad.05f5h).

Brazil and Peru

The Brazilian Ministry of Health produces, via the Sistema de Informação de Agravos de Notificação (***SINAN, 2013***) reporting network, records of infections at the municipality level. This allowed for a more thorough evidence consensus to be performed at district level (termed Admin 2 ***FAO, 2008***) within Brazil. As above, WHO and GIDEON status as well as peer-reviewed literature score were recorded, both aggregated to Admin 1 provincial level. Case data were then defined by the presence of a municipality reporting leishmaniasis between 2008 and 2011 inclusive, with positive reports scoring +6 and absence scoring −6. The overall consensus score was then calculated as above. In addition, provincial level case data for Peru was replaced by Ministry of Health information as it was more contemporary than that listed by ***Alvar et al. (2012)***.

Occurrence records

Two separate searches using PubMed and Web of Knowledge were undertaken using the search parameter "leish*," and including articles up to December 2012, and their respective abstracts, were filtered for relevance. From these searches, 4845 articles were collated, with data recorded at the resolution of either a point or Admin 1 or 2 polygon. These were then geo-positioned using Google Maps (https://maps.google.co.uk/). Each entry was evaluated to ensure that non-autochthonous cases and duplicate entries were eliminated. Each occurrence was assigned a start and end date based upon the content of the paper, used to define the time period over which occurrences were reported.

In addition to this resource, reports were taken from the HealthMap database (http://healthmap.org/en/). HealthMap is an online based infectious disease surveillance system that compiles data from informal data sources ranging from online news articles to ProMED reports (***Freifeld et al., 2008***). It parses information from these sources searching for relevant keywords, and then, using crowdsourcing and automated processes, geopositions those relating to the disease of interest. As of December 2012, a total of 690 leishmaniasis relevant articles were archived.

Searches were also performed on GenBank accessions, searching for archived genetic information from *Leishmania* spp. known to infect humans (**WHO, 2010**). If the host was identified as human, geographic indicators were assigned either as point, Admin 1 or Admin 2, based upon the information in the location tag. Tags at the national level were filtered out of the data set. In total, 563 accessions were associated with sub-national location details and added to the database.

Finally, data were provided from the curated strain archives of the Centre National de Référence des Leishmanioses (CNR-L) in Montpellier, France. In total, information about 3465 strains isolated from humans was provided, collected from between 1954 and 2013.

All data were geopositioned as precisely as possible, which resulted in both point-level data (referring to cities, towns or villages) as well as polygon-level data (provinces or districts) with area no greater than one square decimal degree. All data that had been manually geopositioned were checked to ensure coordinates were plausible and then occurrences were standardised annually to remove intra-annual duplicates, so that each individual record used in our model represented an occurrence of leishmaniasis infections in a given 5 km × 5 km location or administrative unit for one given year. As a result, the occurrence data were independent of burden; a location with 200 cases in one year has equal weighting in the model as a location with just one reported case, since it was only the presence of the disease being modelled.

Environmental correlates

Leishmania spp. are known to have anthroponotic, zoonotic, or sylvatic transmission cycles in nature (**WHO, 2010**; **Ready, 2013**) which is apparent in the focal nature of the disease; however, there are some key features of the environment that are important in determining the distribution of disease across the globe. Numerous models have been constructed for local transmission scenarios implicating various environmental features from temperature and precipitation to socioeconomic factors relating to standards of living in villages in endemic foci. For the modelling process, a suite of global gridded environmental, biologically plausible, correlates was generated.

Precipitation

Humidity and moisture, whether from rainfall or in the soil, have often been identified as important for the sandfly, with humidity influencing breeding and resting (**Ready, 2013**). Whilst relatively little is known about these breeding sites, of the few that have been identified, high humidity seems to be a common trait, including moist Amazonian soils, caves, animal burrows, and select human dwellings (**Killick-Kendrick, 1999**; **Feliciangeli, 2004**). Studies have indicated soil type and their moisture profiles as determinants of sandfly distribution (**Bhunia et al., 2010**; **Elnaiem, 2011**). Precipitation represents a good global proxy measure for moisture, and has been shown to play a prominent role in shaping disease distribution in previous leishmaniasis modelling efforts (**Thomson et al., 1999**; **Elnaiem et al., 2003**; **Bhunia et al., 2010**; **Chamaille et al., 2010**; **Gonzalez et al., 2010**, **2011**; **Elnaiem, 2011**; **Hartemink et al., 2011**; **Malaviya et al., 2011**).

Estimates of precipitation were obtained from the WorldClim database (www.worldclim.org). This resource, which is freely available online, provides data spanning from 1950 to 2000, describing monthly averages over this time, at a 1 km × 1 km resolution (**Hijmans et al., 2005**). Using this baseline, interpolated global climate surfaces were produced using ANUSPLIN-SPLINA software (**Hutchinson, 1995**). With the use of temporal Fourier analysis, seasonal and inter-annual variation in precipitation patterns, taken from the interpolated global surface, were used to calculate minimum and maximum monthly precipitation averages (**Rogers et al., 1996**; **Scharlemann et al., 2008**).

Temperature

Temperature influences both the development of the infecting *Leishmania* parasite in the sandfly (**Hlavacova et al., 2013**) as well as the life cycle of the sandfly vectors. On one hand, studies have shown that with increasing temperatures, the metabolism of the sandfly increases, influencing oviposition, defecation, hatching, and adult emergence rates (**Kasap and Alten, 2005**; **Benkova and Volf, 2007**; **Guzman and Tesh, 2000**). On the other hand, higher temperatures have also been shown to increase mortality rates of adults (**Benkova and Volf, 2007**; **Guzman and Tesh, 2000**). Studies have integrated the effects of temperature on sandfly biting rates, sandfly mortality, and extrinsic incubation periods to produce maps of how the basic reproductive number of canine leishmaniasis varied spatially (**Hartemink et al., 2011**). Multiple studies have also implicated temperature (including maximum, minimum, and mean temperatures) as being an important explanatory variable for both sandfly and

disease distribution (**Thomson et al., 1999**; **Gebre-Michael et al., 2004**; **Bhunia et al., 2010**; **Chamaille et al., 2010**; **Fischer et al., 2010**; **Galvez et al., 2011**; **Fernandez et al., 2012**; **Branco et al., 2013**).

Using a similar methodology to generating precipitation surfaces, minimum, maximum, and mean monthly temperature values were generated (**Hijmans et al., 2005**).

Normalised difference vegetation index (NDVI) and land cover

Vegetation provides many roles in sandfly habitat and survival, ranging from maintaining the necessary moisture profile for both immature stages and adults, to a sugar resource for both male and female sandflies (**Killick-Kendrick, 1999**; **Feliciangeli, 2004**; **Ready, 2013**). Moreover, vegetation is an important resource for many mammals that sandflies feed on, and that potentially are *Leishmania* reservoirs. The importance of considering NDVI was demonstrated with respect to the distribution of the reservoir *Psammomys obesus* (sand rat) and the distribution of its primary food, chenopods (**Toumi et al., 2012**). NDVI has been implicated as a key explanatory variable in the distribution of leishmaniasis cases in several studies (**Cross et al., 1996**; **Thomson et al., 1999**; **Elnaiem et al., 2003**; **Gebre-Michael et al., 2004**; **Elnaiem, 2011**; **Hartemink et al., 2011**; **Bhunia et al., 2012**; **Toumi et al., 2012**; **de Oliveira et al., 2012**).

The Advanced Very High Resolution Radiometer (AVHRR) NDVI product uses the spectral reflectance of AVHRR channels 1 and 2 (visible red and near infrared wavelength) to quantitatively assess the level of photosynthesising vegetation in a region (**Hay et al., 2006**). Using this data, compiled over multiple time intervals, patterns of NDVI were extracted for each gridded 1 km × 1 km cell.

Poverty

Neglected tropical diseases and poverty are often found to be linked and the use of a purely economic variable was chosen to act as a proxy for a variety of important global risk factors for disease, including malnutrition, housing quality, and living with domesticated animals (**Bern et al., 2010**; **Boelaert et al., 2009**; **Herrero et al., 2009**; **Malafaia, 2009**; **Zeilhofer et al., 2008**).

The G-Econ database (gecon.yale.edu) takes economic data, at the smallest administrative division available, and spatially rescales these data to create a 1° × 1° gridded surface of the globe (**Nordhaus, 2006**, **2008**). This rescaling estimates the gross cell product of each grid cell, conceptually similar to gross domestic product, referring to the total market value of all final goods and services produced within 1 year, and can be considered as an indicator of overall standard of living within that area. Some cells provided multiple data; in these scenarios the best-quality information, as outlined by the quality field associated with the data, was used to select one value. All gross cell product values were then adjusted using purchasing power parity in US$ for the years 1990, 1995, 2000, and 2005, using national aggregates estimated by the World Bank (**Nordhaus, 2006**) and computed the mean across all years for each gridded cell globally. This adjusted measure was used as the indicator of poverty in the model.

Urbanisation

Over the last few decades, there has been a tendency for the leishmaniases having a sylvatic/zoonotic transmission cycle to transition into the urban and peri-urban environment in response to increasing urbanisation trends (**Harhay et al., 2011**). The increasing overlap in habitat between suitable human and animal hosts and multiple available resting sites for adults can allow for transmission of disease to occur relatively easily (**Singh et al., 2008**; **Poche et al., 2011**; **Uranw et al., 2013**).

The Gridded Population of the World version 3 (GPW3) population density database projected for 2010 was used. The core Global Rural–Urban Mapping Project Urban Extents surface used night-time light satellite imagery to differentiate urban areas (**Balk et al., 2006**); GPW3 is a revision which updates the criteria for urban areas to those areas where population density is greater than or equal to 1000 people per km^2. Using the most up-to-date national censuses available and other demographic data resolved to the smallest available administrative unit, a gridded surface of 5 km × 5 km cells was generated. Each pixel could then be classified as urban, peri-urban, or rural.

Modelling with boosted regression trees

The boosted regression trees (BRT) methodology employed for mapping the leishmaniases is a variant of the model used in a previous analysis of dengue (**Bhatt et al., 2013**). Boosted regression tree modelling combines both regression trees, which build a set of decision rules on the predictor variables by portioning the data into successively smaller groups with binary splits (**De'ath, 2007**;

Elith et al., 2008), and boosting, which selects the tree that minimises the loss of function, to best capture the variables that define the distribution of the input data. The core BRT setup followed standard protocol already defined elsewhere (***Elith et al., 2008***; ***Bhatt et al., 2013***).

Pseudo-data generation

As BRT requires both the presence and absence data, the latter which is often hard to collate in an unbiased manner, pseudo-data had to be generated (***Elith et al., 2008***). There is no general consensus on how best to generate pseudo-data (***Bhatt et al., 2013***); however, several factors of the generation process are known to influence the predicted distribution and thus can be sources of potential bias (***Phillips et al., 2009***; ***Van Der Wal et al., 2009***; ***Phillips and Elith, 2011***; ***Barbet-Massin et al., 2012***). In order to minimise such effects, pseudo-absence selection was directly related to the evidence consensus layer and restricted to a maximum distance (μ) from any occurrence point. Pseudo-presence data was also incorporated, again informed by the evidence consensus layer, to compensate for poor surveillance capacity in low prevalence regions. As in ***Bhatt et al. (2013)*** points were randomly located in regions above an evidence consensus threshold of −25, with regional placement probability weighted by evidence consensus scores, so that regions with higher evidence consensus contained more pseudo-presences than lower scoring areas. Since the occurrence data set is from a wide range of sources and institutions, this procedure aims to mitigate sampling bias. By referencing the evidence consensus layer for pseudo-data selection, detection bias was also mitigated.

'Ensemble' analysis

There is no definitive procedure for choosing the best number of pseudo-data points to generate the most accurate predictive map. To account for the impact that these parameters might have on the model predictions, an ensemble BRT model was constructed with multiple BRT submodels fitted using pseudo-data points generated using different combinations of parameters n_a, n_p, and μ. The numbers of pseudo-absences (n_a) and pseudo-presences (n_p) were defined as a proportion of the total number of actual data occurrence records (6426 and 6137 for CL and VL). The proportions used for generating pseudo-absences were 2:1, 4:1, 6:1, 8:1, and 10:1, and pseudo-presences were 0.025:1, 0.05:1 and 0.1:1. The pseudo-data were also generated within a restricted maximum distance (μ) from any actual presence point, and μ was varied through 5 distances: 5, 10, 15, 20, and 25 arc degrees. All combinations of these parameter values resulted in a total of 75 ($5n_a \times 3n_p \times 5\mu$) individual input data sets and BRT submodels (making up the BRT ensemble).

For each disease, the 75 BRT submodels were used to predict a range of different risk maps (each at 5 km × 5 km resolution), and these were combined to produce a single mean ensemble risk map for each disease, also allowing for computation of the associated range of uncertainty in these predictions for every 5 km × 5 km pixel as shown in ***Figure 1—figure supplement 1***, ***Figure 2—figure supplement 1***, ***Figure 3—figure supplement 1***, ***Figure 4—figure supplement 1***. For both diseases, the New World (the Americas) and Old World (Eurasia and Africa) were modelled separately in order to account for and explore any differences in the epidemiology of the diseases between these regions. This was done to differentiate the potential effect that the different vectors namely *Lutzomyia* spp. in the New World and *Phlebotomus* spp. in the Old World and their varying life histories, might have on the distribution of the diseases within these regions.

Summarising the BRT model

The relative importance of predictor variables was quantified for the final BRT ensemble. Relative importance is defined as the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split and averaged over all trees (***Friedman, 2001***). These contributions are scaled to sum to 100, with a higher number indicating a greater effect on the response. To evaluate the ensemble's predictive performance, we used the area under the receiver operator curve (AUC) (***Fleiss et al., 2003***)—the area under a plot of the true positive rate versus false positive rate, reflecting the ability to discriminate between the presence and absence. An AUC value of 0.5 indicates no discriminative ability, and a value of 1 indicates perfect discrimination.

It is important to note that this distribution modelling technique assesses pixel level risk, rather than population level risk. As such, the ensemble evaluates the likelihood of leishmaniasis presence based upon the covariates supplied. In reality, some other factors, such as national healthcare provisioning and standards of living will influence the true observed burden. Therefore, whilst these two levels of

risk are inherently related, additional information, namely incidence data from many different populations, is required in order to assess the link quantitatively (***Bhatt et al., 2013***).

Estimation of population living in areas of environmental risk

Population living in areas of risk was estimated by using a threshold probability to reclassify the probabilistic risk maps into a binary risk map, then extracting the total human population in the 'at risk' areas using a gridded data set of human population density from 2010 (***Balk et al., 2006***; ***CIESIN/IFPRI/WB/CIAT, 2007***). The threshold value was set such that 95% of the point occurrence records fell within the at risk area. 5% of occurrence points were allowed to fall outside the predicted risk area to account for errors which could have arisen either from errors in the occurrence data set or from inaccuracies in the predicted risk maps.

For external validation, this population at risk information was compared to national reported annual cases (***Alvar et al., 2012***) to produce ***Figure 4—figure supplement 4***. In these figures, the points represent the mean value of the estimated annual incidence reported taking into account the authors estimates of underreporting rates (***Alvar et al., 2012***). The upper and lower limits to these estimates are reflected by the bars around each point. Note that these figures use a log-scale on each axis and that only countries with non-zero estimates by ***Alvar et al. (2012)*** are included.

The threshold probabilities of occurrence used to define 'at risk' were as follows: NW CL—0.22, OW CL—0.19, NW VL—0.42, OW VL—0.19.

Acknowledgements

DMP is funded by a Sir Richard Southwood Graduate Scholarship from the Department of Zoology at the University of Oxford. SIH is funded by a Senior Research Fellowship from the Wellcome Trust (#095066) which also supports KAD and KEB. NG is funded by a grant from the Bill & Melinda Gates Foundation (#OPP1053338). PWG is a Medical Research Council (UK) Career Development Fellow (#K00669X) and receives support from the Bill and Melinda Gates Foundation (#OPP1068048) which also supports SB. OJB is funded by a BBSRC studentship. JPM is funded by, and SIH and OJB acknowledge the support of, the International Research Consortium on Dengue Risk Assessment Management and Surveillance (IDAMS, European Commission 7th Framework Programme (#21803) http://www.idams.eu). JSB, CF and SRM acknowledge funding from NIH National Library of Medicine (#R01LM010812). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Additional information

Funding

Funder	Grant reference number	Author
University Of Oxford	Department of Zoology, Sir Richard Southwood Graduate Scholarship	David M Pigott
Wellcome Trust	095066	Kirsten A Duda, Katherine E Battle, Simon I Hay
Bill and Melinda Gates Foundation	#OPP1053338	Nick Golding
Biotechnology and Biological Sciences Research Council	Studentship	Oliver J Brady
European Commission	#21803	Oliver J Brady, Jane P Messina, Simon I Hay
National Institutes of Health	R01LM010812	John S Brownstein, Clark C Freifeld, Sumiko R Mekaru
Bill and Melinda Gates Foundation	#OPP1068048	Samir Bhatt, Peter W Gething

Funder	Grant reference number	Author
Medical Research Council	#K00669X	Peter W Gething

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

DMP, SIH, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; SB, NG, PWG, DBG, Conception and design, Analysis and interpretation of data, Drafting or revising the article; KAD, KEB, YB, FP, JSB, CF, SRM, MFM, Acquisition of data, Drafting or revising the article; OJB, PB, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; JPM, RR, Analysis and interpretation of data, Drafting or revising the article

Author ORCIDs

David M Pigott, http://orcid.org/0000-0002-6731-4034
Simon I Hay, http://orcid.org/0000-0002-0611-7272

Additional files

Major dataset

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset ID and/or URL	Database, license, and accessibility information
Pigott DM, Bhatt S, Golding N, Duda KA, Battle KE, Brady OJ, Messina JP, Balard Y, Bastien P, Pratlong F, Brownstein JS, Freifeld C, Mekaru SR, Gething PW, George DB, Myers MF, Reithinger R	2014	Data from: Global Distribution Maps of the Leishmaniases	10.5061/dryad.05f5h	Available at Dryad Digital Repository under a CC0 Public Domain Dedication.

References

Alvar J, Velez ID, Bern C, Herrero M, Desjeux P, Cano J, Jannin J, den Boer M, WHO Leishmaniasis Control Team. 2012. Leishmaniasis worldwide and global estimates of its incidence. *PLOS ONE* **7**:e35671. doi: 10.1371/journal.pone.0035671.

Ashford RW. 1996. Leishmaniasis reservoirs and their significance in control. *Clinics in Dermatology* **14**:523–532. doi: 10.1016/0738-081x(96)00041-7.

Balk DL, Deichmann U, Yetman G, Pozzi F, Hay SI, Nelson A. 2006. Determining global population distribution: methods, applications and data. *Advances in Parasitology* **62**:119–156. doi: 10.1016/S0065-308X(05)62004-0.

Banuls AL, Bastien P, Pomares C, Arevalo J, Fisa R, Hide M. 2011. Clinical pleiomorphism in human leishmaniases, with special mention of asymptomatic infection. *Clinical Microbiology and Infection* **17**:1451–1461. doi: 10.1111/j.1469-0691.2011.03640.x.

Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution* **3**:327–338. doi: 10.1111/j.2041-210X.2011.00172.x.

Benkova I, Volf P. 2007. Effect of temperature on metabolism of *Phlebotomus papatasi* (Diptera : Psychodidae). *Journal of Medical Entomology* **44**:150–154. doi: 10.1603/0022-2585(2007)44[150:EOTOMO]2.0.CO;2.

Bern C, Courtenay O, Alvar J. 2010. Of cattle, sand flies and men: a systematic review of risk factor analyses for South Asian visceral leishmaniasis and implications for elimination. *PLOS Neglected Tropical Diseases* **4**:e599. doi: 10.1371/journal.pntd.0000599.

Bern C, Maguire JH, Alvar J. 2008. Complexities of assessing the disease burden attributable to leishmaniasis. *PLOS Neglected Tropical Diseases* **2**:e313. doi: 10.1371/journal.pntd.0000313.

Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, Drake JM, Brownstein JS, Hoen AG, Sankoh O, Myers MF, George DB, Jaenisch T, Wint GR, Simmons CP, Scott TW, Farrar JJ, Hay SI. 2013. The global distribution and burden of dengue. *Nature* **496**:504–507. doi: 10.1038/Nature12060.

Bhunia GS, Kesari S, Chatterjee N, Mandal R, Kumar V, Das P. 2012. Seasonal relationship between normalized difference vegetation index and abundance of the *Phlebotomus* kala-azar vector in an endemic focus in Bihar, India. *Geospatial Health* **7**:51–62.

Bhunia GS, Kumar V, Kumar AJ, Das P, Kesari S. 2010. The use of remote sensing in the identification of the eco-environmental factors associated with the risk of human visceral leishmaniasis (kala-azar) on the Gangetic plain, in north-eastern India. *Annals of Tropical Medicine and Parasitology* **104**:35–53. doi: 10.1179/136485910x12607012373678.

Boelaert M, Meheus F, Sanchez A, Singh SP, Vanlerberghe V, Picado A, Meessen B, Sundar S. 2009. The poorest of the poor: a poverty appraisal of households affected by visceral leishmaniasis in Bihar, India. *Tropical Medicine & International Health* **14**:639–644. doi: 10.1111/j.1365-3156.2009.02279.x.

Brady OJ, Gething PW, Bhatt S, Messina JP, Brownstein JS, Hoen AG, Moyes CL, Farlow AW, Scott TW, Hay SI. 2012. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLOS Neglected Tropical Diseases* **6**:e1760. doi: 10.1371/Journal.Pntd.0001760.

Branco S, Alves-Pires C, Maia C, Cortes S, Cristovão JM, Gonçalves L, Campino L, Afonso MO. 2013. Entomological and ecological studies in a new potential zoonotic leishmaniasis focus in Torres Novas municipality, Central region, Portugal. *Acta Tropica* **125**:339–348. doi: 10.1016/j.actatropica.2012.12.008.

Cardo LJ. 2006. *Leishmania*: risk to the blood supply. *Transfusion* **46**:1641–1645. doi: 10.1111/j.1537-2995.2006.00941.x.

Chamaille L, Tran A, Meunier A, Bourdoiseau G, Ready P, Dedet JP. 2010. Environmental risk mapping of canine leishmaniasis in France. *Parasites & Vectors* **3**:31. doi: 10.1186/1756-3305-3-31.

CIESIN/IFPRI/WB/CIAT. 2007. Global rural urban mapping project (GRUMP): gridded population of the world, version 3. Available: http://sedac.ciesin.columbia.edu/gpw. Accessed: December 2013.

Cross ER, Newcomb WW, Tucker CJ. 1996. Use of weather data and remote sensing to predict the geographic and seasonal distribution of *Phlebotomus papatasi* in southwest Asia. *The American Journal of Tropical Medicine and Hygiene* **54**:530–536.

De'ath G. 2007. Boosted trees for ecological modeling and prediction. *Ecology* **88**:243–251. doi: 10.1890/0012-9658(2007)88[243:Btfema]2.0.Co;2.

de Oliveira EF, Silva EAE, Fernandes CED, Paranhos AC, Gamarra RM, Ribeiro AA, Brazil RP, Oliveira AG. 2012. Biotic factors and occurrence of *Lutzomyia longipalpis* in endemic area of visceral leishmaniasis, Mato Grosso do Sul, Brazil. *Memorias Do Instituto Oswaldo Cruz* **107**:396–401. doi: 10.1590/S0074-02762012000300015.

Dedet JP, Pratlong F. 2009. Leishmaniasis. In: Cook GC, Zumla AI, editors. *Manson's tropical diseases*. 22nd edition. London: Saunders Elsevier. p. 1341–1365.

Desjeux P, Alvar J. 2003. Leishmania/HIV co-infections: epidemiology in Europe. *Annals of Tropical Medicine and Parasitology* **97**:3–15. doi: 10.1179/000349803225002499.

Edberg SC. 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. doi: 10.1086/426549.

Elith J, Graham CH, Anderson RP, Dudik M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton JM, Peterson AT, Phillips SJ, Richardson K, Scachetti-Pereira R, Schapire RE, Soberon J, Williams S, Wisz MS, Zimmerman NE. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**:129–151. doi: 10.1111/j.2006.0906-7590.04596.x.

Elith J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. *The Journal of Animal Ecology* **77**:802–813. doi: 10.1111/j.1365-2656.2008.01390.x.

Elnaiem DA, Schorscher J, Bendall A, Obsomer V, Osman ME, Mekkawi AM, Connor SJ, Ashford RW, Thomson MC. 2003. Risk mapping of visceral leishmaniasis: the role of local variation in rainfall and altitude on the presence and incidence of kala-azar in eastern Sudan. *The American Journal of Tropical Medicine and Hygiene* **68**:10–17.

Elnaiem DEA. 2011. Ecology and control of the sand fly vectors of *Leishmania donovani* in East Africa, with special emphasis on *Phlebotomus orientalis*. *Journal of Vector Ecology* **36**:S23–S31. doi: 10.1111/j.1948-7134.2011.00109.x.

FAO. 2008. *The Global Administrative Unit Layers (GAUL): Technical Aspects*. Rome: Food and Agriculture Organization of the United Nations, EC-FAO Food Security Programme (ESTG).

Feliciangeli MD. 2004. Natural breeding places of phlebotomine sandflies. *Medical and Veterinary Entomology* **18**:71–80. doi: 10.1111/j.0269-283X.2004.0487.x.

Fernandez MS, Lestani EA, Cavia R, Salomon OD. 2012. Phlebotominae fauna in a recent deforested area with american tegumentary leishmaniasis transmission (Puerto Iguazu, Misiones, Argentina): seasonal distribution in domestic and peridomestic environments. *Acta Tropica* **122**:16–23. doi: 10.1016/j.actatropica.2011.11.006.

Fischer D, Thomas SM, Beierkuhnlein C. 2010. Temperature-derived potential for the establishment of phlebotomine sandflies and visceral leishmaniasis in Germany. *Geospatial Health* **5**:59–69.

Fleiss JL, Levin B, Paik MC. 2003. *Statistical methods for rates and proportions*. Hoboken, New Jersey: John Wiley & Sons. p. 800.

Freifeld CC, Mandl KD, Ras BY, Bronwnstein JS. 2008. HealthMap: global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association* **15**:150–157. doi: 10.1197/Jamia.M2544.

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29**:1189–1232. doi: 10.1214/aos/1013203451.

Galvez R, Descalzo MA, Guerrero I, Miro G, Molina R. 2011. Mapping the current distribution and predicted spread of the leishmaniosis sand fly vector in the Madrid region (Spain) based on environmental variables and expected climate change. *Vector Borne and Zoonotic Diseases* **11**:799–806. doi: 10.1089/vbz.2010.0109.

Gebre-Michael T, Malone JB, Balkew M, Ali A, Berhe N, Hailu A, Herzi AA. 2004. Mapping the potential distribution of *Phlebotomus martini* and *P. orientalis* (Diptera : Psychodidae), vectors of kala-azar in East Africa by use of geographic information systems. *Acta Tropica* **90**:73–86. doi: 10.1016/j.actatropica.2003.09.021.

Gonzalez C, Rebollar-Tellez EA, Ibanez-Bernal S, Becker-Fauser I, Martinez-Meyer E, Peterson AT, Sánchez-Cordero V. 2011. Current knowledge of *Leishmania* vectors in Mexico: how geographic distributions of species relate to transmission areas. *The American Journal of Tropical Medicine and Hygiene* **85**:839–846. doi: 10.4269/ajtmh.2011.10-0452.

Gonzalez C, Wang O, Strutz SE, Gonzalez-Salazar C, Sanchez-Cordero V, Sarkar S. 2010. Climate change and risk of leishmaniasis in North America: predictions from ecological niche models of vector and reservoir species. *PLOS Neglected Tropical Diseases* **4**:e585. doi: 10.1371/journal.pntd.0000585.

Guzman H, Tesh R. 2000. Effects of temperature and diet on growth and longevity of phlebotomine sand flies (Diptera: Psychodidae). *Biomedica: revista del Instituto Nacional de Salud* **20**:190–199.

Harhay MO, Olliaro PL, Costa DL, Costa CHN. 2011. Urban parasitology: visceral leishmaniasis in Brazil. *Trends in Parasitology* **27**:403–409. doi: 10.1016/j.pt.2011.04.001.

Hartemink N, Vanwambeke SO, Heesterbeek H, Rogers D, Morley D, Pesson B, Davies C, Mahamdallie S, Ready P. 2011. Integrated mapping of establishment risk for merging vector-borne infections: a case study of canine leishmaniasis in southwest France. *PLOS ONE* **6**:e20817. doi: 10.1371/journal.pone.0020817.

Hay SI, Battle KE, Pigott DM, Smith DL, Moyes CL, Bhatt S, Brownstein JS, Collier N, Myers MF, George DB, Gething PW. 2013. Global mapping of infectious disease. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences* **368**:20120250. doi: 10.1098/rstb.2012.0250.

Hay SI, Tatem AJ, Graham AJ, Goetz SJ, Rogers DJ. 2006. Global environmental data for mapping infectious disease distribution. *Advances in Parasitology* **62**:37–77. doi: 10.1016/S0065-308X(05)62002-7.

Herrero M, Orfanos G, Argaw D, Mulugeta A, Aparicio P, Parreño F, Bernal O, Rubens D, Pedraza J, Lima MA, Flevaud L, Palma PP, Bashaye S, Alvar J, Bern C. 2009. Natural history of a visceral leishmaniasis outbreak in highland Ethiopia. *The American Journal of Tropical Medicine and Hygiene* **81**:373–377.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**:1965–1978. doi: 10.1002/Joc.1276.

Hlavacova J, Votypka J, Volf P. 2013. The effect of temperature on *Leishmania* (Kinetoplastida: Trypanosomatidae) development in sand flies. *Journal of Medical Entomology* **50**:955–958. doi: 10.1603/Me13053.

Hutchinson MF. 1995. Interpolating mean rainfall using thin-plate smoothing splines. *International Journal of Geographical Information Systems* **9**:385–403. doi: 10.1080/02693799508902045.

Ives A, Ronet C, Prevel F, Ruzzante G, Fuertes-Marraco S, Schutz F, Zangger H, Revaz-Breton M, Lye LF, Hickerson SM, Beverley SM, Acha-Orbea H, Launois P, Fasel N, Masina S. 2011. *Leishmania* RNA virus controls the severity of mucocutaneous leishmaniasis. *Science* **331**:775–778. doi: 10.1126/science.1199326.

Kasap OE, Alten B. 2005. Laboratory estimation of degree-day developmental requirements of *Phlebotomus papatasi* (Diptera : Psychodidae). *Journal of Vector Ecology* **30**:328–333.

Killick-Kendrick R. 1999. The biology and control of phlebotomine sand flies. *Clinics in Dermatology* **17**:279–289. doi: 10.1016/S0738-081x(99)00046-2.

Lima BS, Dantas-Torres F, de Carvalho MR, Marinho JF, de Almeida EL, Brito ME, Gomes F, Brandão-Filho SP. 2013. Small mammals as hosts of *Leishmania* spp. in a highly endemic area for zoonotic leishmaniasis in north-eastern Brazil. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **107**:592–597. doi: 10.1093/trstmh/trt062.

Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham J, Adair T, Aggarwal R, Ahn SY, Alvarado M, Anderson HR, Anderson LM, Andrews KG, Atkinson C, Baddour LM, Barker-Collo S, Bartels DH, Bell ML, Benjamin EJ, Bennett D, Bhalla K, Bikbov B, Bin Abdulhak A, Birbeck G, Blyth F, Bolliger I, Boufous S, Bucello C, Burch M, Burney P, Carapetis J, Chen H, Chou D, Chugh SS, Coffeng LE, Colan SD, Colquhoun S, Colson KE, Condon J, Connor MD, Cooper LT, Corriere M, Cortinovis M, de Vaccaro KC, Couser W, Cowie BC, Criqui MH, Cross M, Dabhadkar KC, Dahodwala N, De Leo D, Degenhardt L, Delossantos A, Denenberg J, Des Jarlais DC, Dharmaratne SD, Dorsey ER, Driscoll T, Duber H, Ebel B, Erwin PJ, Espindola P, Ezzati M, Feigin V, Flaxman AD, Forouzanfar MH, Fowkes FG, Franklin R, Fransen M, Freeman MK, Gabriel SE, Gakidou E, Gaspari F, Gillum RF, Gonzalez-Medina D, Halasa YA, Haring D, Harrison JE, Havmoeller R, Hay RJ, Hoen B, Hotez PJ, Hoy D, Jacobsen KH, James SL, Jasrasaria R, Jayaraman S, Johns N, Karthikeyan G, Kassebaum N, Keren A, Khoo JP, Knowlton LM, Kobusingye O, Koranteng A, Krishnamurthi R, Lipnick M, Lipshultz SE, Ohno SL, Mabweijano J, MacIntyre MF, Mallinger L, March L, Marks GB, Marks R, Matsumori A, Matzopoulos R, Mayosi BM, McAnulty JH, McDermott MM, McGrath J, Mensah GA, Merriman TR, Michaud C, Miller M, Miller TR, Mock C, Mocumbi AO, Mokdad AA, Moran A, Mulholland K, Nair MN, Naldi L, Narayan KM, Nasseri K, Norman P, O'Donnell M, Omer SB, Ortblad K, Osborne R, Ozgediz D, Pahari B, Pandian JD, Rivero AP, Padilla RP, Perez-Ruiz F, Perico N, Phillips D, Pierce K, Pope CA III, Porrini E, Pourmalek F, Raju M, Ranganathan D, Rehm JT, Rein DB, Remuzzi G, Rivara FP, Roberts T, De León FR, Rosenfeld LC, Rushton L, Sacco RL, Salomon JA, Sampson U, Sanman E, Schwebel DC, Segui-Gomez M, Shepard DS, Singh D, Singleton J, Sliwa K, Smith E, Steer A, Taylor JA, Thomas B, Tleyjeh IM, Towbin JA, Truelsen T, Undurraga EA, Venketasubramanian N, Vijayakumar L, Vos T, Wagner GR, Wang M, Wang W, Watt K, Weinstock MA, Weintraub R, Wilkinson JD, Woolf AD, Wulf S, Yeh PH, Yip P, Zabetian A, Zheng ZJ, Lopez AD, Murray CJ, AlMazroa MA, Memish ZA. 2012. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**:2095–2128. doi: 10.1016/S0140-6736(12)61728-0.

Malafaia G. 2009. Protein-energy malnutrition as a risk factor for visceral leishmaniasis: a review. *Parasite Immunology* **31**:587–596. doi: 10.1111/j.1365-3024.2009.01117.x.

Malaviya P, Picado A, Singh P, Hasker E, Singh RP, Boelaert M, Sundar S. 2011. Visceral leishmaniasis in Muzaffarpur district, Bihar, India from 1990 to 2008. *PLOS ONE* **6**:e14751. doi: 10.1371/journal.pone.0014751.

Murray CJL, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, Ezzati M, Shibuya K, Salomon JA, Abdalla S, Aboyans V, Abraham J, Ackerman I, Aggarwal R, Ahn SY, Ali MK, Alvarado M, Anderson HR, Anderson LM, Andrews KG, Atkinson C, Baddour LM, Bahalim AN, Barker-Collo S, Barrero LH, Bartels DH, Basáñez MG, Baxter A, Bell ML, Benjamin EJ, Bennett D, Bernabé E, Bhalla K, Bhandari B, Bikbov B, Bin Abdulhak A, Birbeck G, Black JA, Blencowe H, Blore JD, Blyth F, Bolliger I, Bonaventure A, Boufous S, Bourne R, Boussinesq M, Braithwaite T, Brayne C, Bridgett L, Brooker S, Brooks P, Brugha TS, Bryan-Hancock C, Bucello C, Buchbinder R, Buckle G, Budke CM, Burch M, Burney P, Burstein R, Calabria B, Campbell B, Canter CE, Carabin H, Carapetis J, Carmona L, Cella C, Charlson F, Chen H, Cheng AT, Chou D, Chugh SS, Coffeng LE, Colan SD, Colquhoun S, Colson KE, Condon J, Connor MD, Cooper LT, Corriere M, Cortinovis M, de Vaccaro KC, Couser W, Cowie BC, Criqui MH, Cross M, Dabhadkar KC, Dahiya M, Dahodwala N, Damsere-Derry J, Danaei G, Davis A, De Leo D, Degenhardt L, Dellavalle R, Delossantos A, Denenberg J, Derrett S, Des Jarlais DC, Dharmaratne SD, Dherani M, Diaz-Torne C, Dolk H, Dorsey ER, Driscoll T, Duber H, Ebel B, Edmond K, Elbaz A, Ali SE, Erskine H, Erwin PJ, Espindola P, Ewoigbokhan SE, Farzadfar F, Feigin V, Felson DT, Ferrari A, Ferri CP, Fèvre EM, Finucane MM, Flaxman S, Flood L, Foreman K, Forouzanfar MH, Fowkes FG, Fransen M, Freeman MK, Gabbe BJ, Gabriel SE, Gakidou E, Ganatra HA, Garcia B, Gaspari F, Gillum RF, Gmel G, Gonzalez-Medina D, Gosselin R, Grainger R, Grant B, Groeger J, Guillemin F, Gunnell D, Gupta R, Haagsma J, Hagan H, Halasa YA, Hall W, Haring D, Haro JM, Harrison JE, Havmoeller R, Hay RJ, Higashi H, Hill C, Hoen B, Hoffman H, Hotez PJ, Hoy D, Huang JJ, Ibeanusi SE, Jacobsen KH, James SL, Jarvis D, Jasrasaria R, Jayaraman S, Johns N, Jonas JB, Karthikeyan G, Kassebaum N, Kawakami N, Keren A, Khoo JP, King CH, Knowlton LM, Kobusingye O, Koranteng A, Krishnamurthi R, Laden F, Lalloo R, Laslett LL, Lathlean T, Leasher JL, Lee YY, Leigh J, Levinson D, Lim SS, Limb E, Lin JK, Lipnick M, Lipshultz SE, Liu W, Loane M, Ohno SL, Lyons R, Mabweijano J, MacIntyre MF, Malekzadeh R, Mallinger L, Manivannan S, Marcenes W, March L, Margolis DJ, Marks GB, Marks R, Matsumori A, Matzopoulos R, Mayosi BM, McAnulty JH, McDermott MM, McGill N, McGrath J, Medina-Mora ME, Meltzer M, Mensah GA, Merriman TR, Meyer AC, Miglioli V, Miller M, Miller TR, Mitchell PB, Mock C, Mocumbi AO, Moffitt TE, Mokdad AA, Monasta L, Montico M, Moradi-Lakeh M, Moran A, Morawska L, Mori R, Murdoch ME, Mwaniki MK, Naidoo K, Nair MN, Naldi L, Narayan KM, Nelson PK, Nelson RG, Nevitt MC, Newton CR, Nolte S, Norman P, Norman R, O'Donnell M, O'Hanlon S, Olives C, Omer SB, Ortblad K, Osborne R, Ozgediz D, Page A, Pahari B, Pandian JD, Rivero AP, Patten SB, Pearce N, Padilla RP, Perez-Ruiz F, Perico N, Pesudovs K, Phillips D, Phillips MR, Pierce K, Pion S, Polanczyk GV, Polinder S, Pope CA III, Popova S, Porrini E, Pourmalek F, Prince M, Pullan RL, Ramaiah KD, Ranganathan D, Razavi H, Regan M, Rehm JT, Rein DB, Remuzzi G, Richardson K, Rivara FP, Roberts T, Robinson C, De Leòn FR, Ronfani L, Room R, Rosenfeld LC, Rushton L, Sacco RL, Saha S, Sampson U, Sanchez-Riera L, Sanman E, Schwebel DC, Scott JG, Segui-Gomez M, Shahraz S, Shepard DS, Shin H, Shivakoti R, Singh D, Singh GM, Singh JA, Singleton J, Sleet DA, Sliwa K, Smith E, Smith JL, Stapelberg NJ, Steer A, Steiner T, Stolk WA, Stovner LJ, Sudfeld C, Syed S, Tamburlini G, Tavakkoli M, Taylor HR, Taylor JA, Taylor WJ, Thomas B, Thomson WM, Thurston GD, Tleyjeh IM, Tonelli M, Towbin JA, Truelsen T, Tsilimbaris MK, Ubeda C, Undurraga EA, van der Werf MJ, van Os J, Vavilala MS, Venketasubramanian N, Wang M, Wang W, Watt K, Weatherall DJ, Weinstock MA, Weintraub R, Weisskopf MG, Weissman MM, White RA, Whiteford H, Wiebe N, Wiersma ST, Wilkinson JD, Williams HC, Williams SR, Witt E, Wolfe F, Woolf AD, Wulf S, Yeh PH, Zaidi AK, Zheng ZJ, Zonies D, Lopez AD, AlMazroa MA, Memish ZA. 2012. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**:2197–2223. doi: 10.1016/S0140-6736(12)61689-4.

Murray HW, Berman JD, Davies CR, Saravia NG. 2005. Advances in leishmaniasis. *Lancet* **366**:1561–1577. doi: 10.1016/S0140-6736(05)67629-5.

Nordhaus W. 2008. New metrics for environmental economics: gridded economic data. *Integrated Assessment* **8**:73–84.

Nordhaus WD. 2006. Geography and macroeconomics: new data and new findings. *Proceedings of the National Academy of Sciences of the United States of America* **103**:3510–3517. doi: 10.1073/pnas.0509842103.

Novais FO, Carvalho LP, Graff JW, Beiting DP, Ruthel G, Roos DS, Betts MR, Goldschmidt MH, Wilson ME, de Oliveira CI, Scott P. 2013. Cytotoxic T cells mediate pathology and metastasis in cutaneous leishmaniasis. *PLOS Pathogens* **9**:e1003504. doi: 10.1371/journal.ppat.1003504.

Peterson AT. 2008. Biogeography of diseases: a framework for analysis. *Die Naturwissenschaften* **95**:483–491. doi: 10.1007/s00114-008-0352-5.

Peterson AT, Soberon J, Pearson RG, Anderson RP, Martinez-Meyer E, Nakamura M, Araujo MB. 2011. *Ecological niches and geographic distributions*. Princeton: Princeton University Press. p. 314.

Phillips SJ, Dudik M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications* **19**:181–197. doi: 10.1890/07-2153.1.

Phillips SJ, Elith J. 2011. Logistic methods for resource selection functions and presence-only species distribution models. In: AAAI, editor. *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. San Francisco: AAAI (Association for the Advancement of Artificial Intelligence). p. 1384–1389.

Poche D, Garlapati R, Ingenloff K, Remmers J, Poche R. 2011. Bionomics of phlebotomine sand flies from three villages in Bihar, India. *Journal of Vector Ecology* **36**:S106–S117. doi: 10.1111/j.1948-7134.2011.00119.x.

Ready PD. 2013. Biology of phlebotomine sand flies as vectors of disease agents. *Annual Review of Entomology* **58**:227–250. doi: 10.1146/annurev-ento-120811-153557.

Reithinger R. 2008. Leishmaniases' burden of disease: ways forward for getting from speculation to reality. *PLOS Neglected Tropical Diseases* **2**:e285. doi: 10.1371/journal.pntd.0000285.

Reithinger R, Dujardin JC. 2007. Molecular diagnosis of leishmaniasis: current status and future applications. *Journal of Clinical Microbiology* **45**:21–25. doi: 10.1128/Jcm.02029-06.
Reithinger R, Dujardin JC, Louzir H, Pirmez C, Alexander B, Brooker S. 2007. Cutaneous leishmaniasis. *The Lancet Infectious Diseases* **7**:581–596. doi: 10.1016/S1473-3099(07)70209-8.
Rogers DJ, Hay SI, Packer MJ. 1996. Predicting the distribution of tsetse flies in west Africa using temporal Fourier processed meteorological satellite data. *Annals of Tropical Medicine and Parasitology* **90**:225–241.
Scharlemann JPW, Benz D, Hay SI, Purse BV, Tatem AJ, Wint GR, Rogers DJ. 2008. Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. *PLOS ONE* **3**:e1408. doi: 10.1371/Journal.Pone.0001408.
SINAN. 2013. Sistema de Informação de Agravos de Notificação. Available: http://dtr2004.saude.gov.br/sinanweb/index.php.
Singh R, Lal S, Saxena VK. 2008. Breeding ecology of visceral leishmaniasis vector sandfly in Bihar state of India. *Acta Tropica* **107**: 117–120. doi: 10.1016/j.actatropica.2008.04.025.
Sinka ME, Bangs MJ, Manguin S, Chareonviriyaphap T, Patil AP, Temperley WH, Gething PW, Elyazar IR, Kabaria CW, Harbach RE, Hay SI. 2011. The dominant *Anopheles* vectors of human malaria in the Asia-Pacific region: occurrence data, distribution maps and bionomic precis. *Parasites & Vectors* **4**:89. doi: 10.1186/1756-3305-4-89.
Sinka ME, Bangs MJ, Manguin S, Coetzee M, Mbogo CM, Hemingway J, Patil AP, Temperley WH, Gething PW, Kabaria CW, Okara RM, Van Boeckel T, Godfray HC, Harbach RE, Hay SI. 2010. The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic precis. *Parasites & Vectors* **3**:117. doi: 10.1186/1756-3305-3-117.
Sinka ME, Rubio-Palis Y, Manguin S, Patil AP, Temperley WH, Gething PW, Van Boeckel T, Kabaria CW, Harbach RE, Hay SI. 2010. The dominant *Anopheles* vectors of human malaria in the Americas: occurrence data, distribution maps and bionomic precis. *Parasites & Vectors* **3**:72. doi: 10.1186/1756-3305-3-72.
Syfert MM, Smith MJ, Coomes DA. 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLOS ONE* **8**:e55158. doi: 10.1371/journal.pone.0055158.
Thomson MC, Elnaiem DA, Ashford RW, Connor SJ. 1999. Towards a kala azar risk map for Sudan: mapping the potential distribution of *Phlebotomus orientalis* using digital data of environmental variables. *Tropical Medicine & International Health* **4**:105–113. doi: 10.1046/j.1365-3156.1999.00368.x.
Toumi A, Chlif S, Bettaieb J, Ben Alaya N, Boukthir A, Ahmadi ZE, Ben Salah A. 2012. Temporal dynamics and impact of climate factors on the incidence of zoonotic cutaneous leishmaniasis in central Tunisia. *PLOS Neglected Tropical Diseases* **6**:e1633. doi: 10.1371/Journal.Pntd.0001633.
Uranw S, Hasker E, Roy L, Meheus F, Das ML, Bhattarai NR, Rijal S, Boelaert M. 2013. An outbreak investigation of visceral leishmaniasis among residents of Dharan town, eastern Nepal, evidence for urban transmission of Leishmania donovani. *BMC Infectious Diseases* **13**:21. doi: 10.1186/1471-2334-13-21.
Van Der Wal J, Shoo LP, Graham C, William SE. 2009. Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? *Ecological Modelling* **220**:589–594. doi: 10.1016/j.ecolmodel.2008.11.010.
WHO. 2009. *Neglected tropical diseases, hidden successes, emerging opportunities*. Geneva: World Health Organization. p. 59.
WHO. 2010. *Control of the Leishmaniases. Report of a Meeting of the WHO Expert Committee on the Control of Leishmaniases*. Geneva, 22–26 March 2010: World Health Organization. p. 186.
WHO. 2011. *World Health Statistics 2011*. Geneva: World Health Organization. p. 170.
Zeilhofer P, Kummer OP, dos Santos ES, Ribeiro ALM, Missawa NA. 2008. Spatial modelling of *Lutzomyia (Nyssomyia) whitmani sensu lato* (Antunes & Coutinho, 1939) (Diptera: Psychodidae: Phlebotominae) habitat suitability in the state of Mato Grosso, Brazil. *Memorias Do Instituto Oswaldo Cruz* **103**:653–660. doi: 10.1590/S0074-02762008000700005.

Chapter 4

A framework for mapping zoonotic disease: Ebola virus disease.

The West African outbreak of Ebola virus disease (EVD) was unexpected and not predicted; the number of people infected is unprecedented in the history of EVD. In response, this work applies species distribution modelling to Ebola virus, incorporating information from both animal and human sources, in order to define areas environmentally suitable for the transmission of the virus from reservoir hosts into humans with the intention of revising estimates of areas of potential risk. This work has been published in *eLife* and is included here in its final form. The additional information referred to in this chapter is included in the Appendix of this thesis. Also included in the Appendix is a more detailed paper discussing the index case geopositioning protocol accompanying discussion of the spatial progress of each historical outbreak.

eLIFE elifesciences.org

RESEARCH ARTICLE

Mapping the zoonotic niche of Ebola virus disease in Africa

David M Pigott[1†], Nick Golding[1†], Adrian Mylne[1], Zhi Huang[1], Andrew J Henry[1], Daniel J Weiss[1], Oliver J Brady[1], Moritz UG Kraemer[1], David L Smith[1,2], Catherine L Moyes[1], Samir Bhatt[1], Peter W Gething[1], Peter W Horby[3], Isaac I Bogoch[4,5], John S Brownstein[6,7], Sumiko R Mekaru[8], Andrew J Tatem[9,10,13], Kamran Khan[4,11], Simon I Hay[1,12]*

[1]Spatial Ecology and Epidemiology Group, Department of Zoology, University of Oxford, Oxford, United Kingdom; [2]Sanaria Institute for Global Health and Tropical Medicine, Rockville, United States; [3]Epidemic Diseases Research Group, Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, United Kingdom; [4]Department of Medicine, Division of Infectious Diseases, University of Toronto, Toronto, Canada; [5]Divisions of Internal Medicine and Infectious Diseases, University Health Network, Toronto, Toronto, Canada; [6]Department of Pediatrics, Harvard Medical School, Boston, United States; [7]Children's Hospital Informatics Program, Boston Children's Hospital, Boston, United States; [8]Children's Hospital Informatics Program, Boston Children's Hospital, Boston, United States; [9]Department of Geography and Environment, University of Southampton, Southampton, United Kingdom; [10]Fogarty International Center, National Institutes of Health, Bethesda, United States; [11]Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, Canada; [12]Fogarty International Center, National Institutes of Health, Bethseda, United States; [13]Flowminder Foundation, Stockholm, Sweden

Abstract Ebola virus disease (EVD) is a complex zoonosis that is highly virulent in humans. The largest recorded outbreak of EVD is ongoing in West Africa, outside of its previously reported and predicted niche. We assembled location data on all recorded zoonotic transmission to humans and Ebola virus infection in bats and primates (1976–2014). Using species distribution models, these occurrence data were paired with environmental covariates to predict a zoonotic transmission niche covering 22 countries across Central and West Africa. Vegetation, elevation, temperature, evapotranspiration, and suspected reservoir bat distributions define this relationship. At-risk areas are inhabited by 22 million people; however, the rarity of human outbreaks emphasises the very low probability of transmission to humans. Increasing population sizes and international connectivity by air since the first detection of EVD in 1976 suggest that the dynamics of human-to-human secondary transmission in contemporary outbreaks will be very different to those of the past.
DOI: 10.7554/eLife.04395.001

***For correspondence:** simon.hay@zoo.ox.ac.uk

†These authors contributed equally to this work

Competing interests: See page 20

Funding: See page 21

Received: 18 August 2014
Accepted: 31 August 2014
Published: 08 September 2014

Reviewing editor: Prabhat Jha, University of Toronto, Canada

Introduction

Ebola viruses have for the last forty years been responsible for a number of outbreaks of Ebola virus disease (EVD) in humans (***Pattyn et al., 1977***), with high case fatality rates typically around 60–70%, but potentially reaching as high as 90% (***Feldmann and Geisbert, 2011***). The most recent outbreak began in Guinea in December 2013 (***Baize et al., 2014***; ***Bausch and Schwarz, 2014***) and has subsequently spread to Liberia, Sierra Leone and Nigeria (***ECDC, 2014***). The unprecedented size and scale of this ongoing outbreak has the potential to destabilise already fragile economies and healthcare

eLife digest Since the first outbreaks of Ebola virus disease in 1976, there have been numerous other outbreaks in humans across Africa with fatality rates ranging from 50% to 90%. Humans can become infected with the Ebola virus after direct contact with blood or bodily fluids from an infected person or animal. The virus also infects and kills other primates—such as chimpanzees or gorillas—though Old World fruit bats are suspected to be the most likely carriers of the virus in the wild.

The largest recorded outbreak of Ebola virus disease is ongoing in West Africa: more people have been infected in this current outbreak than in all previous outbreaks combined. The current outbreak is also the first to occur in West Africa—which is outside the previously known range of the Ebola virus.

Pigott et al. have now updated predictions about where in Africa wild animals may harbour the virus and where the transmission of the virus from these animals to humans is possible. As such, the map identifies the regions that are most at risk of a future Ebola outbreak. The data behind these new maps include the locations of all recorded primary cases of Ebola in human populations—the 'index' cases—many of which have been linked to animal sources. The data also include the locations of recorded cases of Ebola virus infections in wild bats and primates from the last forty years. The maps, which were modelled using more flexible methods than previous predictions, also include new information—collected using satellites—about environmental factors and new predictions of the range of wild fruit bats.

Pigott et al. report that the transmission of Ebola virus from animals to humans is possible in 22 countries across Central and West Africa—and that 22 million people live in the areas at risk. However, outbreaks in human populations are rare and the likelihood of a human getting the disease from an infected animal still remains very low. The updated map does not include data about how infections spread from one person to another, so the next challenge is to use existing data on human-to-human transmission to better understand the likely size and extent of current and future outbreaks. As more people live in, and travel to and from, the at-risk regions than ever before, Pigott et al. note that new outbreaks of Ebola virus disease are likely to be very different to those of the past.

DOI: 10.7554/eLife.04395.002

systems (***Fauci, 2014***), and fears of international spread of a Category A Priority Pathogen (***NIH, 2014***) have made this a massive focus for international public health (***Chan, 2014***). This has led to the current outbreak being declared a Public Health Emergency of International Concern on the 8 August 2014 (***Briand et al., 2014***; ***Gostin et al., 2014***; ***WHO, 2014d***).

The *Filoviridae*, of which *Ebolavirus* is a constituent genus, belong to the order *Mononegavirales*. Two other genera complete the family: *Marburgvirus*, itself responsible for a number of outbreaks of haemorrhagic fever across Africa (***Gear et al., 1975***; ***Conrad et al., 1978***; ***Smith et al., 1982***; ***Towner et al., 2006***) and *Cuevavirus*, recently isolated from bats in northern Spain (***Negredo et al., 2011***). Five species of *Ebolavirus* have been isolated to date (***Kuhn et al., 2010***; ***King et al., 2012***); the earliest recognised outbreaks of EVD were reported in Zaire (now the Democratic Republic of the Congo [DRC]) and Sudan in 1976 (***International Commission, 1978***; ***WHO International Study Team, 1978***). The causative viruses were isolated (***Pattyn et al., 1977***) and later identified to be distinct species, *Zaire ebolavirus* (EBOV) and *Sudan ebolavirus* (SUDV). A third species of *Ebolavirus*, *Reston ebolavirus*, was isolated from Cynomologus monkeys imported from the Philippines to a facility in the United States, where they experienced severe haemorrhaging (***Jahrling et al., 1990***). Whilst serological evidence of infection with this species has been reported in individuals in the Philippines (***Miranda et al., 1991***), no pathogenicity has been reported beyond primates and porcids (***Barrette et al., 2009***; ***Feldmann and Geisbert, 2011***). In 1994 a fourth species, *Tai Forest ebolavirus* was isolated from a veterinarian who had autopsied a chimpanzee in Côte d'Ivoire (***Le Guenno et al., 1995***), though the virus has not been detected subsequently. The final species, *Bundibugyo ebolavirus*, was responsible for an outbreak of EVD in Uganda in 2007 (***Towner et al., 2008***), as well as a more recent outbreak in the DRC (***WHO, 2012b***).

Initial analysis suggested that the viruses isolated from the current outbreak, originating in Guinea, formed a separate clade within the five *Ebolavirus* species (***Baize et al., 2014***). Subsequent re-analysis

of the same sequences however, indicated that these isolates instead nest within the *Zaire ebolavirus* lineage (**Dudas and Rambaut, 2014**), and diverged from Central Africa strains approximately ten years ago (**Gire et al., 2014**).

Which reservoir species are responsible for maintaining Ebola transmission between outbreaks is not well understood (**Peterson et al., 2004b**), but over the last decade significant progress has been made in narrowing down the list of likely hosts (**Peterson et al., 2007**) (**Figure 1**). Primates have long been known to harbour filoviral infections, with the first Marburg strains identified in African green monkeys in 1967 (**Siegert et al., 1967**; **Beer et al., 1999**). Significant mortality has also been reported in wild primate populations across Africa, most notably in gorilla (*Gorilla gorilla*) and chimpanzee (*Pan troglodytes*) populations (**Formenty et al., 1999**; **Rouquet et al., 2005**; **Bermejo et al., 2006**). The high case fatality rates recorded in the great apes combined with their declining populations and limited geographical range, indicate they are likely dead-end hosts for the virus and not reservoir species (**Groseth et al., 2007**). A large survey of small mammals in and around Gabon identified three species of bats which were infected with Ebola viruses—*Hypsignathus monstrosus*, *Epomops franqueti* and *Myonycetris torquata* (**Leroy et al., 2005**). Subsequent serological surveys (**Pourrut et al., 2009**; **Hayman et al., 2010**) and evidence linking the potential source of human outbreaks to bats (**Leroy et al., 2009**) lend support to the hypothesis of a bat reservoir. This, coupled with repeated detection of *Marburgvirus* in the fruit bat *Rousettus aegypticus* (**Towner et al., 2009**) and the only isolations of *Cuevavirus* also from bats (specifically *Llovia virus* [**Negredo et al., 2011**]), all support the suspicion that Chiroptera play an important role in the natural life-cycle of the filoviruses.

Humans represent a dead-end host for the virus, with only stuttering chains of transmission reported between humans in the majority of previous outbreaks (**Chowell et al., 2004**; **Legrand et al., 2007**) and no indication that humans can reintroduce the virus back into reservoir species (**Karesh et al., 2012**). The incubation period in humans ranges from two days to three weeks, after which a variety of clinical symptoms arise, affecting multiple organs of the body. At the peak of illness, haemorrhaging shock and widespread tissue damage can occur and can eventually lead to death within 6–16 days (**Feldmann and Geisbert, 2011**). Human-to-human transmission is mainly through direct unprotected contact with infected individuals and cadavers, with infectious particles detected in a number of different body fluids (**Feldmann and Geisbert, 2011**). The typical outbreak profile is defined by an index individual that has recently come into contact with the blood of another mammal through either hunting or the butchering of animal carcasses (**Pourrut et al., 2005**). Whilst it has been difficult to identify the zoonotic source for the index cases of some outbreaks, a recurring theme of hunting and handling bushmeat is suspected (**Table 1**; **Boumandouki et al., 2005**; **Nkoghe et al., 2005**, **2011**; **Leroy et al., 2009**). For some outbreaks, including the most recent, the initial source of zoonotic transmission has not been identified. In subsequent human-to-human transmission, the highest risk activities are those that bring humans into close contact with infected individuals. These include medical settings where insufficient infection control precautions have been taken, as well as home care and funeral preparations carried out by families or close friends (**Baron et al., 1983**; **Georges et al., 1999**; **Boumandouki et al., 2005**). As the conditions required for transmission are culturally and contextually dependent, opportunities for sustained transmission are highly heterogeneously distributed. Typically, chains of infection do not exceed three or four sequential transmission events, although occasionally (and particularly in

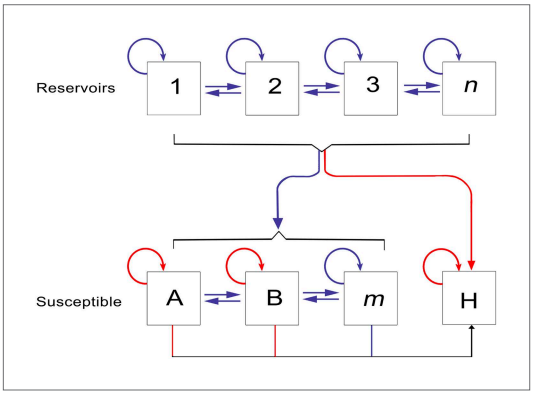


Figure 1. The epidemiology of Ebola virus transmission in Africa. Of the suspected reservoir species, 1, 2 and 3 represent the three bat species from which Ebola virus has been isolated (*Hypsignathus monstrosus*, *Myonycteris torquata* and *Epomops franqueti*) and *n* represents unknown reservoirs of the disease yet to be discovered. Of the susceptible species, A represents *Pan troglodytes*, B *Gorilla gorilla* and *m* represents other organisms susceptible to the disease, such as duikers. H represents humans. Blue arrows indicate unknown transmission cycles or infection routes and red arrow routes have been confirmed or are suspected. Adapted from **Groseth et al. (2007)**.

DOI: 10.7554/eLife.04395.003

Table 1. Locations of outbreaks of Ebola virus disease in humans

Outbreak	Countries	Date range	Location	Species	Reference
1	South Sudan	Jun–Nov 1976	Nzara	SUDV	(***WHO International Study Team, 1978***)
2	DRC	Sep–Oct 1976	Yambuku	EBOV	(***International Commission, 1978***)
3	DRC	Jun 1977	Bonduni	EBOV	(***Heymann et al., 1980***)
4	South Sudan	Jul–Oct 1979	Nzara	SUDV	(***Baron et al., 1983***)
5	Côte d'Ivoire	Nov 1994	Tai Forest	TAFV	(***Le Guenno et al., 1995***; ***Formenty et al., 1999***)
6	Gabon	Nov 1994–Feb 1995	Mekouka and Andock mining camps	EBOV	(***Amblard et al., 1997***; ***Georges et al., 1999***; ***Milleliri et al., 2004***)
7	DRC	Jan–Jul 1995	Mwembe Forest	EBOV	(***Muyembe and Kipasa, 1995***; ***Khan et al., 1999***)
8	Gabon	Jan–Mar 1996	Mayibout 2	EBOV	(***Georges et al., 1999***; ***Milleliri et al., 2004***)
9	Gabon	Jul 1996–Jan 1997	Booue	EBOV	(***Georges et al., 1999***; ***Milleliri et al., 2004***)
10	Uganda	Oct 2000–Feb 2001	Rwot-Obillo	SUDV	(***WHO, 2001***; ***Okware et al., 2002***; ***Lamunu et al., 2004***)
11	Gabon & ROC	Oct 2001–Mar 2002	Memdemba Entsiami, Abolo and Ambomi	EBOV	(WHO, 2003; ***Milleliri et al., 2004***; ***Nkoghe et al., 2005***; ***Pourrut et al., 2005***)
			Ekata		
			Oloba		
			Etakangaye		
			Grand Etoumbi		
12	ROC	Dec 2002–Apr 2003	Yembelangoye	EBOV	(WHO, 2003; ***Pourrut et al., 2005***)
			Nearby hunting camp		
			Mvoula		
13	ROC	Oct–Dec 2003	Mbandza	EBOV	(***Boumandouki et al., 2005***)
14	South Sudan	Apr–Jun 2004	Forests bordering Yambio	SUDV	(***WHO, 2005***; ***Onyango et al., 2007***)
15	ROC	Apr–May 2005	Odzala National Park	EBOV	(***Nkoghe et al., 2011***)
16	DRC	May–Nov 2007	Mombo Mounene 2 market	EBOV	(***Leroy et al., 2009***)
17	Uganda	Aug–Dec 2007	Kabango	BDBV	(***Towner et al., 2008***; ***MacNeil et al., 2010***; ***Wamala et al., 2010***)
18	DRC	Nov 2008–Feb 2009	Luebo	EBOV	(***Grard et al., 2011***)
19	Uganda	May 2011	Nakisamata	SUDV	(***Shoemaker et al., 2012***)
20	DRC	July–Nov 2012	Isiro	BDBV	(***CDC, 2014***; ***WHO, 2012b***)
21	Uganda	July–Oct 2012	Nyanswiga	SUDV	(***CDC, 2014***; ***WHO, 2012a***)
22	Uganda	Nov 2012–Jan 2013	Luwero District	SUDV	(***WHO, 2012c***; ***CDC, 2014***)
23	Guinea	Dec 2013 -	Meliandou	EBOV	(***Baize et al., 2014***; ***Bausch and Schwarz, 2014***)

DRC = Democratic Republic of the Congo, ROC = Republic of Congo.

DOI: 10.7554/eLife.04395.004

the early stages of infection) a single individual may be responsible for directly infecting a large number of others (**Brady et al., 2014**). In the outbreak in Gabon in 1996, a single person was responsible for infecting ten other individuals (**Milleliri et al., 2004**) whilst in the 1995 outbreak in the DRC, thirty five cases resulted from one individual (**Khan et al., 1999**). Secondary transmission can be restricted by effective case detection and isolation measures (**Shoemaker et al., 2012**; **WHO, 2014c**). Where this cannot be achieved, either due to a lack of infrastructure, poor understanding of the disease, or distrust of medical practices, secondary cases can continue to occur (**Khan et al., 1999**; **Larkin, 2003**; **Hewlett et al., 2005**). As the number of infections grows, the ability of healthcare systems to control the further spread diminishes and the risk of a large outbreak increases.

The recent outbreak in Guinea and surrounding countries indicate that the previous paradigm for Ebola outbreaks is shifting (**Briand et al., 2014**; **Chan, 2014**). The last 40 years of EVD outbreaks were accompanied by considerable changes in demographic patterns throughout Africa. There has been a large increase in population size coupled with increasing urbanisation (**Cohen, 2004**; **Seto et al., 2012**; **Linard et al., 2013**). African populations have also become better connected internally and internationally (**Linard et al., 2012**; **Huang and Tatem, 2013**). Only recently have we begun to understand the dynamic nature of these travel patterns (**Garcia et al., 2014**; **Gonzalez et al., 2008**; **Simini et al., 2012**; **Wesolowski et al., 2013**, **2012**) which have been clearly demonstrated to influence disease transmission over different temporal and spatial scales (**Hufnagel et al., 2004**; **Yang et al., 2008**; **Stoddard et al., 2009**; **Talbi et al., 2010**; **Brockmann and Helbing, 2013**; **Pindolia et al., 2014**). Changes in land use and penetration into previously remote areas of rainforest bring humans into contact with potential new reservoirs (**Daszak, 2000**), while changes in human mobility and connectivity will likely have profound impacts on the dispersion of Ebola cases during outbreaks. These conditions are thought to have a major role in setting the stage for the current outbreak.

This paper aims to define the areas suitable for zoonotic transmission of *Ebolavirus* (i.e., those routes defined in **Figure 1** excluding human-to-human transmission) through species distribution modelling techniques. The fundamental niche of a species can be conceptualised as the confluence of environmental conditions that support its presence in a particular location (**Franklin, 2009**). Species distribution models quantitatively describe this niche based on known occurrence records of the organism and their associated environmental conditions, enabling predictions of the likely geographic distribution of the species in other regions (**Elith and Leathwick, 2009**). The era of satellites and geographical information systems has made high resolution global data on environmental conditions increasingly available (**Hay et al., 2006**; **Weiss et al., 2014b**). Species distribution modelling using flexible machine learning approaches have been successfully applied to map the global distributions of disease vectors (**Sinka et al., 2012**) and pathogens such as dengue (**Bhatt et al., 2013**), influenza (**Gilbert et al., 2014**) and leishmaniasis (**Pigott et al., 2014**).

Previous studies applied the GARP (Genetic Algorithm for Rule-set Production) species distribution modelling approach (**Stockwell and Peters, 1999**) to the locations of 12 Ebola outbreaks in humans between 1976 and 2002 to map the likely distribution of Ebola viruses (**Peterson et al., 2004a**) and as a mechanism to identify potential reservoir hosts (**Peterson et al., 2004b**; **Peterson et al., 2007**). Here we update and improve the maps of the zoonotic transmission niche of EVD by: (i) incorporating more recent outbreak data from outside the formerly predicted niche of EVD; (ii) integrating for the first time data on outbreaks in primates and the occurrence of the virus in the suspected Old World fruit bat (OWFB) reservoirs; (iii) using new satellite-derived information on bespoke environmental covariates from Africa, including new distribution maps of the OWFB; and (iv) using new increasingly flexible niche mapping techniques in the modelling framework. To elucidate the relevance of these maps for transmission, we have also calculated the population at risk of primary spillover outbreaks from the zoonotic niche of EVD in Africa, and we investigated the changing nature of the populations within this niche.

Results

Reported EVD outbreaks

In total, 23 outbreaks of Ebola virus were identified in humans across Africa, consisting of a hypothesised 30 independent primary infection events (**Table 1**; **Figure 2**). These outbreaks span the last 40 years from the first outbreaks in 1976 to the five outbreaks that have occurred since 2010 (**Table 1**). The locations of the index cases span from West Africa, with the most westerly outbreak ongoing in Guinea, to Gabon, the Republic of Congo (ROC), the DRC, South Sudan and Uganda. Before December

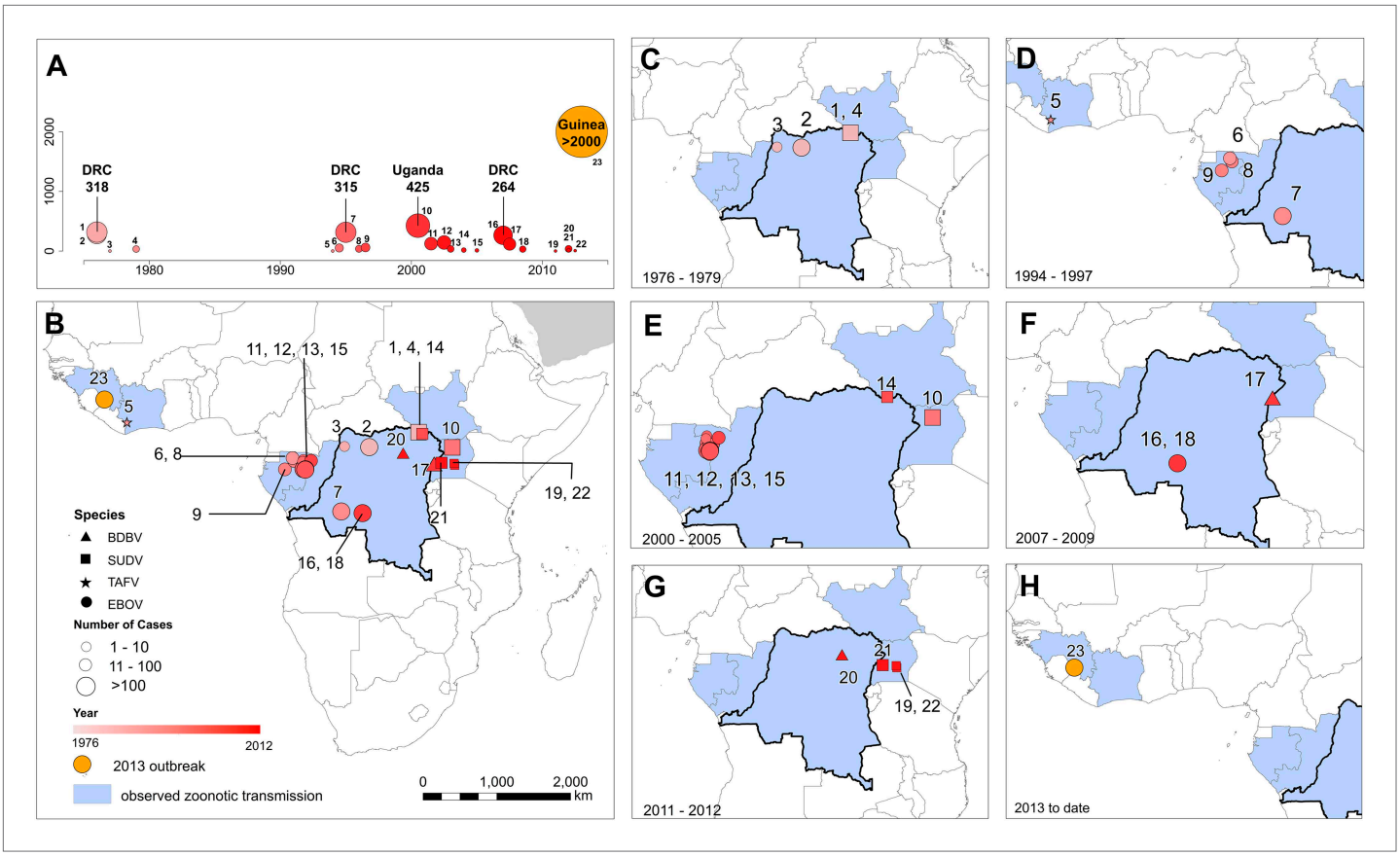


Figure 2. The locations of Ebola virus disease outbreaks in humans in Africa. (**A**) Illustrates the 23 reported outbreaks of Ebola virus disease through time, with the area of each circle and its position along the y-axis representing the number of cases. The onset year is represented by the colour as per (**B**). (**B**) Shows a map of the index cases for each of these outbreaks. (**C**–**H**) Show these outbreaks over a series of time periods. Numbers refer to outbreaks as listed in ***Table 1***. In (**B**–**H**) the species of Ebola virus responsible for the outbreak is illustrated by the symbol shape, the number of resulting cases and onset date by symbol colour. The most recent outbreak (#23) is indicated in orange. Countries in which zoonotic transmission to humans has been reported or is assumed to have occurred are coloured in blue. In each map the Democratic Republic of Congo is outlined for reference.
DOI: 10.7554/eLife.04395.005

2013, a total of 2322 cases had occurred from *Ebolavirus* infections, a number already overtaken by the likely underreported current case count of the ongoing outbreak >2250 (***WHO, 2014a***) (***Figure 2A***). Of the four viruses circulating in Africa, *Zaire ebolavirus* has been responsible for the most outbreaks (13), followed by *Sudan ebolavirus* (7) and *Bundigbuyo ebolavirus* with just two outbreaks in 2007/8 and 2012. Tai Forest has caused one confirmed infection in humans, from which the patient recovered (***Le Guenno et al., 1995***; ***Formenty et al., 1999***). Although outbreaks have been reported since 1976, there was an absence of reported outbreaks in humans for 15 years between 1979 and 1994 (although antibodies in humans were identified over the period [***Kuhn, 2008***]) and the frequency of outbreaks has increased substantially post 2000 (***Figure 2A***).

Reported Ebola virus infections in animals

A total of 51 surveyed locations reporting infections in animals were identified in the literature since the discovery of the disease (***Table 2***; ***Figure 3***). These comprised 17 infections in gorillas (*Gorilla gorilla*), nine infections in chimpanzees (*Pan troglodytes*), 18 in OWFB and 2 in duikers (*Cephalophus* spp.). A large proportion of the great ape cases originated from the ROC/Gabon border, coinciding with the main known distributions of both chimpanzees and gorillas (***Petter and Desbordes, 2013***) and representing a period of well-documented great ape Ebola outbreaks in and around the Lossi Animal Sanctuary (***Rouquet et al., 2005***; ***Bermejo et al., 2006***; ***Walsh et al., 2009***). All animal isolations of Ebola viruses have come from countries that have also reported index cases of human outbreaks, with the exception of several seropositive bats from a survey in southern Ghana.

Table 2. Locations of reported infections with Ebola virus in animals

Site	Country	Date range	Location	Species	Diagnosis	Reference
1	Côte d'Ivoire	Oct–Nov 1994	Tai Forest	Chimpanzee	Serology	(**Formenty et al., 1999**)
2	Gabon	Jan 1996	Mayiboth 2	Chimpanzee	PCR	(**Lahm et al., 2007**)
3	Gabon	Jul 1996	Near Booue	Chimpanzee	Serology	(**Georges-Courbot et al., 1997**)
4	Gabon	Sept 1996	Lope National Park	Chimpanzee	PCR	(**Lahm et al., 2007**)
5	Gabon & ROC	Aug 2001	Mendemba/Lossi Animal Sanctuary	Chimpanzee	PCR	(**Lahm et al., 2007**)
6	Gabon & ROC	Aug 2001	Mendemba/Lossi Animal Sanctuary	Gorilla	PCR	(**Lahm et al., 2007**)
7	Gabon & ROC	Aug 2001	Mendemba/Lossi Animal Sanctuary	*Cephalophus dorsalis*	PCR	(**Lahm et al., 2007**)
8	Gabon	Nov 2001	Zadie	Gorilla	PCR	(**Rouquet et al., 2005**)
9	Gabon	Nov 2001	Ekata	Gorilla	PCR	(**Wittmann et al., 2007**)
10	Gabon	Dec 2001	Medemba and neighbouring villages	Chimpanzee and Gorilla	PCR	(**Leroy et al., 2002**)
11	Gabon	Feb 2002	Zadie	Gorilla	PCR	(**Rouquet et al., 2005**)
12	Gabon	Feb 2002	Ekata	Various bat species	Serology	(**Leroy et al., 2005**)
13	Gabon	Mar 2002	Zadie	Gorilla	PCR	(**Rouquet et al., 2005**)
14	Gabon	Mar 2002	Grand Etoumbi	Gorilla	PCR	(**Wittmann et al., 2007**)
15	Gabon	Apr 2002	Ekata	Gorilla	PCR	(**Wittmann et al., 2007**)
16	ROC	May 2002	Oloba	Chimpanzee	PCR	(**Lahm et al., 2007**)
17	ROC	Dec 2002	Lossi Animal Sanctuary	Gorilla	PCR	(**Rouquet et al., 2005**)
18	ROC	Dec 2002	Lossi Animal Sanctuary	Gorilla	PCR	(**Rouquet et al., 2005**)
19	ROC	Dec 2002	Lossi Animal Sanctuary	Chimpanzee	Serology	(**Rouquet et al., 2005**)
20	ROC	Dec 2002	Lossi Animal Sanctuary	Gorilla	PCR	(**Rouquet et al., 2005**)
21	ROC	Dec 2002	Lossi Animal Sanctuary	Gorilla	PCR	(**Rouquet et al., 2005**)
22	ROC	Dec 2002	Lossi Animal Sanctuary	*Cephalophus* spp.	PCR	(**Rouquet et al., 2005**)
23	Gabon	Feb 2003	Mbomo	Various bat species	PCR	(**Leroy et al., 2005**)
24	ROC	Feb 2003	Lossi Animal Sanctuary	Gorilla	Serology	(**Rouquet et al., 2005**)
25	Gabon	Feb 2003	Lossi Animal Sanctuary	Chimpanzee	PCR	(**Wittmann et al., 2007**)
26	Gabon	Jun 2003	Mbomo	Various bat species	PCR and serology	(**Leroy et al., 2005**)
27	ROC	Jun 2003	Near Mbomo and Ozala National Park	*Epomops franqueti*	Serology	(**Pourrut et al., 2009**)
28	ROC	Jun 2003	Near Mbomo and Ozala National Park	*Hypsignathus monstrosus*	Serology	(**Pourrut et al., 2009**)
29	ROC	Jun 2003	Near Mbomo and Ozala National Park	*Myonycteris torquata*	Serology	(**Pourrut et al., 2009**)

Table 2. Continued on next page

Table 2. Continued

Site	Country	Date range	Location	Species	Diagnosis	Reference
30	ROC	Jun 2003	Mbanza	Gorilla	PCR	(**Rouquet et al., 2005**)
31	ROC	Jan–Jun 2004	Lokoué	Gorilla	Reported	(**Caillaud et al., 2006**)
32	ROC	May 2004	Lokoué	Gorilla	PCR	(**Wittmann et al., 2007**)
33	Gabon	Feb 2005	Near Franceville	*Epomops franqueti*	Serology	(**Pourrut et al., 2009**)
34	Gabon	Feb 2005	Near Franceville	*Myonycteris torquata*	Serology	(**Pourrut et al., 2009**)
35	Gabon	Apr 2005	Near Lambarene	*Epomops franqueti* and *Hypsignathus monstrosus*	Serology	(**Pourrut et al., 2007**)
36	ROC	May 2005	Near Mbomo and Ozala National Park	*Epomops franqueti*	Serology	(**Pourrut et al., 2009**)
37	ROC	May 2005	Near Mbomo and Ozala National Park	*Hypsignathus monstrosus*	Serology	(**Pourrut et al., 2009**)
38	ROC	May 2005	Near Mbomo and Ozala National Park	*Myonycteris torquata*	Serology	(**Pourrut et al., 2009**)
39	ROC	Jun 2005	Odzala National Park	Gorilla	PCR	(**Wittmann et al., 2007**)
40	Gabon	Feb 2006	Near Tchibanga	Various bat species	Serology	(**Pourrut et al., 2009**)
41	ROC	May 2006	Near Mbomo and Ozala National Park	*Epomops franqueti*	Serology	(**Pourrut et al., 2009**)
42	ROC	May 2006	Near Mbomo and Ozala National Park	*Hypsignathus monstrosus*	Serology	(**Pourrut et al., 2009**)
43	ROC	May 2006	Near Mbomo and Ozala National Park	*Myonycteris torquata*	Serology	(**Pourrut et al., 2009**)
44	Gabon	Oct 2006	Near Franceville	*Epomops franqueti*	Serology	(**Pourrut et al., 2009**)
45	Ghana	May 2007	Sagyimase	*Epomops franqueti*	Serology	(**Hayman et al., 2012**)
46	Ghana	May 2007	Sagyimase	*Hypsignathus monstrosus*	Serology	(**Hayman et al., 2012**)
47	Ghana	May 2007	Adoagyir	*Epomophorus gambianus*	Serology	(**Hayman et al., 2012**)
48	Ghana	May 2007	Adoagyir	*Epomops franqueti*	Serology	(**Hayman et al., 2012**)
49	Ghana	Jun 2007	Oyibi	*Epomophorus gambianus*	Serology	(**Hayman et al., 2012**)
50	Ghana	Jan 2008	Accra	*Eidolon helvum*	Serology	(**Hayman et al., 2010**)
51	Gabon	Mar 2008	Near Franceville	*Epomops franqueti*	Serology	(**Pourrut et al., 2009**)

ROC = Republic of Congo.

DOI: 10.7554/eLife.04395.006

Predicted distribution of suspected reservoir species of bats

Three species of bats, *Hypsignathus monstrosus*, *Myonycteris torquata* and *Epomops franqueti*, were identified as the most likely candidates to be reservoir species for Ebola viruses due to high

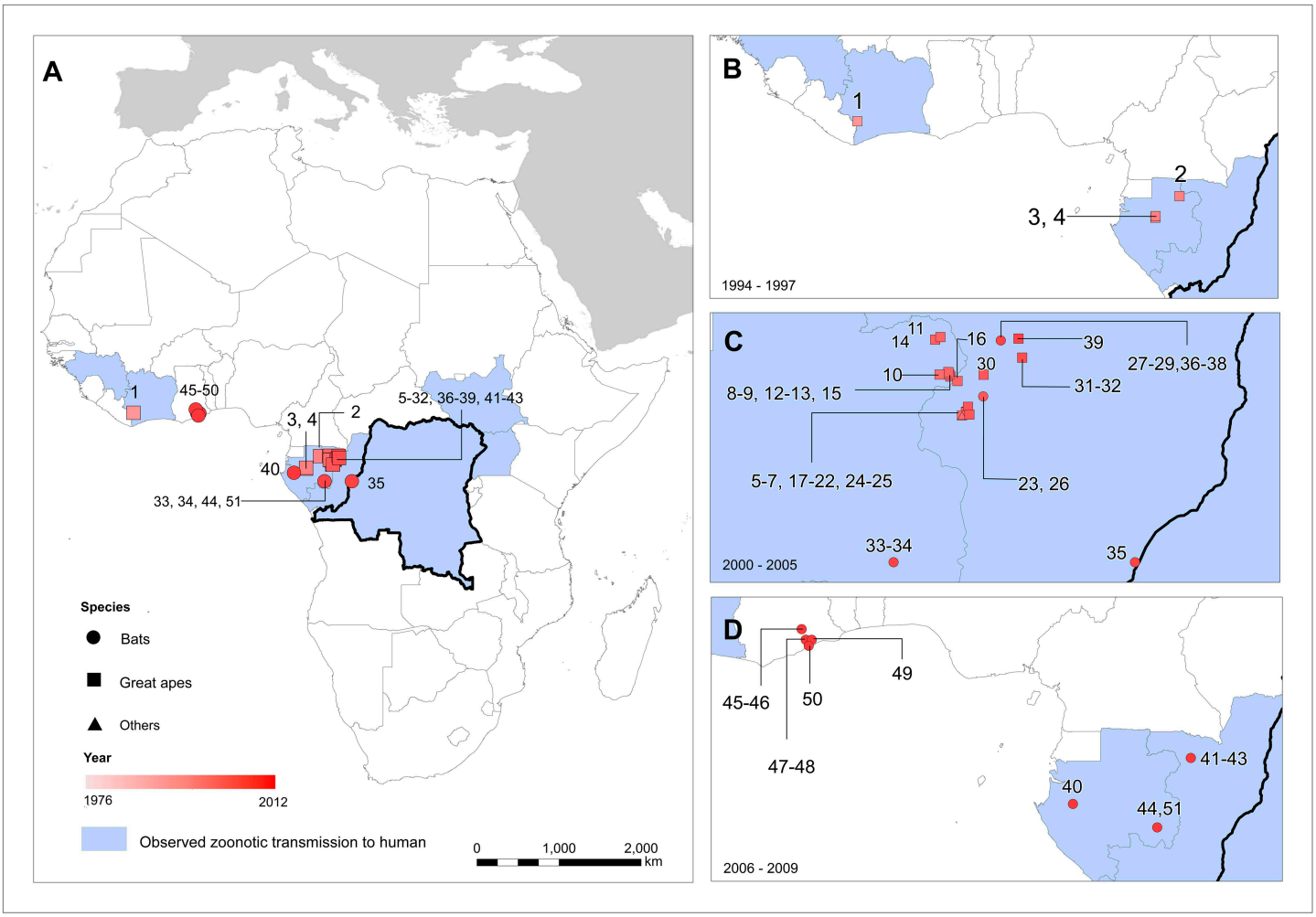


Figure 3. The locations of reported Ebola virus infection in animals in Africa. (**A**) Shows the locations of reported Ebola virus infection in animals. (**B**–**D**) Show these records in animals over three different time periods. Numbers refer to records as listed in ***Table 2***. In all panels, the species in which infection was detected is given by symbol shape and the year recorded by symbol colour. Blue countries represent locations where zoonotic transmission to humans has been reported or is assumed to have occurred. In each map the Democratic Republic of Congo is outlined for reference.

DOI: 10.7554/eLife.04395.007

seroprevalence and the isolation of RNA closely related to *Zaire ebolavirus* (***Leroy et al., 2005***; ***Olival and Hayman, 2014***). In total, 239 locations were identified from the Global Biodiversity Information Facility (GBIF) (***GBIF, 2014***): 67 for *H. monstrosus* (***Figure 4A***), 52 for *M. torquata* (***Figure 4B***) and 120 for *E. franqueti* (***Figure 4C***). Distribution models for all three species demonstrated predictive skill (indicated by an area under the curve (AUC) greater than 0.5) as follows: *H. monstrosus* AUC 0.63 ± 0.04; *M. torquata* AUC = 0.59 ± 0.04; *E. franqueti* AUC = 0.58 ± 0.03, n = 50 submodels for all three species. In addition, each species was broadly predicted within its considered expert opinion range (***Figure 4A–C***) (***Schipper et al., 2008***). The marginal effect plots (not shown) were strongly influenced by land surface temperature (LST) and vegetation (as measured by the enhanced vegetation index [EVI]). The predicted combined distribution of these species (***Figure 4D***), covers West and Central Africa, specifically the moist forests of the northeastern, western and central Congo basin, and Guinea, as well as the Congolian coastal forest ecoregions (***WWF, 2014***).

Predicted environmental suitability for zoonotic transmission of Ebola

The predicted environmental niche for zoonotic transmission of EVD is shown in ***Figure 5***. All countries with observed index cases of EVD (n = 7, hereafter Set 1) have areas of the highest environmental suitability (see list in ***Table 1***). In addition, areas of high environmental suitability for zoonotic

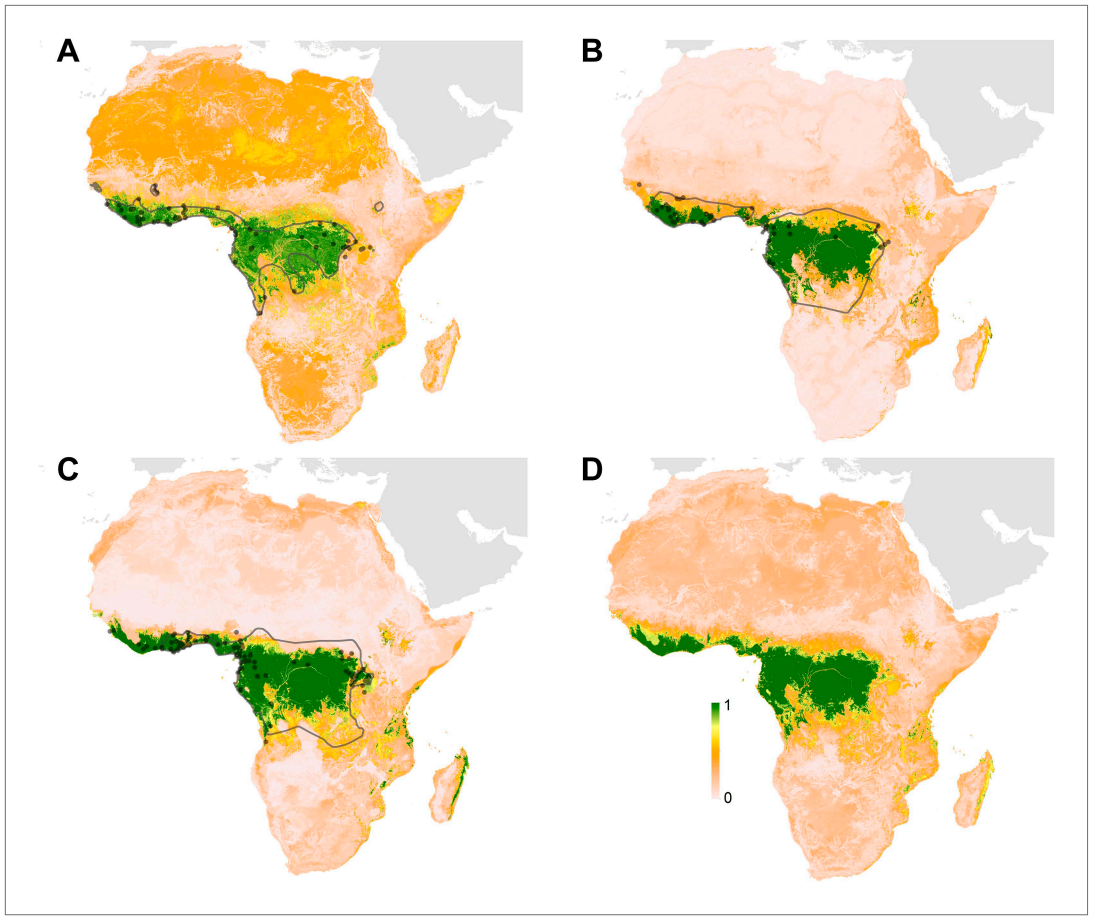


Figure 4. Predicted geographical distribution of the three species of Megachiroptera suspected to reservoir Ebola virus. (**A**) Shows the distribution of the hammer-headed bat (*Hypsignathus monstrosus*), (**B**) The little collared fruit bat (*Myonycteris torquata*) and (**C**) Franquet's epauletted fruit bat (*Epomops franqueti*). In each map, the locations of reported observations of each species, extracted and curated from the Global Biodiversity Information Facility (***GBIF, 2014***) and used to train each model are given as grey points (*H. monstrosus*, n = 67; *E. franqueti*, n = 120 and *M. torquata*, n = 52). Expert opinion maps of the known range of each species, generated by the IUCN (***Schipper et al., 2008***), are outlined in grey. The colour legend represents a scale of the relative probability that the species occurs in that location from 0 (white, low) to 1 (green, high). Area under the curve statistics, calculated under a stringent ten-fold cross validation procedure, are 0.63 ± 0.04, 0.59 ± 0.04 and 0.58 ± 0.03 for *H. monstrosus*, *M. torquata* and *E. franqueti* respectively. (**D**) Is a composite distribution map giving the mean, relative probability of occurrence from (**A**–**C**).

DOI: 10.7554/eLife.04395.008

transmission are predicted in a further 15 countries where, to date, index cases of the four African species of *Ebolavirus* have not been recorded. These are Nigeria, Cameroon, Central African Republic (CAR), Ghana, Liberia, Sierra Leone, Angola, Tanzania, Togo, Ethiopia, Mozambique, Burundi, Equatorial Guinea, Madagascar and Malawi (hereafter Set 2).

The AUC for the Ebola model was relatively high (AUC = 0.85 ± 0.04, n = 500 submodels) indicating that the model could strongly distinguish regions of environmental suitability for EVD. Enhanced vegetation index had the greatest impact on the distribution (relative contribution [RC] of 65.3%) followed by elevation (RC = 11.7%), night-time land surface temperature (LST) (RC = 7.7%), potential evapotranspiration (PET) (RC = 5.7%) and combined bat distribution (RC = 3.8%). Marginal effect plots are presented in ***Figure 5—figure supplement 2***.

In total, 22.2 million people are predicted to live in areas suitable for zoonotic transmission of Ebola. The vast majority, 21.7 million (approximately 97%), live in rural areas, as opposed to urban or peri-urban areas (***CIESIN/IFPRI/WB/CIAT, 2007***; ***WorldPop, 2014***). Of these, 15.2 million are in Set 1 and 7 million are in Set 2. In terms of ranked populations at risk, DRC, Guinea and Uganda are highest

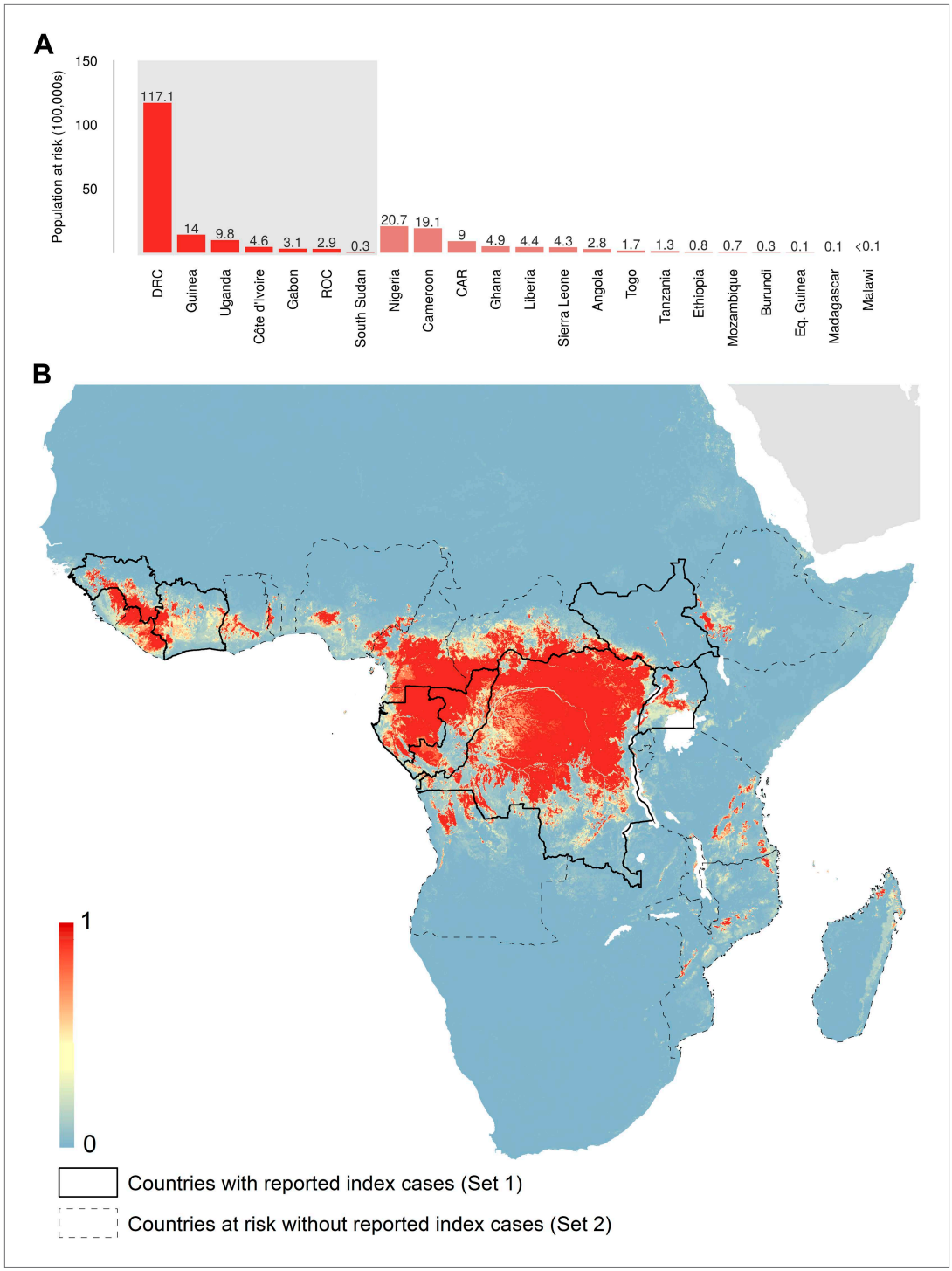


Figure 5. Predicted geographical distribution of the zoonotic niche for Ebola virus. (**A**) Shows the total populations living in areas of risk of zoonotic transmission for each at-risk country. The grey rectangle highlights countries in which index cases of Ebola virus disease have been reported (Set 1); the remainder are countries in which risk of zoonotic transmission is predicted, but in which index cases of Ebola have not been reported (Set 2). These countries are ranked by population at risk within each set. The population at risk Figure in 100,000 s is given above each bar. (**B**) Shows the predicted distribution of zoonotic Ebola virus. The scale reflects the relative probability that zoonotic transmission of Ebola virus could occur at these locations; areas closer to 1 (red) are more likely to harbour zoonotic transmission than those closer to 0 (blue). Countries with borders outlined are those which are predicted to contain at-risk areas for zoonotic transmission based on a thresholding approach (see 'Materials and methods').
Figure 5. Continued on next page

Figure 5. Continued

The area under the curve statistic, calculated under a stringent 10-fold cross-validation procedure is 0.85 ± 0.04. Solid lines represent Set 1 whilst dashed lines delimit Set 2. Areas covered by major lakes have been masked white.
DOI: 10.7554/eLife.04395.009

The following figure supplements are available for figure 5:

Figure supplement 1. Covariates used in predicting zoonotic transmission niche of Ebola.
DOI: 10.7554/eLife.04395.010

Figure supplement 2. Marginal effect plots for each covariate used in the Ebola virus distribution model.
DOI: 10.7554/eLife.04395.011

Figure supplement 3. Comparison of predictions for zoonotic niche of Ebola virus excluding the Guinea outbreak.
DOI: 10.7554/eLife.04395.012

in Set 1 and Nigeria, Cameroon and CAR are top in Set 2. For a full listing of these populations living in areas of risk, see the stacked bar plot in ***Figure 5A***.

National level demographic and mobility changes

Over the 40 year period since discovery of EVD, the total population living in those countries predicted to be within the zoonotic niche has nearly tripled (from 230 million to 639 million) and the proportion of the population in these countries living in an urban (rather than rural) setting has changed from 25.5% to 59.2% (***Figure 6***).

Data on the connectivity of human populations over this period were not available. We can infer however, intuitively, empirically and theoretically (***Zipf, 1946***; ***Simini et al., 2012***) that rates of population movement within a country will scale directly in proportion to population growth.

International connectivity by airline traffic

Records of passenger seat capacity are available since 2000 and show substantive increases over the period in Set 1 (from 2.96 to 4.77 million, a fractional change of 1.61) and Set 2 (from 5.6 to 15.6 million, a change of 2.8) (***Figure 7A***). More specific data on passenger volumes show almost universally similar increases since 2005 with Set 1 nations changing from 2 million to 2.5 million, a fractional change of 1.22 and Set 2 changing from 5 million to 7.9 million, a change of 1.57 (***Figure 7B***).

Global analysis of airline passenger volumes demonstrates that international connectivity has increased amongst all global regions and national income strata (***Figure 8***). Total passenger volumes

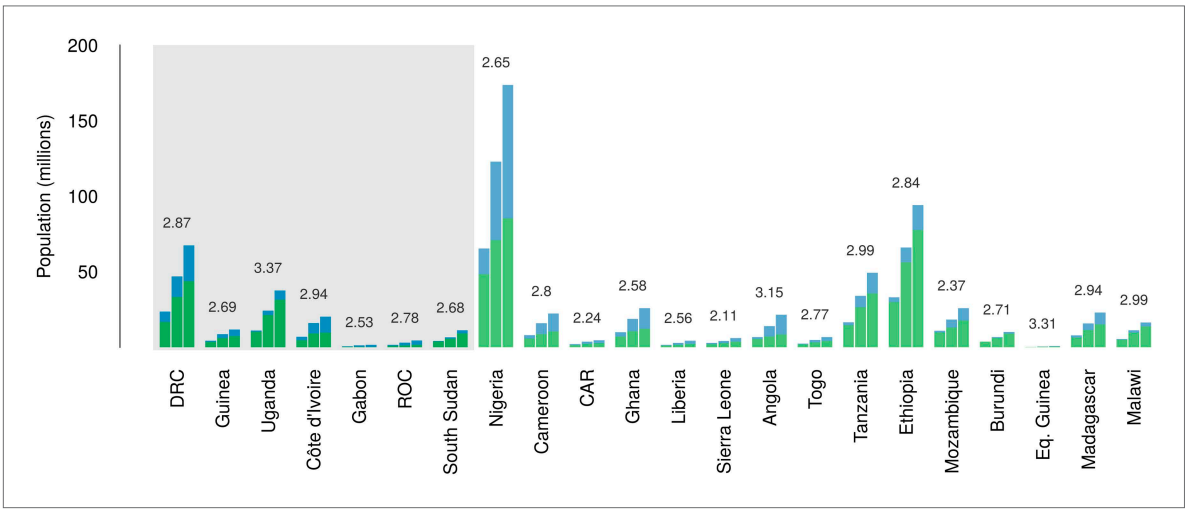


Figure 6. Changes in national population for countries predicted to contain areas at-risk of zoonotic Ebola virus transmission. For each country the population (in millions) is presented for three time periods (1976, 2000 and 2014) as three bars. Each stacked bar gives the rural (green) and urban (blue) populations of the country. The grey rectangle highlights countries in which index cases of Ebola virus diseases have been reported (Set 1); the remainder are countries in which risk of zoonotic transmission is predicted, but where index cases have not been reported (Set 2). The fractional change in population between 1976 and 2014 is given above each set of bars.
DOI: 10.7554/eLife.04395.013

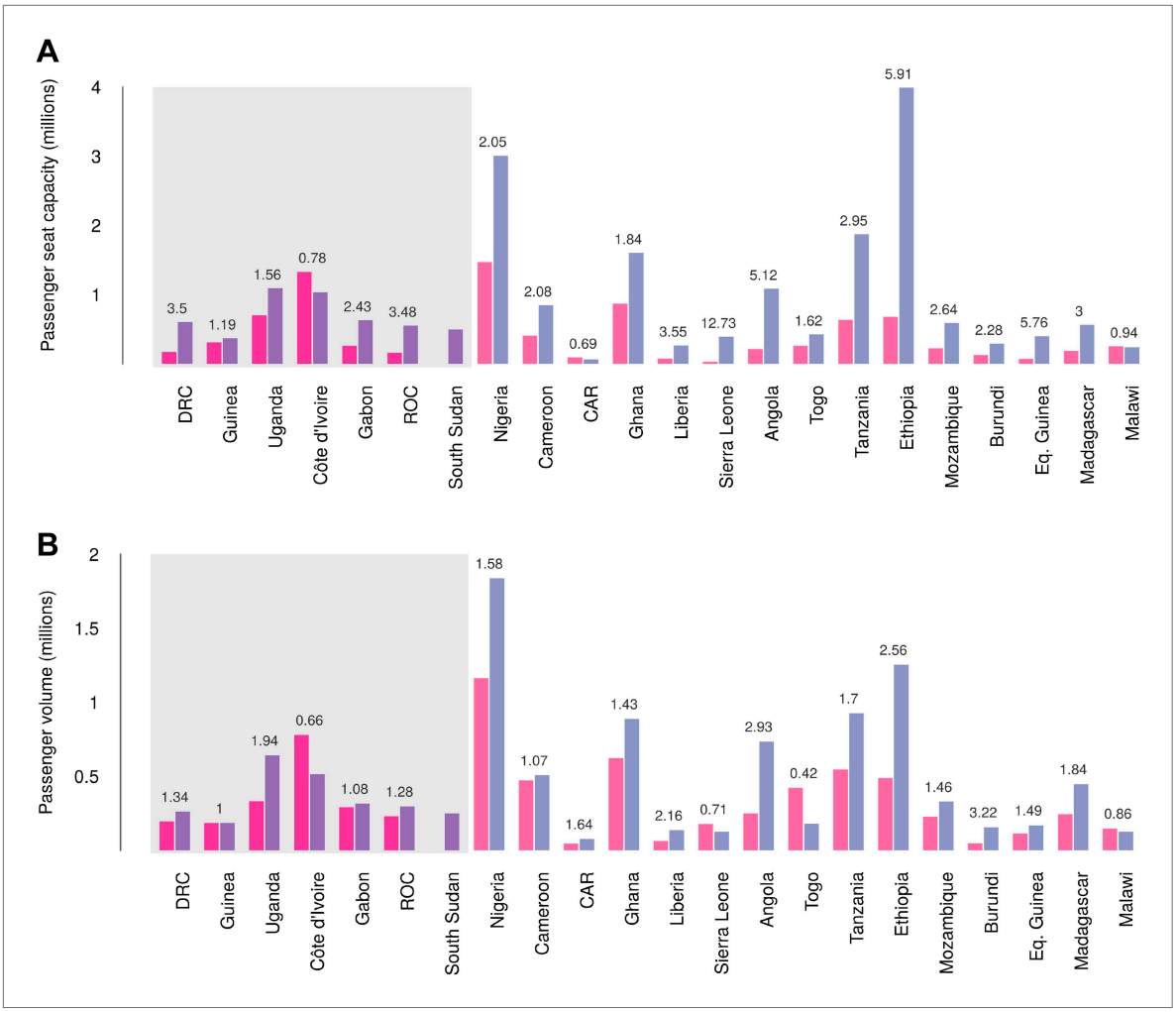


Figure 7. Changes in international flight capacity and traveller volumes for countries predicted to contain areas at-risk of zoonotic Ebola virus transmission. The grey rectangle highlights countries in which index cases of EVD have been reported (Set 1). The remainder are countries in which risk of zoonotic transmission is predicted, but where index cases have not been reported (Set 2). (**A**) Shows changes in annual outbound international seat capacity (between 2000 in red and 2013 in blue). (**B**) Depicts changes in annual outbound international passenger volume by country (between 2005 in red and 2012 in blue). For each country, the fractional change in volume is given above each set of bars. Note that only one bar is presented for South Sudan as data for this region prior to formation of the country in 2011 were unavailable.

DOI: 10.7554/eLife.04395.014

have increased by a third from 9.5 to over 14 million during the eight year window (2005–2012) where records are available. The largest increases have occurred in WHO regions (***WHO, 2014b***) outside of the sub-Saharan African region (AFRO) (***Figure 8A,B***). In 2012, almost half of the final destinations of those travelling from these at-risk countries were to other AFRO nations (47%). Other frequent destinations were in Europe (EURO; 27%) and the Eastern Mediterranean (EMRO; 13%). Similarly, analysis of passenger volumes by World Bank national income groupings (***WHO, 2014b***) (***Figure 8C,D***) show that in 2012 40% of all passenger final destinations were to low or low-middle income countries.

Discussion

Summary of the main findings

We have re-evaluated the zoonotic niche for EVD in Africa. In doing so we have (i) used all existing outbreaks to assemble an inventory of index cases (n = 30); (ii) added to this all confirmed records of Ebola virus in animals (n = 51); (iii) assembled more accurate and contemporary environmental covariates including new maps of the distribution of confirmed OWFB reservoirs of the disease; and (iv) used

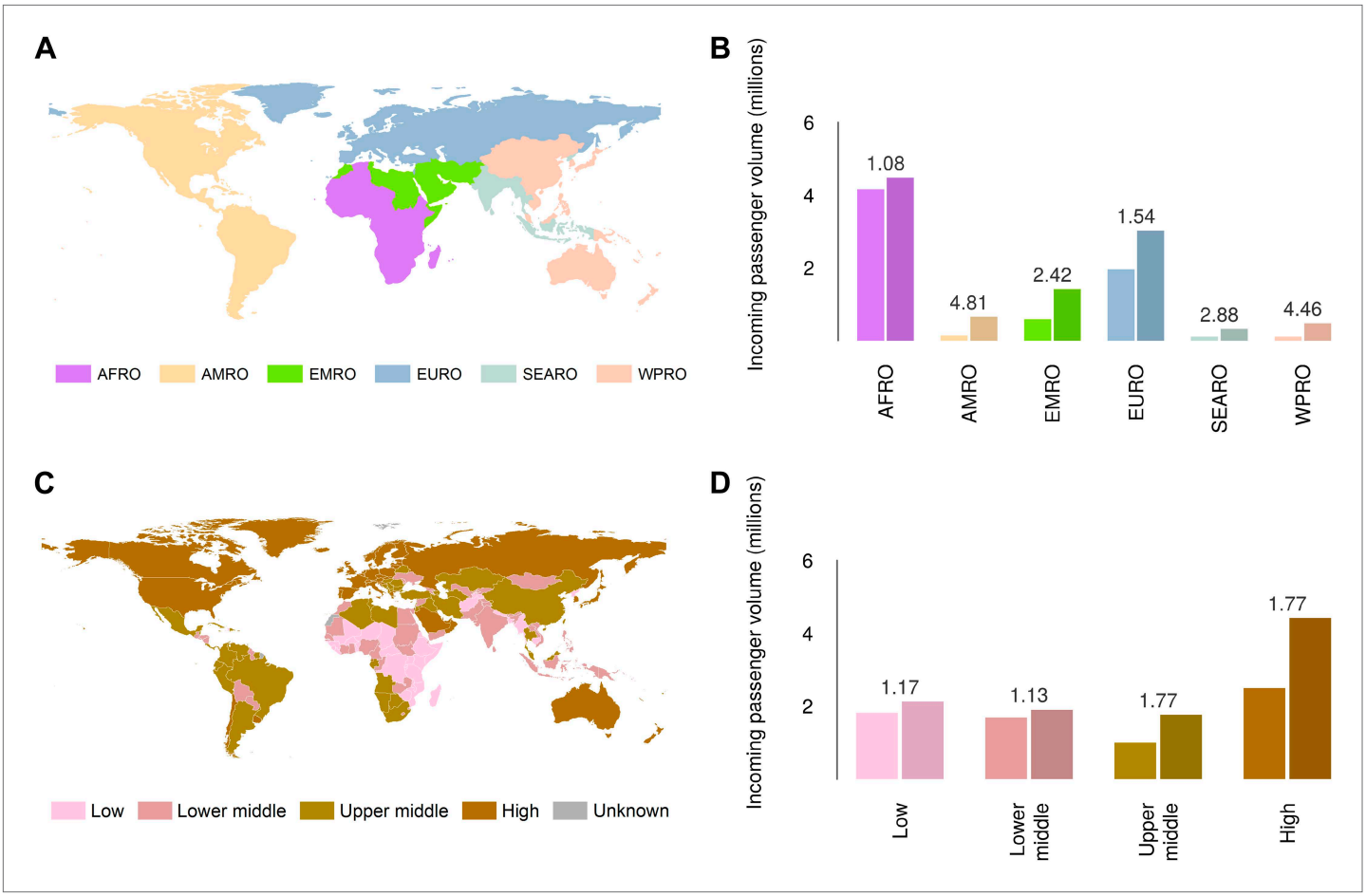


Figure 8. Numbers of airline passengers arriving from at-risk countries to other countries stratified by major geographic regions and national income groups. (**A**) Shows the locations of WHO regions (AFRO–African Region; AMRO–Region of the Americas; EMRO–Eastern Mediterranean Region; EURO–European Region; SEARO–South-East Asian Region; WPRO–Western Pacific Region). (**B**) Displays the numbers of passengers arriving in each of these regions from countries predicted to contain areas at risk of zoonotic Ebola virus transmission (Sets 1 and 2) in 2005 and 2012. (**C**) Shows the income tiers of all countries as defined by the World Bank. (**D**) Displays the total numbers of passengers arriving in countries in each of these income strata from at-risk countries in 2005 and 2012. The number above each pair of bars indicates the fractional change in these numbers of incoming passengers between 2005 and 2012.

DOI: 10.7554/eLife.04395.015

the latest niche modelling techniques to predict the geographic distribution of potential zoonotic transmission of the disease. Using these predictions we have estimated the populations at risk of EVD both in countries which have confirmed index cases (Set 1, n = 7) and those for which we predict strong environmental suitability for outbreaks (Set 2, n = 15). In all countries at risk we show that since the discovery of EVD in 1976, urban and rural populations have increased and have become more interconnected both within and across national borders. During the last 40 years the increasing size and connectivity of these populations may have facilitated the subsequent spread of EVD outbreaks. These factors underline a change in the way in which EVD interacts with human populations.

Interpreting the zoonotic niche

The remote and isolated nature of Ebola zoonotic transmission events, paired with the relatively poor diagnostics and understanding of the disease transmission routes in early outbreaks, mean that under-reporting of previous outbreaks is probable. An increasing understanding and description of a broader range of symptoms used in case definitions of EVD (***Leroy et al., 2000***; ***Feldmann and Geisbert, 2011***) also increase the possibility that past outbreaks may have been misattributed to different diseases (***Tignor et al., 1993***). This poor detectability of EVD also clearly limits capacity to accurately

identify the locations and transmission routes of index cases (**Heymann et al., 1980**; **Baize et al., 2014**). We must assume, as has been done previously (**Peterson et al., 2004a**; **Jones et al., 2008**), that the first reported cases are representative of the true location of the index cases. Where possible we have represented this geographic uncertainty by attributing the index case to a wide-area polygon which then incorporated this uncertainty into the mapping process (see 'Materials and methods').

The relationship between the EVD niche and the environmental covariates (**Figure 5—figure supplement 2**), particularly the high relative contribution of the vegetation index, underscore that there are clear environmental limits to transmission of the virus from animals to humans, and that ecoregions dominated by rainforest are the primary home of such zoonotic cycles. Our analysis has shown that the zoonotic niche of the pathogen is more widespread than previously predicted or appreciated (**Peterson et al., 2004a**), most notably in West Africa.

This analysis used information from all human outbreaks and animal infections to delineate the likely zoonotic niche of the disease. Further analysis, excluding the existing outbreak focussed in Guinea from the dataset used to train the model (**Figure 5—figure supplement 3**), still resulted in prediction of high suitability in this region, with the presumed index village located within 5 km of an at-risk pixel. This implies that the eco-epidemiological situation in Guinea is very similar to that in past outbreaks, mirroring phylogenetic similarity in the causative viruses (**Dudas and Rambaut, 2014**; **Gire et al., 2014**). The ecological similarity between the past and current outbreaks also lends support to the notion that the scale of this outbreak is more heavily influenced by patterns of human-to-human transmission than any expansion of the zoonotic niche.

Interpreting population at risk

It is important to appreciate that this zoonotic niche map delineates areas in which populations are at-risk of zoonotic transmission of EVD (**Figure 5B**). It does not predict the likelihood of EVD spillover, the likelihood of an outbreak establishing, or its subsequent rate of spread within a population. Increasing human encroachment and certain cultural practices sometimes linked with poverty, such as bushmeat hunting, result in increasing exposure of humans to animals which may harbour diseases including Ebola (**Daszak, 2000**; **Wolfe et al., 2005**, **2007**). Increasing human population may accelerate the degree of risk through these processes but spatially refined information on these factors is not available comprehensively. It is hoped that as the understanding of the risk factors for zoonotic transmission of *Ebolavirus* to humans increases, it will be possible to incorporate this information into future risk mapping assessments.

Previous considerations of the geographic distribution of EVD have used human outbreaks alone. We have updated this work to include the last decade of outbreaks, as well as disaggregated outbreaks where evidence suggests multiple independent zoonotic transmission events overlap in space and time. Furthermore, our modelling process accommodates uncertainty in geopositioning of these index cases by utilising both point and polygon data. In addition, we include occurrence of infection in wildlife, important to the wider scale of zoonotic transmission (**Figure 1**), which in total has increased the dataset used in the model to 81 occurrences. The rareness of EVD outbreaks and the prevalence of detectable Ebola virus in reservoir species suggests that there will always be a limited set of observation data when compared to mapping of more prevalent zoonoses (**Pigott et al., 2014**). The results demonstrate predictive skill using a stringent validation procedure, however, indicating strong model performance even with this relatively limited observation dataset.

A broad zoonotic niche is predicted across 22 countries in Central and West Africa. Whilst several of these countries have reported index cases of EVD, others have not, although serological evidence in some regions points to possible underreporting of small-scale outbreaks (**Kuhn, 2008**). With improved ecological understanding, particularly with improvements to our knowledge of specific reservoir species and their distributions, it may be possible to delineate areas not at risk due to the absence of these species.

Despite relatively a large population living in areas of risk and the widespread practice of bushmeat hunting in these predicted areas (**Wolfe et al., 2005**; **Mfunda and Røskaft, 2010**; **Brashares et al., 2011**; **Kamins et al., 2011**), *Ebolavirus* is rare both in suspected animal reservoirs (**Leroy et al., 2005**; **Olival and Hayman, 2014**) and in terms of human outbreaks (**Table 1**). There is some indication however, that the frequency of Ebola outbreaks has increased since 2000, as shown in **Figure 2A**. We have shown that the human population living within this niche is larger, more mobile and better internationally connected than when the pathogen was first observed. As a result, when spillover events do

occur, the likelihood of continued spread amongst the human population is greater, particularly in areas with poor healthcare infrastructure (**Briand et al., 2014**; **Fauci, 2014**).

Whilst rare in comparison to other high burden diseases prevalent in this region, such as malaria (**Gething et al., 2011**; **Murray et al., 2012**), Ebola outbreaks can have a considerable economic and political impact, and the subsequent destabilisation of basic health care provisioning in affected regions increases the toll of unrecorded morbidity and mortality of more common infectious diseases (**Murray et al., 2014**; **Wang et al., 2014**), throughout and after the epidemic period. The number of concurrent infections during the present outbreak represents a significant strain on healthcare systems that are already poorly provisioned (**Briand et al., 2014**; **Chan, 2014**; **Fauci, 2014**) and many other Set 1 and Set 2 countries rank amongst the lowest per capita healthcare spenders. These considerations should be paramount when international organizations debate the financing requirements for the improvement of healthcare needed in the region and the urgency with which new therapeutics and vaccines can be brought into production (**Brady et al., 2014**; **Goodman, 2014**).

Together, these considerations necessitate prioritisation of efforts to reinforce and improve existing surveillance and control, and encourage the development of therapeutics and vaccines. The national population at risk estimates presented here would be a strong rationale for improving, prioritising and stratifying surveillance for EVD outbreaks and diagnostic capacity in these countries. We believe it would be prudent to test OWFB species in Set 2 countries for Ebola virus (**Hayman et al., 2012**), particularly during the bat breeding season to maximise chances of isolation in order to clarify the outbreak risk in these countries. In all Set 1 and Set 2 countries, raising awareness about the risk presented by reservoir bats and incidental primate hosts and the modes of transmission of this disease could be of value. Finally, increasing our capacity to rapidly map ever changing biological threats is also a core need (**Hay et al., 2013b**).

Interpreting International connectivity

The increasing connectedness of the Africa region means that EVD is now a problem of international concern. While most EVD secondary transmission occurs locally and is likely transported via ground transit (**Francesconi et al., 2003**), the potential for international spread of infection is possible, as demonstrated by the importation of the disease from Liberia to Nigeria, culminating in further secondary transmission in Lagos (**WHO, 2014a**). The aetiology of EVD infection and disease progression means that an international outbreak propagated by air travel remains unlikely, particularly in high-income countries better able to handle EVD cases (**Fauci, 2014**). Nevertheless, a non-negligible threat remains, particularly in the low and middle income destinations and the rapid increase in global connectivity of these at-risk regions indicates that international airports could see more imported cases (**Chan, 2014**).

Future work

We have focussed on reanalysing the zoonotic niche for EVD transmission and the characterisation of the populations at risk to improve the landscape in which future risk and impact of EVD outbreaks can be discussed. During the current emergency much of the work will concentrate on routes of secondary transmission in the human population—conceptually the red arrow of the H box in **Figure 1**. An important task is to stratify the risk of EVD spread both within and between countries and identify the most likely pathways of spread for characterisation and surveillance. Our next priority therefore is to investigate aspects of secondary human-to-human transmission by documenting the rate of geographic spread of EVD during the past and ongoing epidemics to help understand changes in these patterns in the historical record. Simulating these movements in a real landscape of population movement patterns, inferred from population movements assessed by mobile phones and other data (**Garcia et al., 2014**), as well as parametric movement models (**Simini et al., 2013**) is a logical next step, and can be used in future targeting of interventions and potential new treatments for both the current and future outbreaks (**Brady et al., 2014**; **Goodman, 2014**).

As previously discussed, whilst there is the risk of human travel during the latent phase of infection, and therefore potential for international spread, the high pathogenicity during infectiousness (immobilising infected persons) and the likely rapid and effective isolation measures implemented in regions with strong health care systems, limit the pandemic potential of EVD. Nevertheless, improvement of international containment plans and informed discussions of potential risks to airline carriers and populations of other regions will be supported by knowledge of local, regional and international population flows. Assessing these flows by air traffic volumes is an ongoing priority.

There are several other zoonotic viral haemorrhagic fevers (for example *Marburgvirus*, Lassa fever, hantaviral infections and arenaviruses) that are of similar public health and biosecurity concern (**Bannister, 2010**), due to their high virulence and mortality and their potential to cause outbreaks and spread internationally. Despite this their geographical distributions are poorly understood (**Hay et al., 2013a**). Many of the methods applied here can be adapted to these diseases and improve our geographical understanding of the risk presented by these pathogens.

We are in the midst of a public health emergency that will likely last for many more months (**Chan, 2014**) and which has brought EVD to global attention. We emphasise that the maps of zoonotic transmission presented here do not enable assessment of secondary transmission rates in human populations, but they do act as an evidenced-based indicator of locations with potential for future zoonotic transmission and thus outbreaks. Interestingly, early reports of another independent zoonotic outbreak in the DRC (**MSF, 2014**) are in predicted at-risk areas. An improved understanding of the spatial extent of the zoonotic niche can only help future efforts in biosurveillance.

Materials and methods

Methodological overview

A boosted regression tree (BRT) modelling framework was used to generate predictive risk maps of the zoonotic Ebola virus niche in Africa. This methodology combines regression trees, where trees are built according to optimal decision rules based on how binary decisions best accommodate a given dataset (**De'ath, 2007**; **Elith et al., 2008**), and boosting, which selects the tree that minimises the loss function. In doing so, a parameter space is defined which captures the greatest amount of variation present in the dataset. In order to train the model, four component datasets were compiled: (i) a comprehensive dataset of the reported locations of Ebola virus transmission from a zoonotic reservoir to a human; (ii) a dataset of the locations of Ebola virus infections in suspected reservoir and (non-human) susceptible host species (iii) a suite of ecologically relevant environmental covariates for Africa, including predicted distribution maps of suspected reservoir bat species and (iv) background (or pseudo-absence) records representing locations where zoonotic Ebola virus has not been reported. This study was limited to the African continent since no natural outbreaks of EVD have occurred outside the continent (**CDC, 2014**). Only *Reston ebolavirus* has a distribution reported outside of Africa, focussed in the Philippines, but has never been reported as pathogenic in humans; as a result this species was not included in the analysis.

Identifying index cases and reconstructing zoonotic transmission events in space and time

Tables detailing proven outbreaks of Ebola virus, initially sourced from the scientific literature (**Kuhn, 2008**) and from health reporting organisations (**CDC, 2014**), were used to coordinate searches of the formal scientific literature using Web of Science and PubMed for each specific outbreak. Relevant papers were abstracted and where possible outbreak-specific epidemiological surveys were sourced. The citations in these references were obtained in order to reconstruct the outbreak in detail and to identify the most probable index case. Index cases were defined as any human infection resulting from interaction with non-human sources of the disease. Some of these cases arose from presumed interactions with zoonotic reservoirs or hosts, such as primates and other mammals during hunting trips (**Boumandouki et al., 2005**; **Nkoghe et al., 2005**, **2011**; WHO, 2003) or butchering of bats (**Leroy et al., 2009**). Any cases arising from existing human infections are considered as secondary infections rather than index cases. Similar to methodology employed elsewhere (**Messina et al., 2014**), the site, or supposed site, of this zoonotic transmission event was geopositioned using Google Earth. For locations where precise geographic information (e.g., geographic coordinates of a hunting camp) was provided by the authors, these were used. Where precise geographic information could not be accurately geopositioned, a geographic area (termed a polygon) was defined covering the reported region. In several cases only the first reported patient could be identified, with the source of infection unknown. With these outbreaks the location of the first patient was geopositioned under the assumption that an initial zoonotic spillover event occurred in the vicinity of this location. In two outbreaks multiple independent zoonotic transmission events were identified (WHO, 2003; **Nkoghe et al., 2005**; **Pourrut et al., 2005**), and each unique event was geopositioned and included in the model as separate entities. ***Table 1*** catalogues the outbreaks used in this study.

Assembling a database of reported infections in animals

A literature search was conducted in Web of Science using the search term "Ebola" that returned 8635 citations. The abstracts were examined and for those that contained possible data on animal Ebola infection, the full text was obtained. The sampling site or location of the animal in the study was identified and geopositioned using Google Maps. These locations were recorded either as precise locations or as polygons, as with human index cases. Records for which local transmission of Ebola virus was deemed unlikely (e.g., seropositive primates tested in containment facilities several years after their capture) were excluded from the study. The non-human Ebola virus occurrence data collected are detailed in ***Table 2***, including the diagnostic methods used.

GenBank isolates

The open access sequence database GenBank (***NCBI, 2014***) was searched using MESH Umbrella search terms for Ebola virus, returning 181 results. These were then manually cross-referenced with the existing human and animal Ebola information, collected above, and 30 duplicates were removed. For the remaining isolates, original references and GenBank information fields were examined, but as there was insufficient information to establish precise location of isolation and/or whether the isolate represented an index case for any of these data sources, they were excluded from subsequent analyses.

Covariates assembled and used in the analyses

A suite of ecologically relevant gridded environmental covariates for Africa was compiled, each having a nominal resolution of 5 km × 5 km. The environmental covariates used in this analysis were: elevation (from the shuttle radar topography mission [***ORNL DAAC, 2000***]); the mean value, and a measure of spatial variation (range, described below) between 2000 and 2012 of Enhanced Vegetation Index (EVI), daytime Land Surface Temperature (LST) and night-time LST; and mean potential evapotranspiration from 1950–2000 (***Trabucco and Zomer, 2009***) (***Figure 5—figure supplement 1***).

The EVI and LST datasets were derived from satellite imagery collected by NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) remote sensing platform (***Tatem et al., 2004***). EVI is a metric designed to characterise vegetation density and vigour based on the ratio of absorbed photosynthetically active radiation to near infrared radiation (***Huete et al., 2002***). LST is a modelled product derived from emissivity as measured by the MODIS thermal sensor (***Wan and Li, 1997***), which is correlated, though not synonymous with air temperature, and effective for differentiating landscapes based on a combination of thermal energy and properties of surface types. The MODIS datasets utilized in this research (EVI was derived from the MCD43B4 product and the MOD11A2 LST product was used directly) were acquired as composite datasets created using imagery collected over multiple days, a procedure that results in products with 8-day temporal resolutions. Despite compositing, the EVI and LST datasets contained gaps due to persistent cloud cover found in forested equatorial regions, and these gaps were filled using a previously described approach (***Weiss et al., 2014a***). The EVI and LST datasets were then aggregated from their native 1 km × 1 km spatial resolution to a final 5 km × 5 km resolution, calculating both the mean and the range of the values of the subpixels making up each larger pixel. These spatial summaries therefore characterise both the mean temperature in each location as well as the degree of spatial heterogeneity within that pixel. This is of interest as humans and susceptible species are more likely to come into contact in transitional areas (e.g., boundary areas between areas of highly suitable susceptible species habitat and areas heavily utilised by humans). The final covariate production step consisted of summarising temporally across the 13-year data archive to produce synoptic datasets devoid of annual or seasonal anomalies (***Weiss et al., 2014a***).

Implicated bat reservoir distributions

Over recent years, significant research has been undertaken in investigating the role bats have to play in the transmission cycle of Ebola viruses (***Olival and Hayman, 2014***) and evidence of asymptomatic infection in fruit bats has been documented to varying extents (***Leroy et al., 2005***; ***Pourrut et al., 2007***, ***2009***; ***Hayman et al., 2010***; ***Hayman et al., 2012***). In order to incorporate this potential driver of Ebola virus transmission into the model we developed predicted distribution maps for three species of fruit bat implicated as primary reservoirs of the disease: *Hypsignathus monstrosus*, *Epomops franqueti* and *Myonycteris torquata*. The evidence was strongest for these three species having a reservoir role

as Ebola virus RNA (all nested within the *Zaire ebolavirus* phylogeny [**Leroy et al., 2005**]) has been detected in all three. Whilst a handful of other bat species have been found to be seropositive, no further viral isolations have been recorded (**Olival and Hayman, 2014**).

Whilst expert opinion range maps for these species exist (**Schipper et al., 2008**), there is some disagreement with independently-sourced occurrence data (all archived in the Global Biodiversity Information Facility). As a result, a predictive modelling approach was used to create a continuous surface of habitat suitability for these species which we then included as a predictor in the model. Occurrence data for all Megachiroptera in Africa was extracted from GBIF (**GBIF, 2014**) using the R packages dismo (**Hijmans et al., 2014**) and taxize (**Chamberlain et al., 2014**). To remove apparently erroneous records in the GBIF archive all occurrence records more than 100 km from the species known ranges, as determined by expert-opinion range maps (**Schipper et al., 2008**), were excluded, as were duplicate records and those located in the ocean. This resulted in a total dataset of 1341 unique occurrence records.

The occurrence database was then used to train separate boosted regression tree species distribution models (**Elith et al., 2008**) to predict the likely distribution of each of these suspected reservoir species. For each model, occurrence records for the target species (*H. monstrosus*, n = 67; *E. franqueti*, n = 120; and *M. torquata*, n = 52) were considered presence records and occurrence records of all other species were used as background records. This procedure is designed to account for the potentially biasing effect of spatial variation in recording of Megachiroptera occurrences (**Phillips et al., 2009**).

For each species we ran fifty submodels each trained to a randomly selected bootstrap of this dataset, subject to the constraint that each bootstrap contained a minimum of 10 occurrence and 10 background records. Each submodel was fitted using the gbm.step subroutine (**Elith et al., 2008**) in the dismo R package. In each submodel the background records were down weighted so that the weighted sum of presence records equalled the weighted sum of background records (**Barbet-Massin et al., 2012**) in order to maximise the discrimination capacity of the model. We generated a prediction map from each of these submodels and calculated both the mean prediction and 95% confidence interval around the prediction for each 5 km × 5 km pixel for each species.

Model accuracy was assessed by calculating the mean area under the curve (AUC) statistic for each submodel under a stringent 10-fold cross validation for each submodel and obtaining the mean and standard deviation across all 50 submodels. Under this procedure the dataset was split into ten subsets, each containing approximately equal numbers of presence and background points. The ability of a model trained on each subset to predict the distribution of the other 90% of records was assessed by AUC and the mean value taken. As so few presence records were used to train each fold model (i.e., around 5 presence records for *M. torquata* up to 12 for *E. franqueti*), this represents a very stringent test of the model's predictive capacity. Additionally, to prevent inflation of the accuracy statistics due to spatial sorting bias, these statistics were estimated using a pairwise distance sampling procedure (**Hijmans, 2012**). Consequently, the AUC statistics presented here are lower than would be returned by standard procedures but gives a more realistic quantification of the model's ability to extrapolate predictions to new regions (**Wenger and Olden, 2012**). We also generated marginal effect plots with associated uncertainty intervals and relative contribution statistics (how often each covariate was selected during the model fitting process) as quantification of the sensitivity of the model to the different covariates. These allow us to make inferences about the ecological relationship between each species and its environment as well as to identify where this relationship is most uncertain.

To generate a single surface describing the distribution of the bat reservoir species to be used as a covariate in the subsequent Ebola modelling, the three mean prediction distribution maps were merged by taking the average habitat suitability for each of the three bat species at each pixel.

Ebola distribution modelling

The Ebola virus occurrence dataset was supplemented with a background record dataset generated by randomly sampling 10,000 locations across Africa, biased towards more populous areas as a proxy for reporting bias (**Phillips et al., 2009**). We fitted 500 submodels to bootstraps of this dataset. To account for uncertainty in the geographic location of those occurrences reported as polygons, for each submodel one point was randomly selected from each of these occurrence polygons. This Monte

Carlo procedure enabled the model to efficiently integrate over the environmental uncertainty associated with imprecise geographic data. A bootstrap sample was then taken from each of these datasets and used to train the BRT model using the same procedure and weighting of background records as for the bat distribution models. Similarly, we generated a prediction map from each of these models and calculated both the mean prediction and corresponding 95% confidence intervals for each pixel and analysed prediction accuracy using the same stringent cross validation and sensitivity analysis procedure as for the bat distribution models (detailed above).

The predicted distribution map produced by this approach represents the environmental suitability of each pixel for zoonotic Ebola virus transmission. This may be interpreted as a relative probability of presence in the sense that more suitable pixels are more likely to contain zoonotic transmission than less suitable pixels, though not an absolute probability that transmission occurs in a given pixel. Similarly, the presence of zoonotic transmission increases the risk of transmission to a human, though this is also contingent on how humans interact with these zoonotic pools, through hunting or other activities.

Population living in areas of environmental suitability for zoonotic transmission

In order to identify areas which are likely to be at risk of transmission of *Ebolavirus* from zoonotic reservoir hosts to humans, the continuous map of the predicted environmental suitability for zoonotic transmission (shown in ***Figure 5***) was converted into a binary map classifying pixels as either at risk or not at risk. A pixel was assumed to be at risk if its predicted environmental suitability for zoonotic Ebola virus transmission was greater than 0.673, the lowest suitability value predicted at the locations of known transmission to humans (point records of human index cases). Countries containing at least one at-risk pixel are shown in ***Figure 5B***—those countries that previously report an index case were defined as Set 1; countries with at least one at-risk pixel with no previous index cases of EVD were categorised as Set 2. The number of people living in at-risk areas in each of these countries was calculated by summing the estimated population of at-risk pixels using population density maps from the AfriPop project (***Linard et al., 2012***; ***WorldPop, 2014***) and the proportion of those living in urban, periurban and rural areas was evaluated using the Global Rural Urban Mapping Project (***CIESIN/IFPRI/WB/CIAT, 2007***).

The R code used for all of the analysis has been made available on an open source basis (https://github.com/SEEG-Oxford/ebola_zoonotic).

National level demographic and mobility data

For three separate years (1976, 2000 and 2014), total national populations were retrieved and the proportion of rural to urban populations noted from World Bank statistics (***World Bank, 2014***). To describe global air travel patterns from Set 1 and Set 2 countries, flight schedules data from the Official Airline Guide, reflecting an estimated 95% of all commercial flights worldwide, were analysed between 2000 and 2013 to calculate the annual volume of seats on direct flights that depart from each predicted country and which have an international destination. Complementing these seat capacity data, worldwide data on anonymised, individual passenger flight itineraries from the International Air Transport Association (2012) (***IATA, 2014***) were analysed between 2005 and 2012 to calculate the annual volume of international passenger departures out of each Set 1 and Set 2 country. The IATA dataset represents an estimated 93% of the world's commercial air traffic at the passenger level and includes points of departure and arrival and final destination information for travellers as well as their connecting flights.

Acknowledgements

We thank Katherine Battle, Maria Devine and Kirsten Duda for proof-reading and Jane Messina for creating ***Figure 1***. We also thank Andrew Rambaut for his comments on the final draft.

Additional information

Competing interests

SIH: Reviewing editor, *eLife*. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
University Of Oxford	Sir Richard Southwood Graduate Scholarship	David M Pigott
Bill and Melinda Gates Foundation	OPP1053338	Nick Golding
Medical Research Council	K00669X	Peter W Gething
Biotechnology and Biological Sciences Research Council	Studentship	Oliver J Brady
German Academic Exchange Service	Graduate Scholarship	Moritz UG Kraemer
U.S. National Library of Medicine	R01LM010812	John S Brownstein, Sumiko R Mekaru
7th Framework Programme for Research and Technological Development (EU FP7)	602525	Peter W Horby
Canadian Institutes of Health Research		Isaac I Bogoch, Kamran Khan
Wellcome Trust	095066	Adrian Mylne, Simon I Hay
RAPIDD program of the Science & Technology Directorate		Andrew J Tatem, Simon I Hay
Fogarty International Center		Andrew J Tatem, Simon I Hay
Bill and Melinda Gates Foundation	OPP1106023	Zhi Huang, Andrew J Henry, Catherine L Moyes
Bill and Melinda Gates Foundation	OPP1068048	Daniel J Weiss, Samir Bhatt
Bill and Melinda Gates Foundation	OPP1093011	Catherine L Moyes, John S Brownstein, Sumiko R Mekaru, Simon I Hay
National Institute of Allergy and Infectious Diseases	U19AI089674	Andrew J Tatem
Bill and Melinda Gates Foundation	OPP1106427	Andrew J Tatem
Bill and Melinda Gates Foundation	OPP1032350	Andrew J Tatem

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

DMP, Advised and assembled the outbreak and animal infection data, Wrote the article; NG, Extracted the bat data, Conducted all of the analysis, Drafting or revising the article; AM, Assembled and geo-positioned the outbreak and animal infection data, Drafting or revising the article; ZH, Geo-positioned the outbreak and animal infection data, Provided all maps, Drafting or revising the article; AJH, Analysed international flight data, Helped assemble maps, Drafting or revising the article; IIB, Assisted with international transportation analysis, Edited the final draft of the article; SRM, Assisted in geopositioning, Drafting or revising the article; AJT, Provided information on urban change and migration data, Drafting or revising the article; KK, Provided data, Conducted all analyses on international air traffic patterns, Drafting or revising the article; DJW, Assembled the covariate layers, Drafting or revising the article; SB, Extracted Ebola information from GenBank, Drafting or revising the article; OJB, Screened GenBank data, Provided ***Figure 2A***, Drafting or revising the article; MUGK, DLS, CLM, Analysis and interpretation of data, Drafting or revising the article; PWG, Assembled the covariate layers, Drafting or revising the article; PWH, JSB, Advised on international public health context, Edited the final draft of the article; SIH, Conceived the work and analysis, Wrote content and edited the article at all stages of development, Acts as guarantor of the paper

Author ORCIDs
David M Pigott, http://orcid.org/0000-0002-6731-4034
Nick Golding, http://orcid.org/0000-0001-8916-5570
David L Smith, http://orcid.org/0000-0003-4367-3849
Simon I Hay, http://orcid.org/0000-0002-0611-7272

References

Amblard J, Obiang P, Edzang S, Prehaud C, Bouloy M, Guenno BL. 1997. Identification of the Ebola virus in Gabon in 1994. *Lancet* **349**:181–182. doi: 10.1016/S0140-6736(05)60984-1.

Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, Soropogui B, Sow MS, Keïta S, De Clerck H, Tiffany A, Dominguez G, Loua M, Traoré A, Kolié M, Malano ER, Heleze E, Bocquin A, Mély S, Raoul H, Caro V, Cadar D, Gabriel M, Pahlmann M, Tappe D, Schmidt-Chanasit J, Impouma B, Diallo AK, Formenty P, Van Herp M, Günther S. 2014. Emergence of Zaire Ebola virus disease in Guinea - preliminary report. *The New England Journal of Medicine* doi: 10.1056/NEJMoa1404505.

Bannister B. 2010. Viral haemorrhagic fevers imported into non-endemic countries: risk assessment and management. *British Medical Bulletin* **95**:193–225. doi: 10.1093/Bmb/Ldq022.

Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution* **3**:327–338. doi: 10.1111/j.2041-210X.2011.00172.x.

Baron RC, McCormick JB, Zubeir OA. 1983. Ebola virus disease in southern Sudan: hospital dissemination and intrafamilial spread. *Bulletin of the World Health Organization* **61**:997–1003.

Barrette RW, Metwally SA, Rowland JM, Xu LZ, Zaki SR, Nichol ST, Rollin PE, Towner JS, Shieh WJ, Batten B, Sealy TK, Carrillo C, Moran KE, Bracht AJ, Mayr GA, Sirios-Cruz M, Catbagan DP, Lautner EA, Ksiazek TG, White WR, McIntosh MT. 2009. Discovery of swine as a host for the Reston ebolavirus. *Science* **325**:204–206. doi: 10.1126/science.1172705.

Bausch DG, Schwarz L. 2014. Outbreak of Ebola virus disease in Guinea: where ecology meets economy. *PLOS Neglected Tropical Diseases* **8**:e3056. doi: 10.1371/journal.pntd.0003056.

Beer B, Kurth R, Bukreyev A. 1999. Characteristics of Filoviridae: Marburg and Ebola viruses. *Die Naturwissenschaften* **86**:8–17. doi: 10.1007/s001140050562.

Bermejo M, Rodríguez-Teijeiro JD, Illera G, Barroso A, Vilà C, Walsh PD. 2006. Ebola outbreak killed 5000 gorillas. *Science* **314**:1564. doi: 10.1126/science.1133105.

Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, Drake JM, Brownstein JS, Hoen AG, Sankoh O, Myers MF, George DB, Jaenisch T, Wint GR, Simmons CP, Scott TW, Farrar JJ, Hay SI. 2013. The global distribution and burden of dengue. *Nature* **496**:504–507. doi: 10.1038/Nature12060.

Boumandouki P, Formenty P, Epelboin A, Campbell P, Atsangandoko C, Allarangar Y, Leroy EM, Kone ML, Molamou A, Dinga-Longa O, Salemo A, Kounkou RY, Mombouli V, Ibara JR, Gaturuku P, Nkunku S, Lucht A, Feldmann H. 2005. [Clinical management of patients and deceased during the Ebola outbreak from October to December 2003 in Republic of Congo]. *Bulletin de la Société de Pathologie Exotique* **98**:218–223.

Brady OJ, Hay SI, Horby P. 2014. Estimating vaccine and drug requirements for Ebola outbreaks. *Nature* **512**:7514. doi: 10.1038/512233a.

Brashares JS, Golden CD, Weinbaum KZ, Barrett CB, Okello GV. 2011. Economic and geographic drivers of wildlife consumption in rural Africa. *Proceedings of the National Academy of Sciences of USA* **108**:13931–13936. doi: 10.1073/pnas.1011526108.

Briand S, Bertherat E, Cox P, Formenty P, Kieny MP, Myhre JK, Roth C, Shindo N, Dye C. 2014. The international Ebola emergency. *The New England Journal of Medicine* doi: 10.1056/NEJMp1409858.

Brockmann D, Helbing D. 2013. The hidden geometry of complex, network-driven contagion phenomena. *Science* **342**:1337–1342. doi: 10.1126/science.1245200.

Caillaud D, Levréro F, Cristescu R, Gatti S, Dewas M, Douadi M, Gautier-Hion A, Raymond M, Ménard N. 2006. Gorilla susceptibility to Ebola virus: the cost of sociality. *Current Biology* **16**:R489–R491. doi: 10.1016/j.cub.2006.06.017.

CDC. 2014. Chronology of Ebola hemorrhagic fever outbreaks. Accessed: August 2014. Available: http://www.cdc.gov/vhf/ebola/resources/outbreak-table.html.

Chamberlain S, Szocs E, Boettiger C, Ram K, Bartomeus I, Baumgartner J. 2014. taxize: taxonomic information from around the web. *R package*. Accessed: August 2014. Available: https://github.com/ropensci/taxize.

Chan M. 2014. Ebola virus disease in West Africa - no early end to the outbreak. *The New England Journal of Medicine* online. doi: 10.156/NEJMp1409859.

Chowell G, Hengartner NW, Castillo-Chavez C, Fenimore PW, Hyman JM. 2004. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *Journal of Theoretical Biology* **229**:119–126. doi: 10.1016/j.jtbi.2004.03.006.

CIESIN/IFPRI/WB/CIAT. 2007. Global Rural Urban Mapping Project (GRUMP): gridded population of the world, version 3. Accessed: December 2013. Available: http://sedac.ciesin.columbia.edu/gpw.

Cohen B. 2004. Urban growth in developing countries: a review of current trends and a caution regarding existing forecasts. *World Development* **32**:23–51. doi: 10.1016/j.worlddev.2003.04.008.

Conrad JL, Isaacson M, Smith EB, Wulff H, Crees M, Geldenhuys P, Johnston J. 1978. Epidemiologic investigation of Marburg virus disease, Southern Africa, 1975. *The American Journal of Tropical Medicine and Hygiene* **27**:1210–1215.

Daszak P. 2000. Emerging infectious diseases of wildlife - threats to biodiversity and human health. *Science* **287**:443–449. doi: 10.1126/science.287.5452.443.
De'ath G. 2007. Boosted trees for ecological modeling and prediction. *Ecology* **88**:243–251. doi: 10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2.
Dudas G, Rambaut A. 2014. Phylogenetic analysis of Guinea 2014 EBOV ebolavirus outbreak. *PLOS Currents* **6**:ecurrents.outbreaks.84eefe85ce43ec89dc80bf0670f0677b0678b0417d. doi: 10.1371/currents.outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d.
ECDC. 2014. *Outbreak of Ebola virus disease in West Africa*. Stockholm: ECDC. p. 17.
Elith J, Leathwick JR. 2009. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* **40**:677–697. doi: 10.1146/annurev.ecolsys.110308.120159.
Elith J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. *The Journal of Animal Ecology* **77**:802–813. doi: 10.1111/j.1365-2656.2008.01390.x.
Fauci AS. 2014. Ebola - underscoring the global disparities in health care resources. *The New England Journal of Medicine* doi: 10.1056/NEJMp1409494.
Feldmann H, Geisbert TW. 2011. Ebola haemorrhagic fever. *Lancet* **377**:849–862. doi: 10.1016/S0140-6736(10)60667-8.
Formenty P, Boesch C, Wyers M, Steiner C, Donati F, Dind F, Walker F, Le Guenno B. 1999. Ebola virus outbreak among wild chimpanzees living in a rain forest of Cote d'Ivoire. *The Journal of Infectious Diseases* **179**(suppl 1): S120–S126. doi: 10.1086/514296.
Francesconi P, Yoti Z, Declich S, Onek PA, Fabiani M, Olango J, Andraghetti R, Rollin PE, Opira C, Greco D, Salmaso S. 2003. Ebola hemorrhagic fever transmission and risk factors of contacts, Uganda. *Emerging Infectious Diseases* **9**:1430–1437. doi: 10.3201/eid0911.030339.
Franklin J. 2009. *Mapping species distributions*. Cambridge: Cambridge University Press. p. 320.
Garcia AJ, Pindolia DK, Lopiano KK, Tatem AJ. 2014. Modeling internal migration flows in sub-Saharan Africa using census microdata. *Migration Studies* in press. doi: 10.1093/migration/mnu036.
GBIF. 2014. Global biodiversity information facility. Accessed: August 2014. Available: http://www.gbif.org/.
Gear JS, Cassel GA, Gear AJ, Trappler B, Clausen L, Meyers AM, Kew MC, Bothwell TH, Sher R, Miller GB, Schneider J, Koornhof HJ, Gomperts ED, Isaäcson M, Gear JH. 1975. Outbreake of Marburg virus disease in Johannesburg. *British Medical Journal* **4**:489–493. doi: 10.1136/bmj.4.5995.489.
Georges AJ, Leroy EM, Renaut AA, Benissan CT, Nabias RJ, Ngoc MT, Obiang PI, Lepage JP, Bertherat EJ, Bénoni DD, Wickings EJ, Amblard JP, Lansoud-Soukate JM, Milleliri JM, Baize S, Georges-Courbot MC. 1999. Ebola hemorrhagic fever outbreaks in Gabon, 1994-1997: epidemiologic and health control issues. *The Journal of Infectious Diseases* **179**:S65–S75. doi: 10.1086/514290.
Georges-Courbot MC, Sanchez A, Lu CY, Baize S, Leroy E, Lansout-Soukate J, Tévi-Bénissan C, Georges AJ, Trappier SG, Zaki SR, Swanepoel R, Leman PA, Rollin PE, Peters CJ, Nichol ST, Ksiazek TG. 1997. Isolation and phylogenetic characterization of Ebola viruses causing different outbreaks in Gabon. *Emerging Infectious Diseases* **3**:59–62. doi: 10.3201/eid0301.970107.
Gething PW, Patil AP, Smith DL, Guerra CA, Elyazar IR, Johnston GL, Tatem AJ, Hay SI. 2011. A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malaria Journal* **10**:378. doi: 10.1186/1475-2875-10-378.
Gilbert M, Golding N, Zhou H, Wint GR, Robinson TP, Tatem AJ, Lai S, Zhou S, Jiang H, Guo D, Huang Z, Messina JP, Xiao X, Linard C, Van Boeckel TP, Martin V, Bhatt S, Gething PW, Farrar JJ, Hay SI, Yu H. 2014. Predicting the risk of avian influenza A H7N9 infection in live-poultry markets across Asia. *Nature Communications* **5**:4116. doi: 10.1038/ncomms5116.
Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang PP, Nekoui M, Colubri A, Coomber MR, Fonnie M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, Saffa S, Sellu J, Jalloh AA, Kovoma A, Koninga J, Mustapha I, Kargbo K, Foday M, Yillah M, Kanneh F, Robert W, Massally JLB, Chapman SB, Bochicchio J, Murphy C, Nusbaum C, Young S, Birren BW, Grant DS, Scheiffelin JS, Lander ES, Happi C, Gevao SM, Gnirke A, Rambaut A, Garry RF, Khan SH, Sabeti PC. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**:1369–1372. doi: 10.1126/science.1259657.
Gonzalez MC, Hidalgo CA, Barabasi AL. 2008. Understanding individual human mobility patterns. *Nature* **453**:779–782. doi: 10.1038/Nature06958.
Goodman JL. 2014. Studying "secret serums" - toward safe, effective Ebola treatments. *The New England Journal of Medicine*.
Gostin LO, Lucey D, Phelan A. 2014. The ebola epidemic: a global health emergency. *JAMA* E1–E2. doi: 10.1001/jama.2014.11176.
Grard G, Biek R, Tamfum JJM, Fair J, Wolfe N, Formenty P, Paweska J, Leroy E. 2011. Emergence of divergent Zaire Ebola virus strains in Democratic Republic of the Congo in 2007 and 2008. *The Journal of Infectious Diseases* **204**:S776–S784. doi: 10.1093/infdis/jir364.
Groseth A, Feldmann H, Strong JE. 2007. The ecology of Ebola virus. *Trends in Microbiology* **15**:408–416. doi: 10.1016/j.tim.2007.08.001.
Hay SI, Battle KE, Pigott DM, Smith DL, Moyes CL, Bhatt S, Brownstein JS, Collier N, Myers MF, George DB, Gething PW. 2013a. Global mapping of infectious disease. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **368**:20120250. doi: 10.1098/rstb.2012.0250.

Hay SI, George DB, Moyes CL, Brownstein JS. 2013b. Big data opportunities for global infectious disease surveillance. *PLOS Medicine* **10**:e1001413. doi: 10.1371/journal.pmed.1001413.
Hay SI, Tatem AJ, Graham AJ, Goetz SJ, Rogers DJ. 2006. Global environmental data for mapping infectious disease distribution. *Advances in Parasitology* **62**:37–77. doi: 10.1016/S0065-308x(05)62002-7.
Hayman DT, Emmerich P, Yu M, Wang LF, Suu-Ire R, Fooks AR, Cunningham AA, Wood JL. 2010. Long-term survival of an urban fruit bat seropositive for Ebola and Lagos Bat viruses. *PLOS ONE* **5**:e11978. doi: 10.1371/journal.pone.0011978.
Hayman DT, Yu M, Crameri G, Wang LF, Suu-Ire R, Wood JL, Cunningham AA. 2012. Ebola virus antibodies in fruit bats, Ghana, West Africa. *Emerging Infectious Diseases* **18**:1207–1209. doi: 10.3201/eid1807.111654.
Hewlett BS, Epelboin A, Hewlett BL, Formenty P. 2005. Medical anthropology and Ebola in Congo: cultural models and humanistic care. *Bulletin de la Société de Pathologie Exotique* **98**:230–236.
Heymann DL, Weisfeld JS, Webb PA, Johnson KM, Cairns T, Berquist H. 1980. Ebola hemorrhagic fever: Tandala, Zaire, 1977-1978. *The Journal of Infectious Diseases* **142**:372–376. doi: 10.1093/infdis/142.3.372.
Hijmans RJ. 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology* **93**:679–688. doi: 10.1890/11-0826.1.
Hijmans RJ, Phillips S, Leathwick J, Elith J. 2014. dismo. R package. Accessed: August 2014. Available: http://cran.r-project.org/web/packages/dismo/dismo.pdf.
Huang ZJ, Tatem AJ. 2013. Global malaria connectivity through air travel. *Malaria Journal* **12**:269. doi: 10.1186/1475-2875-12-269.
Huete A, Didan K, Miura T, Rodriguez EP, Gao X, Ferreira LG. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment* **83**:195–213. doi: 10.1016/S0034-4257(02)00096-2.
Hufnagel L, Brockmann D, Geisel T. 2004. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences of USA* **101**:15124–15129. doi: 10.1073/pnas.0308344101.
IATA. 2014. International Air Transport Association. Accessed: August 2014. Available: http://www.iata.org/Pages/default.aspx.
International Commission. 1978. Ebola haemorrhagic fever in Zaire, 1976. *Bulletin of the World Health Organization* **56**:271–293.
Jahrling PB, Geisbert TW, Dalgard DW, Johnson ED, Ksiazek TG, Hall WC, Peters CJ. 1990. Preliminary report: isolation of Ebola virus from monkeys imported to USA. *Lancet* **335**:502–505. doi: 10.1016/0140-6736(90)90737-P.
Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P. 2008. Global trends in emerging infectious diseases. *Nature* **451**:990–994. doi: 10.1038/Nature06536.
Kamins AO, Restif O, Ntiamoa-Baidu Y, Suu-Ire R, Hayman DT, Cunningham AA, Wood JL, Rowcliffe JM. 2011. Uncovering the fruit bat bushmeat commodity chain and the true extent of fruit bat hunting in Ghana, West Africa. *Biological Conservation* **144**:3000–3008. doi: 10.1016/j.biocon.2011.09.003.
Karesh WB, Dobson A, Lloyd-Smith JO, Lubroth J, Dixon MA, Bennett M, Aldrich S, Harrington T, Formenty P, Loh EH, Machalaba CC, Thomas MJ, Heymann DL. 2012. Ecology of zoonoses: natural and unnatural histories. *Lancet* **380**:1936–1945. doi: 10.1016/S0140-6736(12)61678-X.
Khan AS, Tshioko FK, Heymann DL, Le Guenno B, Nabeth P, Kerstiëns B, Fleerackers Y, Kilmarx PH, Rodier GR, Nkuku O, Rollin PE, Sanchez A, Zaki SR, Swanepoel R, Tomori O, Nichol ST, Peters CJ, Muyembe-Tamfum JJ, Ksiazek TG. 1999. The reemergence of Ebola hemorrhagic fever, Democratic Republic of the Congo, 1995. *The Journal of Infectious Diseases* **179**:S76–S86. doi: 10.1086/514306.
King AMQ, Adams MJ, Carstens EB, Lefkowitz E. 2012. *Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses*. London: Elsevier. p. 1338.
Kuhn JH. 2008. Filoviruses: a compendium of 40 years of epidemiological, clinical, and laboratory studies. Vienna: Springer-Verlag. p. 413.
Kuhn JH, Becker S, Ebihara H, Geisbert TW, Johnson KM, Kawaoka Y, Lipkin WI, Negredo AI, Netesov SV, Nichol ST, Palacios G, Peters CJ, Tenorio A, Volchkov VE, Jahrling PB. 2010. Proposal for a revised taxonomy of the family *Filoviridae*: classification, names of taxa and viruses, and virus abbreviations. *Archives of Virology* **155**:2083–2103. doi: 10.1007/s00705-010-0814-x.
Lahm SA, Kombila M, Swanepoel R, Barnes RF. 2007. Morbidity and mortality of wild animals in relation to outbreaks of Ebola haemorrhagic fever in Gabon, 1994-2003. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **101**:64–78. doi: 10.1016/j.trstmh.2006.07.002.
Lamunu M, Lutwama JJ, Kamugisha J, Opio A, Nambooze J, Ndayimirije N, Okware S. 2004. Containing a haemorrhagic fever epidemic: the Ebola experience in Uganda (October 2000-January 2001). *International Journal of Infectious Diseases* **8**:27–37. doi: 10.1016/j.ijid.2003.04.001.
Larkin M. 2003. Ebola outbreak in the news. *The Lancet Infectious Diseases* **3**:255. doi: 10.1016/S1473-3099(03)00584-X.
Le Guenno B, Formenty P, Wyers M, Gounon P, Walker F, Boesch C. 1995. Isolation and partial characterization of a new strain of Ebola virus. *Lancet* **345**:1271–1274. doi: 10.1016/S0140-6736(95)90925-7.
Legrand J, Grais RF, Boelle PY, Valleron AJ, Flahault A. 2007. Understanding the dynamics of Ebola epidemics. *Epidemiology and Infection* **135**:610–621. doi: 10.1017/S0950268806007217.
Leroy EM, Baize S, Volchkov VE, Fisher-Hoch SP, Georges-Courbot MC, Lansoud-Soukate J, Capron M, Debré P, McCormick JB, Georges AJ. 2000. Human asymptomatic Ebola infection and strong inflammatory response. *Lancet* **355**:2210–2215. doi: 10.1016/S0140-6736(00)02405-3.

Leroy EM, Epelboin A, Mondonge V, Pourrut X, Gonzalez JP, Muyembe-Tamfum JJ, Formenty P. 2009. Human Ebola outbreak resulting from direct exposure to fruit bats in Luebo, Democratic Republic of Congo, 2007. *Vector Borne and Zoonotic Diseases* **9**:723–728. doi: 10.1089/vbz.2008.0167.

Leroy EM, Kumulungui B, Pourrut X, Rouquet P, Hassanin A, Yaba P, Délicat A, Paweska JT, Gonzalez JP, Swanepoel R. 2005. Fruit bats as reservoirs of Ebola virus. *Nature* **438**:575–576. doi: 10.1038/438575a.

Leroy EM, Souquière S, Rouquet P, Drevet D. 2002. Re-emergence of ebola haemorrhagic fever in Gabon. *Lancet* **359**:712–712. doi: 10.1016/S0140-6736(02)07796-6.

Linard C, Gilbert M, Snow RW, Noor AM, Tatem AJ. 2012. Population distribution, settlement patterns and accessibility across Africa in 2010. *PLOS ONE* **7**:e31743. doi: 10.1371/journal.pone.0031743.

Linard C, Tatem AJ, Gilbert M. 2013. Modelling spatial patterns of urban growth in Africa. *Applied Geography* **44**:23–32. doi: 10.1016/j.apgeog.2013.07.009.

MacNeil A, Farnon EC, Wamala J, Okware S, Cannon DL, Reed Z, Towner JS, Tappero JW, Lutwama J, Downing R, Nichol ST, Ksiazek TG, Rollin PE. 2010. Proportion of deaths and clinical features in Bundibugyo Ebola virus infection, Uganda. *Emerging Infectious Diseases* **16**:1969–1972. doi: 10.3201/eid1612.100627.

Messina JP, Brady OJ, Pigott DM, Brownstein JS, Hoen AG, Hay SI. 2014. A global compendium of human dengue virus occurrence. *Scientific Data* **1**:140004. doi: 10.1038/sdata.2014.4.

Mfunda IM, Røskaft E. 2010. Bushmeat hunting in Serengeti, Tanzania: an important economic activity to local people. *International Journal of Biodiversity and Conservation* **2**:263–272.

Milleliri JM, Tevi-Benissan C, Baize S, Leroy E, Georges-Courbot MC. 2004. [Epidemics of Ebola haemorrhagic fever in Gabon (1994-2002). Epidemiologic aspects and considerations on control measures]. *Bulletin de la Société de Pathologie Exotique* **97**:199–205.

Miranda ME, White ME, Dayrit MM, Hayes CG, Ksiazek TG, Burans JP. 1991. Seroepidemiological study of filovirus related to Ebola in the Philippines. *Lancet* **337**:425–426. doi: 10.1016/0140-6736(91)91199-5.

MSF. 2014. Ebola epidemic confirmed in Democratic Republic of Congo: MSF sends specialists and material to the epicentre. Accessed: August 2014. Available: http://www.msf.org/article/ebola-epidemic-confirmed-democratic-republic-congo-msf-sends-specialists-and-material.

Murray CJ, Ortblad KF, Guinovart C, Lim SS, Wolock TM, Roberts DA, Dansereau EA, Graetz N, Barber RM, Brown JC, Wang H, Duber HC, Naghavi M, Dicker D, Dandona L, Salomon JA, Heuton KR, Foreman K, Phillips DE, Fleming TD, Flaxman AD, Phillips BK, Johnson EK, Coggeshall MS, Abd-Allah F, Abera SF, Abraham JP, Abubakar I, Abu-Raddad LJ, Abu-Rmeileh NM, Achoki T, Adeyemo AO, Adou AK, Adsuar JC, Agardh EE, Akena D, Al Kahbouri MJ, Alasfoor D, Albittar MI, Alcalá-Cerra G, Alegretti MA, Alemu ZA, Alfonso-Cristancho R, Alhabib S, Ali R, Alla F, Allen PJ, Alsharif U, Alvarez E, Alvis-Guzman N, Amankwaa AA, Amare AT, Amini H, Ammar W, Anderson BO, Antonio CA, Anwari P, Arnlöv J, Arsenijevic VS, Artaman A, Asghar RJ, Assadi R, Atkins LS, Badawi A, Balakrishnan K, Banerjee A, Basu S, Beardsley J, Bekele T, Bell ML, Bernabe E, Beyene TJ, Bhala N, Bhalla A, Bhutta ZA, Abdulhak AB, Binagwaho A, Blore JD, Bose D, Brainin M, Breitborde N, Castañeda-Orjuela CA, Catalá-López F, Chadha VK, Chang JC, Chiang PP, Chuang TW, Colomar M, Cooper LT, Cooper C, Courville KJ, Cowie BC, Criqui MH, Dandona R, Dayama A, De Leo D, Degenhardt L, Del Pozo-Cruz B, Deribe K, Des Jarlais DC, Dessalegn M, Dharmaratne SD, Dilmen U, Ding EL, Driscoll TR, Durrani AM, Ellenbogen RG, Ermakov SP, Esteghamati A, Faraon EJ, Farzadfar F, Fereshtehnejad SM, Fijabi DO, Forouzanfar MH, Fra Paleo U, Gaffikin L, Gamkrelidze A, Gankpé FG, Geleijnse JM, Gessner BD, Gibney KB, Ginawi IA, Glaser EL, Gona P, Goto A, Gouda HN, Gugnani HC, Gupta R, Gupta R, Hafezi-Nejad N, Hamadeh RR, Hammami M, Hankey GJ, Harb HL, Haro JM, Havmoeller R, Hay SI, Hedayati MT, Pi IB, Hoek HW, Hornberger JC, Hosgood HD, Hotez PJ, Hoy DG, Huang JJ, Iburg KM, Idrisov BT, Innos K, Jacobsen KH, Jeemon P, Jensen PN, Jha V, Jiang G, Jonas JB, Juel K, Kan H, Kankindi I, Karam NE, Karch A, Karema CK, Kaul A, Kawakami N, Kazi DS, Kemp AH, Kengne AP, Keren A, Kereselidze M, Khader YS, Khalifa SE, Khan EA, Khang YH, Khonelidze I, Kinfu Y, Kinge JM, Knibbs L, Kokubo Y, Kosen S, Defo BK, Kulkarni VS, Kulkarni C, Kumar K, Kumar RB, Kumar GA, Kwan GF, Lai T, Balaji AL, Lam H, Lan Q, Lansingh VC, Larson HJ, Larsson A, Lee JT, Leigh J, Leinsalu M, Leung R, Li Y, Li Y, De Lima GM, Lin HH, Lipshultz SE, Liu S, Liu Y, Lloyd BK, Lotufo PA, Machado VM, Maclachlan JH, Magis-Rodriguez C, Majdan M, Mapoma CC, Marcenes W, Marzan MB, Masci JR, Mashal MT, Mason-Jones AJ, Mayosi BM, Mazorodze TT, Mckay AC, Meaney PA, Mehndiratta MM, Mejia-Rodriguez F, Melaku YA, Memish ZA, Mendoza W, Miller TR, Mills EJ, Mohammad KA, Mokdad AH, Mola GL, Monasta L, Montico M, Moore AR, Mori R, Moturi WN, Mukaigawara M, Murthy KS, Naheed A, Naidoo KS, Naldi L, Nangia V, Narayan KM, Nash D, Nejjari C, Nelson RG, Neupane SP, Newton CR, Ng M, Nisar MI, Nolte S, Norheim OF, Nowaseb V, Nyakarahuka L, Oh IH, Ohkubo T, Olusanya BO, Omer SB, Opio JN, Orisakwe OE, Pandian JD, Papachristou C, Caicedo AJ, Patten SB, Paul VK, Pavlin BI, Pearce N, Pereira DM, Pervaiz A, Pesudovs K, Petzold M, Pourmalek F, Qato D, Quezada AD, Quistberg DA, Rafay A, Rahimi K, Rahimi-Movaghar V, Rahman SU, Raju M, Rana SM, Razavi H, Reilly RQ, Remuzzi G, Richardus JH, Ronfani L, Roy N, Sabin N, Saeedi MY, Sahraian MA, Samonte GM, Sawhney M, Schneider IJ, Schwebel DC, Seedat S, Sepanlou SG, Servan-Mori EE, Sheikhbahaei S, Shibuya K, Shin HH, Shiue I, Shivakoti R, Sigfusdottir ID, Silberberg DH, Silva AP, Simard EP, Singh JA, Skirbekk V, Sliwa K, Soneji S, Soshnikov SS, Sreeramareddy CT, Stathopoulou VK, Stroumpoulis K, Swaminathan S, Sykes BL, Tabb KM, Talongwa RT, Tenkorang EY, Terkawi AS, Thomson AJ, Thorne-Lyman AL, Towbin JA, Traebert J, Tran BX, Dimbuene ZT, Tsilimbaris M, Uchendu US, Ukwaja KN, Vallely AJ, Vasankari TJ, Venketasubramanian N, Violante FS, Vlassov VV, Waller S, Wallin MT, Wang L, Wang SX, Wang Y, Weichenthal S, Weiderpass E, Weintraub RG, Westerman R, White RA, Wilkinson JD, Williams TN, Woldeyohannes SM, Wong JQ, Xu G, Yang YC, Yano Y, Yip P, Yonemoto N, Yoon SJ, Younis M, Yu C, Jin KY, El Sayed Zaki M, Zhao Y, Zheng Y, Zhou M, Zhu J, Zou XN, Lopez AD, Vos T. 2014. Global, regional, and national incidence and mortality for HIV, tuberculosis, and malaria during 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*. doi: 10.1016/S0140-6736(14)60844-8.

Murray CJL, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, Ezzati M, Shibuya K, Salomon JA, Abdalla S, Aboyans V, Abraham J, Ackerman I, Aggarwal R, Ahn SY, Ali MK, Alvarado M, Anderson HR, Anderson LM, Andrews KG, Atkinson C, Baddour LM, Bahalim AN, Barker-Collo S, Barrero LH, Bartels DH, Basáñez MG, Baxter A, Bell ML, Benjamin EJ, Bennett D, Bernabé E, Bhalla K, Bhandari B, Bikbov B, Bin Abdulhak A, Birbeck G, Black JA, Blencowe H, Blore JD, Blyth F, Bolliger I, Bonaventure A, Boufous S, Bourne R, Boussinesq M, Braithwaite T, Brayne C, Bridgett L, Brooker S, Brooks P, Brugha TS, Bryan-Hancock C, Bucello C, Buchbinder R, Buckle G, Budke CM, Burch M, Burney P, Burstein R, Calabria B, Campbell B, Canter CE, Carabin H, Carapetis J, Carmona L, Cella C, Charlson F, Chen H, Cheng AT, Chou D, Chugh SS, Coffeng LE, Colan SD, Colquhoun S, Colson KE, Condon J, Connor MD, Cooper LT, Corriere M, Cortinovis M, de Vaccaro KC, Couser W, Cowie BC, Criqui MH, Cross M, Dabhadkar KC, Dahiya M, Dahodwala N, Damsere-Derry J, Danaei G, Davis A, De Leo D, Degenhardt L, Dellavalle R, Delossantos A, Denenberg J, Derrett S, Des Jarlais DC, Dharmaratne SD, Dherani M, Diaz-Torne C, Dolk H, Dorsey ER, Driscoll T, Duber H, Ebel B, Edmond K, Elbaz A, Ali SE, Erskine H, Erwin PJ, Espindola P, Ewoigbokhan SE, Farzadfar F, Feigin V, Felson DT, Ferrari A, Ferri CP, Fèvre EM, Finucane MM, Flaxman S, Flood L, Foreman K, Forouzanfar MH, Fowkes FG, Fransen M, Freeman MK, Gabbe BJ, Gabriel SE, Gakidou E, Ganatra HA, Garcia B, Gaspari F, Gillum RF, Gmel G, Gonzalez-Medina D, Gosselin R, Grainger R, Grant B, Groeger J, Guillemin F, Gunnell D, Gupta R, Haagsma J, Hagan H, Halasa YA, Hall W, Haring D, Haro JM, Harrison JE, Havmoeller R, Hay RJ, Higashi H, Hill C, Hoen B, Hoffman H, Hotez PJ, Hoy D, Huang JJ, Ibeanusi SE, Jacobsen KH, James SL, Jarvis D, Jasrasaria R, Jayaraman S, Johns N, Jonas JB, Karthikeyan G, Kassebaum N, Kawakami N, Keren A, Khoo JP, King CH, Knowlton LM, Kobusingye O, Koranteng A, Krishnamurthi R, Laden F, Lalloo R, Laslett LL, Lathlean T, Leasher JL, Lee YY, Leigh J, Levinson D, Lim SS, Limb E, Lin JK, Lipnick M, Lipshultz SE, Liu W, Loane M, Ohno SL, Lyons R, Mabweijano J, MacIntyre MF, Malekzadeh R, Mallinger L, Manivannan S, Marcenes W, March L, Margolis DJ, Marks GB, Marks R, Matsumori A, Matzopoulos R, Mayosi BM, McAnulty JH, McDermott MM, McGill N, McGrath J, Medina-Mora ME, Meltzer M, Mensah GA, Merriman TR, Meyer AC, Miglioli V, Miller M, Miller TR, Mitchell PB, Mock C, Mocumbi AO, Moffitt TE, Mokdad AA, Monasta L, Montico M, Moradi-Lakeh M, Moran A, Morawska L, Mori R, Murdoch ME, Mwaniki MK, Naidoo K, Nair MN, Naldi L, Narayan KM, Nelson PK, Nelson RG, Nevitt MC, Newton CR, Nolte S, Norman P, Norman R, O'Donnell M, O'Hanlon S, Olives C, Omer SB, Ortblad K, Osborne R, Ozgediz D, Page A, Pahari B, Pandian JD, Rivero AP, Patten SB, Pearce N, Padilla RP, Perez-Ruiz F, Perico N, Pesudovs K, Phillips D, Phillips MR, Pierce K, Pion S, Polanczyk GV, Polinder S, Pope CA III, Popova S, Porrini E, Pourmalek F, Prince M, Pullan RL, Ramaiah KD, Ranganathan D, Razavi H, Regan M, Rehm JT, Rein DB, Remuzzi G, Richardson K, Rivara FP, Roberts T, Robinson C, De Leòn FR, Ronfani L, Room R, Rosenfeld LC, Rushton L, Sacco RL, Saha S, Sampson U, Sanchez-Riera L, Sanman E, Schwebel DC, Scott JG, Segui-Gomez M, Shahraz S, Shepard DS, Shin H, Shivakoti R, Singh D, Singh GM, Singh JA, Singleton J, Sleet DA, Sliwa K, Smith E, Smith JL, Stapelberg NJ, Steer A, Steiner T, Stolk WA, Stovner LJ, Sudfeld C, Syed S, Tamburlini G, Tavakkoli M, Taylor HR, Taylor JA, Taylor WJ, Thomas B, Thomson WM, Thurston GD, Tleyjeh IM, Tonelli M, Towbin JA, Truelsen T, Tsilimbaris MK, Ubeda C, Undurraga EA, van der Werf MJ, van Os J, Vavilala MS, Venketasubramanian N, Wang M, Wang W, Watt K, Weatherall DJ, Weinstock MA, Weintraub R, Weisskopf MG, Weissman MM, White RA, Whiteford H, Wiebe N, Wiersma ST, Wilkinson JD, Williams HC, Williams SR, Witt E, Wolfe F, Woolf AD, Wulf S, Yeh PH, Zaidi AK, Zheng ZJ, Zonies D, Lopez AD, AlMazroa MA, Memish ZA. 2012. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**: 2197–2223. doi: 10.1016/S0140-6736(12)61689-4.

Muyembe T, Kipasa M. 1995. Ebola haemorrhagic fever in Kikwit, Zaire. International Scientific and Technical Committee and WHO Collaborating Centre for Haemorrhagic Fevers. *Lancet* **345**:1448. doi: 10.1016/S0140-6736(95)92640-2.

NCBI. 2014. GenBank - ebola. Accessed: August 2014. Available: http://www.ncbi.nlm.nih.gov/genbank/.

Negredo A, Palacios G, Vazquez-Moron S, Gonzalez F, Dopazo H, Molero F, Juste J, Quetglas J, Savji N, de la Cruz Martínez M, Herrera JE, Pizarro M, Hutchison SK, Echevarría JE, Lipkin WI, Tenorio A. 2011. Discovery of an ebolavirus-like filovirus in Europe. *PLOS Pathogens* **7**:e1002304. doi: 10.1371/journal.ppat.1002304.

NIH. 2014. NIAID emerging infectious diseases/pathogens. Available: Accessed: August 2014.

Nkoghe D, Formenty P, Leroy EM, Nnegue S, Edou SY, Ba JI, Allarangar Y, Cabore J, Bachy C, Andraghetti R, de Benoist AC, Galanis E, Rose A, Bausch D, Reynolds M, Rollin P, Choueibou C, Shongo R, Gergonne B, Koné LM, Yada A, Roth C, Mve MT. 2005. [Multiple Ebola virus haemorrhagic fever outbreaks in Gabon, from October 2001 to April 2002]. *Bulletin de la Société de Pathologie Exotique* **98**:224–229.

Nkoghe D, Kone ML, Yada A, Leroy E. 2011. A limited outbreak of Ebola haemorrhagic fever in Etoumbi, Republic of Congo, 2005. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **105**:466–472. doi: 10.1016/j.trstmh.2011.04.011.

Okware SI, Omaswa FG, Zaramba S, Opio A, Lutwama JJ, Kamugisha J, Rwaguma EB, Kagwa P, Lamunu M. 2002. An outbreak of Ebola in Uganda. *Tropical Medicine & International Health* **7**:1068–1075. doi: 10.1046/j.1365-3156.2002.00944.x.

Olival KJ, Hayman DT. 2014. Filoviruses in bats: current knowledge and future directions. *Viruses* **6**:1759–1788. doi: 10.3390/V6041759.

Onyango CO, Opoka ML, Ksiazek TG, Formenty P, Ahmed A, Tukei PM, Sang RC, Ofula VO, Konongoi SL, Coldren RL, Grein T, Legros D, Bell M, De Cock KM, Bellini WJ, Towner JS, Nichol ST, Rollin PE. 2007. Laboratory diagnosis of Ebola hemorrhagic fever during an outbreak in Yambio, Sudan, 2004. *The Journal of Infectious Diseases* **196**:S193–S198. doi: 10.1086/520609.

ORNL DAAC. 2000. Shuttle radar topography mission near-global digital elevation models. Accessed: August 2014. Available: http://webmap.ornl.gov/wcsdown/wcsdown.jsp?dg_id=10008_1.

Pattyn S, Jacob W, Vandergroen G, Piot P, Courteille G. 1977. Isolation of Marburg-like virus from a case of hemorrhagic fever in Zaire. *Lancet* **1**:573–574. doi: 10.1016/S0140-6736(77)92002-5.
Peterson AT, Bauer JT, Mills JN. 2004a. Ecologic and geographic distribution of filovirus disease. *Emerging Infectious Diseases* **10**:40–47. doi: 10.3201/eid1001.030125.
Peterson AT, Carroll DS, Mills JN, Johnson KM. 2004b. Potential mammalian filovirus reservoirs. *Emerging Infectious Diseases* **10**:2073–2081. doi: 10.3201/eid1012.040346.
Peterson AT, Papes M, Carroll DS, Leirs H, Johnson KM. 2007. Mammal taxa constituting potential coevolved reservoirs of filoviruses. *Journal of Mammalogy* **88**:1544–1554. doi: 10.1644/06-Mamm-a-280r1.1.
Petter JJ, Desbordes F. 2013. *Primates of the world*. Princeton: Princeton University Press. p. 186.
Phillips SJ, Dudik M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications* **19**:181–197. doi: 10.1890/07-2153.1.
Pigott DM, Bhatt S, Golding N, Duda KA, Battle KE, Brady OJ, Messina JP, Balard Y, Bastien P, Pratlong F, Brownstein JS, Freifeld C, Mekaru SR, Gething PW, George DB, Myers MF, Reithinger R, Hay SI. 2014. Global distribution maps of the leishmaniases. *eLife* e02851. doi: 10.7554/eLife.02851.
Pindolia DK, Garcia AJ, Huang ZJ, Fik T, Smith DL, Tatem AJ. 2014. Quantifying cross-border movements and migrations for guiding the strategic planning of malaria control and elimination. *Malaria Journal* **13**:169. doi: 10.1186/1475-2875-13-169.
Pourrut X, Delicat A, Rollin PE, Ksiazek TG, Gonzalez JP, Leroy EM. 2007. Spatial and temporal patterns of Zaire ebolavirus antibody prevalence in the possible reservoir bat species. *The Journal of Infectious Diseases* **196**:S176–S183. doi: 10.1086/520541.
Pourrut X, Kumulungui B, Wittmann T, Moussavou G, Délicat A, Yaba P, Nkoghe D, Gonzalez JP, Leroy EM. 2005. The natural history of Ebola virus in Africa. *Microbes and Infection* **7**:1005–1014. doi: 10.1016/j.micinf.2005.04.006.
Pourrut X, Souris M, Towner JS, Rollin PE, Nichol ST, Gonzalez JP, Leroy E. 2009. Large serological survey showing cocirculation of Ebola and Marburg viruses in Gabonese bat populations, and a high seroprevalence of both viruses in Rousettus aegyptiacus. *BMC Infectious Diseases* **9**:e159. doi: 10.1186/1471-2334-9-159.
Rouquet P, Froment JM, Bermejo M, Kilbourn A, Karesh W, Reed P, Kumulungui B, Yaba P, Délicat A, Rollin PE, Leroy EM. 2005. Wild animal mortality monitoring and human Ebola outbreaks, Gabon and Republic of Congo, 2001-2003. *Emerging Infectious Diseases* **11**:283–290. doi: 10.3201/eid1102.040533.
Schipper J, Chanson JS, Chiozza F, Cox NA, Hoffmann M, Katariya V, Lamoreux J, Rodrigues AS, Stuart SN, Temple HJ, Baillie J, Boitani L, Lacher TE Jnr, Mittermeier RA, Smith AT, Absolon D, Aguiar JM, Amori G, Bakkour N, Baldi R, Berridge RJ, Bielby J, Black PA, Blanc JJ, Brooks TM, Burton JA, Butynski TM, Catullo G, Chapman R, Cokeliss Z, Collen B, Conroy J, Cooke JG, da Fonseca GA, Derocher AE, Dublin HT, Duckworth JW, Emmons L, Emslie RH, Festa-Bianchet M, Foster M, Foster S, Garshelis DL, Gates C, Gimenez-Dixon M, Gonzalez S, Gonzalez-Maya JF, Good TC, Hammerson G, Hammond PS, Happold D, Happold M, Hare J, Harris RB, Hawkins CE, Haywood M, Heaney LR, Hedges S, Helgen KM, Hilton-Taylor C, Hussain SA, Ishii N, Jefferson TA, Jenkins RK, Johnston CH, Keith M, Kingdon J, Knox DH, Kovacs KM, Langhammer P, Leus K, Lewison R, Lichtenstein G, Lowry LF, Macavoy Z, Mace GM, Mallon DP, Masi M, McKnight MW, Medellín RA, Medici P, Mills G, Moehlman PD, Molur S, Mora A, Nowell K, Oates JF, Olech W, Oliver WR, Oprea M, Patterson BD, Perrin WF, Polidoro BA, Pollock C, Powel A, Protas Y, Racey P, Ragle J, Ramani P, Rathbun G, Reeves RR, Reilly SB, Reynolds JE III, Rondinini C, Rosell-Ambal RG, Rulli M, Rylands AB, Savini S, Schank CJ, Sechrest W, Self-Sullivan C, Shoemaker A, Sillero-Zubiri C, De Silva N, Smith DE, Srinivasulu C, Stephenson PJ, van Strien N, Talukdar BK, Taylor BL, Timmins R, Tirira DG, Tognelli MF, Tsytsulina K, Veiga LM, Vié JC, Williamson EA, Wyatt SA, Xie Y, Young BE. 2008. The status of the world's land and marine mammals: diversity, threat, and knowledge. *Science* **322**:225–230. doi: 10.1126/science.1165115.
Seto KC, Guneralp B, Hutyra LR. 2012. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proceedings of the National Academy of Sciences of USA* **109**:16083–16088. doi: 10.1073/pnas.1211658109.
Shoemaker T, MacNeil A, Balinandi S, Campbell S, Wamala JF, McMullan LK, Downing R, Lutwama J, Mbidde E, Ströher U, Rollin PE, Nichol ST. 2012. Reemerging Sudan ebola virus disease in Uganda, 2011. *Emerging Infectious Diseases* **18**:1480–1483. doi: 10.3201/eid1809.111536.
Siegert R, Shu H-L, Slenczka W, Peters D, Mueller G. 1967. Zur Ätiologie einer unbekannten, von Affen ausgegangenen menschlichen Infektionskrankheit. *Deutsche Medizinische Wochenschrift* **92**:2341–2343. doi: 10.1055/s-0028-1106144.
Simini F, Gonzalez MC, Maritan A, Barabasi AL. 2012. A universal model for mobility and migration patterns. *Nature* **484**:96–100. doi: 10.1038/nature10856.
Simini F, Maritan A, Neda Z. 2013. Human mobility in a continuum approach. *PLOS ONE* **8**:e60069. doi: 10.1371/journal.pone.0060069.
Sinka ME, Bangs MJ, Manguin S, Rubio-Palis Y, Chareonviriyaphap T, Coetzee M, Mbogo CM, Hemingway J, Patil AP, Temperley WH, Gething PW, Kabaria CW, Burkot TR, Harbach RE, Hay SI. 2012. A global map of dominant malaria vectors. *Parasites & Vectors* **5**:69. doi: 10.1186/1756-3305-5-69.
Smith DH, Isaacson M, Johnson KM, Bagshawe A, Johnson BK, Swanapoel R, Killey M, Siongok T, Keruga WK. 1982. Marburg virus disease in Kenya. *Lancet* **1**:816–820. doi: 10.1016/S0140-6736(82)91871-2.
Stockwell D, Peters D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* **13**:143–158. doi: 10.1080/136588199241391.
Stoddard ST, Morrison AC, Vazquez-Prokopec GM, Soldan VP, Kochel TJ, Kitron U, Elder JP, Scott TW. 2009. The role of human movement in the transmission of vector-borne pathogens. *PLOS Neglected Tropical Diseases* **3**:e481. doi: 10.1371/journal.pntd.0000481.

Talbi C, Lemey P, Suchard MA, Abdelatif E, Elharrak M, Nourlil J, Faouzi A, Echevarría JE, Vazquez Morón S, Rambaut A, Campiz N, Tatem AJ, Holmes EC, Bourhy H. 2010. Phylodynamics and human-mediated dispersal of a zoonotic virus. *PLOS Pathogens* **6**:e1001166. doi: 10.1371/journal.ppat.1001166.

Tatem AJ, Goetz SJ, Hay SI. 2004. Terra and aqua: new data for epidemiology and public health. *International Journal of Applied Earth Observation and Geoinformation* **6**:33–46. doi: 10.1016/j.jag.2004.07.001.

Tignor GH, Casals J, Shope RE. 1993. The Yellow Fever epidemic in Ethiopia, 1961-1962-retrospective serological evidence for concomitant Ebola or Ebola-like virus infection. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **87**:162–162. doi: 10.1016/0035-9203(93)90471-2.

Towner JS, Amman BR, Sealy TK, Carroll SAR, Comer JA, Kemp A, Swanepoel R, Paddock CD, Balinandi S, Khristova ML, Formenty PB, Albarino CG, Miller DM, Reed ZD, Kayiwa JT, Mills JN, Cannon DL, Greer PW, Byaruhanga E, Farnon EC, Atimnedi P, Okware S, Katongole-Mbidde E, Downing R, Tappero JW, Zaki SR, Ksiazek TG, Nichol ST, Rollin PE. 2009. Isolation of genetically diverse Marburg viruses from Egyptian fruit bats. *PLOS Pathogens* **5**:e1000536. doi: 10.1371/journal.ppat.1000536.

Towner JS, Khristova ML, Sealy TK, Vincent MJ, Erickson BR, Bawiec DA, Hartman AL, Comer JA, Zaki SR, Ströher U, Gomes da Silva F, del Castillo F, Rollin PE, Ksiazek TG, Nichol ST. 2006. Marburgvirus genomics and association with a large hemorrhagic fever outbreak in Angola. *Journal of Virology* **80**: 6497–6516. doi: 10.1128/JVI.00069-06.

Towner JS, Sealy TK, Khristova ML, Albarino CG, Conlan S, Reeder SA, Quan PL, Lipkin WI, Downing R, Tappero JW, Okware S, Lutwama J, Bakamutumaho B, Kayiwa J, Comer JA, Rollin PE, Ksiazek TG, Nichol ST. 2008. Newly discovered Ebola virus associated with hemorrhagic fever outbreak in Uganda. *PLOS Pathogens* **4**:e1000212. doi: 10.1371/journal.ppat.1000212.

Trabucco A, Zomer RJ. 2009. Global Aridity Index (Global-Aridity) and global potential Evapo-Transpiration (Global-PET) Geospatial database. Accessed: August 2014. Available: http://www.csi.cgiar.org.

Walsh PD, Bermejo M, Rodriguez-Teijeiro JD. 2009. Disease avoidance and the evolution of primate social connectivity: Ebola, bats, gorillas, and chimpanzees. In: Huffman MA, Chapman CA, editors. *Primate parasite ecology: the dynamics and study of host-parasite relationships*. Cambridge: Cambridge University Press. p. 183–197.

Wamala JF, Lukwago L, Malimbo M, Nguku P, Yoti Z, Musenero M, Amone J, Mbabazi W, Nanyunja M, Zaramba S, Opio A, Lutwama JJ, Talisuna AO, Okware SI. 2010. Ebola hemorrhagic fever associated with novel virus strain, Uganda, 2007-2008. *Emerging Infectious Diseases* **16**:1087–1092. doi: 10.3201/eid1607.091525.

Wan ZM, Li ZL. 1997. A physics-based algorithm for retrieving land-surface emissivity and temperature from EOS/MODIS data. *IEEE Transactions on Geoscience and Remote Sensing* **35**:980–996. doi: 10.1109/36.602541.

Wang H, Liddell CA, Coates MM, Mooney MD, Levitz CE, Schumacher AE, Apfel H, Iannarone M, Phillips B, Lofgren KT, Sandar L, Dorrington RE, Rakovac I, Jacobs TA, Liang X, Zhou M, Zhu J, Yang G, Wang Y, Liu S, Li Y, Ozgoren AA, Abera SF, Abubakar I, Achoki T, Adelekan A, Ademi Z, Alemu ZA, Allen PJ, Almazroa MA, Alvarez E, Amankwaa AA, Amare AT, Ammar W, Anwari P, Cunningham SA, Asad MM, Assadi R, Banerjee A, Basu S, Bedi N, Bekele T, Bell ML, Bhutta Z, Blore J, Basara BB, Boufous S, Breitborde N, Bruce NG, Bui LN, Carapetis JR, Cárdenas R, Carpenter DO, Caso V, Castro RE, Catalá-Lopéz F, Cavlin A, Che X, Chiang PP, Chowdhury R, Christophi CA, Chuang TW, Cirillo M, da Costa Leite I, Courville KJ, Dandona L, Dandona R, Davis A, Dayama A, Deribe K, Dharmaratne SD, Dherani MK, Dilmen U, Ding EL, Edmond KM, Ermakov SP, Farzadfar F, Fereshtehnejad SM, Fijabi DO, Foigt N, Forouzanfar MH, Garcia AC, Geleijnse JM, Gessner BD, Goginashvili K, Gona P, Goto A, Gouda HN, Green MA, Greenwell KF, Gugnani HC, Gupta R, Hamadeh RR, Hammami M, Harb HL, Hay S, Hedayati MT, Hosgood HD, Hoy DG, Idrisov BT, Islami F, Ismayilova S, Jha V, Jiang G, Jonas JB, Juel K, Kabagambe EK, Kazi DS, Kengne AP, Kereselidze M, Khader YS, Khalifa SE, Khang YH, Kim D, Kinfu Y, Kinge JM, Kokubo Y, Kosen S, Defo BK, Kumar GA, Kumar K, Kumar RB, Lai T, Lan Q, Larsson A, Lee JT, Leinsalu M, Lim SS, Lipshultz SE, Logroscino G, Lotufo PA, Lunevicius R, Lyons RA, Ma S, Mahdi AA, Marzan MB, Mashal MT, Mazorodze TT, McGrath JJ, Memish ZA, Mendoza W, Mensah GA, Meretoja A, Miller TR, Mills EJ, Mohammad KA, Mokdad AH, Monasta L, Montico M, Moore AR, Moschandreas J, Msemburi WT, Mueller UO, Muszynska MM, Naghavi M, Naidoo KS, Narayan KV, Nejjari C, Ng M, de Dieu Ngirabega J, Nieuwenhuijsen MJ, Nyakarahuka L, Ohkubo T, Omer SB, Caicedo AJ, Wyk VP, Pope D, Prabhakaran D, Rahman SU, Rana SM, Reilly RQ, Rojas-Rueda D, Ronfani L, Rushton L, Saeedi MY, Salomon J, Sampson U, Santos IS, Sawhney M, Schmidt JC, Shakh-Nazarova M, She J, Sheikhbahaei S, Shibuya K, Shin HH, Shishani K, Shiue I, Sigfusdottir ID, Singh JA, Skirbekk V, Sliwa K, Soshnikov SS, Sposato LA, Stathopoulou VK, Stroumpoulis K, Tabb KM, Talongwa RT, Teixeira CM, Terkawi AS, Thomson AJ, Thorne-Lyman AL, Toyoshima H, Dimbuene ZT, Uwaliraye P, Uzun SB, Vasankari TJ, Vasconcelos AM, Vlassov VV, Vollset SE, Vos T, Waller S, Wan X, Weichenthal S, Weiderpass E, Weintraub RG, Westerman R, Wilkinson JD, Williams HC, Yang YC, Yentur GK, Yip P, Yonemoto N, Younis M, Yu C, Jin KY, El Sayed Zaki M, Zhu S, Lopez AD, Murray CJ. 2014. Global, regional, and national levels of neonatal, infant, and under-5 mortality during 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*:doi: 10.1016/S0140-6736(14)60497-9.

Weiss DJ, Atkinson PM, Bhatt S, Mappin B, Hay SI, Gething PW. 2014a. An effective approach for gap-filling continental scale remotely sensed time-series. *ISPRS Journal of Photogrammetry and Remote Sensing* accepted manuscript.

Weiss DJ, Bhatt S, Mappin B, Van Boeckel TP, Smith DL, Hay SI, Gething PW. 2014b. Air temperature suitability for Plasmodium falciparum malaria transmission in Africa 2000-2012: a high-resolution spatiotemporal prediction. *Malaria Journal* **13**:171. doi: 10.1186/1475-2875-13-171.

Wenger SJ, Olden JD. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution* **3**:260–267. doi: 10.1111/j.2041-210X.2011.00170.x.

Wesolowski A, Buckee CO, Pindolia DK, Eagle N, Smith DL, Garcia AJ, Tatem AJ. 2013. The use of census migration data to approximate human movement patterns across temporal scales. *PLOS ONE* **8**:e52971. doi: 10.1371/journal.pone.0052971.
Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO. 2012. Quantifying the impact of human mobility on malaria. *Science* **338**:267–270. doi: 10.1126/science.1223467.
WHO. 2001. Outbreak of Ebola haemorrhagic fever, Uganda, August 2000-January 2001. *Weekly Epidemiological Record* **76**:41–46.
WHO. 2003a. Outbreak(s) of Ebola haemorrhagic fever, Congo and Gabon, October 2001-July 2002. *Weekly Epidemiological Record* **78**:223–228.
WHO. 2003b. Outbreak(s) of Ebola haemorrhagic fever in the Republic of the Congo, January-April 2003. *Weekly Epidemiological Record* **78**:285–289.
WHO. 2005. Outbreak of Ebola haemorrhagic fever in Yambio, south Sudan, April - June 2004. *Weekly Epidemiological Record* **80**:370–375.
WHO. 2012a. Ebola in Uganda - 29 July 2012. Accessed: August 2014. Available: http://www.who.int/csr/don/2012_07_29/en/.
WHO. 2012b. Ebola outbreak in democratic Republic of Congo - 17 August 2012. Accessed: August 2014. Available: http://www.who.int/csr/don/2012_08_18/en/.
WHO. 2012c. Ebola in Uganda - 17 November 2012. Accessed: August 2014. Available: http://www.who.int/csr/don/2012_11_17/en/.
WHO. 2014a. Ebola virus disease update - west Africa. Accessed: August 2014. Available: http://www.who.int/csr/don/2014_08_15_ebola/en/.
WHO. 2014b. Health statistics and information systems - definition of region groupings. Accessed: August 2014. Available: http://www.who.int/healthinfo/global_burden_disease/definition_regions/en/.
WHO. 2014c. *Interim manual - Ebola and Marburg virus disease epidemics: preparedness, alert, control, and evaluation*. Geneva: World Health Organization. P. 123.
WHO. 2014d. WHO statement on the meeting of the International Health Regulations Emergency Committee regarding the 2014 Ebola outbreak in West Africa. Accessed: August 2014. Available: http://www.who.int/mediacentre/news/statements/2014/ebola-20140808/en/.
WHO International Study Team. 1978. Ebola haemorrhagic fever in Sudan, 1976. *Bulletin of the World Health Organization* **56**:247–270.
Wittmann TJ, Biek R, Hassanin A, Rouquet P, Reed P, Yaba P, Pourrut X, Real LA, Gonzalez JP, Leroy EM. 2007. Isolates of Zaire ebolavirus from wild apes reveal genetic lineage and recombinants. *Proceedings of the National Academy of Sciences of USA* **104**:17123–17127. doi: 10.1073/pnas.0704076104.
Wolfe ND, Daszak P, Kilpatrick AM, Burke DS. 2005. Bushmeat Hunting, deforestation, and prediction of zoonotic disease emergence. *Emerging Infectious Diseases* **11**:1822–1827. doi: 10.3201/eid1112.040789.
Wolfe ND, Dunavan CP, Diamond J. 2007. Origins of major human infectious diseases. *Nature* **447**:279–283. doi: 10.1038/Nature05775.
World Bank. 2014. World Bank open dataset. Accessed: July 2014. Available: http://data.worldbank.org/.
WorldPop. 2014. WorldPop project. Accessed: August 2014. Available: http://www.worldpop.org.uk/.
WWF. 2014. List of terrestrial ecoregions. Accessed: August 2014. Available: http://wwf.panda.org/about_our_earth/ecoregions/ecoregion_list/.
Yang Y, Atkinson P, Ettema D. 2008. Individual space-time activity-based modelling of infectious disease transmission within a city. *Journal of the Royal Society, Interface* **5**:759–772. doi: 10.1098/rsif.2007.1218.
Zipf GK. 1946. The P_1 P_2/D hypothesis: on the intercity movement of persons. *American Sociological Review* **11**:677–686. doi: 10.2307/2087063.

Chapter 5

Understanding risk maps: Marburg virus disease.

Marburg virus disease is a rare filoviral disease that can cause severe haemorrhagic symptoms in individuals and can be transmitted from human to human. This chapter applies species distribution models to understand the risk this virus poses in Africa and provides a more detailed look at how such maps should be used to inform policy and surveillance as well as identifying key information gaps. This work has been published in *Transactions of the Royal Society of Tropical Medicine and Hygiene* and is included here in its final form. The additional information referred to in this chapter is included in the Appendix of this thesis.

Trans R Soc Trop Med Hyg 2015; **109**: 366–378
doi:10.1093/trstmh/trv024 Advance Access publication 27 March 2015

ORIGINAL ARTICLE

Mapping the zoonotic niche of Marburg virus disease in Africa

David M. Pigott[a,*], Nick Golding[a], Adrian Mylne[a], Zhi Huang[a], Daniel J. Weiss[a], Oliver J. Brady[a], Moritz U. G. Kraemer[a] and Simon I. Hay[a,b]

[a]*Spatial Ecology & Epidemiology Group, Department of Zoology, University of Oxford, Oxford, UK;*
[b]*Fogarty International Center, National Institutes of Health, Bethesda, Maryland, USA*

*Corresponding author: Tel: +44 1865 271137; E-mail: david.pigott@zoo.ox.ac.uk

Received 19 December 2014; revised 20 February 2015; accepted 23 February 2015

Background: Marburg virus disease (MVD) describes a viral haemorrhagic fever responsible for a number of outbreaks across eastern and southern Africa. It is a zoonotic disease, with the Egyptian rousette (*Rousettus aegyptiacus*) identified as a reservoir host. Infection is suspected to result from contact between this reservoir and human populations, with occasional secondary human-to-human transmission.

Methods: Index cases of previous human outbreaks were identified and reports of infection in animals recorded. These data were modelled within a species distribution modelling framework in order to generate a probabilistic surface of zoonotic transmission potential of MVD across sub-Saharan Africa.

Results: Areas suitable for zoonotic transmission of MVD are predicted in 27 countries inhabited by 105 million people. Regions are suggested for exploratory surveys to better characterise the geographical distribution of the disease, as well as for directing efforts to communicate the risk of practices enhancing zoonotic contact.

Conclusions: These maps can inform future contingency and preparedness strategies for MVD control, especially where secondary transmission is a risk. Coupling this risk map with patient travel histories could be used to guide the differential diagnosis of highly transmissible pathogens, enabling more rapid response to outbreaks of haemorrhagic fever.

Keywords: Boosted regression trees, Filovirus, Marburg virus disease, *Rousettus aegyptiacus*, Species distribution models, Viral haemorrhagic fever

Introduction

In 1967, outbreaks of a previously undescribed disease in workers of three laboratories in West Germany and Yugoslavia were reported, characterised by high fever, haemorrhaging and organ failure.[1] A novel virus, named Marburg virus (MARV), the first described in the *Filoviridae* family, was subsequently identified as the causative pathogen.[2] In 1975, the first recognised case of the disease outside of a laboratory occurred in Rhodesia (now Zimbabwe), with one case in 1980 due to MARV and another in 1987 due to Ravn virus (RAVV), another marburgvirus.[3] Not until 1998, when a series of fatal haemorrhagic cases were identified in the vicinity of Durba, Democratic Republic of the Congo (DRC), was a large-scale outbreak reported. A total of 154 cases were reported, with the source of infection traced back to bat colonies in local gold mines.[4] While a large number of cases were reported between 1998 and 2000, it was found that multiple introductions of the virus from the same zoonotic pool were responsible for the continued outbreak rather than only human-to-human transmission, as more commonly reported with Ebola virus disease (EVD).[4–6] In 2004 however, a large outbreak in Uige province, Angola, occurred where, unlike in Durba, continued cases were driven by subsequent human-to-human transmission rather than repeated introductions from a natural source.[7] More recent outbreaks have been smaller in comparison (Figure 1).[8–12]

The wider epidemiology of Marburg virus disease (MVD) remains relatively unknown (Figure 2). While non-human primates are susceptible to the disease, there have been no reported transmission events from primates to humans outside of a laboratory setting. Furthermore, no significant epizootics have been reported among non-human primates, unlike the closely related ebolaviruses.[13,14] Past outbreaks have strongly implicated bats as the origin of initial index cases in humans. Serological and molecular surveys conducted in caves and

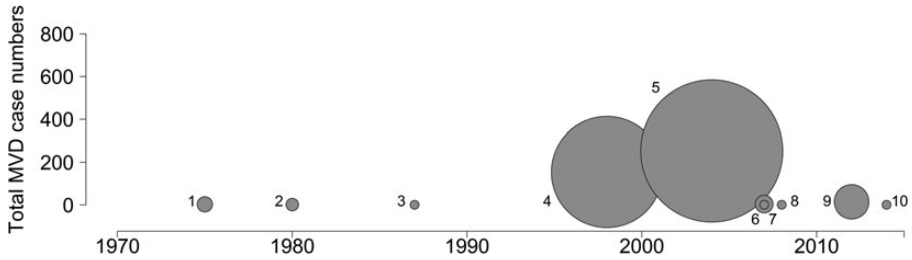


Figure 1. Case numbers in previous Marburg virus disease outbreaks. The size of each circle is proportional to the number of cases of the disease in a given outbreak. Outbreaks are labelled as per Table 1.

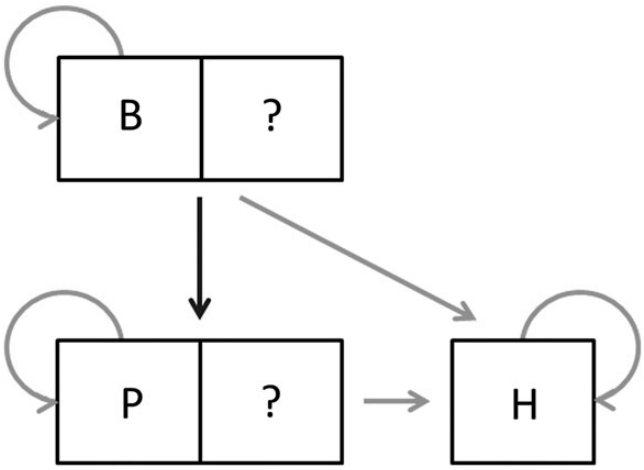


Figure 2. The epidemiology of marburgvirus transmission in Africa. B represents suspected bat reservoirs (including Egyptian rousettes). Susceptible animals include non-human primates, such as the monkeys responsible for the 1967 outbreaks (P). H represents humans. Question marks indicate potential animals of other species. All routes have been confirmed or are suspected to occur apart from transmission between bats and primates, which remains unknown. Adapted from Laminger and Prinz and Groseth et al.[15,16]

mines visited by the infected individuals, have identified the virus in the Egyptian rousette (*Rousettus aegyptiacus*).[17,18] Colonies of bats have also been reported in the vicinity of the supposed index case in other outbreaks.[19,20]

In order to better understand the nature of MVD risk, this study attempts to define those areas where zoonotic transmission of MVD may occur in order to identify people at potential risk of zoonotic spillover. Such a methodology has previously been employed with EVD to identify 22 western and equatorial African nations where ebolavirus transmission may occur.[21] Ecological niche modelling of MVD has previously been undertaken and this work seeks to update these efforts by including more recent outbreaks, records of infection in animals, improved environmental covariate layers and recent advances in modelling techniques.[21–25] The need for such information is critical, not only to assist in differential diagnosis of fevers across Africa, but also to increase awareness of the potential risk of more widespread outbreaks that could arise from a delay in the response to initial cases.[26]

Materials and methods

Methodological overview

A species distribution model, specifically an ensemble boosted regression trees (BRT) framework, was used to model the zoonotic niche of MVD. This model optimally builds ensembles of trees based upon binary decisions used to classify suitable environmental covariates in reference to a database of known occurrence locations.[27,28] Areas which are environmentally similar to those with reported zoonotic transmission of MVD are predicted to be at higher levels of risk. To perform this analysis, we obtained four key information components: 1. a database of cases where MVD has been transmitted from animals to humans; 2. reported infections of MARV and RAVV in animals; 3. a collection of spatially gridded environmental variables that are likely to be correlates of disease presence; and 4. background (pseudo-absence) data indicating locations where MVD has not been reported. The model was restricted to the African continent since there have been no reported natural outbreaks, in humans or animals, outside this region.

Identifying human and animal infections with marburgviruses

Outbreaks of MVD in humans were identified from review articles and by sourcing original references.[29] Where possible, index cases (individuals infected by animal reservoir species) were located and the supposed location of animal to human transfer of MARV and RAVV was geopositioned using Google Earth. When an accurate site location could not be determined, a geographic area (termed a polygon) was defined covering the reported region, identified using the source articles (e.g., a specified landmark, or an area referenced in relation to another directly identifiable site); otherwise a precise, site-specific latitude and longitude was recorded. For larger settlements, the centroid of the site was recorded. In some instances, only the first reported patient could be identified, with little information on the initial route of infection. In these instances we assumed that the index case occurred where the zoonotic transmission event took place. For some outbreaks there was sufficient evidence to suggest multiple independent zoonotic transmission events. For these outbreaks, each individual transmission event was separately positioned.

To obtain a comprehensive database of MARV infections in animals, a literature search was conducted in Web of Science using the search term 'Marburg reservoir OR Marburg monkey OR Marburg bat OR Marburg primate'. This procedure returned 1544

unique citations. Abstracts for these citations were processed and where they indicated that the article might contain spatial information on Marburg infections in animals the full articles were obtained. Once identified, the references of these articles (as well as more general review articles discussing MVD reservoirs) were followed up in case relevant articles were omitted from the initial literature search. Locations of infected animals were geopositioned using the same methodology as for human index cases.

Covariates used in the analyses

A suite of ecologically relevant gridded environmental covariates for Africa was compiled, each having a nominal resolution of 5 km × 5 km. A number of environmental covariates thought to potentially influence MVD distribution were selected for inclusion in this analysis, namely range and mean values of enhanced vegetation index (EVI) and land surface temperature (LST) (day and night) derived from satellite data and parsed through gap-filling algorithms, as well as elevation and potential evapotranspiration (PET).[21,25,30,31] Many of these have been considered in previous investigations.[24] In addition, distance to the nearest Karst formation was included as a covariate.[32] Karst landscapes typically form when soluble rocks dissolve and can create expansive cave networks and as such were used in the model as a proxy for the subterranean roosting habitat of the supposed disease reservoir, the Egyptian rousette.[33,34] Previous work mapping the zoonotic niche of EVD utilised a bat distribution covariate layer. While attempts were made to replicate this approach for MVD, the lack of sufficiently detailed data available from the Global Biodiversity Information Facility to allow for differentiation between roosting and foraging sites meant that the niche modelling approaches were unable to produce reliable results and therefore these outputs could not be included in the final analysis.

Marburg distribution modelling

An ensemble boosted regression trees model was used to define areas environmentally suitable for zoonotic MARV transmission. The model requires both presence and background information to generate a prediction, the latter of which is often hard to collect systematically and in an unbiased manner. As a result, randomly generated background records are often supplied. For this study, a background record dataset was generated by randomly sampling 10 000 locations across Africa, biased towards more populous areas as a proxy for reporting bias.[22] This sampling allows for comparison of factors influencing presence and likely absence locations for MVD by the model. In total, 500 submodels were used. Each submodel was fitted using the gbm.step subroutine in the dismo package in the R statistical programming environment.[28,35,36] Given the limited number of records available, we reduced the number of cross-validation folds used to fit the model to three, from the default of 10. All other tuning parameters of the algorithm were held at their default values (tree complexity=4, learning rate=0.005, bag fraction=0.75, step size=10). For each polygon in the occurrence dataset, one point was randomly selected from within the defined area for each submodel. This Monte Carlo procedure enabled the model to efficiently integrate over the environmental uncertainty associated with imprecise geographic data. A bootstrap sample was then taken from each of these datasets and used to train the BRT model. For each submodel, weightings were applied to the background dataset so that the sum of the weighted background data equalled the weighted sum of the occurrence records.[37] This was done in order to improve the discrimination capacity of the model. Each submodel predicts environmental suitability on a continuous scale from 0 to 1. An ensemble final prediction map was generated by combining the predictions from these submodels, calculating the mean prediction as well as the 5% and 95% confidence intervals around this for each 5 km × 5 km pixel.

Two models were constructed. Model 1 used only records of human index cases and model 2 used both human index cases and reported infection in animals. This was done in order to augment the relatively small number of index case records available and to evaluate the influence of including animal data on the model.

The area under the curve (AUC) statistic was used to assess model accuracy. The statistic was calculated for each submodel using a three-fold cross validation, and then summarised across all the submodels to generate a mean and standard deviation for this value. This procedure divided the dataset into three subsets that had approximately equal numbers of presence records and background data. Due to the small number of presence records used to train each submodel, this approach represents a very thorough test of the model's predictive ability. In order to prevent inflation of the accuracy statistics due to spatial sorting bias, a pairwise distance sampling procedure was used.[38] As a result, these AUC statistics are lower than typical outputs, but give a more realistic evaluation of the ability of the model to predict for different regions.[39] Uncertainty in the prediction was evaluated by considering the difference between the 5% and 95% confidence intervals.

The final outputs represent the environmental suitability for zoonotic transmission of MARV for each 5 km × 5 km pixel which allows for relative comparison of risk across Africa.

Population living in areas of environmental suitability for zoonotic transmission

Estimates of population living in areas at risk of zoonotic transmission were derived by converting the continuous surface of transmission risk into a binary at-risk/not-at-risk classification for each pixel. The threshold for this classification was based upon the minimum environmental suitability value at the locations of the occurrence records. To calculate this value, the risk estimate for each point occurrence and the mean probability of each area/polygon occurrence were assessed. Countries were classified into two categories of risk. Set 1 are countries where index cases of MVD have been reported and set 2 are countries where no index cases have been previously reported and have more than 100 pixels (i.e., approximately 2500 km^2) at risk. The number of people living in these pixels was calculated from existing population surfaces for Africa.[40,41]

Contiguous areas of risk within each country were visually identified and the latitude and longitude for the approximate midpoint for these areas were recorded, suggesting areas of potential interest for further prospective epidemiological investigation.

The R code used for all of the analysis is freely available via https://github.com/SEEG-Oxford/marburg_zoonotic.

Results

Reported infections in humans and animals

A total of 10 distinct outbreaks of MVD were identified, ranging in size from single reported cases to community-wide outbreaks with hundreds of reported cases (Table 1). Five countries have confirmed or suspected instances of animal-to-human zoonotic transmission, namely Kenya, Uganda, Zimbabwe, Angola and the DRC (Figure 3). For the majority of these outbreaks, caves or mines have been singled out as the likely venue for their spill-over events. Some records were of individuals who had subsequently travelled elsewhere before becoming symptomatic, for which efforts were made to identify the original site of infection.[3,9,10]

All available animal infection records were from bat populations, often sampled in response to human cases, with the exception of one reported infection in grivets (the same animals responsible for the 1967 laboratory-based outbreak) (Table 2). These monkeys were trapped near Kidera and Namsale in Uganda where they were assumed to have been originally infected.[46] Where epidemiological surveys of nearby potential animal reservoirs were undertaken during, or shortly after outbreaks in humans, PCR identification of MARV was often performed.[17,18] A serological survey of Gabonese bat populations reported positivity in Egyptian rousettes and other bats.[47] This is the only evidence of animal infection occurring outside the recognised range of reported human populations, all other animal infections have been reported in the vicinity of human outbreaks (Figure 3).

Predicted environmental suitability for zoonotic transmission of marburgviruses

Due to the relative paucity of data, two model variants were used in order to test various assumptions about the poorly understood MVD epidemiology. Model 1, which only included human index case data, identified geological features (elevation and distance to Karst formation) and vegetation indices (both EVI mean and range) as the main predictors of suitability for zoonotic transmission (Table 3). Model 2, which included the entire dataset of MARV infections, implied a broader spatial extent, with environmental factors (EVI, LST and PET) playing a more important role in prediction compared to elevation. The AUC values were 0.64 ± 0.12 and 0.62 ± 0.08 for models 1 and 2, respectively, indicating that both the models demonstrated similar predictive skill. Note however that as these statistics were calculated using different evaluation datasets, they are not directly comparable. Uncertainty maps for the predicted surfaces for MVD are presented in Supplementary Figures 1 and 2.

Both models predict high suitability for zoonotic transmission in the set 1 countries. In total, model 2 predicts 27 countries to be at-potential-risk (set 1 and 2) of zoonotic transmission of MARV with 105 million people living in at-risk areas. Model 1 predicts 19 countries at risk with 75 million individuals living in at-risk areas. These 19 countries are consistently predicted to be at-risk in both models 1 and 2.

Discussion

This work utilises all known outbreaks of MVD in humans and reported infections in animals in order to understand the nature of risk posed by this disease (Figures 4 and 5). Previous assessments have indicated that a much broader region is at-risk of zoonotic transmission than those countries that have reported transmission to-date.[24] Our analysis, reinforced by new outbreak reports and environmental covariate information, is in concordance with previous ecological modelling investigations of MVD, identifying temperature and vegetation indices as key determinants of its spatial distribution.[23,24,50] In addition, we identify the potential importance of geological features in influencing areas of potential MARV risk. The majority of at-risk populations live in areas that have previously reported outbreaks, mainly Uganda, Kenya and the DRC. Amongst countries yet to see human infection (set 2), the most notable are Ethiopia, Cameroon and Zambia, in which large areas are predicted to be at-risk.

Table 1. Locations of natural outbreaks of Marburg virus disease in humans

Outbreak	Date range	Countries	Location	Cases/deaths	Reference
1[a]	Feb 1975	Zimbabwe	Chinoyi caves	3/1	3,24
2	Jan 1980	Kenya	Nzoia	2/1	20
3	Aug 1987	Kenya	Kitum cave	1/1	19
4	Oct 1998–Aug 2000	DRC	Durba	154/128	4,42
5	Oct 2004–Jul 2005	Angola	Uige province	252/227	7,43,44,45
6	Jun 2007–Sept 2007	Uganda	Kitaka gold mine	4/1	8
7[a]	Dec 2007–Jan 2008	Uganda	Python cave	1/0	9
8[a]	Jul 2008	Uganda	Python cave	1/1	10
9	Aug 2012–Oct 2012	Uganda	Ibanda district	15/14	11
10	Sep 2014	Uganda	Mpigi/Mawokota district	1/1	12

DRC: Democratic Republic of the Congo.
[a] Indicates case imported elsewhere.

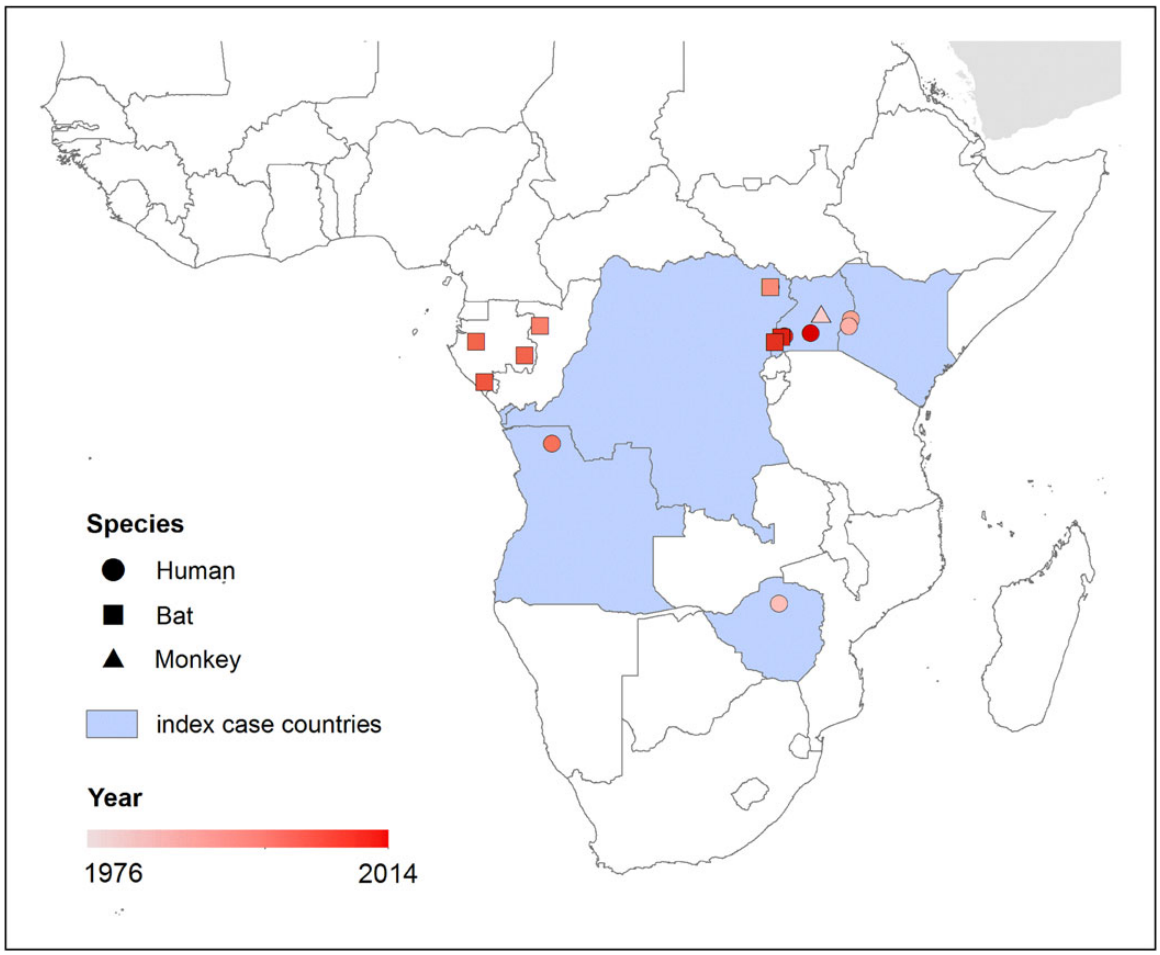


Figure 3. The locations of marburgvirus disease outbreaks in humans and reported animal infections across Africa. This figure is available in black and white in print and in colour at Transactions online.

Table 2. Locations of reported infections with Marburg virus in animals

Site	Date range	Country	Location	Species	Diagnosis	Reference
1	Aug–Sep 1967	Uganda	Kidera and Namasale	*Chlorocebus aethiops*	Serology	42
2	May–Oct 1999	DRC	Durba	*Miniopterus inflatus*	PCR	16
3	May–Oct 1999	DRC	Durba	*Rhinolophus eloquens*	PCR	16
4	May–Oct 1999	DRC	Durba	*Rousettus aegyptiacus*	PCR	16
5	Jun 2003–May 2006	ROC	Mbomo district	Various bat species	Serology	43
6	Feb 2005–Mar 2008	Gabon	Haut-Ogooue district	Various bat species	Serology	43
7	Apr 2005	Gabon	Moyen-Ogooue district	Various bat species	Serology	43
8	Feb 2006	Gabon	Nyanga district	Various bat species	Serology	43
9	Aug 2007	Uganda	Kitaka gold mine	*Rousettus aegyptiacus*	PCR	15
10	Aug 2007	Uganda	Kitaka gold mine	*Hipposideros* spp.	PCR	15
11	Apr 2008	Uganda	Kitaka gold mine	*Rousettus aegyptiacus*	PCR	15
12	Aug 2009	Uganda	Python cave	*Rousettus aegyptiacus*	PCR	48
13	Nov 2009	Uganda	Python cave	*Rousettus aegyptiacus*	PCR	48
14	Nov 2012	Uganda	Kitaka gold mine	*Rousettus aegyptiacus*	PCR	49

DRC: Democratic Republic of the Congo; ROC: Republic of Congo.

Table 3. Summary statistics for model outputs

Statistic	Model 1: human data	Model 2: human and animal data
AUC (± standard deviation)	0.64±0.12	0.62±0.08
1st predictor	Elevation: 49.1%	Mean EVI: 49.1%
2nd predictor	Mean EVI: 30.1%	Night-time mean LST: 19.5%
3rd predictor	EVI range: 7.8%	Elevation: 10.9%
4th predictor	Distance to Karst: 4.9%	Mean PET: 7.5%
5th predictor	Night-time mean LST: 3.1%	Day-time mean LST: 5.0%

Relative contributions for each of the top five predictors are reported as a percentage.
AUC: area under the curve; EVI: enhanced vegetation index; LST: land surface temperature; PET: potential evapotranspiration.

As with any model-based approach, an awareness of the limitations of the data and the assumptions made by the model is important. Limited datasets, particularly those where definitive identification of zoonotic transmission sites is unlikely, will hinder predictive capability. However, this study attempts to be as comprehensive as possible by including all reports of infections, as well as considering uncertainty in geopositioning ability. Further information can only help to improve these predictions. Similarly, the model is only able to assess environmental suitability for MVD, therefore in order to translate this into true outbreak risk, additional information on how humans and animal reservoirs interact, as well as how the disease is transmitted within these populations is required. Bearing these caveats in mind, we hope that these results will act as a springboard for further research to better understand the epidemiology and characterise the risk of this disease.

The two model iterations ('human only' versus 'animal and human data') illustrate the need for further research into MARV hosts and their potential for zoonotic transmission to humans. Areas predicted at-risk in model 1 are consistently identified at-risk in model 2 (although the absolute probability of transmission is altered); the inclusion of animal data however, expands the areas of potential risk to include countries across western and central Africa not indicated as being at-risk by model 1. Figure 6 visualises the differences between these two models. As a result, given the limited data availability, model 2 would currently be the most sensible option when discussing the potential risk posed by MVD. In addition, while no reported cases of MVD have been recorded in set 2 countries, a number have seen serological evidence of past exposure in humans.[29] Seropositive individuals have been reported in locations identified as at-risk in model 2 in West Africa, Cameroon, Central African Republic, Nigeria and South Sudan.[48,49,51–56]

Since many MVD spillover events have only resulted in a handful of cases, the likelihood of outbreaks going unrecognised is a possibility.[57] It is therefore also possible that spatial variation in the probability of cases being identified may have biased our models. While we strived to account for such an observation bias in our analysis by weighting pseudo-absence records to areas where infection might be more likely to be detected, we cannot rule out the presence of residual bias. The true nature of zoonotic transmission potential within these countries can only be elucidated by additional surveys.

Knowledge on the animal reservoir for MVD is limited. Egyptian rousettes have consistently been identified as PCR positive for the virus, however animals of a number of other species have also been seropositive.[15,43,58,59] The maps presented here can be used to target key sites for future surveys of bats to better understand the true nature of risk within those areas where no previous outbreak has been reported.

There is considerable overlap between the reported distribution of Egyptian rousettes (Figure 7) and areas of highest risk. Evidence suggests that there are various subspecies of Egyptian rousette across Africa.[33] All but one outbreak of MVD in humans occurred within the known range of members of *R. aegyptiacus leachi*; the outbreak in Uige Province, Angola, however occurred outside the range of bats of this subspecies, but was within the reported range of *R. aegyptiacus unicolor*. It remains unclear whether these populations differ in disease transmission cycles and the nature of the connectivity between bats of these two potential subspecies has important implications for potential disease transmission, either restricting the likely areas of risk to eastern and southern Africa, or including much of central and west Africa (Figure 7B). Similarly, it is possible that bats of subspecies present in north Africa and the Middle East could also be potential reservoirs for MARV. The inclusion of bat distributions in future models would allow for a better understanding of the relationship between MVD and Egyptian rousettes, with the possibility of identifying regions where other bats may be more likely reservoir hosts.

In addition, further surveys for MARV infection in bats in these regions therefore would not only help to better understand the ecology of these bats but also the nature of the risk posed to human populations. As with EVD, spillover of MARV into humans is rare and infection in bat populations also appears uncommon. Understanding the nature of infection within these bat and other potential reservoirs, is crucial in identifying the true nature of risk to human populations, not just for MVD, but also a variety of other viral pathogens.[61] Table 4 identifies the main regions within each at-risk country where such surveillance activities would be of greatest benefit. The output maps, to allow for national survey placement, are freely available from the following link: http://goo.gl/0qTOfe

Our risk maps provide a baseline estimate for the extent of the zoonotic niche of MVD, which can subsequently be enhanced through more specific research. While an area may have the potential for zoonotic transmission of MVD, if humans rarely interact with these animal hosts, spillover events are unlikely to occur. As a result, in spite of a large number of individuals living in areas where transmission is possible, a considerably smaller number will be at-risk of encountering an infected reservoir and subsequently being infected. Surveys and ethnographic assessments can help better understand the true nature of risk within these regions, particularly important if a quantitative

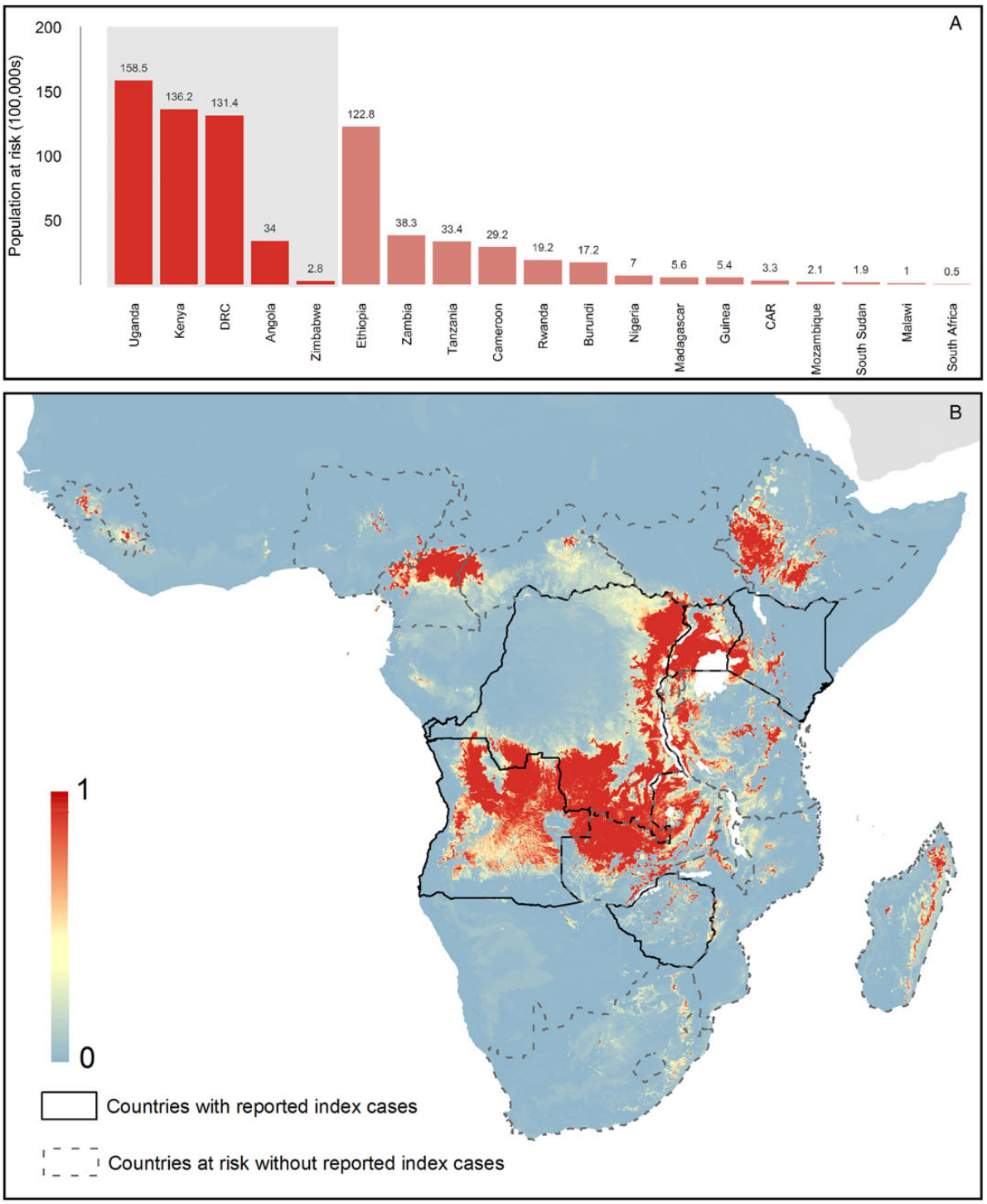


Figure 4. Predicted geographical distribution of the zoonotic niche for marburgviruses using model 1 – human index cases only. Panel A shows the total populations living in areas of risk of zoonotic transmission for each at-risk country. The grey rectangle highlights countries in which index cases of disease have been reported (set 1); the remainder are countries in which risk of zoonotic transmission is predicted, but in which index cases of Marburg virus disease have not been reported and have more than 100 at-risk pixels (set 2). These countries are ranked by population-at-risk within each set. The population-at-risk figure in 100 000 s is given above each bar. Panel B shows the predicted distribution of zoonotic marburgviruses. The scale reflects the relative probability that zoonotic transmission of marburgviruses could occur at these locations; areas closer to 1 (red) are more likely to harbour zoonotic transmission than those closer to 0 (blue). Countries with borders outlined are those which are predicted to contain at-risk areas for zoonotic transmission based on a thresholding approach (see Methods). The area under the curve statistic, calculated under a stringent cross-validation procedure is 0.64 ± 0.12. Solid lines represent set 1 whilst dashed lines delimit set 2. Areas covered by major lakes have been masked white.

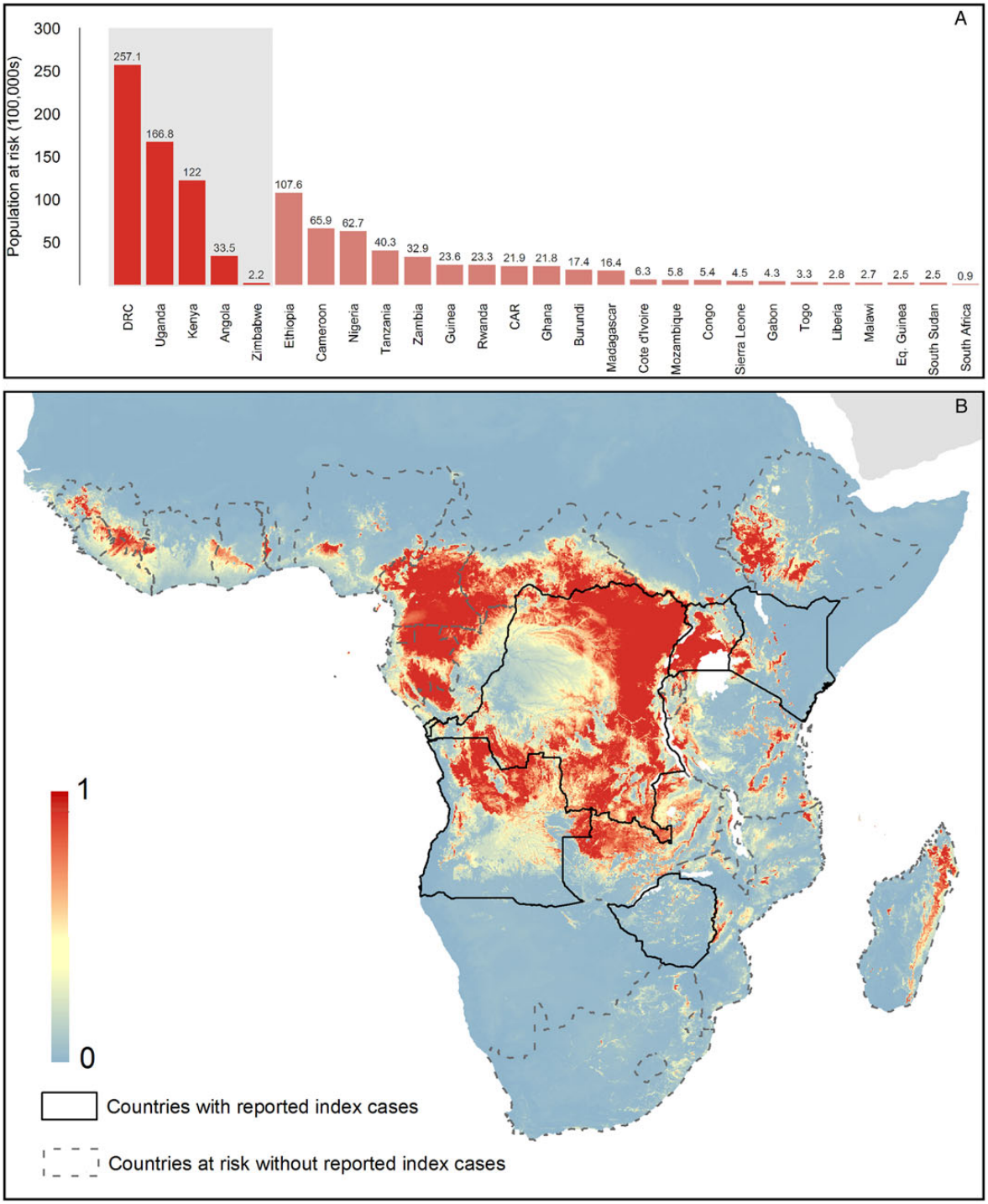


Figure 5. Predicted geographical distribution of the zoonotic niche for marburgviruses using model 2–both human index cases and infections in animals. Panel A shows the total populations living in areas of risk of zoonotic transmission for each at-risk country. The grey rectangle highlights countries in which index cases of Marburg virus disease have been reported (set 1); the remainder are countries in which risk of zoonotic transmission is predicted, but in which index cases of Marburg have not been reported and have more than 100 at-risk pixels (set 2). These countries are ranked by population-at-risk within each set. The population-at-risk figure in 100 000 s is given above each bar. Panel B shows the predicted distribution of zoonotic marburgviruses. The scale reflects the relative probability that zoonotic transmission of marburgviruses could occur at these locations; areas closer to 1 (red) are more likely to harbour zoonotic transmission than those closer to 0 (blue). Countries with borders outlined are those which are predicted to contain at-risk areas for zoonotic transmission based on a thresholding approach (see Methods). The area under the curve statistic, calculated under a stringent cross-validation procedure, is 0.62 ± 0.08. Solid lines represent set 1 whilst dashed lines delimit set 2. Areas covered by major lakes have been masked white.

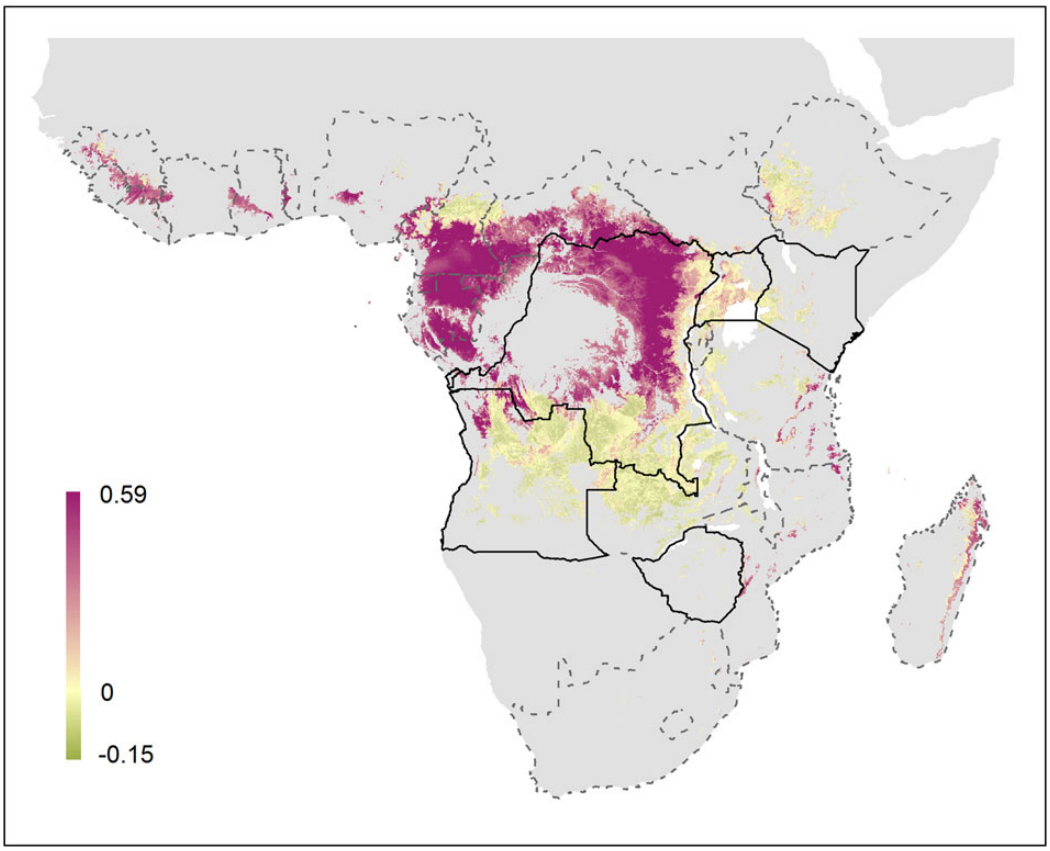


Figure 6. Difference between model predictions with animal data omitted. The difference between outputs for model 2 and model 1 are presented. Pixels in purple represent those regions predicted at higher risk in model 2; regions in green indicate areas where model 1 predicts higher risk. Yellow pixels represent areas with consistent probabilities. Pixels predicted not to be at-risk are in grey.

assessment of outbreak likelihood is wanted. While environmental triggers have been linked to outbreaks of MVD, equally important in determining outbreak potential is an understanding of the dynamics of the virus within reservoir populations, which has also been shown to be highly variable.[44,58,62] In attempting to predict outbreaks it is therefore crucial to understand the interplay between environmental factors, human pressures and reservoir host dynamics.[63]

The differences in areas predicted to be at-risk of infection by models 1 and 2 may in fact reflect the manner in which humans interact with bats. Since the majority of human outbreaks have arisen from contact in caves or other underground systems (rather than in the forest foraging sites of the bats, where animal infections have been reported), the risk map derived from model 1 could be a spatial representation of this transmission risk, as opposed to reflecting the broader distribution of infection in animal populations. Further infection surveys and ethnographic research can help to elucidate and map the risk of animal-human transmission within the at-risk region we have identified. Such an analysis would be particularly important in order to produce an absolute, rather than relative estimate of the likelihood of an outbreak in humans.

Nevertheless, while the true nature of risk to humans is likely to be a function of a variety of different factors, it is still important to gauge how and where potential spillover events could occur. The west African outbreak of EVD has shown that it is critical to understand the potential for such outbreaks in geographically distinct areas, and the subsequent need for other causes to be included in the differential diagnosis to facilitate rapid detection. This is all the more important where the potential causes of disease have varying potential for nosocomial transmission, as is the case with viral haemorrhagic fevers. Failure to rapidly and accurately diagnose these diseases can lead to uncontrolled chains of secondary transmission in certain scenarios.[64] Maps such as ours can therefore be used to shape clinical recommendations for diagnosing haemorrhagic fever cases presenting in hospital. MVD has seen a number of significant geographic translocation of cases, with individuals becoming symptomatic far from the original infection site.[9,10] The most recent outbreak of EVD in west Africa has demonstrated the role that both local and global connectivity can play in causing disease importation, and as connectivity continues to increase, the likelihood of widespread secondary cases occurring will also increase, particularly if infection reaches densely populated areas.[65–67] Accounting for a range of possible aetiological agents can therefore reduce the risk of further secondary transmission amongst humans in these settings.[3,26]

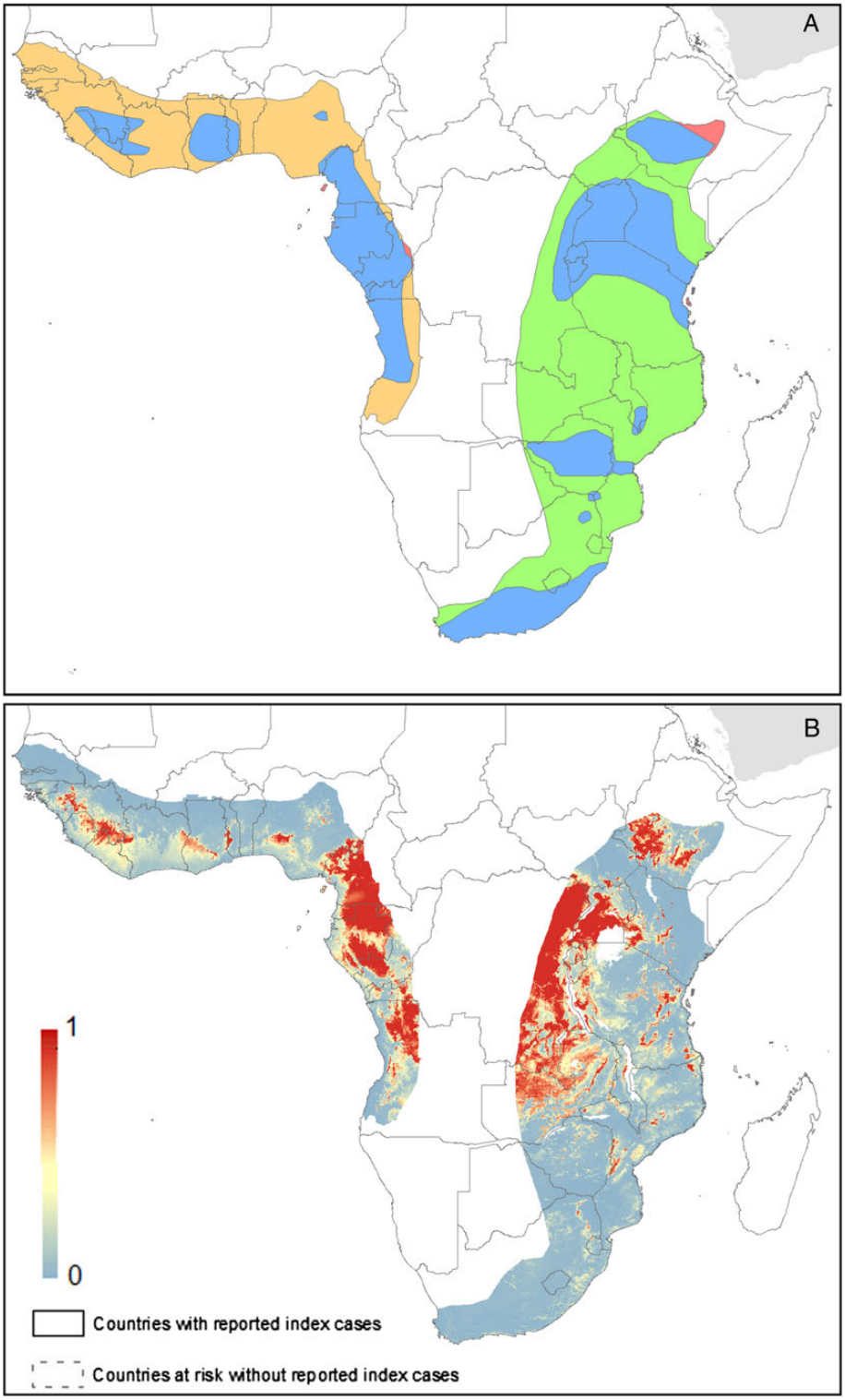


Figure 7. Expert opinion maps for the range of Egyptian rousettes. Panel A is derived from the IUCN and Kwiecinski et al.[33,60] Blue regions are those where both depict Egyptian rousette populations. Red areas are those only indicated in the IUCN dataset. Orange sections are where bats of the subspecies *R. aegyptiacus unicolor* are thought to be present; green shows the distribution of bats of the subspecies *R. aegyptiacus leachi*. Panel B shows the predicted values from model 2, masked by the bat layer.

Table 4. Identification of potential survey sites in at-risk countries

Country	Region	Latitude	Longitude
Angola	Northern Angola	−7.83	16.07
	Moxico	−12.28	23.29
Burundi	Western Burundi	−3.94	29.44
Cameroon	Central and Southern Cameroon	4.73	12.76
CAR	South CAR	5.95	18.91
Côte d'Ivoire	Dix-Huit Montagnes	7.62	−7.89
	Bondoukou	7.96	−3.03
DRC	Eastern Congo Basin	2.04	27.72
	Katanga	−10.70	26.23
	Kasai	−6.42	21.95
	Western DRC	−7.03	18.30
Equatorial Guinea	Kie-Ntem and Wele-Nzas Provinces	1.55	10.83
Ethiopia	Western Oromia	8.29	35.77
	Bale Mountains and Harenna Forest	6.07	39.14
Gabon	Northern Gabon	1.07	12.36
	Southern Gabon	−1.83	11.92
Ghana	Ashanti Uplands and Kwahu Plateau	6.99	−1.61
	Akwapim-Togo Range	8.01	0.51
Guinea	Moyenne Guinea	10.71	−12.47
	Guinea Highlands	8.23	−9.09
Kenya	Lake Victoria	−0.70	34.99
	Mount Kenya	−0.43	37.49
Liberia	Guinea Highlands–Wologizi and Wonegizi Ranges	8.15	−9.87
	Guinea Highlands–Nimba Range	7.41	−8.59
Madagascar	Tsaratanana and Marojejy Nature Reserves	−14.21	49.38
	Ambohijanahay Reserve	−18.38	45.45
Malawi	Lake Malawi	−11.89	33.90
Mozambique	Maeda Plateau	−11.44	39.63
	Mount Mabu	−16.30	36.54
	Inyanga Mountains	−19.64	33.29
Nigeria	Ekiti	7.76	5.09
	Gashaka Gumti National Park	7.36	11.57
ROC	Niari and Lekoumou	−2.92	13.19
	Sangha	1.56	14.53
	Likouala	3.17	16.92
Rwanda	Eastern Rwanda	−2.14	30.37
Sierra Leone	Loma Mountains	8.99	−10.98
South Africa	Woodbush Forest Reserve and Motlatse Canyon	−23.94	30.12
South Sudan	Imatong Mountains	3.95	32.64
	Yambio	4.57	28.29
Tanzania	Tanga	−5.94	37.57
	Mahale Mountain National Park	−6.49	30.30
	Kigoma	−4.49	30.64
	Morogoro	−9.58	35.60
Togo	Akwapim-Togo Range	7.59	0.79
Uganda	Great Lakes Region	0.18	31.01
Zambia	Northwestern Zambia	−12.80	24.76
	Northern Zambia	−9.49	29.23
	Muchinga Escarpment	−11.89	31.84
Zimbabwe	Chinoyi	−17.36	30.13

Sites are identified by country. The latitude and longitude represent the centroid of the proposed survey area.
CAR: Central African Republic; DRC: Democratic Republic of Congo; ROC: Republic of Congo.

Supplementary data

Supplementary data are available at Transactions Online (http://trstmh.oxfordjournals.org).

Authors' disclaimer: Funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Authors' contributions: SIH conceived the work; DMP, AM, OJB and MUGK collected the data; DMP and NG implemented the modelling; DJW generated the environmental covariate layers; ZH geopositioned the outbreak data and provided all maps. All authors read and approved the final manuscript. DMP is the guarantor of the paper.

Acknowledgments: We thank Dr. Jens Kuhn for his comments and Maria Devine, Joshua Longbottom and Freya Shearer for proofreading.

Funding: DMP is funded by a Sir Richard Southwood Graduate Scholarship from the Department of Zoology at the University of Oxford; NG is funded by a grant from the Bill & Melinda Gates Foundation [OPP1053338]; ZH is funded by the Bill & Melinda Gates Foundation [OPP1106023]; DJW is funded by the Bill & Melinda Gates Foundation [OPP1068048]; OJB is funded by a BBSRC studentship; MUGK is funded by the German Academic Exchange Service (DAAD) through a graduate scholarship; SIH is funded by a Senior Research Fellowship from the Wellcome Trust [095066] that also supports AM and a grant from the Bill & Melinda Gates Foundation [OPP1093011]. SIH would also like to acknowledge funding support from the RAPIDD program of the Science & Technology Directorate, Department of Homeland Security and the Fogarty International Center, National Institutes of Health.

Competing interests: None declared.

Ethical approval: Not required.

References

1 Slenczka WG. The Marburg virus outbreak of 1967 and subsequent episodes. Curr Top Microbiol Immunol 1999;235:49–75.

2 Siegert R, Shu H-L, Slenczka W et al. On the eitology of an unknown infection originating in monkeys [in German]. Dtsch med Wochenschr 1967;92:2341–3.

3 Conrad JL, Isaacson M, Smith EB et al. Epidemiologic investigation of Marburg virus disease, Southern Africa, 1975. Am J Trop Med Hyg 1978;27:1210–5.

4 Bausch DG, Nichol ST, Muyembe-Tamfum JJ et al. Marburg hemorrhagic fever associated with multiple genetic lineages of virus. N Engl J Med 2006;355:909–19.

5 Mylne A, Brady OJ, Huang Z et al. A comprehensive database of the geographic spread of past human Ebola outbreaks. Sci Data 2014;1:e140042.

6 Gire SK, Goba A, Andersen KG et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science 2014;345:1369–72.

7 Towner JS, Khristova ML, Sealy TK et al. Marburgvirus genomics and association with a large hemorrhagic fever outbreak in Angola. J Virol 2006;80:6497–516.

8 Adjemian J, Farnon EC, Tschioko F et al. Outbreak of Marburg hemorrhagic fever among miners in Kamwenge and Ibanda districts, Uganda, 2007. J Infect Dis 2011;204:S796–9.

9 Fujita N, Miller A, Miller G et al. Imported case of Marburg hemorrhagic fever - Colorado, 2008. MMWR Morb Mortal Wkly Rep 2009;58: 1377–81.

10 Timen A, Koopmans MPG, Vossen AC et al. Response to imported case of Marburg hemorrhagic fever, the Netherlands. Emerg Infect Dis 2009;15:1171–5.

11 Mbonye A, Wamala J, Winyi-Kaboyo et al. Repeated outbreaks of viral hemorrhagic fevers in Uganda. Afr Health Sci 2012;12:579–83.

12 WHO. Marburg virus disease - Uganda. Geneva: World Health Organization; 2014. http://www.who.int/csr/don/10-october-2014-marburg/en/ [accessed 15 December 2014].

13 Bermejo M, Rodriguez-Teijeiro JD, Illera G et al. Ebola outbreak killed 5000 gorillas. Science 2006;314:1564.

14 Formenty P, Boesch C, Wyers M et al. Ebola virus outbreak among wild chimpanzees living in a rain forest of Côte d'Ivoire. J Infect Dis 1999;179:S120–6.

15 Groseth A, Feldmann H, Strong JE. The ecology of Ebola virus. Trends Microbiol 2007;15:408–16.

16 Laminger F, Prinz A. Bats and other reservoir hosts of Filoviridae. Danger of epidemic on the African continent? a deductive literature analysis [in German]. Wien Klin Wochenschr 2010;122:19–30.

17 Towner JS, Amman BR, Sealy TK et al. Isolation of genetically diverse Marburg viruses from Egyptian fruit bats. PLoS Pathog 2009;5:e1000536.

18 Swanepoel R, Smit SB, Rollin PE et al. Studies of reservoir hosts for Marburg virus. Emerg Infect Dis 2007;13:1847–51.

19 Johnson ED, Johnson BK, Silverstein D et al. Characterization of a new Marburg virus isolated from a 1987 fatal case in Kenya. Arch Virol Suppl 1996;11:101–14.

20 Smith DH, Johnson KM, Saacson M et al. Marburg virus disease in Kenya. Lancet 1982;1:816–20.

21 Pigott DM, Golding N, Mylne A et al. Mapping the zoonotic niche of Ebola virus disease in Africa. eLife 2014;3:e04395.

22 Phillips SJ, Dudik M, Elith J et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl 2009;19:181–97.

23 Peterson AT, Bauer JT, Mills JN. Ecologic and geographic distribution of filovirus disease. Emerg Infect Dis 2004;10:40–7.

24 Peterson AT, Lash RR, Carroll DS, Johnson KM . Geographic potential for outbreaks of Marburg hemorrhagic fever. Am J Trop Med Hyg 2006;75:9–15.

25 Weiss DJ, Atkinson PM, Bhatt S et al. An effective approach for gap-filling continental scale remotely sensed time-series. ISPRS J Photogramm Remote Sens 2014;98:106–18.

26 Bannister B. Viral haemorrhagic fevers imported into non-endemic countries: risk assessment and management. Br Med Bull 2010;95: 193–225.

27 De'ath G. Boosted trees for ecological modeling and prediction. Ecology 2007;88:243–51.

28 Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol 2008;77:802–13.

29 Kuhn JH. Filoviruses: a compendium of 40 years of epidemiological, clinical, and laboratory studies. Vienna: Springer-Verlag; 2008.

30 Wan ZM, Li ZL. A physics-based algorithm for retrieving land-surface emissivity and temperature from EOS/MODIS data. IEEE Trans Geosci Remote Sens 1997;35:980–96.

31 ORNL DAAC Shuttle radar topography mission near-global digital elevation models. http://webmap.ornl.gov/wcsdown/wcsdown.jsp?dg_id=10008_1 [accessed 1 August 2014].

32 USF Geoportal and Data Repository Karst regions of the world. http://arcweb.forest.usf.edu/flex/KarstRegions/ [accessed 6 October 2014].

33 Kwiecinski GG, Griffiths TA. *Rousettus egyptiacus*. J Mammal 1999; 611:1–9.

34 Ford D, Williams P. Karst hydrogeology and geomorphology. Chichester: Wiley; 2007.

35 Hijmans RJ, Phillips S, Leathwick J, Elith J. Package dismo. R package. http://cran.r-project.org/web/packages/dismo/dismo.pdf [accessed 1 August 2014].

36 R Core Team R: a language and environment for statistical computing. www.R-project.org/ [accessed 1 November 2014].

37 Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol Evol 2012;3:327–38.

38 Hijmans RJ. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 2012;93:679–88.

39 Wenger SJ, Olden JD. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods Ecol Evol 2012;3:260–7.

40 WorldPop. WorldPop project. http://www.worldpop.org.uk/ [accessed 1 August 2014].

41 Linard C, Gilbert M, Snow RW et al. Population distribution, settlement patterns and accessibility across Africa in 2010. PLoS One 2012; 7:e31743.

42 Borchert M, Muyembe-Tamfum JJ, Colebunders R et al. A cluster of Marburg virus disease involving an infant. Trop Med Int Health 2002;7:902–6.

43 Jeffs B, Roddy P, Weatherill D et al. The Médecins Sans Frontières intervention in the Marburg hemorrhagic fever epidemic, Uige, Angola, 2005. I. Lessons learned in the hospital. J Infect Dis 2007;196: S154–61.

44 Roddy P, Weatherill D, Jeffs B et al. The Médecins Sans Frontières intervention in the Marburg hemorrhagic fever epidemic, Uige,

Angola, 2005. II. Lessons learned in the community. J Infect Dis 2007;196(Supp 2):S162–7.

45 Roddy P, Thomas SL, Jeffs B et al. Factors associated with Marburg hemorrhagic fever: analysis of patient data from Uige, Angola. J Infect Dis 2010;201:1909–18.

46 Henderson BE, Kissling RE, Williams MC et al. Epidemiological studies in Uganda relating to the Marburg agent. In: Martini GA, Siegert R, editors. Marburg virus disease. Berlin: Springer-Verlag; 1971.

47 Pourrut X, Souris M, Towner JS et al. Large serological survey showing cocirculation of Ebola and Marburg viruses in Gabonese bat populations, and a high seroprevalence of both viruses in *Rousettus aegyptiacus*. BMC Infect Dis 2009;9:159.

48 Gonzalez JP, Nakoune E, Slenczka W et al. Ebola and Marburg virus antibody prevalence in selected populations of the Central African Republic. Microbes Infect 2000;2:39–44.

49 Tomori O, Fabiyi A, Sorungbe A et al. Viral hemorrhagic fever antibodies in Nigerian populations. Am J Trop Med Hyg 1998;38:407–10.

50 Lash RR, Brunsell NA, Peterson AT. Spatiotemporal environmental triggers of Ebola and Marburg virus transmission. Geocarto Int 2008;23:451–66.

51 Schoepp RJ, Rossi CA, Khan SH et al. Undiagnosed acute viral febrile illnesses, Sierra Leone. Emerg Infect Dis 2014;20:1176–82.

52 Neppert J, Gohring S, Schneider W, Wernet P. No evidence of LAV infection in the Republic of Liberia, Africa, in the year 1973. Blut 1986;53:115–7.

53 Bergmann JF, Bouree P. Haemorrhagic fever due to Ebola. A study of 1517 sera from Cameroon. Med Maladies Infect 1982;12:638–42.

54 Saluzzo JF, Gonzalez JP, Georges AJ. Anti-Marburg antibodies in the Republic of Central Africa. Ann Virol 1982;133:129–31.

55 Johnson ED, Gonzalez JP, Georges A. Haemorrhagic fever virus activity in equatorial Africa: distribution and prevalence of filovirus reactive antibody in the Central African Republic. Trans R Soc Trop Med Hyg 1993;87:530–5.

56 Woodruff PWR, Morrill JC, Burans JP et al. A study of viral and rickettsial exposure and causes of fever in Juba, southern Sudan. Trans R Soc Trop Med Hyg 1988;82:761–6.

57 Gire SK, Stremlau M, Andersen KG et al. Emerging disease or diagnosis? Science 2012;338:750–2.

58 Amman BR, Carroll SA, Reed ZD et al. Seasonal pulses of Marburg virus circulation in juvenile *Rousettus aegyptiacus* bats coincide with periods of increased risk of human infection. PLoS Pathog 2012;8: e1002877.

59 Amman BR, Nyakarahuka L, McElroy AK et al. Marburgvirus resurgence in Kitaka mine bat population after extermination attempts, Uganda. Emerg Infect Dis 2014;20:1761–4.

60 Schipper J, Chanson JS, Chiozza F et al. The status of the world's land and marine mammals: diversity, threat, and knowledge. Science 2008;322:225–30.

61 Wood JLN, Leach M, Waldman L et al. A framework for the study of zoonotic disease emergence and its drivers: spillover of bat pathogens as a case study. Philos Trans R Soc Lond B Biol Sci 2012; 367:2881–92.

62 Hayman DT. Biannual birth pulses allow filoviruses to persist in bat populations. Proc R Soc Lond B Biol Sci 2015;282:pii: 20142591.

63 Plowright RK, Eby P, Hudson PJ et al. Ecological dynamics of emerging bat virus spillover. Proc Biol Sci 2015;282:20142124.

64 Farrar JJ, Piot P. The Ebola emergency - immediate action, ongoing strategy. N Engl J Med 2014;371:1545–6.

65 Bogoch II, Creatore MI, Cetron MS et al. Assessment of the potential for international dissemination of Ebola virus via commercial air travel during the 2014 west African outbreak. Lancet 2015;385:29–35.

66 Snyder RE, Marlow MA, Riley LW. Ebola in urban slums: the elephant in the room. Lancet Glob Health 2014;2:e685.

67 Wesolowski A, Buckee CO, Bengtsson L et al. Commentary: containing the Ebola outbreak - the potential and challenge of mobile network data. PLoS Curr 2014;6:pii: ecurrents.outbreaks.0177e7fcf52217b8b 634376e2f3efc5e.

Chapter 6

Mapping a tick-borne zoonotic disease: Crimean-Congo haemorrhagic fever.

Crimean-Congo haemorrhagic fever shares similar epidemiology to Ebola and Marburg virus diseases in having these two spatially distinct processes: the first, involving transmission between reservoir hosts and tick vectors with occasional spillover into humans and the second being subsequent transmission amongst humans. This chapter combines boosted regression trees with an evidence consensus in order to assess the distribution of CCHF across the Old World. Areas of highest risk are identified, as well as regions that would benefit the most from prospective surveillance. This work has been published in *Transactions of the Royal Society of Tropical Medicine* and is included here in its final form. The additional information referred to in this chapter is included in the Appendix of this thesis. Also included in the Appendix is a more detailed paper discussing the occurrence data generation and curation protocols.

Trans R Soc Trop Med Hyg 2015; **109**: 503–513
doi:10.1093/trstmh/trv050 Advance Access publication 4 July 2015

ORIGINAL ARTICLE

The global distribution of Crimean-Congo hemorrhagic fever

Jane P. Messina[a,*], David M. Pigott[a], Nick Golding[b], Kirsten A. Duda[a], John S. Brownstein[c], Daniel J. Weiss[a], Harry Gibson[a], Timothy P. Robinson[d], Marius Gilbert[e,f], G. R. William Wint[a], Patricia A. Nuttall[a], Peter W. Gething[a], Monica F. Myers[a], Dylan B. George[g] and Simon I. Hay[b,g,h]

[a]Department of Zoology, University of Oxford, Oxford, UK; [b]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; [c]Department of Pediatrics, Harvard Medical School and Children's Hospital Informatics Program, Boston Children's Hospital, Boston, MA, USA; [d]Livestock Systems and Environment (LSE), International Livestock Research Institute (ILRI),Nairobi, Kenya; [e]Biological Control and Spatial Ecology, Université Libre de Bruxelles, Brussels, Belgium; [f]Fonds National de la Recherche Scientifique, Brussels, Belgium; [g]Fogarty International Center, National Institutes of Health, Bethesda, MD, USA; [h]Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, USA

*Corresponding author: Tel: +44 (0) 1865 271 137; E-mail: jane.messina@zoo.ox.ac.uk

Received 2 April 2015; revised 19 May 2015; accepted 20 May 2015

Background: Crimean-Congo hemorrhagic fever (CCHF) is a tick-borne infection caused by a virus (CCHFV) from the Bunyaviridae family. Domestic and wild vertebrates are asymptomatic reservoirs for the virus, putting animal handlers, slaughter-house workers and agricultural labourers at highest risk in endemic areas, with secondary transmission possible through contact with infected blood and other bodily fluids. Human infection is characterized by severe symptoms that often result in death. While it is known that CCHFV transmission is limited to Africa, Asia and Europe, definitive global extents and risk patterns within these limits have not been well described.

Methods: We used an exhaustive database of human CCHF occurrence records and a niche modeling framework to map the global distribution of risk for human CCHF occurrence.

Results: A greater proportion of shrub or grass land cover was the most important contributor to our model, which predicts highest levels of risk around the Black Sea, Turkey, and some parts of central Asia. Sub-Saharan Africa shows more focalized areas of risk throughout the Sahel and the Cape region.

Conclusions: These new risk maps provide a valuable starting point for understanding the zoonotic niche of CCHF, its extent and the risk it poses to humans.

Keywords: Crimean-Congo hemorrhagic fever, Crimean-Congo hemorrhagic fever virus, Ecological niche modeling, Infectious diseases, Tick-borne diseases, Vector-borne diseases

Introduction

Crimean-Congo hemorrhagic fever (CCHF) is a tick-borne viral (Nairovirus, family Bunyaviridae) infection first identified in the Crimean region in 1944.[1,2] It was subsequently shown to be the same virus as that causing similar hemorrhagic disease outbreaks in the Congo basin, giving the virus its current name.[3,4] CCHF is one of the most widely distributed arboviral diseases in the world, ranging from southern Russia and the Black Sea region to the southern tip of Africa.[4] The disease is considered as 'emerging' across the globe, with many countries reporting new infections in humans in recent decades, including Albania (2001),[5] Turkey (2002)[6] and Georgia (2009).[7] In some regions, human CCHF infection has also been recently reported after long periods of absence, for example in south-western Russia[8] and Central Africa.[9]

While no apparent disease manifestation occurs in animals,[10] both wild and domesticated animals represent an important link in the disease transmission cycle, acting as reservoirs for continued tick re-infection (Figure 1). Many tick species have been associated with CCHF virus (CCHFV), but members of the genus *Hyalomma* are considered the primary vectors and are the most common ticks known to transmit the virus to humans.[11] These ticks are adapted to hot and dry or semiarid environments, and are found in many parts of Africa, Asia, and Europe.[1,12–15]

Infection of humans is a comparatively rare event, with those living or working in close proximity to livestock (particularly cattle,

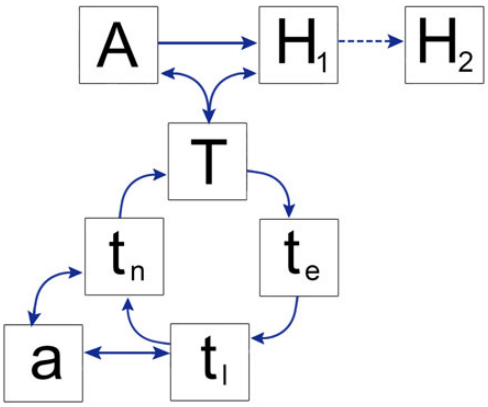


Figure 1. Transmission cycle of Crimean-Congo hemorrhagic fever virus (CCHFV) where t_e, t_l, and t_n represent the eggs, larvae, and nymphs of competent tick vectors, respectively. Nymphs (t_n) transmit CCHFV to small mammals and birds (a), whereas transmission to ruminants and other large animals (A) is by adult ticks (T). Primary human infections (H_1) occur as a result of being directly bitten by adult ticks or squashing ticks between the fingers (T), or through contact with the blood of infected animals, usually livestock (A). The comparatively rarer human-to-human transmission (represented by the dashed line from H_1 to H_2) is typically between infected individuals and healthcare workers or close relatives having exposed to their infectious blood and/or bodily fluids.[88]

sheep and goats) or tick vector habitats being particularly at risk of infectious tick bites, and those working in animal slaughterhouses being at risk for blood-borne exposure.[16,17] Human-to-human transmission is possible, typically amongst healthcare workers or close relatives having close contact and exposure to infectious blood or bodily fluids of those infected with CCHFV.[16,18,19] There is no widely available safe and effective vaccine against CCHF, although a recombinant CCHFV vaccine candidate has shown good in vivo efficacy.[20] Currently, treatment for the potentially fatal disease remains largely supportive.[21,22] Personal protective measures such as the use of pyrethroid acaricides and wearing protective clothing are important, but generally there is little knowledge about these measures in areas where current levels of risk are ill-defined.[11,23]

Currently, spatial analyses of CCHF are few in number compared to those for many other diseases and even other tick-borne viruses.[24,25] Still, several studies have elucidated the chief drivers of CCHF geographic distribution patterns. Strong correlations have been found in Turkey and Bulgaria between CCHF risk and suitable environments for *Hyalomma* ticks, including grass and shrub cover, as well as forested land fragmented by agricultural or shrub cover.[26–29] Non-irrigated agricultural land cover (e.g., pasture and rangeland) has also been found to be associated with CCHF incidence in Turkey and Greece.[26,27,30] The CCHFV infection rate in livestock was found to be a strong positive predictor of CCHF incidence in humans in Iran and Mauritania,[31,32] although in Bulgaria where vaccination coverage is high amongst at-risk populations (e.g., veterinarians and farm workers), livestock density was not a significant driver of CCHF incidence in humans.[29] Climate indicators have also been found as important predictors of CCHF risk. Areas regularly experiencing long periods of low rainfall and humidity were associated with increased occurrence of CCHF in Iran and Senegal,[33,34] and higher temperatures were indicators of CCHF occurrence in Turkey, Bulgaria, and Iran.[28,29,33]

Existing global distribution maps of CCHF are largely in the form of national-level maps of vector presence or reported human cases, such as that provided by WHO.[35] Here we draw upon the findings of several of the country-specific studies to model risk for CCHF infection in humans at a global scale using an ecological niche modeling approach. This approach enables us to better identify at-risk areas by using environmental correlations found in areas of good CCHF reporting to predict risk in those areas where less is known about transmission of the disease. While a preliminary CCHF risk map was produced using a similar statistical approach in the past,[36] the current paper offers a more recent and refined global geographic estimation of the distribution of CCHF. This was made possible by the addition of new data for the locations of disease occurrence, an evidence-based consensus layer for background data sampling, and high-resolution environmental layers alongside newer methodologies. Because the geographic distribution of CCHF is taken into account when patients' travel histories are considered during differential diagnosis of hemorrhagic fevers,[37] an up-to-date and high-resolution map of the global distribution of the disease is essential. As the maps we provide define regions not only where CCHF has been reported but also where transmission is possible, successful identification of both locally acquired and imported cases[38–40] may be expedited, therefore reducing the likelihood of further secondary human-to-human transmission. The recent Ebola virus outbreak in West Africa has highlighted the critical nature of such considerations.[41] In several African countries, the risk of CCHF is poorly defined meaning infection with CCHFV is more likely to go undiagnosed or unreported in this region. An improved understanding of the geographic extents of CCHF and the true level of risk within these extents is vital for increasing awareness about the disease, advocating for improved individual protection from *Hyalomma* tick bites, and promoting safe practices for slaughterhouse and healthcare workers. Finally, this work contributes to a wider initiative to better map the ecological niche of several of the viral hemorrhagic fevers occurring in Africa which not only pose the risk of zoonotic transmission, but also of secondary nosocomial and community-level transmission.[37,42,43]

Methods

We used boosted regression trees (BRT), a method for modeling species distributions, to create maps of environmental suitability for CCHF occurrence. Our particular approach has been successfully employed in similar disease mapping efforts,[44,45] and requires the generation of: a layer assessing the strength of evidence for CCHF presence or absence, termed evidence consensus, at a national and sometimes sub-national level[46] (Figure 2A); a comprehensive database of the locations of CCHF reports in humans (Figure 2B); and a suite of globally gridded environmental and socioeconomic covariates known or hypothesised to affect CCHF transmission. The output map presents a probabilistic surface of environmental suitability of CCHF occurrence ('CCHF risk') within its global geographic extents at a 5 km×5 km spatial resolution (Figure 2C).

Definitive extents

We carried out a process consisting of four components (described below) to evaluate the certainty of presence or absence of CCHF for each country and certain select sub-national

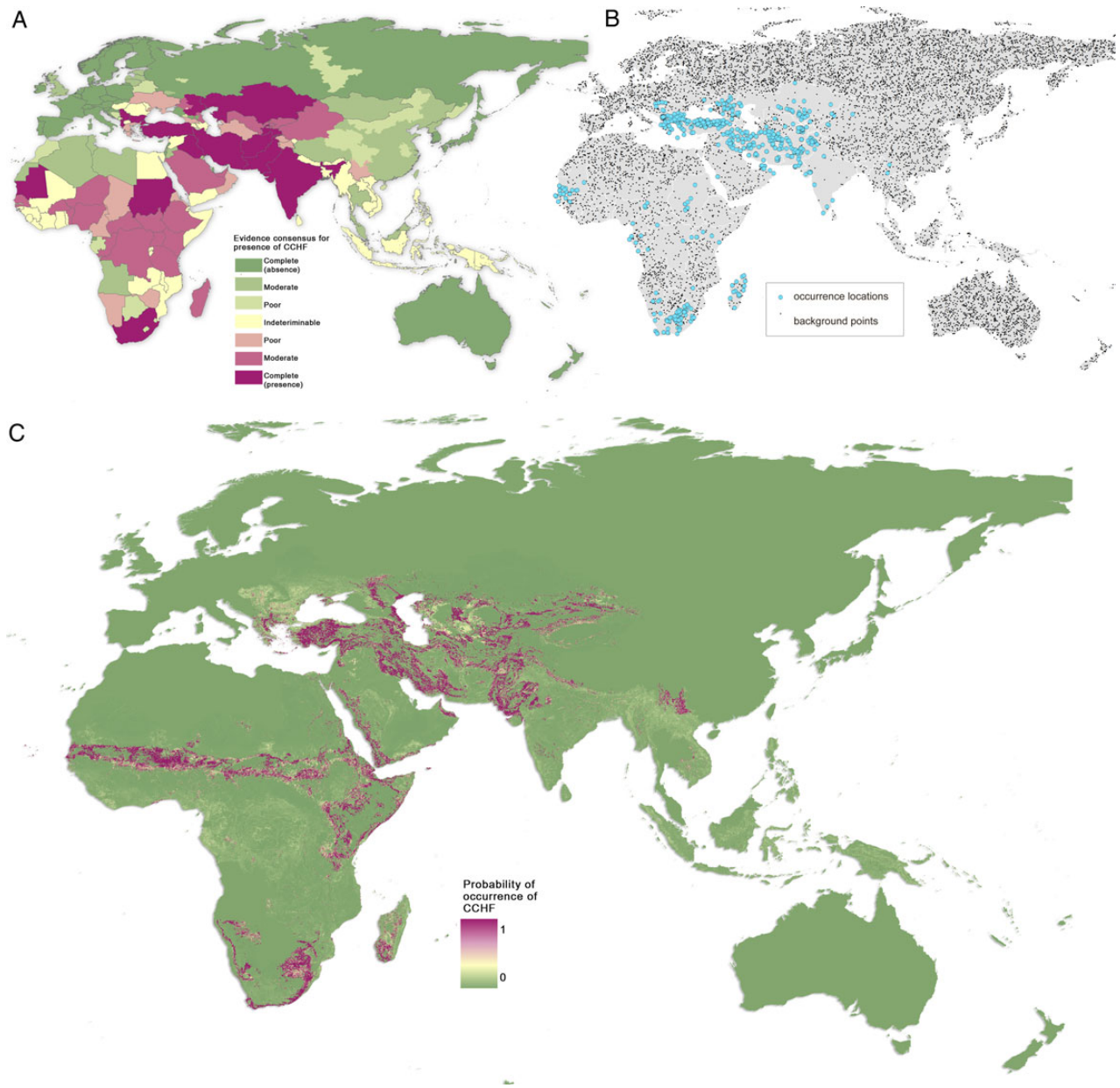


Figure 2. Maps of A: definitive extents as determined by evidence consensus; B: recorded occurrence and generated background points used in the BRT procedure; and C: probability of occurrence of Crimean-Congo haemorrhagic fever (CCHF). A: shows the consensus on CCHF presence globally, ranging from dark green (complete consensus on absence) to purple (complete consensus on presence). Countries in yellow are those where evidence was inconclusive or contradictory for CCHF presence. B: shows the probability of CCHF occurrence in humans. Areas in purple are those most suitable for transmission, with areas in green least suitable.

regions at the edges of its distribution. The methodology used to generate the definitive extents of CCHF was adapted from that used for dengue and is termed evidence consensus.[45,46] The information used to determine the final score for each country or subnational region is provided in Supplementary Table 1.

Health reporting organization evidence (max ±3)

Evidence from WHO[35] and the Global Infectious Diseases and Epidemiology Online Network (GIDEON)[47] was used. WHO places each nation into one of five categories of evidence for CCHF: absent; *Hyalomma* ticks present; CCHF virological or serological evidence and vector present; 5–49 CCHF cases reported per year; and 50 or more CCHF cases reported per year. GIDEON listed each country as either endemic or non-endemic; if the country was not listed, the entry was recorded as unspecified. Quantitative scores for each unique permutation are laid out in Table 1.

Peer-reviewed evidence (max +6)

We conducted a country-specific search on PubMed and Web of Science using the terms '[country] CCHF' or '[country] Crimean

Table 1. Derivation of quantitative scores for health-reporting organization evidence

GIDEON	WHO	Score
Endemic	50+ CCHF cases reported per year	+3
	5–49 CCHF cases reported per year	+2.5
	CCHF virological or serological evidence and vector present	+2
	Hyalomma tick vector present	+1.5
	Absent	0
Unspecified	50+ CCHF cases reported per year	+2
	5–49 CCHF cases reported per year	+1.5
	CCHF virological or serological evidence and vector present	+1
	Hyalomma tick vector present	+0.5
	Absent	−2
Non-endemic	50+ CCHF cases reported per year	−0.5
	5–49 CCHF cases reported per year	−1
	CCHF virological or serological evidence and vector present	−1.5
	Hyalomma tick vector present	−2
	Absent	−3

CCHF: Crimean-Congo hemorrhagic fever; GIDEON: Global Infectious Diseases and Epidemiology Online Network.

Congo Hemorrhagic Fever' or '[country] Crimean Hemorrhagic Fever' or '[country] Congo Hemorrhagic Fever'. Reported cases in the literature were evaluated based upon their contemporariness and diagnostic accuracy. Contemporariness was evaluated in three categories: 2006–2013=3, 1998–2005=2, 1997 and earlier=1. Different diagnostic techniques were scored in a similar banding system, with PCR techniques, or genotyping achieving 3 points, 2 awarded for the use of IgM- and IgG-based ELISA or other serological techniques, and with 1 point for cases that were just reported or referred to an unspecified 'laboratory diagnosis'.

Case data (max ±6)

Data on reported outbreaks of CCHF, with a threshold of five cases, were obtained from GIDEON datasets and the literature. Outbreaks above this threshold were scored for their contemporariness with outbreaks before 1998=+2, 1998–2005=+4, 2006–2013=+6. If there were no reported outbreaks, healthcare expenditure as reported by the WHO was used as a proxy for diagnostic capacity in an attempt to differentiate genuine absence or sporadic cases from inability to adequately diagnose CCHF. Expenditure was stratified annually per capita at average US$ exchanges rates (2011 US$ WHO Health Statistics) with three categories defined: HE Low (<US$100), HE Medium (US$100 to <US$500) and HE High (>US$500). These measures are used as a proxy for the quality of healthcare reporting, as it is unlikely that vast numbers of CCHF infections have gone undetected in countries where healthcare expenditure is high (e.g., in northern European countries).

Supplementary evidence (max ±3)

In cases where contradictory evidence led to uncertainty in the presence or absence of CCHF, supplementary evidence was supplied. Here, seroepidemiological surveys, or the presence of CCHFV in ticks or livestock were evaluated and scored. Country-specific scoring is outlined in Supplementary Table 1 where appropriate.

Assembly of the occurrence database

An occurrence database comprising point (e.g., town or city) or polygon (e.g., county or province) locations of confirmed CCHF infection presence was compiled from peer-reviewed literature, Genbank records, and HealthMap alerts.[48,49] A literature search was conducted on PubMed and Web of Science using the terms 'CCHF' or 'Crimean Congo Hemorrhagic Fever' or 'Crimean Hemorrhagic Fever' or 'Congo Hemorrhagic Fever'. The same terms were used in our Genbank search. An occurrence was defined as one or more laboratory or clinically confirmed infection(s) of CCHF occurring at a unique location (the same administrative area or 5 km×5 km pixel for points) within one calendar year. All occurrence data underwent manual review and quality control to ensure information fidelity and precise geo-positioning. Reports of autochthonous (locally transmitted) cases or outbreaks were entered as an occurrence within the country in which transmission occurred. If imported cases were reported with information about the site of infection, they were geo-positioned to the country where transmission occurred. If imported cases were reported with no information about the site of contagion, they were not entered into the database. In addition, polygons greater than one square degree in area at the equator were removed from the database, as their inclusion in niche modeling would introduce a large amount of bias. This database has been made publicly available for download.[50]

Explanatory covariates

We assembled gridded global data (5 km×5 km) for a set of five explanatory covariates. The covariates were chosen based on factors known or hypothesised to contribute to suitability for CCHF transmission based upon the national-level studies described in the introduction. These included annual mean precipitation interpolated from global meteorological stations[51] and mean land surface temperature derived from NASA's moderate resolution imaging spectrometer (MODIS) imagery,[52] intended to capture the generally warm and arid climate zones where CCHFV is transmitted. We also included a 5 km×5 km resolution measure of the mean annual Enhanced Vegetation Index (EVI; also from MODIS)[53] (computed from the original 1 km×1 km data set), as well as the SD of this mean, which is intended to serve as a proxy for landscape diversity and habitat fragmentation. All these surfaces were parsed through a gap-filling algorithm prior to inclusion in the analysis.[54] The proportion of each 5 km×5 km grid cell covered by shrub or grass land cover types was also derived using the MODIS MCD12Q1 dataset which was originally obtained at 1 km×1 km resolution. To do this, we computed the proportion of 1 km×1 km grid cells within the larger 5 km×5 km cells that were classified as either grass, open shrub or closed shrub. The International Geosphere-Biosphere Programme

(IGBP) land cover classification scheme was utilized.[55] No covariate grids were shown to be adversely affected by multicollinearity and were standardised to ensure identical spatial resolution, extent, and boundaries. Maps for each covariate are provided in Supplementary Figure 1.

Modeling risk for CCHF occurrence

We used a boosted regression tree (BRT) approach to establish a multivariate empirical relationship between the probability of CCHF presence and the environmental conditions (as determined by the set of covariates described above) sampled at each occurrence location. This method combines regression trees[56] with gradient boosting,[57] whereby an initial regression tree is fitted and iteratively improved upon in a forward stagewise manner (boosting) by minimising the variation in the response not explained by the model at each iteration. It has been shown to fit complicated response functions efficiently, while guarding against over-fitting, and has thus been applied in the past for vector and disease distribution mapping.[42,44,45,58–61]

A large proportion of the point-level records in our occurrence database were geo-positioned to urban areas where cases are more likely to have been diagnosed rather than acquired (based on the ecology of CCHF tick vectors). Because of this, there is uncertainty about where in the vicinity CCHFV transmission actually took place. In an effort to reduce spatial bias, we thus assumed that when a location was recorded as a point, transmission could have taken place anywhere within the larger administrative corresponding to the FAO's Admin2-level Global Administrative Unit Layers (GAUL),[62] which typically represents counties or municipalities. We then calculated the mean of all covariates within these polygons and for any records which were originally recorded as polygons. While this approach reduced bias toward urban areas in our models, it is limited in its assumption that Admin2 units are correct for sampling environmental correlates of CCHF and in its disregard for variation in covariate values within polygons.

Like other ecological niche-mapping approaches, the BRT models require not only presence data but also background data defining areas of potentially unsuitable environmental conditions at unsampled locations, since data on absence of disease are rarely reported. No consensus approach has been developed to optimise the generation of background data and we therefore created an evidence-based probabilistic framework for generating pseudo-data. To represent the environmental conditions in locations where the disease has not been reported, 10 000 background points[63–65] were randomly generated and weighted based on a continuous raster surface derived from the national (and sometimes sub-national) CCHF evidence consensus scores. As such, more background points were located in areas with high consensus on absence.

To increase the robustness of model predictions and quantify model uncertainty, we fitted an ensemble[66] of 100 BRT models to separate bootstraps of the data. We then evaluated the central tendency as the mean across all 100 BRT models (see Bhatt et al.[44]). Each of the 100 individual models was fitted using the gbm.step subroutine in the dismo package in the R statistical programming environment.[67] All other tuning parameters of the algorithm were held at their default values (tree complexity=4, learning rate=0.005, bag fraction=0.75, step size=10, cross-validation folds=10). One 5 km×5 km pixel was randomly selected from within each polygon for each individual model in order to account for the environmental uncertainty associated with imprecise geographic data. In order to improve the weighting capacity of each of the 100 models, weightings were applied to the background dataset such that the sum of the weighted background data equalled the weighted sum of the occurrence records.[68] Each of the 100 models predicts environmental suitability on a continuous scale from 0 to 1, with a final prediction map then being generated by calculating the mean prediction across all models for each 5 km×5 km pixel. Cross-validation was applied to each model, whereby 10 subsets of the data comprising 10% of the presence and background observations were assessed based on their ability to predict the distribution of the other 90% of records using the mean area under the curve (AUC) statistic. This AUC value was then averaged across the 10 sub-models and finally across all 100 models in the ensemble in order to derive an overall estimate of goodness-of-fit. Additionally, to avoid AUC inflation due to spatial sorting bias, a pairwise distance sampling procedure was used,[69] resulting in a final AUC which is lower than would be returned by standard procedures but which gives a more realistic quantification of the model's ability to extrapolate predictions to new regions.[70]

Results

In total, 1721 occurrence records were included in our final dataset after performing all quality control procedures. These included 1470 county or district-level occurrences and 251 province-level occurrences spanning 1953 to 2012. We assumed that any recorded location of CCHF occurrence, regardless of the date of the record, represented an environment permissible for the disease.

The evidence consensus map (Figure 2A) showed 47 countries to have an indeterminate status in terms of CCHF presence or absence (scores between −15 and +15), as well as certain parts of China and Russia. Those with particularly poor evidence (score of zero) are mostly located in sub-Saharan Africa, but also include Cambodia, Laos, Myanmar, Vietnam, Nepal and Arunachal Pradesh in Asia, as well as Azerbaijan, Bhutan, Yemen, Moldova and Macedonia in the region spanning eastern Europe to central Asia. More information is needed about the possible occurrence of CCHF in these places, as well as the presence of any ticks that have been proven as competent vectors of CCHFV. Yunnan province in China is classified as having some evidence for CCHF presence due to CCHFV seropositivity having been found in humans,[71] although the overall evidence is weak since no cases of human disease have been reported in the province. According to our evidence consensus measure, the five countries currently having the strongest evidence for CCHF presence are Turkey, Iran, Afghanistan, Tajikistan, and Pakistan.

The average of the ensemble of BRT models predicted high levels of risk for CCHF in the Black Sea region and some parts of central Asia, with more focalized areas of risk being found in the Sahel and Cape regions of Africa (Figure 2C). The countries with the largest areas of high risk for CCHF occurrence are Turkey, Iran, Romania, Moldova and Ukraine, with some parts of southwest Russia, Syria, Iraq and central Asia demonstrating high probabilities of occurrence as well (Figure 3). Although the evidence

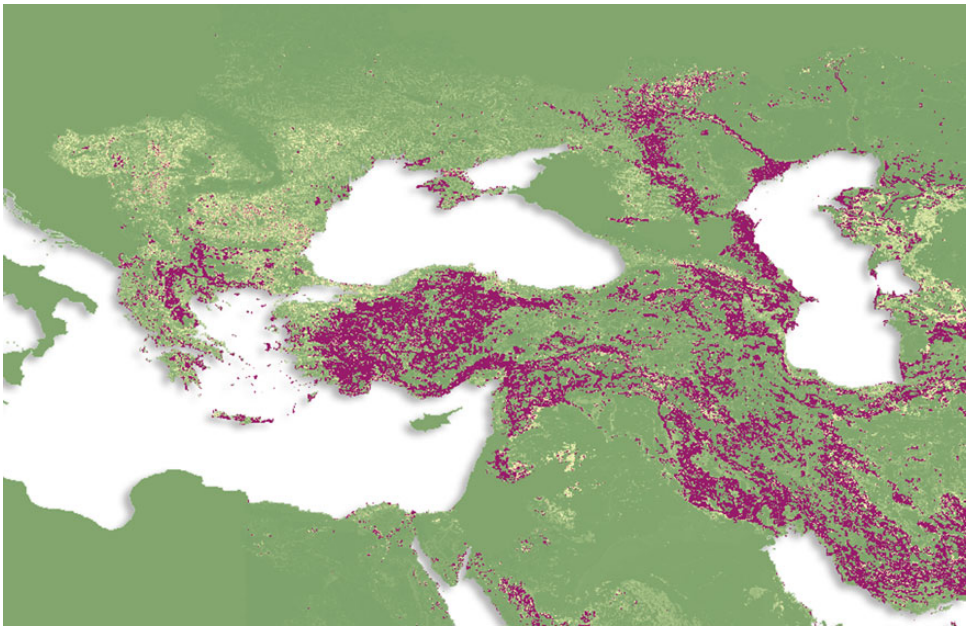


Figure 3. The probability of occurrence of Crimean-Congo haemorrhagic fever (CCHF) in the Balkans region. Areas in purple are those most suitable for transmission, with areas in green least suitable.

consensus for CCHF presence is strong for many countries in sub-Saharan Africa, our model predicts that areas with the highest probability of occurrence are overall much smaller in area and more irregularly distributed in this region than in the Black Sea region (see Figure 4). However, this may be an artefact of the small number of occurrence observations available for CCHF in Africa, as can be seen in Figure 2B (see also Burt and Swanepoel[72]). For example, while CCHF occurrences have been reported in central Africa and therefore consensus on presence is relatively high for some countries such as the Democratic Republic of the Congo, the areas of predicted suitability within this region are actually quite sparsely distributed.

Our models showed CCHF risk to be particularly determined by the proportion of grass and shrub land cover within a 5 km×5 km grid cell, contributing 62% to the ensemble of models. Land surface temperature had the second most important effect, contributing 19% to the models, followed by the standard deviation of mean EVI (8%), mean annual precipitation (6%), and mean EVI (5%). Effect plots for each covariate are provided in Supplementary Figure 2. Validation statistics indicated high predictive performance of the BRT ensemble mean map with area under the receiver operating characteristic (AUC) of 0.923 (±0.051 SD).

Discussion

Temperature, precipitation and moisture indices have been found to be important drivers of CCHF infection in past studies[29,34]; however, in this model, land cover types were more important in predicting the global ecological niche for CCHF transmission to humans. When considered together, these land cover types are generally reflective of the environments where wild and/or domestic herbivore CCHF hosts exist and enable tick survival and virus circulation. While shrub and grass land cover types are effective for delineating global risk patterns, variations in climate and moisture availability may be more important in predicting heterogeneity in finer-scale prevalence patterns. It is also well understood that those living or working in close proximity to livestock are at the greatest risk for infection with CCHF, yet livestock population layers had minimal influence in predicting human disease occurrence. It is, therefore, possible that the abundance of CCHF livestock reservoirs is more important in driving prevalence patterns within endemic areas rather than as a predictor of overall transmission at a global scale.

We have strived to be exhaustive in the assembly of contemporary data on CCHF occurrence and have applied new modeling approaches to maximise the predictive power of these data. The consideration of a range of biogeographic factors alongside disease occurrence information enabled us to infer risk in areas with uncertainty about CCHF transmission, and the resulting map thus presents sub-national refinements of the distribution of risk without relying on national-level reporting systems. Furthermore, the use of an evidence consensus layer allowed us to limit our predictions to within the current CCHF transmission extents in Africa, Europe and Asia. Due to phylogeographic evidence that CCHFV genotypes tend to vary between Africa and Eurasia more so than within each region,[73] we did carry out an additional model which distinguished between these two regions. This distinction had a minimal effect and was thus excluded in our final modeling procedures. Such a finding, however, does highlight that while particular CCHFV strains may vary between Africa and Eurasia, the ecological determinants of its zoonotic niche are consistent between the two regions.

It is possible to highlight areas where surveillance is most needed. The map in Figure 5 was created by defining 'high-risk' pixels as those for which the modelled probability of occurrence (Figure 2C) was greater than or equal to 0.5. We then selected

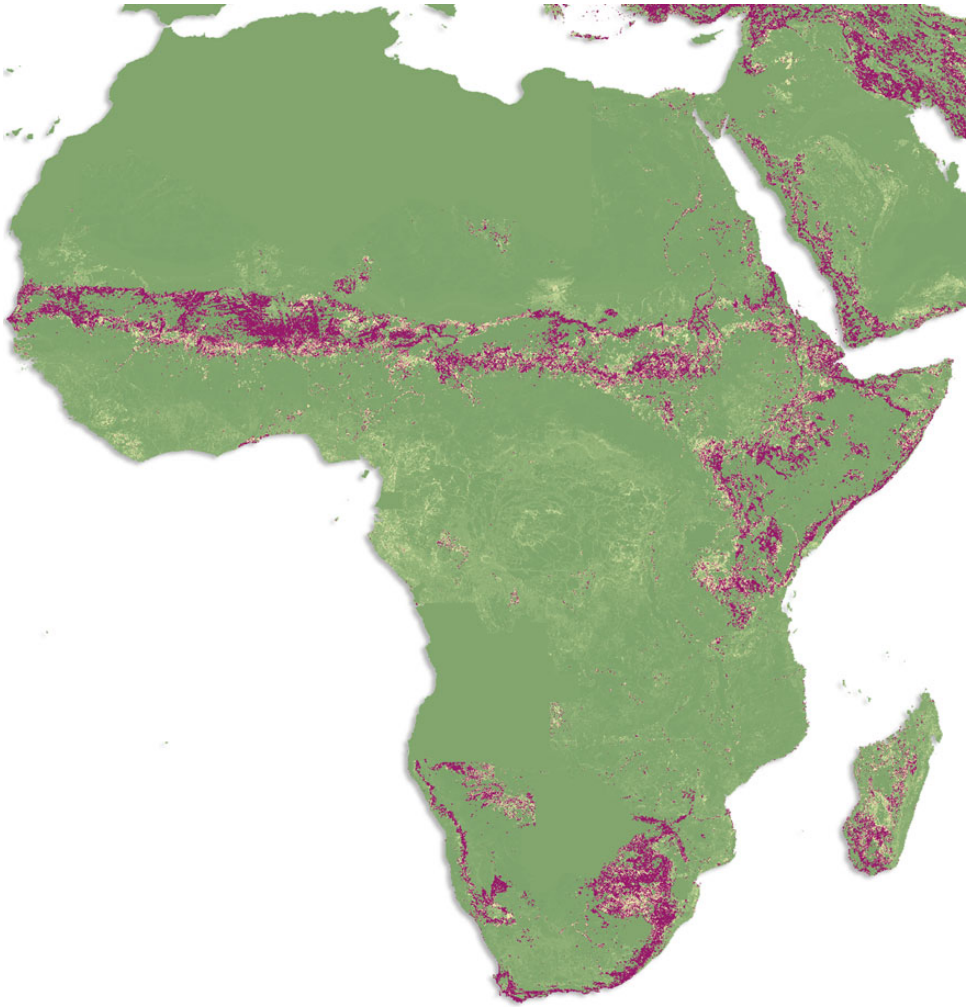


Figure 4. Probability of occurrence of Crimean-Congo haemorrhagic fever in Africa. Areas in purple are those most suitable for transmission, with areas in green least suitable.

those high-risk pixels which fell inside countries with low evidence consensus scores (between −25 and +25). The result is a visualization of areas where humans are predicted to be at potential risk for CCHF yet where evidence is most lacking and thus where surveillance is a priority. While there are many countries in Africa and Eurasia that have small and sparse areas in need of surveillance, several countries in Figure 5 stand out as having large and more continuous areas in need of surveillance. These include Mali, Chad, Somalia, Djibouti and Zimbabwe in Africa; Syria, Macedonia, Azerbaijian, Armenia, Turkmenistan and Yemen in Eastern Europe and Central Asia; and Kashmir, Nepal and the Yunnan province of China in southern Asia.

For all viral hemorrhagic fevers, public health education about disease vectors or reservoirs and behavioural risk factors for secondary infection is required in endemic areas, and many lessons can be learned from past Marburg, Ebola and Lassa fever outbreaks, for example.[79–81] Specific to CCHF is the need for agricultural workers and others working with animals to apply acaricidal repellent to exposed skin and clothing, as well as to wear protective clothing and gloves while dealing with the blood or body fluids of livestock. Crimean-Congo hemorrhagic fever does, however, share with these other viral hemorrhagic fevers the risk for infection through contact with other infected humans, and failure to avoid such contact has led to multiple nosocomial CCHF outbreaks in recent years.[82–86] Such outbreaks indicate that awareness is lacking in many parts of the world about the presence of CCHF, which represents the largest barrier to the rapid diagnosis required for the prevention of nosocomial outbreaks. Healthcare workers in at-risk areas must not only understand the intensive care needs of infected patients, but also understand the precautions required to prevent occupational exposure to CCHF when treating these patients and handling infectious laboratory specimens.[87] Preventing primary transmission of CCHF to humans also requires strategic allocation of vector and/or reservoir control resources, which are limited in many of the settings we have predicted to be at risk for CCHF, particularly in Africa.

While we have emphasized that better information is still needed in several regions, our map provides a baseline for monitoring change in the global distribution of CCHF going forward. Further cartographic refinements are required in order to help

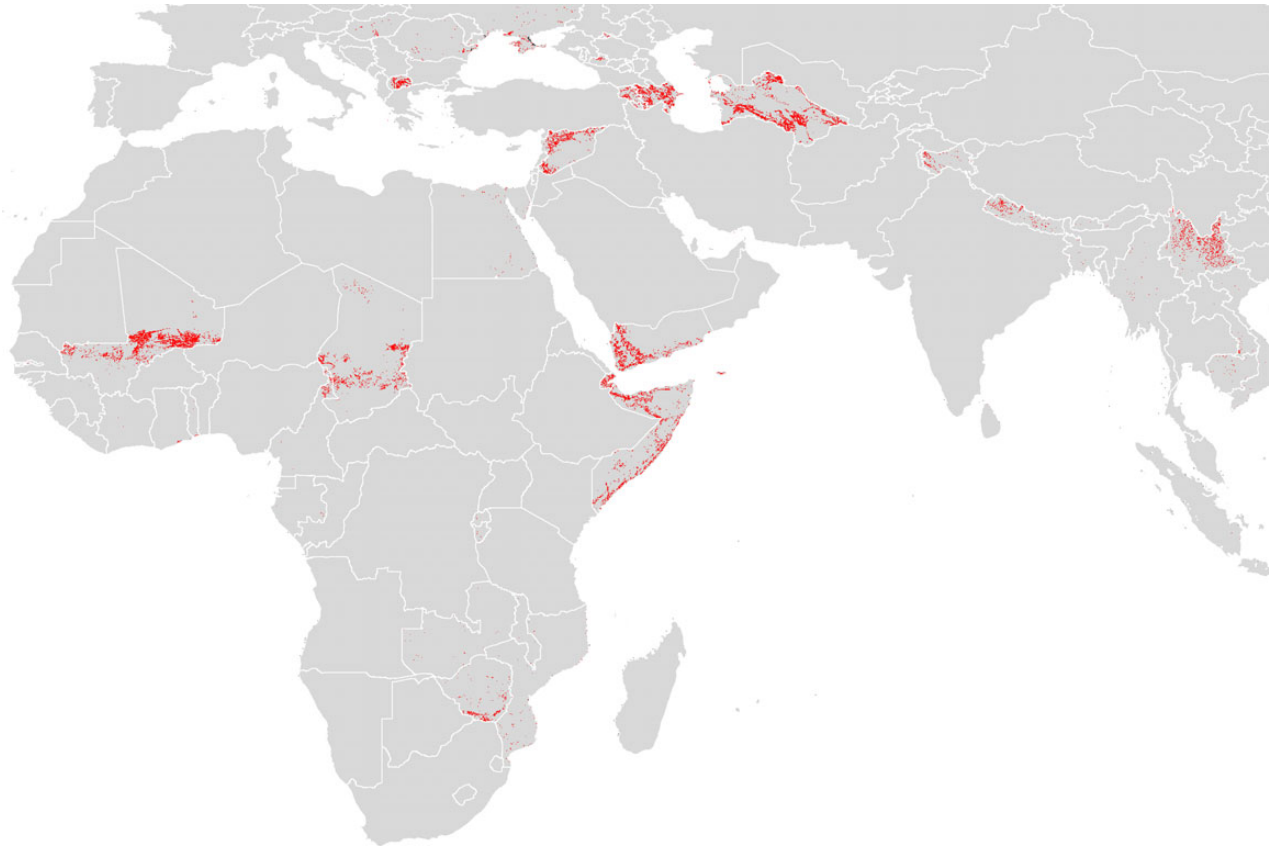


Figure 5. Areas in need of surveillance for Crimean-Congo haemorrhagic fever (CCHF). Red colouring shows areas where our models have predicted high risk for CCHF ($\geq$0.5), but which lie within countries having low evidence consensus (between -25 and $+25$) on disease presence or absence. The red areas thus signify places most in need of CCHF surveillance.

differentiate endemic from epidemic-prone areas, particularly in Africa where there is less certainty about the presence or absence of CCHF overall. The occurrence database used in creating this map can be updated with new information as necessary, and a stronger global evidence base, particularly for the regions highlighted in Figure 5, would improve the accuracy of future iterations of this mapping procedure. The ultimate aim is to provide more useful information in evaluating control and prevention strategies and their impact, and a refined map of the global risk for CCHF is a first step towards reaching this goal.

Although our resulting map is an improvement on those which have previously been produced, the abundance of information about CCHF occurrence still comprises a weaker evidence base than that available, for example, for *Plasmodium falciparum*[74] and *P. vivax* malaria,[75] for which a large amount of prevalence information is available. Records of disease occurrence do not easily translate to population-level metrics, and so as databases of CCHF prevalence become more widespread, future approaches should focus on using geostatistical methods to assess risk,[76] as with many other neglected tropical diseases.[77,78]

Conclusions

In this study, we have refined the map of the geographic extents of CCHF and the level of risk within these extents using an exhaustive assembly of known records of CCHF occurrence worldwide and an ecological niche modeling framework. We hope that our improved estimate of the spatial distribution of CCHF will serve as a starting point for a wider discussion about the global impact of CCHF. Not only can it encourage public health awareness in areas we have defined as having a high probability of risk, but it can also guide targeted distribution should an effective vaccine or treatment become available.

Supplementary data

Supplementary data are available at Transactions online (http://trstmh.oxfordjournals.org/).

Authors' contributions: SIH conceived the research and the modeling approach. JPM drafted the manuscript with editorial input from DMP and SIH. PWG and MFM were involved with the initial database creation. JSB provided HealthMap occurrence data and advice on its provenance. DMP and KAD carried out evidence consensus computation and quality-checked all HealthMap points. TPR, MG, and WGRW provided expertise about Crimean-Congo hemorrhagic fever distribution and the distribution of tick vectors. DJW and HG produced environmental covariate layers for input into the models. NG advised on statistical modeling procedures. JPM executed all modeling created the final maps and wrote the manuscript. All authors discussed the results and contributed to revisions of the final

manuscript. All authors read and approved the final manuscript. SIH is the guarantor of this paper.

Funding: SIH is funded by a Senior Research Fellowship from the Wellcome Trust [#095066], which also supports AM and KAD, and a grant from the Bill & Melinda Gates Foundation [#OPP1093011]. JPM and SIH received funding from the International Research Consortium on Dengue Risk Assessment Management and Surveillance (IDAMS, [#21803], http://www.idams.eu). SIH would also like to acknowledge funding support from the RAPIDD program of the Science & Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health. PWG is a Career Development Fellow [#K00669X] jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement and receives support from the Bill & Melinda Gates Foundation [#OPP1068048, #OPP1106023]. These grants also support DJW and HG. JSB is supported by a research grant from the National Library of Medicine, the National Institutes of Health [5R01LM010812-05]. DMP is funded by a Sir Richard Southwood Graduate Scholarship from the Department of Zoology at the University of Oxford. NG is funded by a grant from the Bill & Melinda Gates Foundation [#OPP1053338]. TPR is funded by the CGIAR Research Programmes on the Humidtropics, Climate Change, Agriculture and Food Security (CCAFS) and Agriculture for Nutrition and Health (A4NH).

Competing interests: None declared.

Ethical approval: Not required.

References

1 Hoogstraal H. The epidemiology of tick-borne Crimean-Congo hemorrhagic fever in Asia, Europe, and Africa. J Med Entomol 1979;15:307–417.

2 Whitehouse CA. Crimean-Congo hemorrhagic fever. Antiviral Res 2004;64:145–60.

3 Han N, Rayner S. Epidemiology and mutational analysis of global strains of Crimean-Congo haemorrhagic fever virus. Virol Sin 2011;26:229–44.

4 Ergonul O. Crimean-Congo haemorrhagic fever. Lancet Infect Dis 2006;6:203–14.

5 Papa A, Bino S, Llagami A et al. Crimean-Congo hemorrhagic fever in Albania, 2001. Eur J Clin Microbiol Infect Dis 2002;21:603–6.

6 Karti SS, Odabasi Z, Korten V et al. Crimean-Congo hemorrhagic fever in Turkey. Emerg Infect Dis 2004;10:1379–84.

7 Zakhashvili K, Tsertsvadze N, Chikviladze T et al. Crimean-Congo hemorrhagic fever in man, Republic of Georgia, 2009. Emerg Infect Dis 2010;16(8):1326–8.

8 Maltezou HC, Andonova L, Andraghetti R et al. Crimean-Congo hemorrhagic fever in Europe: current situation calls for preparedness. Euro Surveill 2010;15:19504.

9 Grard G, Drexler JF, Fair J et al. Re-emergence of Crimean-Congo hemorrhagic fever virus in Central Africa. PLoS Negl Trop Dis 2011;5:e1350.

10 Appannanavar SB, Mishra B. An update on Crimean Congo hemorrhagic fever. J Glob Infect Dis 2011;3:285–92.

11 Maltezou HC, Papa A. Crimean-Congo hemorrhagic fever: risk for emergence of new endemic foci in Europe? Travel Med Infect Dis 2010;8:139–43.

12 Lutomiah J, Musila L, Makio A et al. Ticks and tick-borne viruses from livestock hosts in arid and semiarid regions of the eastern and northeastern parts of Kenya. J Med Entomol 2014;51:269–77.

13 Okello-Onen J, Tukahirwa EM, Perry BD et al. Population dynamics of ticks on indigenous cattle in a pastoral dry to semi-arid rangeland zone of Uganda. Exp Appl Acarol 1999;23:79–88.

14 Aktas M, Altay K, Dumanli N. A molecular survey of bovine Theileria parasites among apparently healthy cattle and with a note on the distribution of ticks in eastern Turkey. Vet Parasitol 2006;138:179–85.

15 Walker AR, Bouattour A, Camicas J-L et al. Ticks of Domestic Animals in Africa: a Guide to Identification of Species. Edinburgh: Bioscience Reports; 2003.

16 Vorou R, Pierroutsakos IN, Maltezou HC. Crimean-Congo hemorrhagic fever. Curr Opin Infect Dis 2007;20:495–500.

17 Deyde VM, Khristova ML, Rollin PE et al. Crimean-Congo hemorrhagic fever virus genomics and global diversity. J Virol 2006;80:8834–42.

18 Izadi S, Salehi M, Holakouie-Naieni K et al. The risk of transmission of Crimean-Congo hemorrhagic fever virus from human cases to first-degree relatives. Jpn J Infect Dis 2008;61:494–6.

19 Maltezou HC, Maltezos E, Papa A. Contact tracing and serosurvey among healthcare workers exposed to Crimean-Congo haemorrhagic fever in Greece. Scand J Infect Dis 2009;41:877–80.

20 Buttigieg KR, Dowall SD, Findlay-Wilson S et al. A novel vaccine against Crimean-Congo haemorrhagic fever protects 100% of animals against lethal challenge in a mouse model. PLoS One 2014;9:e91516.

21 Leblebicioglu H, Bodur H, Dokuzoguz B et al. Case management and supportive treatment for patients with Crimean-Congo hemorrhagic fever. Vector Borne Zoonotic Dis 2012;12:805–11.

22 Yilmaz R, Kundak AA, Ozer S et al. Successful treatment of severe Crimean-Congo hemorrhagic fever with supportive measures without ribavirin and hypothermia. J Clin Virol 2009;44:181–2.

23 Rahnavardi M, Rajaeinejad M, Pourmalek F et al. Knowledge and attitude toward Crimean-Congo haemorrhagic fever in occupationally at-risk Iranian healthcare workers. J Hosp Infect 2008;69:77–85.

24 Rogers DJ, Randolph SE. Climate change and vector-borne diseases. Adv Parasitol 2006;62:345–81.

25 Hay SI, Battle KE, Pigott DM et al. Global mapping of infectious disease. Philos Trans R Soc Lond B Biol Sci 2013;368:20120250.

26 Terzi O, Sisman A, Canbaz S et al. An evaluation of spatial distribution of Crimean-Congo hemorrhagic fever with geographical information systems (GIS), in Samsun and Amasya region. J Med Plants Res 2011;5:848–54.

27 Estrada-Pena A, Vatansever Z, Gargili A et al. The trend towards habitat fragmentation is the key factor driving the spread of Crimean-Congo haemorrhagic fever. Epidemiol Infect 2010;138:1194–203.

28 Estrada-Pena A, Zatansever Z, Gargili A et al. Modeling the spatial distribution of Crimean-Congo hemorrhagic fever outbreaks in Turkey. Vector Borne Zoonotic Dis 2007;7:667–78.

29 Vescio FM, Busani L, Mughini-Gras L et al. Environmental correlates of Crimean-Congo haemorrhagic fever incidence in Bulgaria. BMC Public Health 2012;12:1116.

30 Sargianou M, Panos G, Tsatsaris A et al. Crimean-Congo hemorrhagic fever: seroprevalence and risk factors among humans in Achaia, western Greece. Int J Infect Dis 2013;17:1160–5.

31 Mostafavi E, Haghdoost A, Khakifirouz S et al. Spatial analysis of Crimean Congo hemorrhagic fever in Iran. Am J Trop Med Hyg 2013;89:1135–41.

32 Gonzalez JP, LeGuenno B, Guillaud M et al. A fatal case of Crimean-Congo haemorrhagic fever in Mauritania: virological and serological evidence suggesting epidemic transmission. Trans R Soc Trop Med Hyg 1990;84:573–6.

33 Ansari H, Shahbaz B, Izadi S et al. Crimean-Congo hemorrhagic fever and its relationship with climate factors in southeast Iran: a 13-year experience. J Infect Dev Ctries 2014;8:749–57.

34 Wilson ML, LeGuenno B, Guillaud M et al. Distribution of Crimean-Congo hemorrhagic fever viral antibody in Senegal: environmental and vectorial correlates. Am J Trop Med Hyg 1990;43:557–66.

35 WHO. Global Alert and Response (GAR). Crimean-Congo haemorrhagic fever (CCHF). Geneva: World Health Organization. http://www.who.int/csr/disease/crimean_congoHF [accessed 7 July 2014].

36 Randolph SE, Rogers DJ. Ecology of tick-borne disease and the role of climate. In: Ergonul O, Whitehouse CAs (editors). Crimean-Congo hemorrhagic fever. The Netherlands: Springer; 2010, p. 167–86.

37 Bannister B. Viral haemorrhagic fevers imported into non-endemic countries: risk assessment and management. Br Med Bull 2010;95:193–225.

38 Lumley S, Atkinson B, Dowall S et al. Non-fatal case of Crimean-Congo haemorrhagic fever imported into the United Kingdom (ex Bulgaria), June 2014. Euro Surveill 2014;19.pii: 20864.

39 Atkinson B, Latham J, Chamberlain J et al. Sequencing and phylogenetic characterisation of a fatal Crimean-Congo haemorrhagic fever case imported into the United Kingdom, October 2012. Euro Surveill 2012;17.pii:20327.

40 Chamberlain J, Atkinson B, Logue CH et al. Genome sequence of ex-Afghanistan Crimean-Congo hemorrhagic fever virus SCT strain, from an imported United Kingdom case in October 2012. Genome Announc 2013;1.pii: e00161-13.

41 Frieden TR, Damon I, Bell BP et al. Ebola 2014 – new challenges, new global response and responsibility. N Engl J Med 2014;371:1177–80.

42 Pigott DM, Golding N, Mylne A et al. Mapping the zoonotic niche of Ebola virus disease in Africa. eLife 2014;3:e04395.

43 Pigott DM, Golding N, Mylne A et al. Mapping the zoonotic niche of Marburg virus disease in Africa. Trans R Soc Trop Med Hyg 2015;109:366–78.

44 Bhatt S, Gething PW, Brady OJ et al. The global distribution and burden of dengue. Nature 2013;496:504–7.

45 Pigott DM, Bhatt S, Golding N et al. Global distribution maps of the leishmaniases. eLife 2014;3.

46 Brady OJ, Gething PW, Bhatt S et al. Refining the global spatial limits of dengue virus transmission by evidence–based consensus. PLoS Negl Trop Dis 2012;6:e1760.

47 GIDEON. The world's premier global infectious diseases database. Los Angeles: GIDEON Informatics. http://www.gideononline.com [accessed 12 June 2013].

48 Brownstein JS, Freifeld CC, Reis BY et al. Surveillance sans frontieres: internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med 2008;5:e151.

49 Freifeld CC, Mandl KD, Reis BY et al. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc 2008;15:150–7.

50 Messina JP, Pigott DM, Duda KA et al. A global compendium of human Crimean-Congo haemorrhagic fever virus occurrence. Sci Data 2015:2:150016.

51 Hijmans RJ, Cameron SE, Parra JL et al. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 2005;25:1965–78.

52 Wan ZM, Zhang YL, Zhan QC et al. The MODIS Land-Surface Temperature products for regional environmental monitoring and global change studies. IGARSS 2002: International Geoscience and Remote Sensing Symposium and 24th Canadian Symposium on Remote Sensing, Vols I–VI, Proceedings 2002:3683–5.

53 Lin QH. Enhanced vegetation index using moderate resolution imaging spectroradiometers. 5th International Congress on Image and Signal Processing (CISP) 2012:1043–6.

54 Weiss DJ, Atkinson PM, Bhatt S et al. An effective approach for gap–filling continental scale remotely sensed time–series. ISPRS J Photogramm Remote Sens 2014;98:106–18.

55 McIver DK, Friedl MA. Local estimation of land cover classification quality using machine learning methods. In: Geoscience and Remote Sensing Symposium, 2000. Proceedings. IGARSS 2000. 24–28 July 2000; Honolulu, HI. IEEE 2000 International 7, 3063–5.

56 Breiman L. Classification and Regression Trees. Boca Ratón: Chapman & Hall/CRC Press LLC; 1984.

57 Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–232.

58 Gilbert M, Golding N, Zhou H et al. Predicting the risk of avian influenza A H7N9 infection in live-poultry markets across Asia. Nat Commun 2014;5:4116.

59 Elith J, Graham CH, Anderson RP et al. Novel methods improve prediction of species' distributions from occurrence data. Ecography 2006;29:129–51.

60 Stevens KB, Pfeiffer DU. Spatial modelling of disease using data- and knowledge-driven approaches. Spat Spatiotemporal Epidemiol 2011;2:125–33.

61 Hay SI, Sinka ME, Okara RM et al. Developing global maps of the dominant anopheles vectors of human malaria. PLoS Med 2010;7:e1000209.

62 FAO. Global Administrative Unit Layers (GAUL). 2012. http://data.fao.org/ref/f7e7adb0-88fd-11da-a88f-000d939bc5d8.html?version=1.0 [accessed 20 February 2015].

63 Chefaoui RM, Lobo JM; Assessing the effects of pseudo-absences on predictive distribution model performance. Ecol Modell 2008;210:478–86.

64 Stokland JN, Halvorsen R, Stoa B. Species distribution modelling: effect of design and sample size of pseudo–absence observations. Ecol Modell 2011;222:1800–9.

65 Lobo JM, Tognelli MF. Exploring the effects of quantity and location of pseudo–absences and sampling biases on the performance of distribution models with limited point occurrence data. J Nature Conserv 2011;19:1–7.

66 Araújo MB, New M; Ensemble forecasting of species distributions. Trends Ecol Evol 2007;22:42–7.

67 Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol 2008;77:802–13.

68 Barbet-Massin M, Jiguet F, Albert CH et al. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol Evol 2012;3:327–38.

69 Hijmans RJ. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 2012;93:679–88.

70 Wenger SJ, Olden JD. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods Ecol Evol 2012;3:260–7.

71 Xia H, Li P, Yang J et al. Epidemiological survey of Crimean-Congo hemorrhagic fever virus in Yunnan, China, 2008. Int J Infect Dis 2011;15:e459–63.

72 Burt FJ, Swanepoel R. Molecular epidemiology of African and Asian Crimean-Congo haemorrhagic fever isolates. Epidemiol Infect 2005;133:659–66.

73 Burt FJ, Paweska JT, Ashkettle B et al. Genetic relationship in southern African Crimean-Congo haemorrhagic fever virus isolates: evidence for occurrence of reassortment. Epidemiol Infect 2009;137:1302–8.

74 Gething PW, Patil AP, Smith DL et al. A new world malaria map: *Plasmodium falciparum* endemicity in 2010. Malar J 2011;10:378.

75 Gething PW, Elyazar IRF, Moyes CM et al. A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. PLoS Negl Trop Dis 2012;6:e1814.

76 Anders KL, Hay SI; Lessons from malaria control to help meet the rising challenge of dengue. Lancet Infect Dis 2012;12:977–84.

77 Brooker S, Hotez PJ, Bundy DAP. The global atlas of helminth infection: mapping the way forward in neglected tropical disease control. PLoS Negl Trop Dis 2010;4:e779.

78 Smith J, Brooker S, Haddad D et al. Mapping the global distribution of trachoma: an updated atlas. Am J Trop Med Hyg 2010;83:160.

79 Allaranga Y, Kone ML, Formenty P et al. Lessons learned during active epidemiological surveillance of Ebola and Marburg viral hemorrhagic fever epidemics in Africa. East Afr J Public Health 2010;7:30–6.

80 Mccormick JB. Epidemiology and control of Lassa fever. Curr Top Microbiol Immunol 1987;134:69–78.

81 Fisher-Hoch SP, Tomori O, Nasidi A et al. Review of cases of nosocomial Lassa fever in Nigeria – the high price of poor medical practice. BMJ 1995;311:857–9.

82 Aradaib IE, Erickson BR, Mustafa ME et al. Nosocomial outbreak of Crimean-Congo hemorrhagic fever, Sudan. Emerg Infect Dis 2010;16:837–9.

83 Naderi H, Sheybani F, Bojdi A et al. Fatal nosocomial spread of Crimean-Congo hemorrhagic Fever with very short incubation period. Am J Trop Med Hyg 2013;88:469–71.

84 Naderi HR, Sarvghad MR, Bojdy A et al. Nosocomial outbreak of Crimean-Congo haemorrhagic fever. Epidemiol Infect 2011;139:862–6.

85 Patel AK, Patel KK, Mehta M et al. First Crimean-Congo hemorrhagic fever outbreak in India. J Assoc Physicians India 2011;59:585–9.

86 Mardani M, Keshtkar-Jahromi M, Ataie B et al. Crimean-Congo hemorrhagic fever virus as a nosocomial pathogen in Iran. Am J Trop Med Hyg 2009;81:675–8.

87 Dixon J, Ong E. Clinical management of viral hemorrhagic fevers. In: Current Treatment Options in Infectious Disease. Vol 6: Viral Infections. 2014, p. 245–55

88 Bente DA, Alimonti JB, Shieh WJ et al. Pathogenesis and immune response of Crimean-Congo hemorrhagic fever virus in a STAT-1 knockout mouse model. J Virol 2010;84:11089–100.

Chapter 7

Mapping a rodent-borne zoonotic disease: Lassa fever.

Lassa fever is a rodent-borne viral haemorrhagic fever that shares similar transmission patterns to the previous VHFs discussed in Chapter 4, 5 and 6. Using species distribution models, this paper completes the suite of African viral haemorrhagic fevers that are contagious amongst humans. This work has been published in *Transactions of the Royal Society of Tropical Medicine* and is included here in its final form. The additional information referred to in this chapter is included in the Appendix of this thesis.

Trans R Soc Trop Med Hyg 2015; **109**: 483–492
doi:10.1093/trstmh/trv047 Advance Access publication 17 June 2015

ORIGINAL ARTICLE

Mapping the zoonotic niche of Lassa fever in Africa

Adrian Q. N. Mylne[a,*], David M. Pigott[b,*], Joshua Longbottom[a], Freya Shearer[a], Kirsten A. Duda[b], Jane P. Messina[b], Daniel J. Weiss[b], Catherine L. Moyes[a], Nick Golding[a] and Simon I. Hay[a,c,d]

[a]*Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK;* [b]*Department of Zoology, University of Oxford, Oxford, UK;* [c]*Institute for Health Metrics and Evaluation, University of Washington, Seattle, USA;* [d]*Fogarty International Center, National Institutes of Health, Bethesda, Maryland, USA*

*Corresponding authors: E-mail: adrian.mylne@well.ox.ac.uk; david.pigott@zoo.ox.ac.uk

Received 17 April 2015; revised 28 May 2015; accepted 29 May 2015

Background: Lassa fever is a viral haemorrhagic illness responsible for disease outbreaks across West Africa. It is a zoonosis, with the primary reservoir species identified as the Natal multimammate mouse, *Mastomys natalensis.* The host is distributed across sub-Saharan Africa while the virus' range appears to be restricted to West Africa. The majority of infections result from interactions between the animal reservoir and human populations, although secondary transmission between humans can occur, particularly in hospital settings.

Methods: Using a species distribution model, the locations of confirmed human and animal infections with Lassa virus (LASV) were used to generate a probabilistic surface of zoonotic transmission potential across sub-Saharan Africa.

Results: Our results predict that 37.7 million people in 14 countries, across much of West Africa, live in areas where conditions are suitable for zoonotic transmission of LASV. Four of these countries, where at-risk populations are predicted, have yet to report any cases of Lassa fever.

Conclusions: These maps act as a spatial guide for future surveillance activities to better characterise the geographical distribution of the disease and understand the anthropological, virological and zoological interactions necessary for viral transmission. Combining this zoonotic niche map with detailed patient travel histories can aid differential diagnoses of febrile illnesses, enabling a more rapid response in providing care and reducing the risk of onward transmission.

Keywords: Boosted regression trees, Lassa fever, LASV, *Mastomys natalensis*, Species distribution models, Viral haemorrhagic fever

Introduction

In 1969, a previously undescribed disease with haemorrhagic symptoms was reported in two missionary nurses in the town of Lassa, Nigeria.[1] The virus was subsequently identified as a novel member of the *Arenaviridae* family and named Lassa virus (*Lassa mammarenavirus*). Since then, this virus, which causes Lassa fever, has been reported in many West African countries with notable outbreaks in Guinea, Liberia, Nigeria and Sierra Leone. Evidence also indicates that the virus may have been present in West Africa long before the first detection date in 1969.[2]

The high seroprevelance for Lassa virus (LASV) specific antibodies in those Guinean (55%), Nigerian (21.3%) and Sierra Leonean (52%) populations tested indicates that most infections are mild or asymptomatic and do not require hospitalisation.[3–6] This is supported by findings that more than 80% of persons who developed antibodies did not report a recent febrile illness.[3] Overall mortality may be less than 5%, once mild or asymptomatic infections in the community are taken into account.[3]

LASV causes an acute viral haemorrhagic illness in a small fraction of those infected. The incubation period of Lassa fever ranges from 7–21 days with a wide range of clinical symptoms including headache, myalgia, fever, vascular bleeding and seizures as well as encephalopathy and oedema of the face and neck.[1,7–9] Human-to-human transmission is possible through direct contact with infected blood or bodily fluid, although chains of transmission are often limited,[10–13] especially if simple barrier nursing methods are implemented.[14–17] A graphical representation of the epidemiology of Lassa virus transmission is presented in Figure 1.

The low capacity for transmission between humans suggests a reservoir host is responsible for maintaining viral circulation in the environment. While LASV has been isolated from a number of rodent species, the majority of evidence implicates the Natal

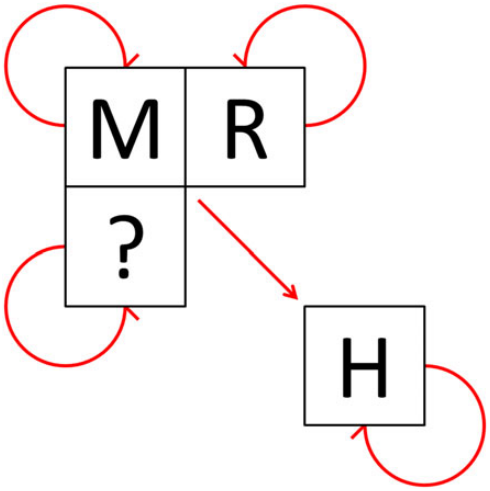


Figure 1. The epidemiology of Lassa virus transmission in West Africa. 'M' represents the suspected Natal multimammate mouse reservoir, *Mastomys natalensis*. 'R' represents other rodent species in which Lassa virus antibodies have been isolated including *Mastomys erythroleucus*, *Rattus rattus* and *Mus minutoides*.[3,18,19] 'H' represents humans. The question mark indicates other potential species. Arrows indicate confirmed or suspected transmission cycles or infection routes. This figure is available in black and white in print and in color at Transactions online.

multimammate mouse, *Mastomys natalensis*, as the primary reservoir species.[3,18,19] Seroprevalence has been reported to be as high as 60–80% in *M. natalensis* populations.[3,6] Human infection can result from exposure to rodent excreta, hypothesised to be aerosolised, or through hunting/butchering of infected rodents for consumption.[3,20]

Recent studies[21] suggest that outbreaks are largely fuelled by independent zoonotic transmission events from infected rodent hosts, whilst approximately 20% of cases result from secondary human-to-human transmission, typically through super-spreader events in hospital settings. This is in contrast to other blood-borne haemorrhagic viruses such as Ebola virus, for which human transmission chains are relatively long and fuel the majority of the outbreak.[22]

Lassa fever represents an importation risk across Africa and beyond,[23] with a number of international cases reported.[13,24–28] The ability of LASV to not only cause local outbreaks but also spread internationally provides a strong rationale for providing high-resolution mapping of Lassa fever risk across West Africa to aid differential diagnosis of viral haemorrhagic fevers.[23,29] This paper aims to identify populations living in areas of environmental suitability for zoonotic transmission of LASV. We improve upon previous modelling efforts[30] by including more recent outbreak data, animal infection records, and more refined environmental covariates.[31] We also take advantage of recent advances in species distribution modelling techniques.[32,33] This completes a series of papers that map the zoonotic niche of key viral haemorrhagic fevers, namely Ebola virus disease,[34,35] Marburg virus disease[36] and Crimean-Congo haemorrhagic fever,[37] using comparable data collection and modelling methods. As a whole, this set of studies improves our understanding of the contemporary geographical distribution of these endemic and internationally important viral haemorrhagic fevers.[23,38]

Materials and methods

Methodological overview

A map defining areas of environmental suitability for zoonotic transmission of Lassa fever was generated using an ensemble boosted regression trees (BRT) species distribution modelling framework. BRT models combine large numbers of regression trees to model probability of species presence based on the values of environmental covariates.[39,40] These models are trained using a spatial database of reported occurrences of infections in humans and animals alongside a set of background (or pseudo-absence) points representing environmental conditions in areas where cases are not reported. Areas that are environmentally similar to locations where zoonotic transmission of Lassa virus has been reported are thus identified. This approach requires a variety of information including: 1. reported index cases of Lassa fever; 2. Lassa virus detection in other mammalian hosts; 3. background (or pseudo-absence) information to represent locations where Lassa fever is likely absent; and 4. gridded surfaces of environmental covariates thought to influence Lassa fever distribution across Africa.

Reported infections in humans and animal hosts

We supplemented existing Lassa fever datasets[30] by searching for the terms 'Lassa' on the literature search engines Web of Science, PubMed and Scopus. In addition, notifications of cases were obtained from ProMED,[41] WHO and Public Health England. After initial assessment of abstracts for relevance, full text versions of articles thought likely to contain spatially explicit information on Lassa virus infection were obtained. When papers referred to articles with additional information not included in the original search, we also obtained these articles. In mapping the full extent of the zoonotic niche of Lassa fever, it is important to differentiate index cases from secondary cases resulting from human-to-human transmission. These two transmission routes should be considered as spatially distinct, with zoonotic transmission likely driven by environmental factors and secondary transmission by human behaviours and contact patterns. We therefore excluded records of infection if there was clear evidence that the case resulted from contact with another infected human (e.g., nosocomial transmission). Similarly, we excluded serosurveys of healthy individuals (due to the possibility of cross-reactivity with other viral agents and the uncertainty regarding time and place of infection) unless this could be linked to a prior fever retrospectively diagnosed as Lassa fever. If there was no indication of human-derived infection, the case was assumed to be of zoonotic origin. ProMED reports that described 'confirmed' cases were assumed to have been diagnosed using at least serological techniques.

Sites of infection were geo-positioned via Google Earth following a variation of existing protocols.[35] The location for geo-positioning was either identified within the article, or assumed to have occurred within the vicinity of the individual's home. If the location was smaller than 5 km × 5 km in area, only the latitude and longitude for the site were recorded (termed a 'point'). The remaining sites were treated as areas or 'polygons' and these were divided into three classes based on size. Areas of up to 10 km at their maximum width were defined by a circle of radius 5 km. Areas between 10 km and 25 km at their maximum width

were defined as a circle of radius 12.5 km. The borders of larger areas were defined using either administrative division boundaries (as defined by the Global Administrative Units Layer, GAUL)[42] or a bespoke polygon generated using geographic information system (GIS) software. Areas larger than one decimal degree squared were excluded. Following existing protocols, an occurrence of a case was defined as one or more cases of Lassa fever, within a specific geographical unit or 5 km × 5 km pixel in a given calendar year.[43,44] In total, 104 articles were used to generate 203 points and 171 polygons. Of these: 62 data points were from the period 1969–1979; 13 from the period 1980–1989; 47 from the period 1990–1999; 171 from the period 2000–2009; 77 from the period 2010–2014 and 4 data points had no start or end date of infection.

Reports of infections in multimammate mice (*M. natalensis*) were confirmed either by serological or genetically based diagnostics. Geo-positioning was performed using the methodology outlined above.

Covariates used in the analysis

A series of 5 km × 5 km gridded surfaces of a variety of environmental correlates thought to influence the distribution of Lassa fever were included as covariates in the model. These included information on the mean and range values for each pixel for land surface temperature (LST) (both night and day), enhanced vegetation indices (EVI), elevation and potential evapotranspiration (PET).[45,46] For those surfaces derived from satellite imagery, gap-filling algorithms were used to correct anomalies caused by cloud cover.[31] The gap-filled data span the years 2000 through 2015 and have a temporal resolution of 8 days. Data used in the models consist of mean values derived from all possible data points, resulting in synoptic surfaces that characterise normal conditions. As satellite data is widely unavailable for the years prior to 2000, we used summary raster values as indicators for the general long term environmental conditions. For additional information on environmental covariates see Weiss et al.[47]

Also included in the model was an estimated distribution of the primary reservoir host, the Natal multimammate mouse, *M. natalensis*. A separate species distribution model was carried out to capture the potential distribution of this rodent, using the same modelling procedure as for LASV (detailed below). Data for all members of the family Muridae were retrieved from the Global Biodiversity Information Facility (GBIF)[48] totalling 2 228 003 records. From this, records of *M. natalensis* were identified and, prior to inclusion in the model, went through a quality control process whereby existing expert-opinion range maps,[49] buffered by 100 km, were used to remove potentially erroneous results. In total 1031 occurrences were included in the analysis. Following existing approaches to deal with bias in observation and collection datasets,[50] all other Muridae occurrences were used as background data, a required component for presence-background modelling approaches.[33] Prior to inclusion in the final model, this prediction was clipped to within 500 km of the expert-opinion range map where the host species has been reported.

Species distribution modelling framework

The occurrence datasets, combined with the covariate data outlined above, were then analysed using an ensemble boosted

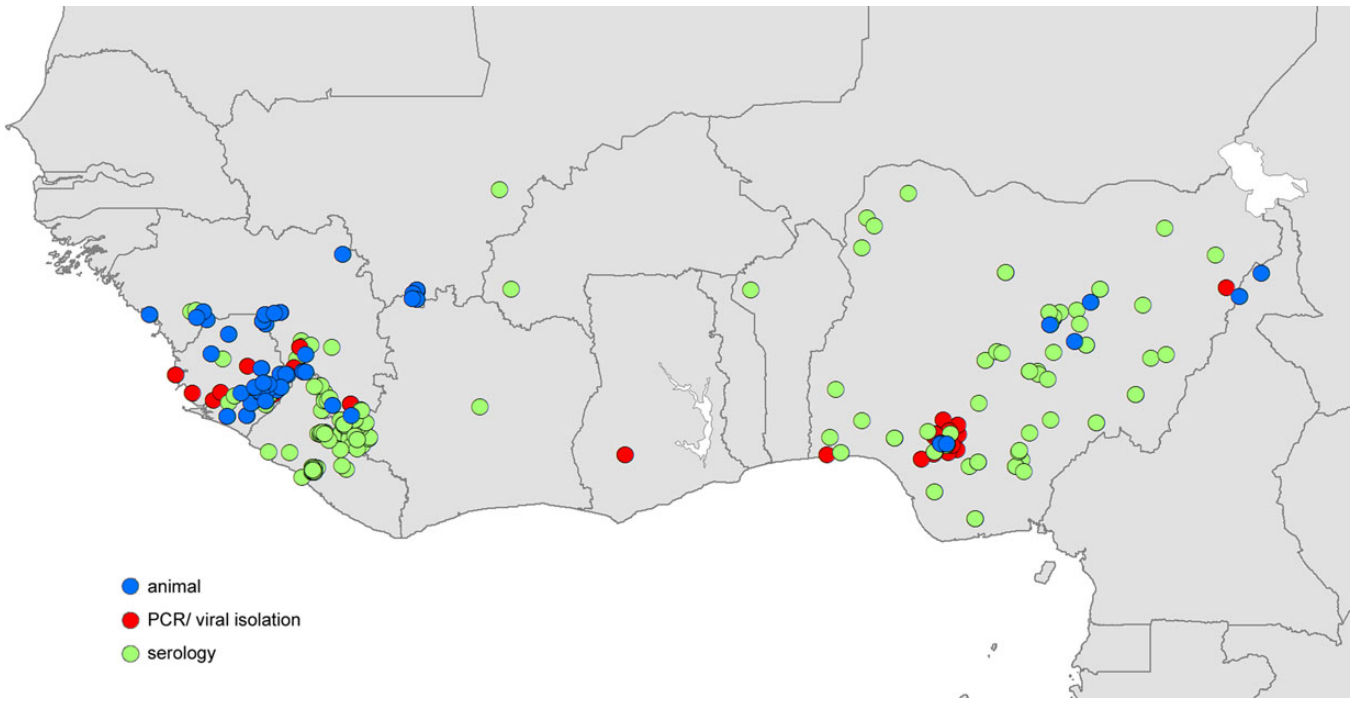


Figure 2. Reported locations of Lassa virus (LASV) infection used to build zoonotic niche maps. Blue circles indicate location mid-points for animal LASV infection surveys. Red circles indicate location mid-points where human cases of Lassa fever were diagnosed using PCR or viral isolation methods. Green circles indicate location mid-points where human cases of Lassa fever were diagnosed using serological methods.

regression trees modelling framework.[40] A total of 120 BRT submodels were run, with each iteration fitted to separate bootstrap-resampled datasets. Each resampled dataset had the same number of records as the full data, sampled randomly with replacement from the full dataset, subject to the constraint that at least five occurrence and five background records were present in each bootstrap. Model fitting was implemented using the gbm.step procedure in the dismo package in R.[51] All tuning parameters of the algorithm were held at their default values (cross-validation folds = 10, tree complexity = 4, learning rate = 0.005, bag fraction = 0.75, step size = 10). In each run, the weighting of background data was adjusted so that the sum of weights of the background points equalled the weighted sum of the presence records to improve discrimination capacity of the model.[52] A prediction map based upon the mean value for each 5 km × 5 km pixel across the 120 submodels was evaluated, as well as a 95% confidence interval around this value.

Accuracy of the models was analysed using the area under the curve (AUC) statistic. For each sub-model AUC was calculated as the mean of the cross-validated AUC across all 10-folds. The validation process divides the dataset into 10 groups of approximately equal presence and background data and assesses the ability of one subset to predict the remaining 90% of the data. These statistics were estimated using a pairwise distance sampling procedure in order to prevent inflation of the accuracy statistics due to spatial sorting bias.[53] The overall mean and standard deviation of the ensemble AUC was then calculated from all the submodels.

Modelling Lassa fever distribution

To incorporate spatial uncertainty in the location of outbreaks associated with polygon occurrence records, for each of the 120 submodels we randomly selected a different single point location within each occurrence polygon and treated this as the occurrence location. Since the final predictions were produced by averaging submodel predictions, this Monte Carlo procedure incorporated geographic uncertainty for some occurrence records in the final model.

These occurrence records of Lassa fever were then supplemented with 10 000 background points that were generated by randomly sampling across Africa, biased towards areas of higher population as a proxy for observation bias, under an assumption that more populous areas will be more likely to detect cases of Lassa fever should an outbreak occur.[34]

Previous investigations have indicated that accounting for differences in diagnostic accuracy can influence predictive ability[32] and therefore several scenarios were considered. Iterations included altering the weighting between human or animal infections diagnosed via PCR and serological tests (ratios 1:1, 2:1, 4:1) as well as an iteration with only PCR diagnosed cases. Finally, a model was run using only human index cases.

Evidence consensus and post-hoc masking

A variety of factors, not just environmental, influence the actual distribution of a species,[54–56] some of which cannot be considered in the modelling framework due to a lack of data at the necessary resolution or an incomplete understanding of what drives the distribution. The evidence consensus framework provides a means by which areas modelled as environmentally suitable but do not have the disease due to, for instance, biogeographic reasons, can be masked out of the final analysis. The evidence consensus system takes information from a variety of sources, considering different aspects of Lassa fever epidemiology to characterise the consensus on the evidence for Lassa fever presence in a country.

Criteria considered included: 1. endemic status as defined by three health reporting organisations (WHO, CDC and the Global Infectious Diseases and Epidemiology Online Network); 2. reported infection in humans archived in peer-reviewed literature and other data sources, assessed for contemporariness and diagnostic accuracy; 3. outbreaks of Lassa fever characterised by size and contemporariness, where cases were absent, the likelihood of missed diagnosis was assessed by considering adjacency to countries with reported infection and healthcare expenditure as a proxy for diagnostic capacity and 4. animal infection information. A full methodology, including how these information sources relate, is outlined in detail in the Supplementary information S1. A threshold for risk was defined at −25 and was applied to Africa to define areas where Lassa fever presence is likely. This threshold was selected as it differentiated areas with high certainty of absence from regions where insufficient information was available to determine absence.

Population living in areas of environmental suitability for Lassa virus transmission

A threshold probability for risk was determined by calculating the probability value that characterises 95% of all occurrences of Lassa fever (both human and animal data) turning the continuous environmental suitability surface into a binary at-risk, not-at-risk layer. Population sizes living in these 5 km × 5 km at-risk pixels were calculated from the WorldPop African population surface.[57,58]

Table 1. Reported locations of Lassa virus infection used to build zoonotic niche maps

Country	Rodent data		Human data	
	PCR/viral isolate	Serology	PCR/viral isolate	Serology
Benin	0	0	0	1
Burkina Faso	0	0	0	1
Cameroon	0	2	0	0
Côte d'Ivoire	0	0	0	1
Ghana	0	0	2	0
Guinea	19	25	8	15
Liberia	0	0	2	69
Mali	8	0	0	1
Nigeria	3	6	49	88
Sierra Leone	12	15	20	27
TOTAL	42	48	81	203

Results

Reported infections in humans and animals

In total, 374 distinct locations were identified as having animal infections or likely index cases of human outbreaks of Lassa fever (Figure 2). Human index cases were reported in nine different countries, mainly focussed in Liberia, Nigeria and Sierra Leone, but with some cases reported also in Benin, Burkina Faso, Côte d'Ivoire, Ghana, Guinea and Mali. Reports of infection in animals were found in four of these countries (Guinea, Mali, Nigeria and Sierra Leone) as well as in Cameroon, where no human index cases have been reported. The majority of human cases were diagnosed used serological techniques, although PCR diagnosis was often used in Nigeria and Sierra Leone (Table 1).

Predicted rodent reservoir distribution

The Natal multimammate mouse, *M. natalensis,* was predicted to have a very broad distribution across sub-Saharan Africa ranging from West Africa, across to the horn of Africa, down to Natal province in eastern South Africa where it was first collected (Figure 3). The model identified vegetation indices (both EVI mean and range

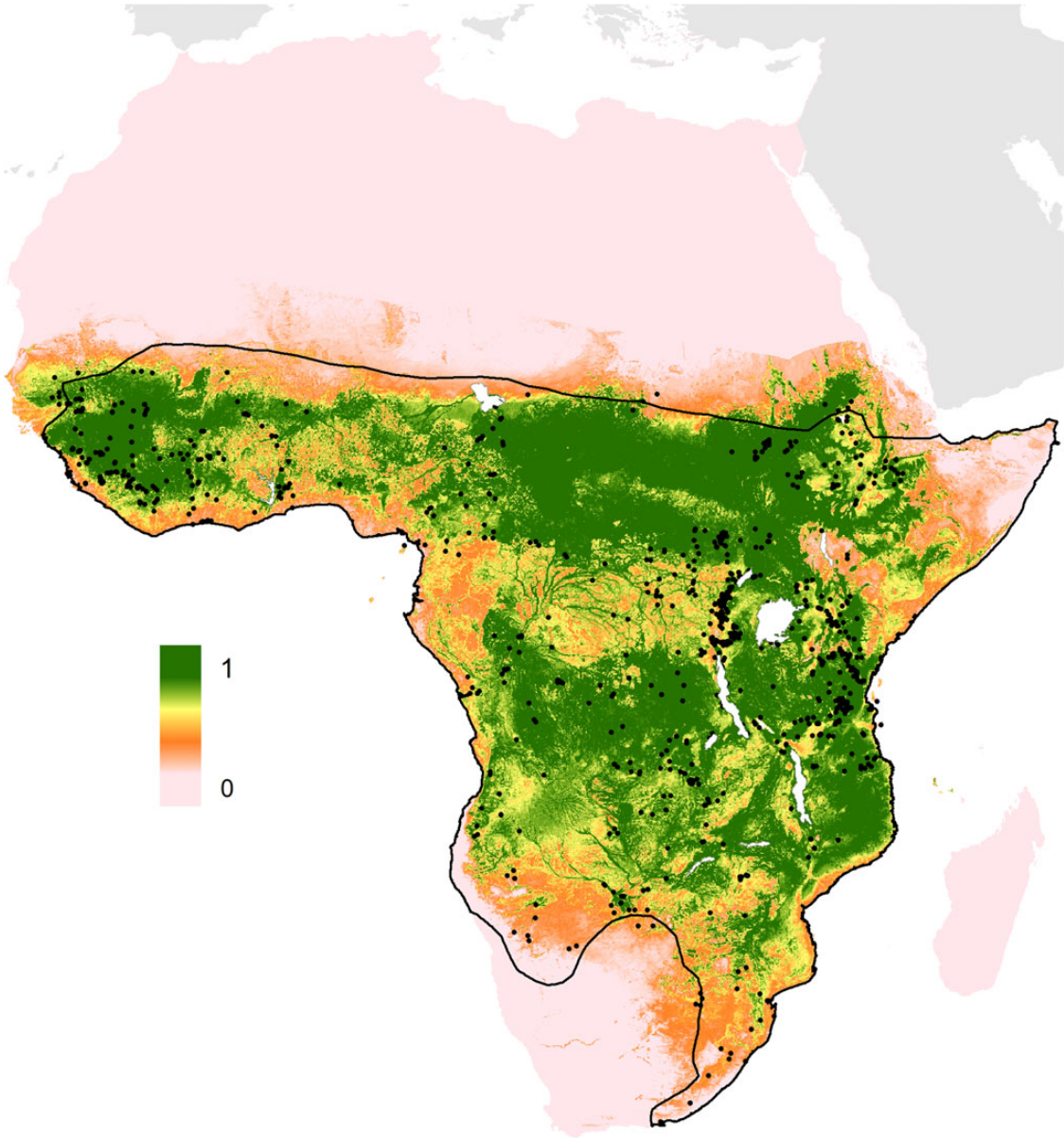


Figure 3. Predicted geographical distribution of the Natal multimammate mouse, *Mastomys natalensis*. The scale reflects the relative suitability of a given pixel for the presence of the Lassa virus zoonotic host, the Natal multimammate mouse, *Mastomys natalensis*. Areas closer to 1 (green) are more likely to harbour the rodent than those closer to 0 (pink). The prediction is clipped to within 500 km of the IUCN expert-opinion range map (solid black line),[49] to remove environmentally suitable areas in which the mouse has never been reported. The black spots represent *M. natalensis* locations as reported by GBIF.

values), potential evapotranspiration, elevation (digital elevations models) and night time land surface temperature as the main predictors of environmental suitability for *M. natalensis* (Table 2). The AUC values were 0.63±0.01 indicating the model demonstrated moderate predictive skill.

Predictions for the zoonotic niche of Lassa virus

The evidence consensus layer defined 13 countries as having consensus values ranging from complete consensus on presence to indeterminate. All countries reporting index cases of Lassa fever had a consensus score ranging from good to complete (above 60%). Togo had a moderate consensus on presence, while the remainder (Niger, Senegal, Guinea-Bissau and Cameroon) had either low consensus or indeterminate status (Figure 4).

All model variants produced broadly consistent predictions across West Africa (see Supplementary information S2). Given these similarities, the equal weighting of occurrences based on diagnosis model was selected as this required the least assumptions and included the maximum amount of data. Large areas of environmental suitability for the zoonotic transmission of Lassa fever were found in Guinea, Liberia, Sierra Leone, Côte d'Ivoire and Nigeria, with smaller regions predicted in Benin, Burkina Faso, Togo, Mali, Senegal, Guinea-Bissau, Niger, Ghana and Cameroon (Figure 4).

The model identified vegetation, night-time land surface temperature, environmental suitability for the host species, elevation and potential evapotranspiration as the main predictors of suitability for zoonotic transmission of LASV (Table 2). The AUC value was 0.79±0.02 indicating the model demonstrated good predictive skill. Environmental covariate partial dependency plots are provided in the Supplementary information S3.

The full prediction surface without the evidence consensus mask is presented in the Supplementary information S2, where uncertainty maps for this prediction surface are also shown.

Population at risk of zoonotic transmission of Lassa virus

The final threshold probability for risk, which captured 95% of all Lassa fever occurrences, was calculated to be equal to or greater than 0.646. In total, approximately 37.7 million individuals in 14 countries live in areas predicted to be environmentally suitable for the zoonotic transmission of LASV. The majority (97.9%) live in countries that have already reported index cases of Lassa fever, with Nigeria accounting for approximately 36% of the total population living in at-risk areas. More information is provided in the Supplementary information S4.

Discussion

These maps present revised estimates of areas environmentally suitable for the zoonotic transmission of LASV and provide an important baseline for guiding Lassa fever surveillance and additional epidemiological investigations. Areas of environmental suitability are defined across a broad area of West Africa, including countries where no cases have been reported. These maps can therefore inform our wider understanding of the disease and aid future differential diagnosis of viral haemorrhagic fevers in areas where two or more viruses are potentially present.[23,38]

As with any model-based study, an awareness of data limitations and model assumptions is essential. Although predictive capability will be hindered by limited datasets where the true site of zoonotic transmission is unlikely to be reported, our study attempts to be as comprehensive as possible in including all known reports of zoonotic infections, as well as considering uncertainty in the location of initial infection. Because our models are only able to assess areas that are environmentally suitable for LASV, more research on how humans and animal reservoirs interact, as well as how the disease is transmitted within these populations, is needed to understand and translate this into true outbreak risk. Even with these limitations, we hope that our results will act as a springboard for further research to better understand the epidemiology of LASV and characterise the risk of this important VHF.

It is important to recognise that the outputs of this study are modelled estimates and are heavily influenced by the data used. Precisely geo-positioning the true site of infection is difficult to achieve. There is often an assumption that infection occurs in the locality of the patient's home address. Even when a patient's place of residence is documented, it may not represent the location where infection took place. Given the resolution of our zoonotic niche maps (5 km × 5 km), point locations cover a relatively broad area, which is likely to include the true infection

Table 2. Summary statistics for model outputs. Relative contributions for each of the top five predictors are reported as a percentage

Statistic	Model 1: *Mastomys natalensis* distribution	Model 2: LASV zoonotic niche
AUC±standard deviation	0.63±0.01	0.79±0.02
1st predictor	Mean EVI: 24.5%	Mean EVI: 26.5%
2nd predictor	Mean PET: 19.7%	Night-time mean LST: 19.2%
3rd predictor	Elevation (DEM): 16.3%	*M. natalensis* distribution: 13.6%
4th predictor	Night-time mean LST: 14.8%	Elevation (DEM): 11.7%
5th predictor	EVI range: 10.5%	Mean PET: 10.6%

AUC: area under the curve; DEM: digital elevations models; EVI: enhanced vegetation index; LASV: Lassa virus; LST: land surface temperature; PET: potential evapotranspiration.

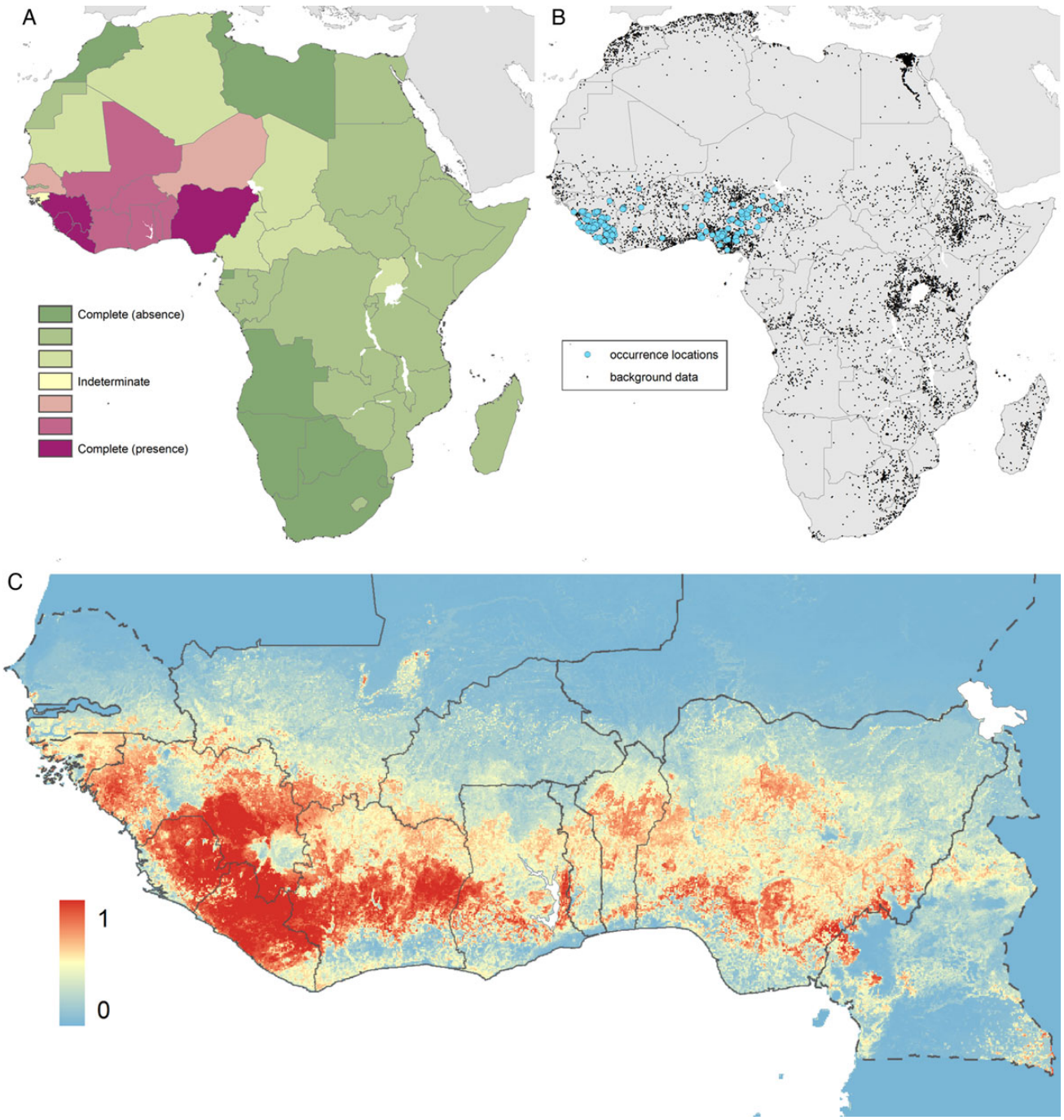


Figure 4. Maps of the: (A) definitive extents as determined by evidence consensus; (B) recorded occurrence and generated background pseudo-absence points used in the BRT procedure and (C) predicted geographical distribution of the zoonotic niche for Lassa virus. Panel A shows the consensus on Lassa fever presence, ranging from dark green (complete consensus on absence) to purple (complete consensus on presence). Countries in yellow are those where evidence was inconclusive or contradictory for Lassa fever presence. Panel B shows the location of data points that went into the model. Panel C shows the environmental suitability for zoonotic transmission of Lassa virus. Areas closer to 1 (red) are more suitable than those closer to 0 (blue). Countries with borders outlined by a solid line are those where cases of LASV have previously been reported. Countries with borders outlined by a dash line have not previously reported LASV cases. The area under the curve statistic, calculated under a stringent cross-validation procedure is 0.79 ± 0.02.

site. For more uncertain locations, larger areas were defined to include the supposed site of infection and the model was adjusted accordingly.

In addition, the need for reliable information on the location of infection excluded hospital cases that did not document home location, because hospitals often serve a much broader area,

particularly referral hospitals. Serosurveys of healthy individuals were also excluded, since seropositivity could reflect prior infection in a range of places and times. With rodent data, however, given their relatively limited dispersal ability,[59] the assumption of infection occurring near trapping sites was valid.

Diagnostic accuracy was a potentially important consideration. We accounted for this factor by including a variety of weighting schema, as well as excluding potentially less reliable serological assays. This was also important given the spatial bias in the availability of specific diagnostic tests and, therefore, in the ability to accurately detect cases of Lassa fever. This spatial bias was reflected in the high proportion of cases seen in Nigeria and Sierra Leone where dedicated surveillance activity and research programmes exist.[60,61] Using alternative diagnostic method weighting schemas to test model predictiveness, however, showed that this factor had little impact on the output maps and the validation and summary statistics.

These maps identify regions of environmental suitability for zoonotic transmission of LASV by defining a set of environmental conditions that best characterise the locations of known infections. Our model indicated mean EVI, night-time mean LST, *M. natalensis* suitability, elevation and PET as key predictors for environmental suitability. While inference of causal relationships cannot be directly evaluated using this approach, all these parameters have plausible influence on the transmission cycle of LASV and likely influence the rodent host population dynamics as well as the nature and frequency of human-rodent interactions. Indeed, other rodent-borne viruses, viral dynamics and corresponding disease risk have been shown to be strongly influenced by the environment.[62] Further investigations must be undertaken to determine the precise nature of any causal relationship.

It is important to note, however, that while the environment has an important impact on influencing viral distribution, other factors will also impact upon its distribution.[54,63] One means by which we accounted for this was by utilising the evidence consensus layer to assess the contemporary endemic status of a country from a variety of different sources. As a consequence, while large areas of sub-Saharan Africa are environmentally suitable for viral transmission, there is a consensus from other evidence,[64] that the virus is not present. On the other hand, a number of countries (Guinea-Bissau, Senegal, Togo, Niger and Cameroon) are more likely to be underreporting cases due to their proximity to other endemic countries and their low healthcare expenditure. Whilst an imperfect process, the use of evidence consensus to define areas of likely under-reporting as opposed to regions of true absence is an important first step. Indeed, as Benin has recently proven, it is possible for outbreaks to occur in areas of previously reported absence.[65,66]

The results of this study act as a useful guide to help refine where potential at-risk populations exist and pinpoint where new surveys and surveillance initiatives would be most beneficial, particularly at the edges of the predicted geographical distribution of LASV (such as Cameroon and Senegal) and in countries where at-risk populations are predicted but have yet to report any cases of LASV (such as Togo). For more information regarding the predicted distribution of at-risk populations, see Supplementary information S4.

Areas suitable for disease transmission may, for other reasons, not result in cases. In addition, regions with similar environments may report very different caseloads. Therefore, translating these environmental suitability layers into actual incidence and prevalence estimates requires more detailed epidemiological information regarding the interactions between virus, host and human, and how the three relate across different spatial scales. The maps presented here act as a baseline for further refinement when additional spatial information becomes available.[67] Evidence is most needed in areas where the host is found but cases have not been reported. For instance, while the primary reservoir host is found across Africa, the virus itself appears to be restricted to West Africa. Could this be related to the seemingly poor dispersal ability of the host, or are there biogeographic barriers preventing mixing of the populations? For example, with bat-borne viruses, the likelihood of infection in bat colonies on the fringes of their distribution is greater than found here for LASV in mice, likely due to the superior dispersal ability of bats.[68,69] Understanding the viral dynamics within the reservoir host, and how reservoir populations spread the virus between themselves, may provide the critical step in understanding true risk. It is also unknown whether all human populations are equally susceptible to infection. Anthropological studies suggest that different groups behave very differently with regards to mammal species and subtle variations in housing conditions, social relations and agricultural practices could have large impacts on how humans and disease reservoirs interact.[70]

Conclusions

The output maps provide an important resource for refining our understanding of the distribution of Lassa fever. Baseline estimates such as these are necessary, not only to aid in selecting locations for initial surveys in areas beyond those that currently report cases and directing both human and animal surveillance activities, but also to inform a wider community of public health officials about the potential risk of Lassa fever. Given the potential for nosocomial transmission of this disease,[10,11,71] and especially given the potential for misdiagnosis as other febrile illnesses,[72,73] incorporating Lassa fever in a differential diagnosis is critical for timely prevention measures to be put in place. In an increasingly connected world, these maps not only inform local risk, but can assist in detection of potential imported cases when a travel history is available. As the recent Ebola virus disease outbreak in West Africa has shown, recognising this risk can be a vital first step in preventing further transmission of this important viral haemorrhagic fever.

Supplementary data

Supplementary data are available at Transactions online (http://trstmh.oxfordjournals.org/).

Authors' disclaimer: The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Authors' contributions: AQNM and DMP contributed equally to this work; SIH conceived the work; AQNM, KAD, JL, FS and DMP collected and geo-positioned the data; AQNM, DMP and NG implemented the modelling; DJW generated the environmental covariate layers; AQNM,

DMP, NG, JPM, CLM and SIH analysed the data outputs; AQNM and DMP generated the first draft of the manuscript. All authors read and approved the final manuscript. AQNM and DMP are guarantors of the paper.

Acknowledgements: We thank Maria Devine for proof-reading this manuscript.

Funding: AQNM, NG and CLM are funded by the Bill & Melinda Gates Foundation [OPP1093011]; FS is funded by a scholarship from the Rhodes Trust; DJW is funded by the Bill & Melinda Gates Foundation [OPP1068048]; DMP is funded by a Sir Richard Southwood Graduate Scholarship from the Department of Zoology; SIH is funded by a Senior Research Fellowship from the Wellcome Trust [0095066] that also supports KAD and JL, and a grant from the Bill & Melinda Gates Foundation [OPP1093011]. SIH would also like to acknowledge funding support from the RAPIDD program of the Science & Technology Directorate, Department of Homeland Security and the Fogarty International Center, National Institutes of Health. JPM is funded by the International Research Consortium on Dengue Risk Assessment Management and Surveillance (IDAMS, European Commission 7th Framework Programme [#21803] http://www.idams.eu).

Competing interests: None declared.

Ethical approval: Not required.

References

1 Frame JD, Baldwin JM Jr., Gocke DJ et al. Lassa fever, a new virus disease of man from West Africa. I. Clinical description and pathological findings. Am J Trop Med Hyg 1970;19:670–6.

2 Rose JR. An outbreak of encephalomyelitis in Sierra Leone. Lancet 1957;273:914–6.

3 McCormick JB, Webb PA, Krebs JW et al. A prospective study of the epidemiology and ecology of Lassa fever. J Infect Dis 1987; 155:437–44.

4 Lukashevich IS, Clegg JC, Sidibe K. Lassa virus activity in Guinea: distribution of human antiviral antibody defined using enzyme-linked immunosorbent assay with recombinant antigen. J Med Virol 1993;40:210–7.

5 Tomori O, Fabiyi A, Sorungbe A et al. Viral hemorrhagic fever antibodies in Nigerian populations. Am J Trop Med Hyg 1988;38:407–10.

6 Keenlyside RA, McCormick JB, Webb PA et al. Case-control study of Mastomys natalensis and humans in Lassa virus-infected households in Sierra Leone. Am J Trop Med Hyg 1983;32:829–37.

7 Mertens PE, Patton R, Baum JJ et al. Clinical presentation of Lassa fever cases during the hospital epidemic at Zorzor, Liberia, March-April 1972. Am J Trop Med Hyg 1973;22:780–4.

8 Wertheim HF, Horby P, Woodall JP. Atlas of human infectious diseases. Chishester: John Wiley & Sons; 2012.

9 McCormick JB, King IJ, Webb PA et al. A case-control study of the clinical diagnosis and course of Lassa fever. J Infect Dis 1987; 155:445–55.

10 Carey DE, Kemp GE, White HA et al. Lassa fever. Epidemiological aspects of the 1970 epidemic, Jos, Nigeria. Trans R Soc Trop Med Hyg 1972;66:402–8.

11 Monath TP, Mertens PE, Patton R et al. A hospital epidemic of Lassa fever in Zorzor, Liberia, March-April 1972. Am J Trop Med Hyg 1973;22:773–9.

12 Singh SK, Ruzek D. Viral hemorrhagic fevers. Boca Raton: CRC Press; 2013.

13 Haas WH, Breuer T, Pfaff G et al. Imported Lassa fever in Germany: surveillance and management of contact persons. Clin Infect Dis 2003;36:1254–8.

14 Cooper CB, Gransden WR, Webster M et al. A case of Lassa fever: experience at St Thomas's Hospital. Br Med J (Clin Res Ed) 1982;285:1003–5.

15 Fisher-Hoch SP, Price ME, Craven RB et al. Safe intensive-care management of a severe case of Lassa fever with simple barrier nursing techniques. Lancet 1985;2:1227–9.

16 Helmick CG, Webb PA, Scribner CL et al. No evidence for increased risk of Lassa fever infection in hospital staff. Lancet 1986;2:1202–5.

17 Holmes GP, McCormick JB, Trock SC et al. Lassa fever in the United States. Investigation of a case and new guidelines for management. N Engl J Med 1990;323:1120–3.

18 Safronetz D, Lopez JE, Sogoba N et al. Detection of Lassa virus, Mali. Emerg Infect Dis 2010;16:1123–6.

19 Wulff H, Fabiyi A, Monath TP. Recent isolations of Lassa virus from Nigerian rodents. Bull World Health Organ 1975;52:609–13.

20 Ter Meulen J, Lukashevich I, Sidibe K et al. Hunting of peridomestic rodents and consumption of their meat as possible risk factors for rodent-to-human transmission of Lassa virus in the Republic of Guinea. Am J Trop Med Hyg 1996;55:661–6.

21 Lo Iacono G, Cunningham AA, Fichet-Calvet E et al. Using modelling to disentangle the relative contributions of zoonotic and anthroponotic transmission: the case of lassa fever. PLoS Negl Trop Dis 2015;9:e3398.

22 Faye O, Boelle PY, Heleze E et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. Lancet Infect Dis 2015;15:320–6.

23 Bannister B. Viral haemorrhagic fevers imported into non-endemic countries: risk assessment and management. Br Med Bull 2010;95:193–225.

24 Hirabayashi Y, Oka S, Goto H et al. The first imported case of Lassa fever in Japan [in Japanese]. Nihon Rinsho 1989;47:71–5.

25 Shlaeffer F, Sikuler E, Keynan A. Lassa fever--first case diagnosed in Israel [in Hebrew]. Harefuah 1988;114:12–4.

26 Communicable Disease Surveillance Centre. Lassa fever imported to England. Commun Dis Rep CDR Wkly 2000;10:99.

27 WHO. Lassa fever, imported case, Netherlands. Wkly Epidemiol Rec 2000;75:265.

28 Centers for Disease Control and Prevention. Imported Lassa fever - New Jersey, 2004. MMWR Morb Mortal Wkly Rep 2004;53:894–7.

29 Gates B. The next epidemic - lessons from Ebola. N Engl J Med 2015;372:1381–4.

30 Fichet-Calvet E, Rogers DJ. Risk maps of Lassa fever in West Africa. PLoS Negl Trop Dis 2009;3:e388.

31 Weiss DJ, Atkinson PM, Bhatt S et al. An effective approach for gap-filling continental scale remotely sensed time-series. ISPRS J Photogramm Remote Sens 2014;98:106–18.

32 Peterson AT, Moses LM, Bausch DG. Mapping transmission risk of lassa fever in West Africa: the importance of quality control, sampling bias, and error weighting. PLoS One 2014;9:e100711.

33 Elith J, Leathwick JR. Species distribution models: ecological explanation and prediction across space and time. Annu Rev Ecol Evol S 2009;40:677–97.

34 Pigott DM, Golding N, Mylne A et al. Mapping the zoonotic niche of Ebola virus disease in Africa. eLife 2014;3:e04395.

35 Mylne A, Brady OJ, Huang Z et al. A comprehensive database of the geographic spread of past human Ebola outbreaks. Sci Data 2014;1:140042.

36 Pigott DM, Golding N, Mylne A et al. Mapping the zoonotic niche of Marburg virus disease in Africa. Trans R Soc Trop Med Hyg 2015; 109:366–78.

37 Messina JP, Pigott DM, Duda KA et al. A global compendium of human Crimean-Congo haemorrhagic fever virus occurrence. Sci Data 2015;2:150016.

38 Moore LS, Moore M, Sriskandan S. Ebola and other viral haemorrhagic fevers: a local operational approach. Br J Hosp Med (Lond) 2014; 75:515–22.

39 De'ath G. Boosted trees for ecological modeling and prediction. Ecology 2007;88:243–51.

40 Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol 2008;77:802–13.

41 Victor LY, Madoff LC. ProMED-mail: an early warning system for emerging diseases. Clin Infect Dis 2004;39:227–32.

42 Food and Agriculture Organization of the United Nations. The Global Administrative Unit Layers (GAUL): Technical Aspects. Rome: Food and Agriculture Organization of the United Nations, EC-FAO Food Security Programme (ESTG); 2008.

43 Pigott DM, Golding N, Messina JP et al. Global database of leishmaniasis occurrence locations, 1960–2012. Sci Data 2014;1:140036.

44 Messina JP, Brady OJ, Pigott DM et al. A global compendium of human dengue virus occurrence. Sci Data 2014;1:140004.

45 Wan Z, Li Z-L. A physics-based algorithm for retrieving land-surface emissivity and temperature from EOS/MODIS data. IEEE Trans Geosci Remote Sens 1997;35:980–96.

46 ORNL DAAC. Shuttle radar topography mission near-global digital elevation models. http://webmap.ornl.gov/wcsdown/wcsdown.jsp?dg_id=10008_1 [accessed 5 August 2014].

47 Weiss DJ, Mappin B, Dalrymple U et al. Re-examining environmental correlates of *Plasmodium falciparum* malaria endemicity: a data-intensive variable selection approach. Malar J 2015;14:68.

48 Global Biodiversity Information Facility. http://www.gbif.org/ [accessed 2 August 2014].

49 Schipper J, Chanson JS, Chiozza F et al. The status of the world's land and marine mammals: diversity, threat, and knowledge. Science 2008;322:225–30.

50 Phillips SJ, Dudik M, Elith J et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl 2009;19:181–97.

51 Hijmans RJ, Phillips S, Leathwick J et al. dismo. R package. http://cran.r-project.org/web/packages/dismo/dismo.pdf [accessed 19 August 2014].

52 Barbet-Massin M, Jiguet F, Albert CH et al. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol Evol 2012;3:327–38.

53 Hijmans RJ. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 2012;93:679–88.

54 Peterson AT, Soberón J, Pearson RG et al. Ecological niches and geographic distributions. Princeton: Princeton University Press; 2011.

55 Soberon J. Grinnellian and Eltonian niches and geographic distributions of species. Ecol Lett 2007;10:1115–23.

56 Soberon J, Nakamura M. Niches and distributional areas: concepts, methods, and assumptions. Proc Natl Acad Sci USA 2009;106(Suppl 2): 19644–50.

57 Linard C, Gilbert M, Snow RW et al. Population distribution, settlement patterns and accessibility across Africa in 2010. PLoS One 2012;7:e31743.

58 WorldPop. WorldPop project. http://www.worldpop.org.uk/ [accessed 7 August 2014].

59 Van Hooft P, Cosson J, Vibe-Petersen S et al. Dispersal in Mastomys natalensis mice: use of fine-scale genetic analyses for pest management. Hereditas 2008;145:262–73.

60 Khan SH, Goba A, Chu M et al. New opportunities for field research on the pathogenesis and treatment of Lassa fever. Antiviral Res 2008;78:103–15.

61 Ehichioya DU, Hass M, Olschlager S et al. Lassa fever, Nigeria, 2005–2008. Emerg Infect Dis 2010;16:1040–1.

62 Carver S, Mills JN, Parmenter CA et al. Toward a mechanistic understanding of environmentally forced zoonotic disease emergence: Sin Nombre Hantavirus. BioScience 2015:biv047.

63 Peterson AT. Biogeography of diseases: a framework for analysis. Naturwissenschaften 2008;95:483–91.

64 Farrar J, Hotez P, Junghanss T et al. Manson's tropical diseases: expert consult-online. Kidlington: Elsevier Health Sciences; 2013.

65 WHO. Ebola virus disease preparedness strengthening team [in French]. Geneva: World Health Organization; 2015. http://apps.who.int/iris/bitstream/10665/145090/1/WHO_EVD_PCV_Mali_14_fre.pdf [accessed 1 February 2015].

66 ProMED-mail. Lassa fever - Benin (02): (Atakora). http://www.promedmail.org/direct.php?id=2992727 [accessed 15 December 2014].

67 Plowright RK, Eby P, Hudson PJ et al. Ecological dynamics of emerging bat virus spillover. Proc Biol Sci 2015;282:20142124.

68 Olival KJ, Hayman DT. Filoviruses in bats: current knowledge and future directions. Viruses 2014;6:1759–88.

69 Dudas G, Rambaut A. Phylogenetic analysis of Guinea 2014 EBOV Ebolavirus outbreak. PLoS Curr 2014;6:pii: ecurrents.outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d.

70 Daszak P, Zambrana-Torrelio C, Bogich TL et al. Interdisciplinary approaches to understanding disease emergence: the past, present, and future drivers of Nipah virus emergence. Proc Natl Acad Sci USA 2013;110(Suppl 1):3681–8.

71 Fisher-Hoch SP, Tomori O, Nasidi A et al. Review of cases of nosocomial Lassa fever in Nigeria: the high price of poor medical practice. Br Med J 1995;311:857–9.

72 Schoepp RJ, Rossi CA, Khan SH et al. Undiagnosed acute viral febrile illnesses, Sierra Leone. Emerg Infect Dis 2014;20:1176–82.

73 Yun NE, Walker DH. Pathogenesis of Lassa fever. Viruses 2012; 4:2031–48.

Chapter 8

Translating maps into policy.

This synoptic chapter draws upon Chapters 4, 5, 6 and 7 to consider the risk that contagious viral haemorrhagic fevers pose to the African continent. Three key phases in a VHF outbreak are identified and their risk is quantified; first the potential for spillover transmission from animal and other reservoir hosts to human populations, secondly the risk of this index case infecting additional humans in a particular location and thirdly the risk of this local outbreak becoming more widespread as people travel around neighbouring regions. This chapter analyses the differences in risk for each of these stages across the continent, for each individual VHF as well as a combined measure. The hope is that this approach highlights areas to be prioritised when future VHF preparedness decisions are made. This chapter is presented as a paper ready for submission.

The co-distribution of contagious viral haemorrhagic fevers in Africa

Pigott, D. M.[1], Golding, N.[2], Messina, J. P.[1], Mylne, A.[2], Shearer, F.[2], Brady, O. J.[2], Kraemer, M. U. G.[1], Bhatt, S. J.[1], Gething, P. W.[1], Weiss, D. J.[1], Moyes, C. L.[2], Tatem, A. J.[3,4,5] and Hay, S.I.[2,5,6]

[1] *Spatial Ecology and Epidemiology Group, Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK*

[2] *Wellcome Trust Centre for Human Genetics, University of Oxford, OX3 7BN, UK*

[3] *Department of Geography, University of Southampton, Southampton, SO17 1BJ, UK*

[4] *Flowminder Foundation, Stockholm, Sweden*

[5] *Fogarty International Center, National Institutes of Health, Bethesda, MD 20892-2220, USA*

[6] *Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, 98121, USA*

Abstract

The West African outbreak of Ebola virus disease represents the largest and most widespread viral haemorrhagic fever outbreak recorded. As the numbers of new cases continues to fall, future attention must focus on how to prevent other such outbreaks from reaching the same scale. This paper considers the risk four contagious viral haemorrhagic diseases (Ebola virus disease, Marburg virus disease, Lassa fever and Crimean-Congo haemorrhagic fever) pose to the African continent. Three important stages in a potential outbreak are identified: the initial viral transmission event from animal and other host species to the index human, the subsequent human-to-human transmission occurring at the initial focus of the outbreak and the final more widespread dissemination of the virus to other neighbouring communities. To quantify the relative risk of each of these stages, we evaluate a variety of different data sources for each nation, using methodology outlined by the Information for Risk Management platform. Data layers incorporating predicted zoonotic niche maps, population surfaces, healthcare facilities and healthcare-seeking behaviour as well as metrics of movement capabilities are integrated into the assessment. The analysis demonstrates the heterogeneities in risk that occur across the continent. West African countries, particularly Nigeria, Guinea and Côte d'Ivoire, are identified as the riskiest areas for widespread outbreaks of these diseases occurring. These maps provide important details to help inform future discussion on how best to combat such infectious disease outbreaks and prioritise areas for improved surveillance, diagnostic and infrastructural developments.

Introduction

The outbreak of Ebola virus disease (EVD) in West Africa is unprecedented in terms of the number of people infected and the extent to which it has spread both in Africa and across the globe *(Heymann et al., 2015; Mylne et al., 2014).* As of June 2015, over 27000 suspected, confirmed or probable cases have been identified *(CDC, 2015)*, approximately 60 times more than reported in any one previous outbreak and ten times more than all previous outbreaks combined *(Mylne et al., 2014)*. Not only has large personnel and financial support been leveraged to assist in the control effort, but also a number of near real-time modelling approaches have been used to aid in answering the key questions related to epidemiology of the outbreak *(WHO Ebola response team, 2014)*, where human movement may cause the disease to spread *(Kraemer and Golding, 2015)* and the genetic heritage of these infections *(Dudas and Rambaut, 2014; Gire et al., 2014)*. As we head towards a local elimination endgame, focus must now be drawn to improving our understanding of EVD and where future outbreaks are most likely to take place.

The West African outbreak shows just how devastating a viral haemorrhagic fever outbreak can be. It is therefore important to recognise that several other diseases, including Marburg virus disease, Lassa fever and Crimean-Congo haemorrhagic fever follow similar transmission patterns *(Bannister, 2010)* and have proven capable of secondary human-to-human transmission both locally *(Altaf et al., 1998; Towner et al., 2006; Troup et al., 1970)* and internationally *(Amorosa et al., 2010; Atkinson et al., 2012; Timen et al., 2009)*.

The risk associated with any zoonotic outbreak can be considered in three key stages (Figure 1). The first considers the risk of virus in reservoir hosts transmitting to human populations. Marburg virus and Ebola virus are thought to have bat reservoir species *(Olival and Hayman, 2014)*, with both viruses experimentally shown to cause asymptomatic infection *(Gonzalez et*

al., 2007), and viral RNA and (for Marburg) DNA have been isolated from individual bats *(Leroy et al., 2005; Towner et al., 2009)*. Previous outbreaks have also implicated local bat populations as the source of infection *(Leroy et al., 2009)*. Similarly, there is a large body of evidence supporting the Natal multimammate mouse, *Mastomys natalensis*, as the reservoir host for Lassa mammarenavirus with most infections thought to occur within the household *(McCormick et al., 1987)*. In contrast, CCHFV is vectored by a variety of tick species, with occasional transmission to other animals, particularly livestock. Infection results from tick bites or contact with contagious particles in the bodily fluid of infected animals *(Ergonul, 2006)*.

The next stage is risk of transmission from this index case to other humans within the local community or in healthcare settings *(Ajayi et al., 2013)*. Historic EVD and MVD infections have seen large number of cases in the first weeks following outbreaks focus around the home of the index individual or where he sought care, either in formal hospital settings or with traditional healers *(Hewlett et al., 2005; Mylne et al., 2014)*. Whilst the spillover stage is often influenced by climatic factors *(Lash et al., 2008; Pigott et al., 2014; Pigott et al., 2015)*, aspects of viral dynamics in animal populations *(Amman et al., 2012)* and how humans interact with reservoir hosts *(Woldehanna and Zimicki, 2015)*, the likelihood of a larger local outbreak is influenced by a variety of human behaviours ranging from healthcare-seeking behaviour to existing infrastructure in place to support individuals presenting with undiagnosed febrile illnesses *(Baron et al., 1983; Boumandouki et al., 2005)*. The West African outbreak, as with many other VHFs before *(Towner et al., 2006)*, has been fuelled entirely by human-to-human transmission *(Gire et al., 2014)*.

The final stage, a more widespread outbreak with cases and chains of transmission occurring in a variety of locations, will be driven by the accessibility of the focus of the epidemic to surrounding areas. Travel, either by land or by air, has been responsible for much of the widespread infections we have seen both in the neighbouring districts areas *(Francesconi et al., 2003)*, as well as internationally *(Bogoch et al., 2014)*. During the current outbreak all districts

in Liberia and Sierra Leone, as well as many in Guinea, have reported cases, driven mainly by overground movement. In addition, the outbreak has caused cases in six further countries across the globe, often related to individuals travelling by air and causing limited local outbreaks.

This analysis sets out to quantify the relative risk of these outbreak transition stages across Africa by utilising a variety of different information sources. By leveraging these diverse metrics related to risk, integrated within a risk management platform *(de Groeve et al., 2015)*, we can attempt to define countries in Africa where VHF outbreaks are of greatest concern. The resulting layers can therefore by used to aid in discussion of future preparedness strategies for not only Ebola virus disease but other similarly contagious pathogens.

Methods

Methodological summary

Using a pre-existing risk assessment framework, we aim to identify regions in Africa at risk from VHFs. The progression from an index case to a widespread outbreak can be characterised into three stages (see Figure 1). Here we use information from a variety of sources to determine a relative ranking of different African countries for each stage of risk for each of the four VHFs identified, as well as a combined measure.

The INFORM framework

INFORM (Information for Risk Management) is a Joint Research Centre of the European Commission (JRC) tool developed to assist in disaster management *(de Groeve et al., 2015)*. Importantly, this approach represents the only quantitative assessment of risk that is comprehensive across the entire globe. Whilst other methods, such as expert opinion questionnaires, can be used to evaluate risk, the open nature of the INFORM database, as well as the wide variety of indicators that it tracks, make the system incredibly powerful. Conceptually, the model considers three dimensions of risk: hazard and exposure, vulnerability, and lack of coping capacity. These three categories therefore consider not only the risk of the disaster occurring in the first instance, but also aspects of the population that will ultimately determine how significant an impact any given disaster will cause. In response to the recent Ebola outbreak in West Africa, the Department for International Development (DFID) worked with the JRC to adapt their existing framework to assess the risk of the current outbreak spreading elsewhere. This paper builds upon that work, extending these concepts of risk to potential outbreaks of contagious VHFs in Africa.

Assessing risk of spillover infection

Previous publications have already described areas of potential zoonotic transmission of contagious viral haemorrhagic fevers across the African continent, specifically Ebola virus disease *(Pigott et al., 2014)*, Marburg virus disease *(Pigott et al., 2015)*, Lassa fever *(Mylne et al., 2015)* and Crimean-Congo haemorrhagic fever *(Messina et al., 2015)*. Each surface consists of 5km x 5km pixel surfaces with values ranging from 0 to 1, where values closer to 1 indicate more environmentally similar locations to those where previous cases have occurred. Each layer was converted into a binary at-risk/not-at-risk surface by deriving a threshold probability for each individual condition which captured 95% of the occurrence data. Each was then converted into a population at risk surface by using gridded global population layers generated from WorldPop 2015 *(WorldPop., 2015)*. In the absence of precise human-reservoir interaction data, it is assumed that all areas of predicted environmental suitability are equally likely to see spillover infections and that the likelihood of this spillover occurring is a function of total population size living in these areas.

For each disease, populations at risk were aggregated to the national level and then $\log_{10}$ transformed. As per INFORM protocol, the 2.5% and 97.5% quantiles were calculated and any national populations that lay beyond these values were reassigned the quantile values. These adjusted values were then rescaled from ten (assigned to the maximum value) to zero (assigned to the minimum value). The relative ranking of the combined risk the four VHFs represent is therefore a function of the risk of a spillover outbreak for any of the four viruses occurring in this country.

Assessing local outbreak risk

A number of indicators were used in this process in order to evaluate the ability of a country to respond to any given case of VHF. Timely intervention and isolation of the index case is critical

for minimising subsequent human-to-human transmission *(Heymann et al., 2015)*. Therefore the indicators included in the vulnerability and lack of coping capacity components assess a variety of different key aspects including: institutional effectiveness (evaluated using measures of government effectiveness, Corruption Perception Index and the Hyogo Framework for Action), communication (measuring adult literacy rates, access to electricity, internet usage and mobile phone subscriptions), sanitation (measured by percentage access to improved water sources and sanitation), remoteness (considered through road density and percentage rural population) and access to health care (evaluated by healthcare expenditure metrics, measles immunisation coverage, density of physicians, treatment seeking behaviour and distance to health facilities). These indicators are considered as proxies for the ability of a national health team to identify and respond to a VHF outbreak in a timely and effective manner.

The relative ranking combined the risk of an initial spillover infection with these national assessments of response effectiveness. The two layers were combined to produce final national level local outbreak risk assessments by taking the geometric mean of the two components with equal weighting, using the following equation, derived from the INFORM system:

$$Local\ outbreak\ risk = Risk\ of\ spillover^{\frac{1}{2}} * Vulnerability\ and\ Lack\ of\ Coping\ Capacity^{\frac{1}{2}}$$

Assessing risk of disease spread to neighbouring districts

The relative risk of an outbreak, once established at a community level, subsequently spreading elsewhere can be evaluated by considering the time taken to travel across the district. It is hypothesised therefore that districts that are easier to travel across will be more likely to seed infections in other districts. An African surface of travel time per pixel was generated [Tatem *et al.* pers. comms.] and national average travel time was calculated by taking the mean pixel

value. This measure is independent of country size as the value references the 5km x 5km pixel gridded surface. The final relative risk of subsequent transmission resulting from a local outbreak was calculated as follows, derived from the INFORM system:

$$Widespread\ outbreak\ risk = Risk\ of\ local\ outbreak^{\frac{1}{2}} * Travel\ time\ across\ pixel^{\frac{1}{2}}$$

Results

Overlapping zoonotic niches of VHFs and risk of index cases

Overlaying the predicted zoonotic niches of the four VHFs identifies a number of areas where several of these viruses could co-occur (Figure 2). Areas with two or more VHFs present are found in Equatorial Africa, primarily Gabon, the Democratic Republic of Congo (DRC), Cameroon and Uganda, as well as several countries in West Africa, most notably around Guinea, Liberia and Sierra Leone. The highest overlap is associated with Marburg virus and Ebola virus, with areas predicted to be at risk of Lassa fever in West Africa, often also at risk of filoviral spillover. CCHF is typically found on the periphery of the other predicted zoonotic niches, but does co-occur with other VHFs in the Horn of Africa and other areas of East Arica. No region is predicted to have all four viruses present.

Introducing population surfaces in relation to these niches allows us to understand the true nature of spillover risk in these areas. Whilst large regions of Equatorial Africa are predicted to be areas where the zoonotic niche is present, they are primarily found in locations with relatively low population numbers. In contrast, we see that Nigeria, in spite of comparatively small regions of risk from each of the diseases, is the highest ranked country in terms of population living in areas of spillover risk due to the much higher population density within these regions (Table 1 and Figure 3).

Local outbreak risk

Figure 4 displays the combined vulnerability and lack of coping capacity as derived from the INFORM index, indicating the areas least capable of preventing continued transmission of a given VHF are Somalia, Chad and South Sudan. Table 2 identifies the countries with the

greatest risk of an index case leading to continued human-to-human transmission from VHFs. Nigeria, Guinea and Côte d'Ivoire have the highest relative risk of a local outbreak of any VHF occurring (Figure 5), with Ethiopia at highest risk from MARV and CCHF, Guinea from LASV and the DRC from EBOV.

Widespread outbreak risk

Figure 6 displays average time to travel across a 5km x 5km pixel aggregated to a district level. When combined with risk evaluations for local outbreaks, we see that Nigeria, Guinea and Côte d'Ivoire remain at the top of the relative ranking, with other, predominantly West African nations also in the top ten (Table 3 and Figure 7). Areas in East Africa similarly rank highly due to a combination of relative ease of travel and high relative ranking of spillover infections occurring in the first place.

Discussion

This paper outlines a broad scale process for understanding the risk that viral haemorrhagic fevers pose to the African continent. Using a variety of indicators, from potential zoonotic transmission maps, to measures of healthcare infrastructure and travel times across regions, this study assesses VHF risk at three key stages: (a) the initial spillover into humans, (b) the potential for a local outbreak and (c) the potential for the outbreak to spread to other geographic locations. This shows that some areas of Africa, should an index case lead to a local outbreak, are inherently more likely to see more widespread infections occurring.

The inclusion of population vulnerability and coping capacity measures demonstrate that for some countries, in spite of having fewer people living in areas of risk, their ability to intervene and minimise human-to-human transmission is less effective compared to others. Somalia and Chad, for instance, do not feature on the top ten relative ranking for CCHF spillover risk, but do rank highly on the risk of a local outbreak. On the other hand, South Africa, in spite of having the largest population living in areas of spillover risk for CCHF has a more robust healthcare system that would likely mitigate some of the risk of further secondary transmission, hence a lower ranking for outbreaks.

Similarly, measures of travel time highlight interesting regional differences. Figure 6 clearly shows that the Mediterranean coast, West and South Africa are very highly connected, with some of the shortest travel times on the continent. Unsurprisingly therefore, West African countries, which have some of the highest risk associated with spillover events and local outbreaks, rank highly on the more widespread risk rating. This can be contrasted with the DRC where we see large areas of the country that are relatively isolated, hence the drop in relative ranking from fourth for local outbreak risk to ninth in more widespread outbreak potential. Comparing and contrasting the two EVD outbreaks of 2014 suggests that this relative isolation

could have significant impacts on the resulting population dynamics *(Maganga et al., 2014; WHO Ebola response team, 2014)*.

It is important to recognise that the need to remain consistent with information sources used at the continental scale resulted in a number of useful metrics not being included. One area of future development is refining the spillover risk layers *(Plowright et al., 2015)*. As it stands, the layers used are the only global surfaces for this quartet of diseases which have been generated using the same methodology. As such they provide a baseline for future refinement and are a key component for this continued analysis. However, it should be noted that a variety of information, relating to host identities *(Ergonul and Whitehouse, 2010)*, reservoir host viral dynamics *(Olival and Hayman, 2014)* and human-host interactions *(Woldehanna and Zimicki, 2015)*, remain critical unknowns in the wider epidemiology of these diseases. Including this information as refining layers will be crucial for future iterations for these maps and will help identify regions in the world where the risk of a zoonotic spillover event occurring is far less. For example, the assumption that risk is a function of population living in that region is likely more nuanced; understanding the proportion of the population living in, or travelling to, these regions and performing risky activities (such as hunting and butchering wildlife) is key *(Wolfe et al., 2005)*.

Similarly, the inclusion of more VHF specific measures, such as availability of personal protective equipment and isolation units or training in their use, would be important to include directly rather than using indirect proxies related to general health infrastructure. In response to the West African outbreak, the World Health Organization (WHO) prepared an EVD preparedness checklist which was used to evaluate a variety of different resources considered necessary for processing EVD patients *(WHO, 2015)*. However, their distribution was limited to those countries deemed at risk during the current outbreak. Integrating questionnaires such as

this across the much wider areas at risk across the continent would allow for a first attempt at better understanding VHF-specific coping capacity in a more formal manner.

Finally, increasingly refined estimates of population connectivity and movement patterns are being developed and would provide a more accurate picture of risk of subsequent spread. Current advances in movement modelling have incorporated a variety of different metrics to assess potential spread, including information related to diaspora data and other migration patterns *(Garcia et al., 2014)*, mobile phone call data records *(Wesolowski et al., 2014)*, air travel patterns *(Bogoch et al., 2014)* and gravity model and radiation model hybrids *(Kraemer and Golding, 2015; Simini et al., 2012; Zipf, 1946)*. A more nuanced approach integrating these data sources will enable a much better understanding of how and why populations move, invaluable in epidemiological contexts where human networks will likely shape subsequent infection patterns *(Wesolowski et al., 2012)*. Also worthy of consideration is the travel patterns that result in emergency situations. Analyses after the 2010 Haiti earthquake showed predictable patterns in human movement resulted but were similar to those observed during Christmas and New Year family celebrations rather than day-to-day travel routines *(Lu et al., 2012)*.

Irrespective of these limitations, all three core outputs (Figures 3, 5 and 7) contain useful information that can provide immediate benefit to national ministries of health, as well as international organisations. Figures 2 and 3, outlining the areas thought to be suitable for zoonotic transmission, can act as an important guide for identifying regions where surveillance for these rare pathogens should be focussed, particularly in already identified reservoir hosts *(Levinson et al., 2013)*. Similarly, Figure 5 identifies countries to be prioritised for distributing diagnostic equipment, stockpiling necessary equipment and potentially in the future considering where to best focus any successful vaccine deployment or other drug therapeutics *(Brady,*

2014). Similarly, these layers can help inform discussions related to general healthcare infrastructure strengthening *(Kieny and Dovlo, 2015)*. Figures 6 and 7 highlight the importance that travel and internal connections can have in future epidemic scenarios, recognising the inherent risk of some regions in comparison to other locations. The potential role that human population patterns could have in driving more widespread outbreaks of directly transmissible diseases should be used to advocate for increased awareness of integrating mobility surfaces into future risk scenarios *(Bengtsson et al., 2011; Kraemer et al., 2015)*.

The outbreak in West Africa may represent a perfect storm of unfortunate events *(Piot, 2014)*, whose true impact could never have been predicted when the first index case became ill. As we try to understand what lessons we can learn from this tragedy, focus must turn to how we consider areas at future risk of VHF outbreaks. This paper outlines a method for rationalising future strategies to prepare for VHF cases and potential outbreaks. Refining and improving upon the data included and the ways of categorising risk from these diseases should be a high priority for the international health community, particularly given the benefits that this method could provide in informing other directly transmitted infectious diseases prevention approaches.

Figures

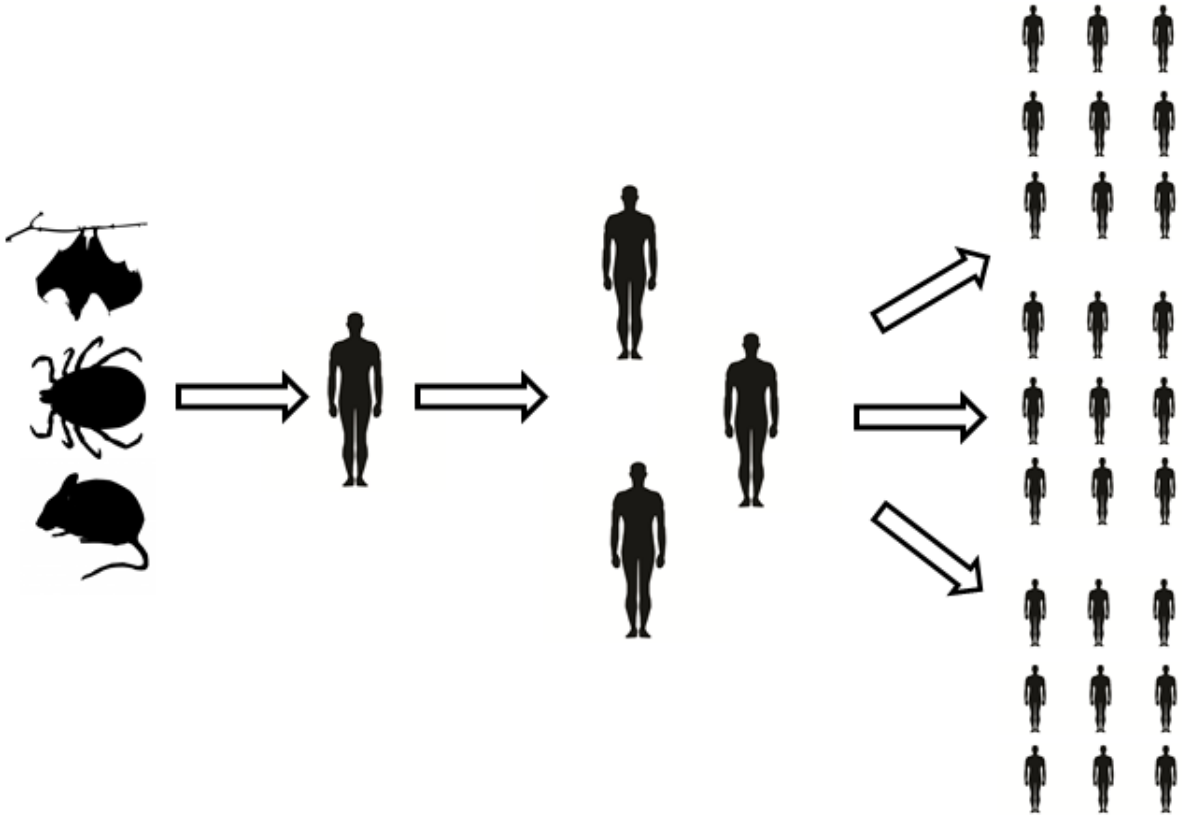


Figure 1: Concepts of risk for a viral haemorrhagic fever outbreak. Each arrow represents an important step in the disease transmission process. The first indicates viral transmission from reservoir hosts (including bats, ticks and rodents) to an index human. Subsequent human-to-human transmission results in a local outbreak (second arrow) with the potential for more widespread transmission occurring in a variety of different geographical locations (third arrow).

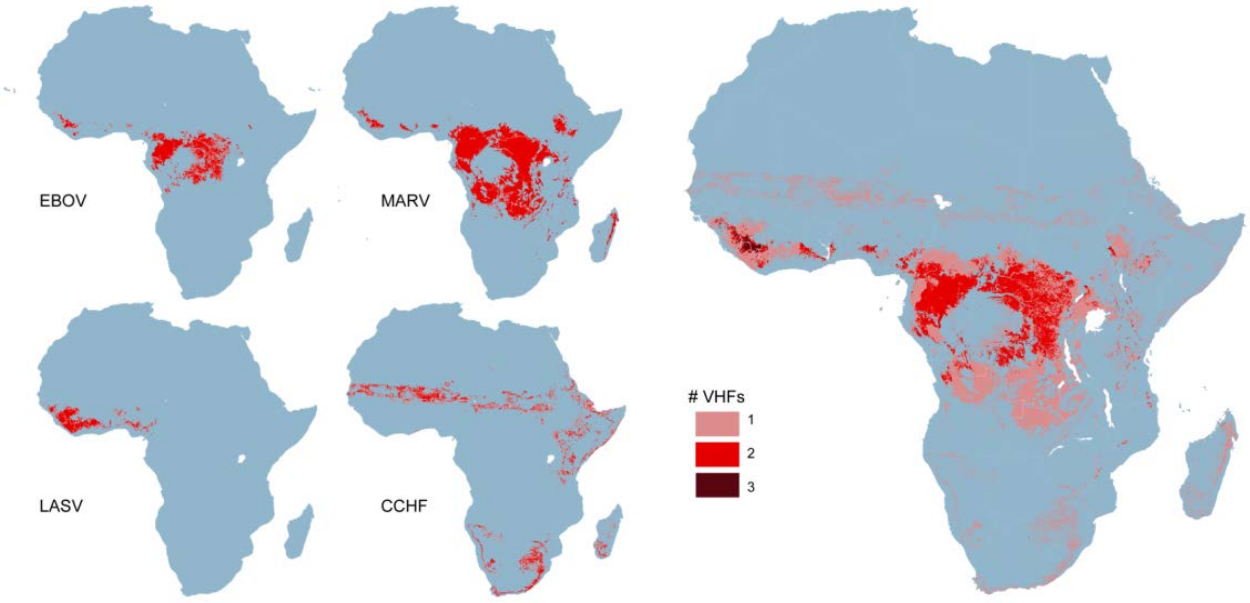


Figure 2: Predicted zoonotic niches for viral haemorrhagic fevers and their co-distribution. The four panels on the left demonstrate the at-risk (in red) / not-at-risk (in blue) binary layers derived from the predicted zoonotic niche maps for Ebola virus disease (EBOV), Marburg virus disease (MVD), Lassa fever (LASV) and Crimean-Congo haemorrhagic fever (CCHF). The right panel shows the overlap of these four maps, with regions with no viral haemorrhagic fevers (VHFs) present in blue, one VHF present in pink, two in red and three different VHFs in brown.

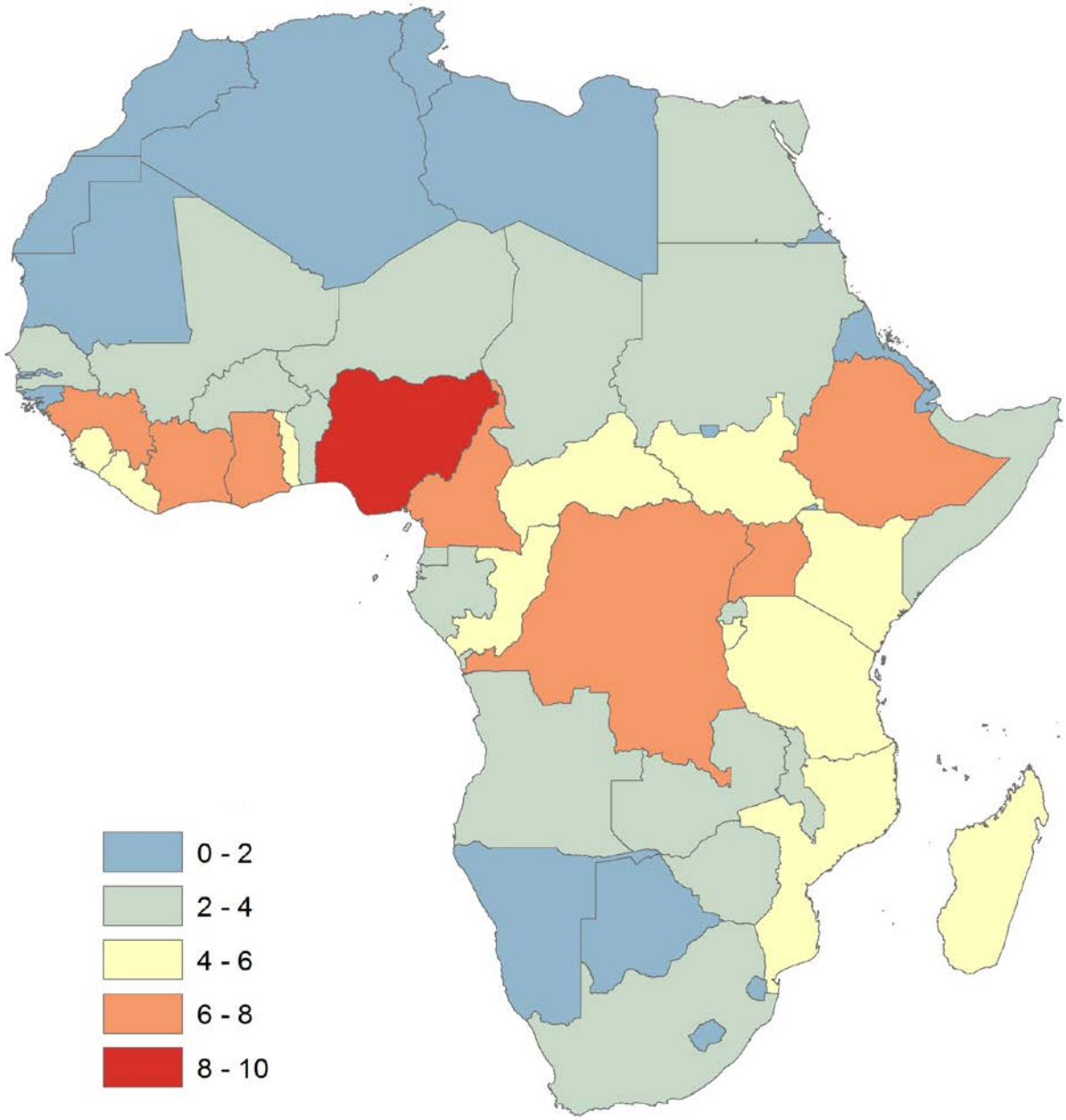


Figure 3: Risk of an initial index case for a VHF arising in an Africa nation. Here relative score when population layers and zoonotic niche maps are combined are displayed. Areas in red have the highest scores, and are therefore most likely to see index cases of viral haemorrhagic fevers. Areas in blue are least likely.

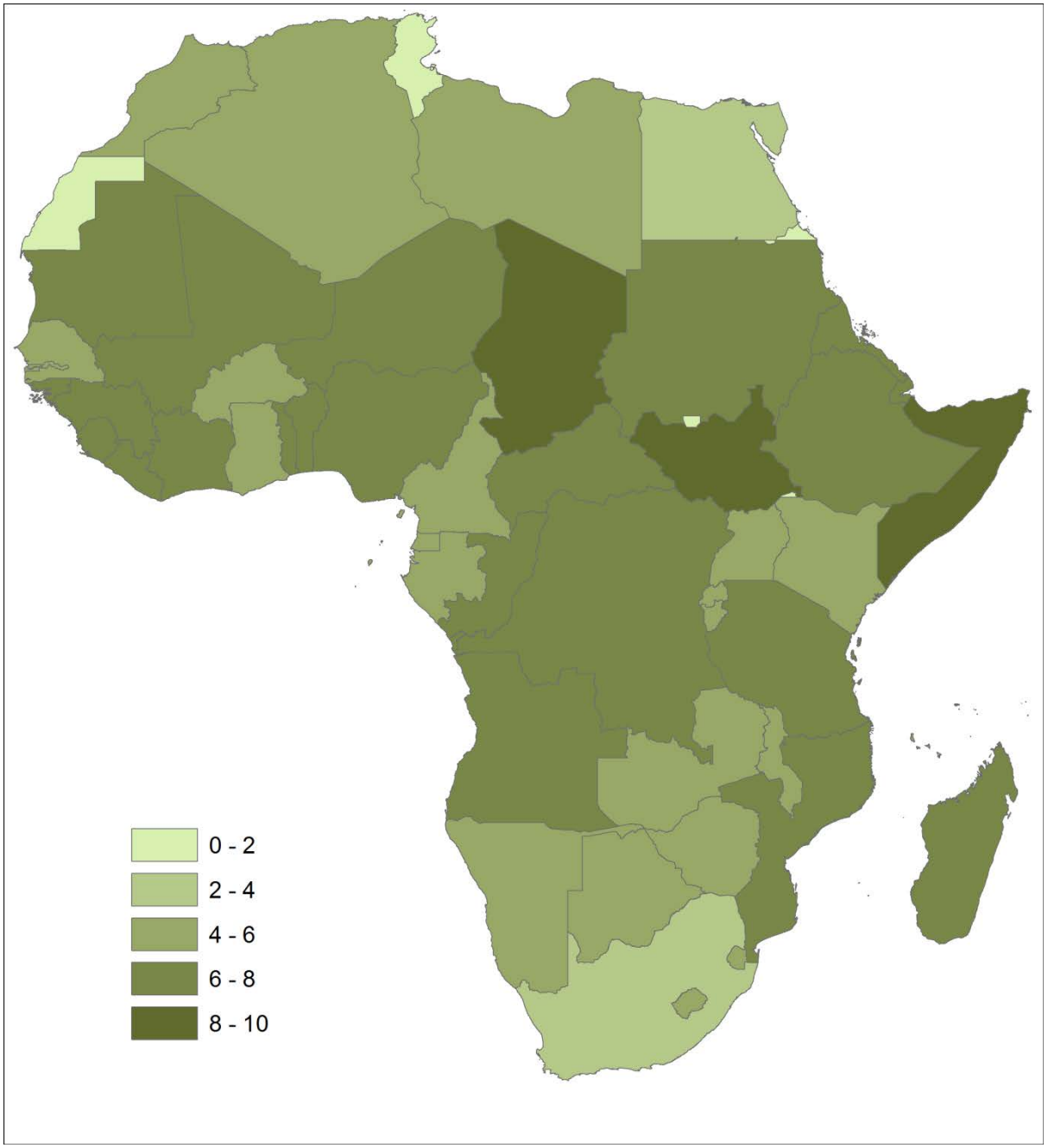


Figure 4: Ranking by vulnerability and lack of coping capacity. Derived from the Index for Risk Management tool *(de Groeve et al., 2015)*, colour indicates the relative score. Areas in dark green have the highest score and represent the most vulnerable countries. Areas of light green are most capable of detecting and adequately treating a viral haemorrhagic fever case.

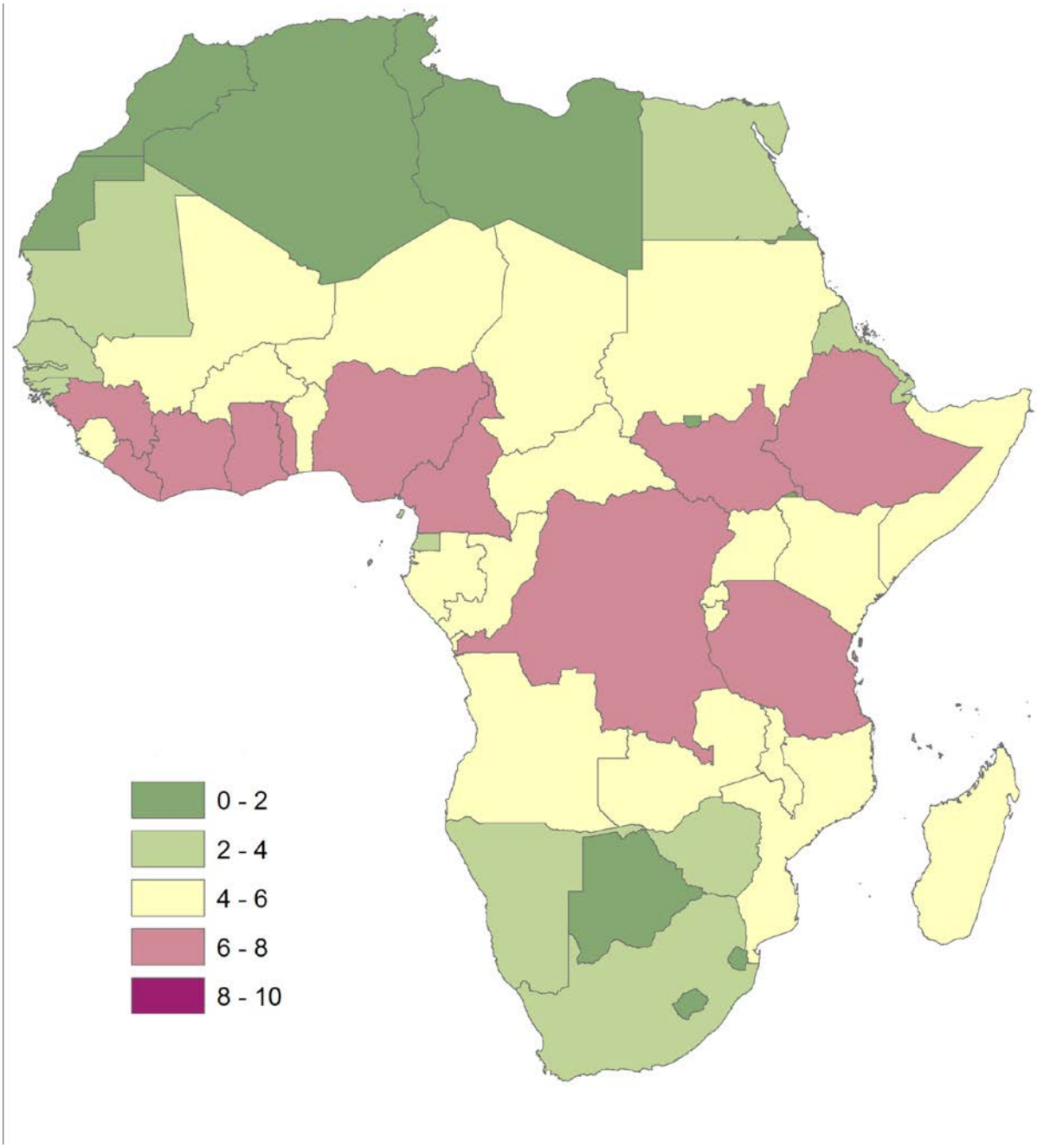


Figure 5: Ranking by local outbreak risk for combined VHFs. Here the vulnerability and lack of coping capacity index is combined with the risk of an initial index case arising. Areas in purple represent the regions with a higher relative score, suggesting these countries are most likely to see a local outbreak resulting from a viral haemorrhagic fever index case. Areas in dark green are least likely to see this transition.

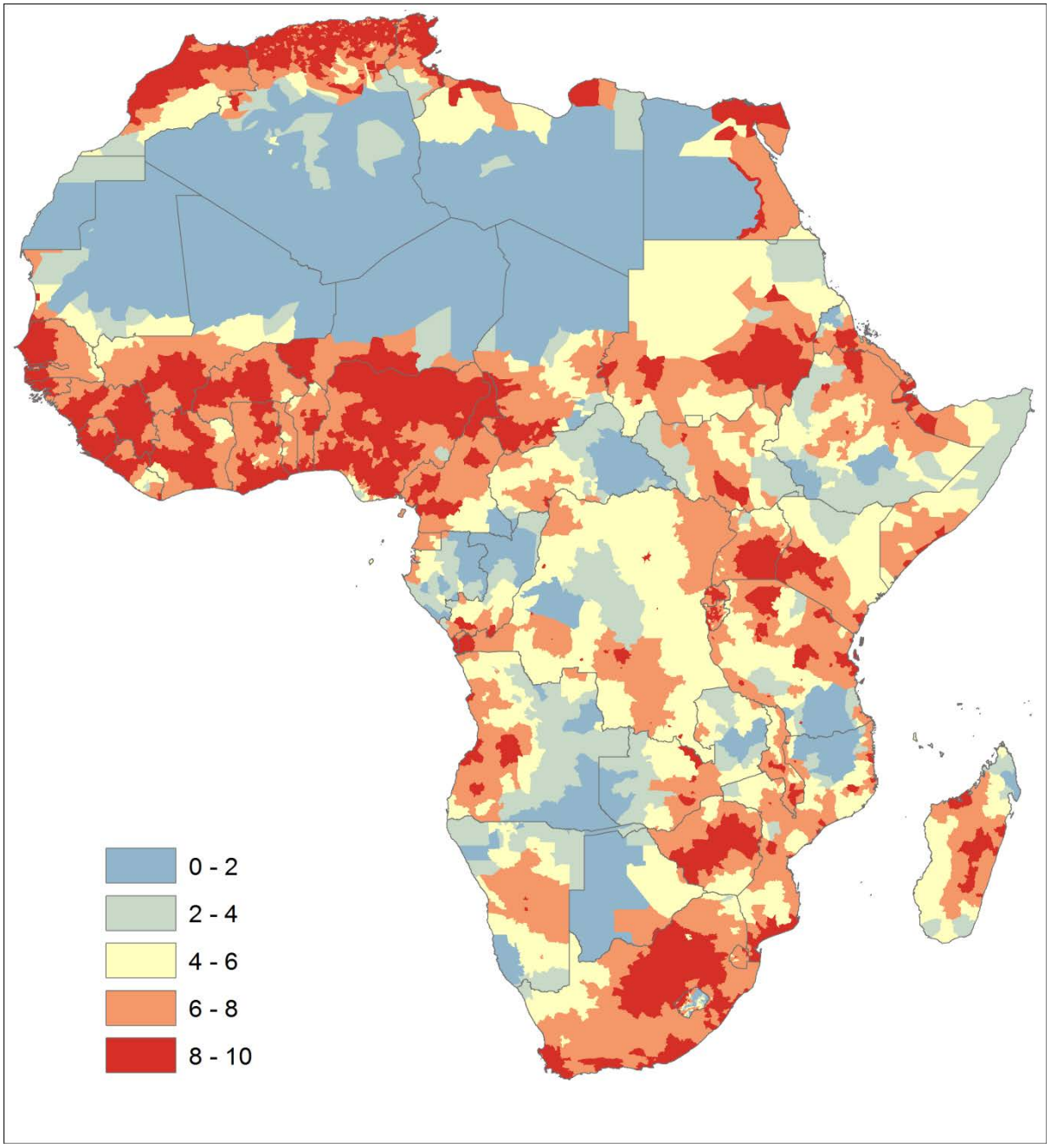


Figure 6 Ranking of districts in Africa based upon average time to cross 5km x 5km pixel.

Africa is divided up into districts and the average time to travel across a pixel is calculated. Regions in red have the shortest travel time, whilst districts in blue indicate much longer travel times in order to cross the same 5km x 5km pixel.

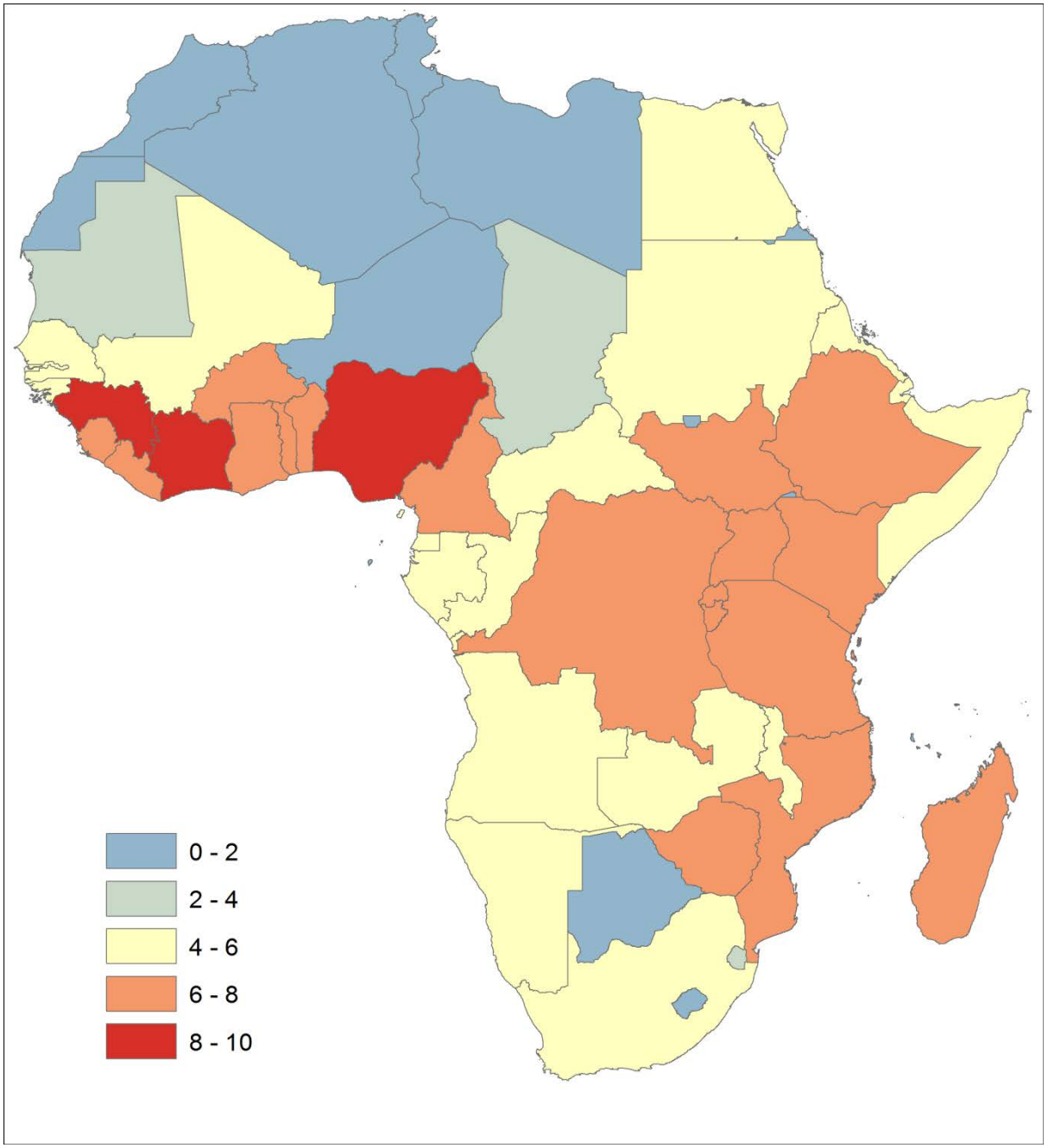


Figure 7 Ranking by widespread outbreak risk for combined VHF. This maps shows the relative score after national mean travel time across a 5km x 5km pixel was combined with risk of local outbreak of a viral haemorrhagic fever occurring. Countries in red have the highest relative scores, suggesting countries more likely to see widespread outbreaks in multiple geographic locations resulting, whilst areas in blue are unlikely to see such outbreaks.

Tables

Table 1. Top ten ranking of African countries by combined VHF spillover risk. Countries in red have seen index cases; countries in blue have reported seropositive humans or reported infection in animals. Numbers in parentheses represent relative score.

EBOV Ranking	MARV Ranking	LASV ranking	CCHF Ranking	Combined Ranking
DRC (10)	DRC (10)	Nigeria (10)	Ethiopia (10)	Nigeria (9.33)
Nigeria (8.82)	Uganda (9.97)	Côte d'Ivoire (9.41)	South Africa (10)	Ghana (7.73)
Côte d'Ivoire (8.65)	Ethiopia (9.43)	Guinea (7.72)	Nigeria (9.97)	Côte d'Ivoire (7.70)
Guinea (8.50)	Kenya (9.38)	Ghana (7.41)	Kenya (9.72)	Cameroon (7.67)
Cameroon (8.42)	Cameroon (8.78)	Sierra Leone (8.52)	Niger (9.69)	Guinea (7.50)
Uganda (8.35)	Nigeria (8.54)	Liberia (8.31)	Sudan (9.64)	DRC (6.96)
CAR (7.75)	Tanzania (8.04)	Benin (6.91)	Tanzania (8.97)	Uganda (6.17)
Liberia (7.10)	Rwanda (7.86)	Cameroon (5.58)	Egypt (8.84)	Ethiopia (6.10)
Sierra Leone (6.83)	Angola (7.86)	Togo (5.57)	Senegal (8.74)	Togo (5.99)
Ghana (6.52)	Guinea (7.72)	Mali (5.46)	Madagascar (8.63)	Liberia (5.77)

Table 2. Top ten ranking of African countries by local outbreak risk. Numbers in parentheses represent relative score.

EBOV Ranking	MARV Ranking	LASV ranking	CCHF Ranking	Combined Ranking
DRC (8.19)	Ethiopia (8.35)	Guinea (8.32)	Ethiopia (8.60)	Nigeria (7.97)
Guinea (8.09)	DRC (8.19)	Nigeria (8.25)	Somalia (8.50)	Guinea (7.60)
CAR (7.77)	Guinea (7.71)	Côte d'Ivoire (7.64)	Niger (8.47)	Côte d'Ivoire (6.91)
Nigeria (7.75)	CAR (7.66)	Sierra Leone (7.61)	Chad (8.26)	DRC (6.83)
Côte d'Ivoire (7.32)	Nigeria (7.62)	Liberia (7.29)	Nigeria (8.23)	Cameroon (6.73)
Cameroon (7.04)	Uganda (7.47)	Benin (6.90)	Sudan (8.21)	Ethiopia (6.72)
Uganda (6.84)	Tanzania (7.39)	Ghana (6.71)	Madagascar (8.04)	Togo (6.61)
Sierra Leone (6.81)	Madagascar (7.27)	Togo (6.38)	Mali (7.93)	Ghana (6.34)
Liberia (6.74)	Cameroon (7.20)	Mali (6.32)	Tanzania (7.81)	South Sudan (6.13)
South Sudan (6.56)	Kenya (7.18)	Cameroon (5.74)	South Sudan (7.53)	Tanzania (6.12)

Table 3. Top ten ranking of African countries by widespread outbreak risk. Numbers in parentheses represent relative score.

EBOV Ranking	MARV Ranking	LASV ranking	CCHF Ranking	Combined Ranking
Nigeria (7.86)	Guinea (8.67)	Nigeria (8.11)	Nigeria (8.10)	Nigeria (8.82)
Guinea (7.84)	Nigeria (8.63)	Guinea (7.95)	Ethiopia (7.60)	Guinea (8.61)
DRC (7.48)	Uganda (8.18)	Côte d'Ivoire (7.27)	DRC (7.03)	Côte d'Ivoire (8.08)
Côte d'Ivoire (7.11)	Rwanda (7.99)	Sierra Leone (6.75)	Tanzania (6.91)	Ghana (7.77)
Cameroon (6.89)	DRC (7.97)	Liberia (6.66)	Togo (6.90)	Togo (7.75)
CAR (6.65)	Ethiopia (7.91)	Togo (6.49)	Guinea (6.79)	Sierra Leone (7.72)
Togo (6.47)	Cameroon (7.90)	Cameroon (6.21)	South Sudan (6.79)	Cameroon (7.63)
Liberia (6.40)	Sierra Leone (7.84)	Benin (5.77)	Cameroon (6.78)	Liberia (7.46)
Sierra Leone (6.38)	Madagascar (7.75)	Mali (5.66)	Côte d'Ivoire (6.66)	DRC (7.28)
Ethiopia (6.34)	Ghana (7.69)	Guinea-Bissau (4.08)	Madagscar (6.66)	Uganda (7.26)

References

Ajayi NA, Nwigwe CG, Azuogu BN, Onyire BN, Nwonwu EU, et al. 2013. Containing a Lassa fever epidemic in a resource-limited setting: outbreak description and lessons learned from Abakaliki, Nigeria (January-March 2012). *Int J Infect Dis* **17**: 1011-1016. doi:10.1016/j.ijid.2013.05.015.

Altaf A, Luby S, Ahmed AJ, Zaidi N, Khan AJ, et al. 1998. Outbreak of Crimean-Congo haemorrhagic fever in Quetta, Pakistan: contact tracing and risk assessment. *Trop Med Int Health* **3**: 878-882. doi:10.1046/j.1365-3156.1998.00318.x.

Amman BR, Carroll SA, Reed ZD, Sealy TK, Balinandi S, et al. 2012. Seasonal pulses of Marburg virus circulation in juvenile *Rousettus aegyptiacus* bats coincide with periods of increased risk of human infection. *PLoS Pathog* **8**: e1002877. doi:10.1371/journal.ppat.1002877.

Amorosa V, MacNeil A, McConnell R, Patel A, Dillon KE, et al. 2010. Imported Lassa fever, Pennsylvania, USA, 2010. *Emerg Infect Dis* **16**: 1598-1600. doi:10.3201/eid1610.100774.

Atkinson B, Latham J, Chamberlain J, Logue C, O'Donoghue L, et al. 2012. Sequencing and phylogenetic characterisation of a fatal Crimean - Congo haemorrhagic fever case imported into the United Kingdom, October 2012. *Euro Surveill* **17**: 7-10.

Bannister B. 2010. Viral haemorrhagic fevers imported into non-endemic countries: risk assessment and management. *Brit Med Bull* **95**: 193-225. doi:10.1093/Bmb/Ldq022.

Baron RC, McCormick JB, Zubeir OA. 1983. Ebola virus disease in southern Sudan - hospital dissemination and intrafamilial spread. *Bull World Health Organ* **61**: 997-1003.

Bengtsson L, Lu X, Thorson A, Garfield R, von Schreeb J. 2011. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. *PLoS Med* **8**: e1001083. doi:10.1371/journal.pmed.1001083.

Bogoch, II, Creatore MI, Cetron MS, Brownstein JS, Pesik N, et al. 2014. Assessment of the potential for international dissemination of Ebola virus via commercial air travel during the 2014 west African outbreak. *Lancet*: 10.1016/S0140-6736(1014)61828-61826. doi:10.1016/S0140-6736(14)61828-6.

Boumandouki P, Formenty P, Epelboin A, Campbell P, Atsangandoko C, et al. 2005. [Clinical management of patients and deceased during the Ebola outbreak from October to December 2003 in Republic of Congo]. *Bull Soc Pathol Exot* **98**: 218-223.

Brady O. 2014. Scale up the supply of experimental Ebola drugs. *Nature* **512**: 233. doi:10.1038/512233a.

CDC. 2014 Ebola outbreak in West Africa - case counts. Available: http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/case-counts.html. Accessed: 5th June 2015

de Groeve T, Poljansek K, Vernaccini L. 2015. Index for Risk Management - INFORM. Concept and Methodology. Ispra: European Commission Joint Research Centre.

Dudas G, Rambaut A. 2014. Phylogenetic analysis of Guinea 2014 EBOV ebolavirus outbreak. *PLoS Curr* **6**: ecurrents.outbreaks.84eefe85ce43ec89dc80bf0670f0677b0678b0417d. doi:10.1371/currents.outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d.

Ergonul O. 2006. Crimean-Congo haemorrhagic fever. *Lancet Infect Dis* **6**: 203-214. doi:10.1016/S1473-3099(06)70435-2.

Ergonul O, Whitehouse CA, editors (2010) Crimean-Congo hemorrhagic fever. A global perspective: Springer. 328 p.

Francesconi P, Yoti Z, Declich S, Onek PA, Fabiani M, et al. 2003. Ebola hemorrhagic fever transmission and risk factors of contacts, Uganda. *Emerg Infect Dis* **9**: 1430-1437. doi:10.3201/eid0911.030339.

Garcia AJ, Pindolia DK, Lopiano KK, Tatem AJ. 2014. Modeling internal migration flows in sub-Saharan Africa using census microdata. *Migrat Stud* **3**: 89-110. doi:10.1093/migration/mnu036.
Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, et al. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**: 1369-1372. doi:10.1126/science.1259657.
Gonzalez JP, Pourrut X, Leroy E. 2007. Ebolavirus and other Filoviruses. *Curr Top Microbiol* **315**: 363-387.
Hewlett BS, Epelboin A, Hewlett BL, Formenty P. 2005. Medical anthropology and Ebola in Congo: cultural models and humanistic care. *Bull Soc Pathol Exot* **98**: 230-236.
Heymann DL, Chen L, Takemi K, Fidler DP, Tappero JW, et al. 2015. Global health security: the wider lessons from the west African Ebola virus disease epidemic. *Lancet* **385**: 1884-1901. doi:10.1016/S0140-6736(15)60858-3.
Kieny MP, Dovlo D. 2015. Beyond Ebola: a new agenda for resilient health systems. *Lancet* **385**: 91-92.
Kraemer MUG, Golding N. Ebola spread in West Africa. Available: http://seeg-oxford.github.io/ebola-spread/contact. Accessed: 5th June 2015
Kraemer MUG, Hay SI, Pigott DM, Smith DL, Wint GRW, et al. 2015. Progress and challenges in infectious disease cartography. *Trends Parasitol*: under review.
Lash RR, Brunsell NA, Peterson AT. 2008. Spatiotemporal environmental triggers of Ebola and Marburg virus transmission. *Geocarto Int* **23**: 451-466.
Leroy EM, Epelboin A, Mondonge V, Pourrut X, Gonzalez JP, et al. 2009. Human Ebola outbreak resulting from direct exposure to fruit bats in Luebo, Democratic Republic of Congo, 2007. *Vector Borne Zoonotic Dis* **9**: 723-728. doi:10.1089/vbz.2008.0167.
Leroy EM, Kumulungui B, Pourrut X, Rouquet P, Hassanin A, et al. 2005. Fruit bats as reservoirs of Ebola virus. *Nature* **438**: 575-576. doi:10.1038/438575a.
Levinson J, Bogich TL, Olival KJ, Epstein JH, Johnson CK, et al. 2013. Targeting surveillance for zoonotic virus discovery. *Emerg Infect Dis* **19**: 743-747. doi:10.3201/eid1905.121042.
Lu X, Bengtsson L, Holme P. 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proc Natl Acad Sci USA* **109**: 11576-11581. DOI 10.1073/pnas.1203882109.
Maganga GD, Kapetshi J, Berthet N, Ilunga BK, Kabange F, et al. 2014. Ebola virus disease in the Democratic Republic of Congo. *N Engl J Med* **371**: 2083-2091. doi:10.1056/Nejmoa1411099.
McCormick JB, Webb PA, Krebs JW, Johnson KM, Smith ES. 1987. A prospective study of the epidemiology and ecology of Lassa fever. *J Infect Dis* **155**: 437-444.
Messina JP, Pigott DM, Golding N, Duda KA, Brownstein JS, et al. 2015. The global distribution of Crimean-Congo hemorrhagic fever. *Trans R Soc Trop Med Hyg*: accepted.
Mylne A, Brady OJ, Huang Z, Pigott DM, Golding N, et al. 2014. A comprehensive database of the geographic spread of past human Ebola outbreaks. *Sci Data* **1**: e140042. doi:10.1038/sdata.2014.42.
Mylne A, Pigott DM, Longbottom J, Shearer F, Duda KA, et al. 2015. Mapping the zoonotic niche of Lassa fever in Africa. *Trans R Soc Trop Med Hyg*: accepted.
Olival KJ, Hayman DTS. 2014. Filoviruses in bats: current knowledge and future directions. *Viruses* **6**: 1759-1788. doi:10.3390/V6041759.
Pigott DM, Golding N, Mylne A, Huang Z, Henry AJ, et al. 2014. Mapping the zoonotic niche of Ebola virus disease in Africa. *eLife* **3**: e04395. doi:10.7554/eLife.04395.
Pigott DM, Golding N, Mylne A, Huang Z, Weiss DJ, et al. 2015. Mapping the zoonotic niche of Marburg virus disease in Africa. *Trans R Soc Trop Med Hyg*: doi:10.1093/trstmh/trv024.
Piot P. 2014. Ebola's perfect storm. *Science* **345**: 1221. doi:10.1126/science.1260695.

Plowright RK, Eby P, Hudson PJ, Smith IL, Westcott D, et al. 2015. Ecological dynamics of emerging bat virus spillover. *Proc R Soc Lond B Biol Sci* **282**: e20142124. doi:10.1098/Rspb.2014.2124.

Simini F, Gonzalez MC, Maritan A, Barabasi AL. 2012. A universal model for mobility and migration patterns. *Nature* **484**: 96-100. doi:10.1038/nature10856.

Timen A, Koopmans MPG, Vossen ACTM, van Doornum GJJ, Gunther S, et al. 2009. Response to imported case of Marburg hemorrhagic fever, the Netherlands. *Emerg Infect Dis* **15**: 1171-1175. doi:10.3201/eid1508.090051.

Towner JS, Amman BR, Sealy TK, Carroll SAR, Comer JA, et al. 2009. Isolation of genetically diverse Marburg viruses from Egyptian fruit bats. *PLoS Pathog* **5**: e1000536. doi:10.1371/journal.ppat.1000536.

Towner JS, Khristova ML, Sealy TK, Vincent MJ, Erickson BR, et al. 2006. *Marburgvirus* genomics and association with a large hemorrhagic fever outbreak in Angola. *J Virol* **80**: 6497-6516. doi:10.1128/JVI.00069-06.

Troup JM, White HA, Fom AL, Carey DE. 1970. An outbreak of Lassa fever on the Jos plateau, Nigeria, in January-February 1970. A preliminary report. *Am J Trop Med Hyg* **19**: 695-696.

Wesolowski A, Buckee CO, Bengtsson L, Wetter E, Lu X, et al. 2014. Commentary: containing the Ebola outbreak - the potential and challenge of mobile network data. *PLoS Currents Outbreaks* **1**: 10.1371/currents.outbreaks.0177e1377fcf52217b52218b634376e634372f634373efc634375e. doi:10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e.

Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, et al. 2012. Quantifying the impact of human mobility on malaria. *Science* **338**: 267-270. doi:10.1126/science.1223467.

WHO. Consolidated Ebola virus disease preparedness checklist. Available: http://www.who.int/csr/resources/publications/ebola/ebola-preparedness-checklist/en/. Accessed: 5th June 2015

WHO Ebola response team. 2014. Ebola virus disease in West Africa - the first 9 months of the epidemic and forward projections. *N Engl J Med* **371**: 1481-1495. doi:10.1056/NEJMoa1411100.

Woldehanna S, Zimicki S. 2015. An expanded One Health model: Integrating social science and One Health to inform study of the human-animal interface. *Soc Sci Med* **129**: 87-95. doi:10.1016/j.socscimed.2014.10.059.

Wolfe ND, Daszak P, Kilpatrick AM, Burke DS. 2005. Bushmeat Hunting, deforestation, and prediction of zoonotic disease emergence. *Emerg Infect Dis* **11**: 1822-1827. doi:10.3201/eid1112.040789.

WorldPop. WorldPop project. Available: http://www.worldpop.org.uk/. Accessed: June 2015

Zipf GK. 1946. The P_1 P_2/D hypothesis: on the intercity movement of persons. *Am Sociol Rev* **11**: 677-686.

Chapter 9

Discussion

This thesis has identified the next key targets for infectious disease cartography, demonstrated the use of species distribution models in evaluating the global distribution of a number of important zoonotic diseases and suggested ways these maps can provide evidence for making health policy decisions at both local and global scales.

These maps provide a useful starting point for framing future surveillance, intervention and policy decisions. In this discussion I will detail how these maps can be used in their current state and what further information is required to improve these outputs, both in terms of accuracy of risk and in what additional information they can provide. I will then outline how these maps can act as a framework for considering a wider array of zoonotic diseases and how such outputs can inform a number of epidemiological questions.

9.1. Chapter summary

The first research chapter of this thesis provided a methodology for considering next targets in infectious disease cartography. Unlike previous methodologies for identifying priority disease targets which incorporate many rounds of expert opinion and consultations within a questionnaire schema (such as Delphi surveys *(Gilsdorf and Krause, 2011; Havelaar et al., 2010; Krause and Prioritisation Working Group, 2008)*), this approach fully utilises quantitative metrics. The integration of Global Burden of Disease estimates *(GBD 2013 Disease*

and Injury Incidence and Prevalence Collaborators, 2015; GBD 2013 Mortality and Causes of Death Collaborators, 2015), h-indices *(McIntyre et al., 2011)*, status as a nationally notifiable disease and inclusion in non-governmental organisation's mission statements allows for complete transparency and for regular updating. This allows for prioritisation of those diseases where maps could have the biggest impact and benefit public health policy the most.

Employing more complex modelling approaches often requires data that are simply not available at a global level. Streamlined household surveys that allow for detailed studies at the population level remain constrained to those diseases that are comparatively well funded (such as malaria *(Pigott et al., 2012)*) or have well-established and organised drug delivery and surveillance programmes in place (such as those neglected tropical diseases where mass-drug administration is possible, including lymphatic filariasis, schistosomiasis and soil-transmitted helminths *(Global Atlas of Helminth Infections, 2014)*). However, for the vast majority of conditions such data sources are not available. Indeed, of the 45 conditions listed in the top 15 clusters to be prioritised, only 24% have sufficient data to allow for model-based geostatistical approaches *(Hay et al., 2013a; Pigott et al., 2015b)*. Chapters 3, 4, 5, 6 and 7 demonstrate how SDMs can use occurrence records, such as the locations of reported cases, rather than population level survey information, to provide estimates of the global distribution of these diseases. Similarly by comparing leishmaniasis with the viral haemorrhagic fevers we can see just how adaptable this modelling framework is and assess its capability for analysing complex multi-vectored disease systems as well as those that involve mammalian reservoir species.

The final research chapter, which ties the distribution maps of the VHFs to indicators of sustained human outbreak risk, demonstrates one of the ways that these outputs can be used to assess future policy options. Accurately predicting the precise location of spillover human infections will always be challenging. By considering risk at three different stages of a viral

haemorrhagic fever outbreak (potential risk of an index case, potential spread amongst humans locally and potential wider spread to neighbouring regions) we can begin to understand where such spillover events may cause the biggest threat and respond accordingly, helping to strengthen weaker health systems or identify areas where better surveillance is needed.

9.2. Increasing the utility of the prioritisation framework

The reliance on quantitative measures is critical to the utility of the prioritisation framework and makes it distinct from previous analyses. Crucially, this quantification allows for more rapid revisions and flexibility to adapt the system than more traditional prioritisation processes reliant on questionnaires assessing aspects such as disease severity and potential economic impact on a five point scale. In addition, the transparency that such an approach allows starkly contrasts with those rankings devised behind closed doors. Obviously, the capacity for expanding this approach is dependent on the availability of globally consistent quantitative information. The Global Burden of Disease (GBD) is one such study whose objectives align with this need *(Murray et al., 2012; Murray and Lopez, 2013).* Additional factors, such as socioeconomic impacts and costs of care associated with these diseases, could be added into future iterations of this analysis with relative ease as and when this information becomes available.

The set of disease clusters defined in Chapter 2 was focussed on mapping but the same approach could equally be applied to other ways of grouping diseases. For example diseases could be clustered according to shared interventions or drug delivery mechanisms to evaluate the likely impacts of different intervention policies. Similarly, the approach could easily be adapted by other users. The availability of sub-national *(Murray et al., 2013; Yang et al., 2013)* and national disability adjusted life year estimates *(GBD 2013 Disease and Injury Incidence*

and Prevalence Collaborators, 2015; GBD 2013 Mortality and Causes of Death Collaborators, 2015) could allow for smaller scale analyses of mapping priorities. In addition, other public health questions could be addressed using this tool enabling comparisons between DALY burden and national healthcare expenditure. I have already generated a more focussed prioritisation for mosquito-borne diseases *(Smith, 2014)*, which provides further refinement to the original mosquito-borne viruses cluster (see Table 9.1). Other alternative analyses that would be of interest include comparing the cost-effectiveness of treatments with the DALYs of the conditions they affect to identify those diseases where the most significant impact could be implemented from further drug or vaccine development.

Ranking	Mosquito-borne disease
1	Malaria
2	Filariasis
3	Dengue
4	Japanese Encephalitis
5	Chikungunya, Rift Valley fever, Sindbis, Venezuelan equine encephalitis, West Nile fever
6	California serogroup viruses, Murray Valley encephalitis, St. Louis encephalitis, Rocio, Western equine encephalitis
7	Ross River virus
8	Barmah Forest disease
9	Yellow fever
10	Eastern equine encephalitis
11	Karelian fever, O'nyong nyong, Ockelbo, Oropouche, Pogosta, Wesselsbron, Zika fever
12	Bunyaviridae infections (misc.)
13	Illheus and Bussuquara
14	Coltiviruses, Mayaro, Spondweni

Table 9.1 Prioritisation amongst mosquito-borne diseases

Whilst not the most complex of protocols, this approach does provide a transparent quantitative rationale that can help inform future public policy decisions. As such, I hope this tool can be

further developed to help investigate these additional epidemiological questions at the policy-health science interface.

9.3. Species distribution models and disease mapping

As with any modelling approach, it is critical to understand the limitations and assumptions of SDMs. The specific limitations of each SDM analysis in this thesis are listed in the relevant chapters. Below summarises the more general limitations of SDMs as applied to disease mapping and key considerations when applying them to different diseases.

Occurrence data is integral to these models. One important consideration therefore is the reliability of the data included, particularly when different time periods and diagnostic methods are involved. When modelling leishmaniasis we assumed that all occurrence data falling within areas of the same evidence consensus were of equivalent validity. In the Lassa fever models (and in part motivated by previous critiques *(Peterson et al., 2014)*) we explicitly considered the effect of diagnostic variation on record reliability, and found little evidence to suggest that this made a significant difference to the final outputs (see Appendix A.9); the predictions of the niche under different data exclusion rules were approximately consistent. As with many approaches, each occurrence database must be considered under its own merits. The leishmaniasis and CCHF databases *(Messina et al., 2015b; Pigott et al., 2014)* (see Appendices A.4 and A.8) mitigate some of these issues due to their larger size, and therefore the increased likelihood of multiple records from separate sources co-occurring and cross-validating. Those databases that are smaller, such as Ebola and Marburg *(Mylne et al., 2014; Pigott et al., 2015a)*, are constrained by the availability of information at the outset. In these instances, every attempt was made to ensure as comprehensive a database as possible. Smaller datasets are still of utility

however. Indeed, there is already precedence for such smaller models being explicitly used in informing wider epidemiological investigations, such as modelling outputs being used to identify potential reservoir species for filoviruses *(Peterson et al., 2004b; Swanepoel et al., 2007)*.

The reliability of geopositioning techniques is also worthy of discussion *(Velasquez-Tibata et al., 2015)*. For the VHFs, identifying where spillover events took place is complex. In many instances estimating the approximate area over which this could have occurred (such as hunters coming across infected animal carcasses in forested areas) can be accommodated in the model by using larger areal estimates as opposed to specific latitudes and longitudes. For some cases it is impossible to accurately determine the site of infection. In the VHF modelling analyses in this thesis an assumption was made that spillover occurred in the vicinity of the individual's home. This has been done in previous studies of this type *(Jones et al., 2008; Peterson et al., 2004a; Peterson et al., 2006)* and is a necessary compromise given the fact that further epidemiological information will be unavailable for historic outbreaks. The work in this thesis was performed at a resolution of 5km x 5km, therefore the most geographically precise occurrence is already treated as occurring anywhere within a $25km^2$ area. In addition, as many covariates are strongly correlated over large distances, error in geopositioning may actually have little impact of the final model outputs. As the resolution of covariates and therefore risk maps increases, this issue of spatial accuracy of occurrences will become of increasing importance. Consequently, the use of areal occurrence records may become more widespread, unless there is a complementary improvement in the geographic resolution at which occurrence records are reported. Some of this uncertainty can be countered by for example using Monte Carlo methods when drawing data from areal estimates as was performed for the VHFs. However, more formally incorporating models that explicitly consider geographic measurement error on a condition-specific basis is worth considering for future iterations *(Velasquez-Tibata et al., 2015)*.

The underlying theory behind SDMs also leads to an interesting discussion as to what these models actually predict. It is important to stress that the mapped surfaces identify areas that share ecological similarities with regions of reported occurrences rather than relating directly to more commonly characterised epidemiological statistics such as prevalence or incidence. One issue raised relates to the reliability of predictions in regions where no cases have been reported. Does a high probability (*i.e.* areas most environmentally suitable) correspond to cases being present? It is not just ecological determinants that influence the distribution of disease *(Peterson, 2006; Peterson et al., 2011)* – some diseases may have historic biogeographical reasons for why they are not present in a particular location. Similarly, factors such as international trade and human movements have been shown to have influenced the distribution of vector-borne diseases such as dengue and chikungunya *(Cauchemez et al., 2014; Nunes et al., 2015; Nunes et al., 2014; Reiter and Sprenger, 1987).* Some of this can be directly accounted for by adding covariates such as socio-economic factors which represent relevant drivers of change, although global surfaces for these factors are still being developed *(Kraemer et al., 2015a).*

The evidence consensus layer, used with leishmaniasis, Lassa fever and CCHF, represents an evidence-based attempt to make this distinction from the outset *(Brady et al., 2012).* For example, leishmaniasis has never been reported in areas in the Far East or Australasia and therefore, with a low evidence consensus value (corresponding to strong evidence for absence of the disease) in these regions, model predictions are not extended to these areas. As a result, areas that immediately appear as false positive predictions can be eliminated. This approach however is limited when multiple sources of information contradict each other or in regions with poor healthcare systems where it is difficult to discern genuine absence from a lack of diagnostic detection capacity. With EVD and MVD, where understanding of the epidemiology

of the disease is incomplete, the model was allowed to predict across a much larger area, with additional covariates such as reservoir host distributions assisting in spatial delimitation. It is crucial, and this is why it was stressed in their discussions in detail, to convey the possibility of false positive predictions when such predictions are allowed. An example of this is characterised by Figure 5.7 which contrasts the expert opinion range of the suspected reservoir host of MVD, the Egyptian rousette *Rousettus aegyptiacus*, with areas predicted to be environmentally suitable for MVD transmission. Here we clearly see that if the Egyptian rousette were to be the only reservoir host for MVD, some areas of predicted environmental suitability in Central Africa are likely not at risk due to the reservoir's absence.

Similarly, those disease systems that require interaction between humans and other host species, will be heavily influenced by spatial variation in these rates of encounter. Anthropological studies have shown considerable differences in human-animal interactions, varying between age groups, genders and ethnicities *(Woldehanna and Zimicki, 2015).* All of this must be considered if we intend to translate these risk surfaces into more accurate representations of where cases are located, and this can only be done as part of a large multidisciplinary investigation *(Wood et al., 2012).*

9.4. Improving existing risk maps

The maps presented in this thesis represent an important first step in understanding the global risk of these zoonotic diseases in a more evidence-based manner. However, they are by no means a complete picture of these very complex epidemiological processes. With good local data, very detailed epidemiological investigations can be considered *(Cuong et al., 2011; de*

Araujo et al., 2013; Stoddard et al., 2013; Stoddard et al., 2014). However, at a regional or global scale, the difficulty in generating this high quality dataset is amplified.

Plowright *et al.* (2015) provides a framework for considering the spatial filters through which potential risk can be translated into actual cases. Using Hendra virus, a bat-borne zoonosis that can infect horses and occasionally humans *(Mahalingam et al., 2012),* as an example the authors demonstrate how a variety of factors influence the spatial distribution of risk. These include the distribution of the reservoir host and where viral shedding can occur, where environmental conditions are viable for viral persistence, where viral spillover into vulnerable species could occur, and where are susceptible hosts located, given the widespread use of equine vaccines (Figure 9.1) *(Plowright et al., 2015).* This survey provides a useful framework for considering the information gaps, as well as the geospatial data layers that must be generated in order to improve upon the existing risk maps. In the sections below I will go into more detail about possible components to refine future iterations of these outputs for the diseases considered in this thesis.

Leishmaniasis – Chapter 3 represents the first attempt at mapping these diseases at a global scale beyond the use of expert opinion range maps *(WHO, 2010).* From this baseline, we can implement new data and approaches to enhance the risk layers. One immediate area for advance is using SDMs to model the global distribution of the diseases' Phlebotomine vector species. Similar studies have carried out mapping of the *Anopheles* vectors of malaria *(Sinka et al., 2011; Sinka et al., 2010a; Sinka et al., 2010b)* and the *Aedes spp.* vectors of dengue and chikungunya *(Kraemer et al., 2015b).* Extending existing modelling efforts for Phlebotomine hosts from national and regional assessments *(Colacicco-Mayhugh et al., 2010; Gonzalez et al., 2011; Gonzalez et al., 2010)* to the global scale will represent an important first step. Interesting analyses have considered risk as the overlap of parasite viability and vector distribution *(Samy*

et al., 2014) and global sandfly surfaces could be used as a key filter for risk, particularly when coupled with bionomic information on the key vectors *(Foley et al., 2012)*. Incorporating mechanistic models of vector survivorship in response to key environmental factors *(Brady et al., 2014; Brady et al., 2013)* could also provide detailed information on vector distributions. An advantage of such a mechanistic modelling approach is that the resulting maps may be formally considered in a full risk model where the basic reproduction number, R_0, can be estimated *(Hartemink et al., 2011)*.

Distinction of disease by clinical grade ignores the fact that a variety of different transmission cycles are in place. This may mask significant variation between these different transmission patterns. In addition, in the Americas, where there is a wide array of *Leishmania spp.* responsible for cutaneous and mucocutaneous leishmaniasis, the actual *Leishmania spp.* influences the progression of disease in humans *(Banuls et al., 2011)*. Being able to model these species separately therefore will not only provide unique epidemiological information, but also will have direct public health policy relevance *(WHO, 2010)*. Differentiating zoonotic transmission cycles from anthroponotic cycles will allow for a more nuanced consideration of the spatial variation in risk at which we consider these syndromes.

A consequence of expanding the database to include species level information will be the need for more stringent data inclusion criteria, which could lead to the exclusion of a number of occurrences in the current dataset where only the broad clinical condition is recorded. New models are currently being developed that will allow for simultaneous modelling of multiple species, which could potentially infer species associated with non-specific records *(Harris, 2015; Pollock et al., 2014)*. In addition, we can expand the existing database to include information on infection in sandfly vectors and additional reservoir hosts, with the assumption that these infections are indicative of more widespread leishmaniasis transmission. Searching

Web of Science and PubMed provides over 17,000 hits of potential vector records and approximately 51,000 reservoir reports for leishmaniasis.

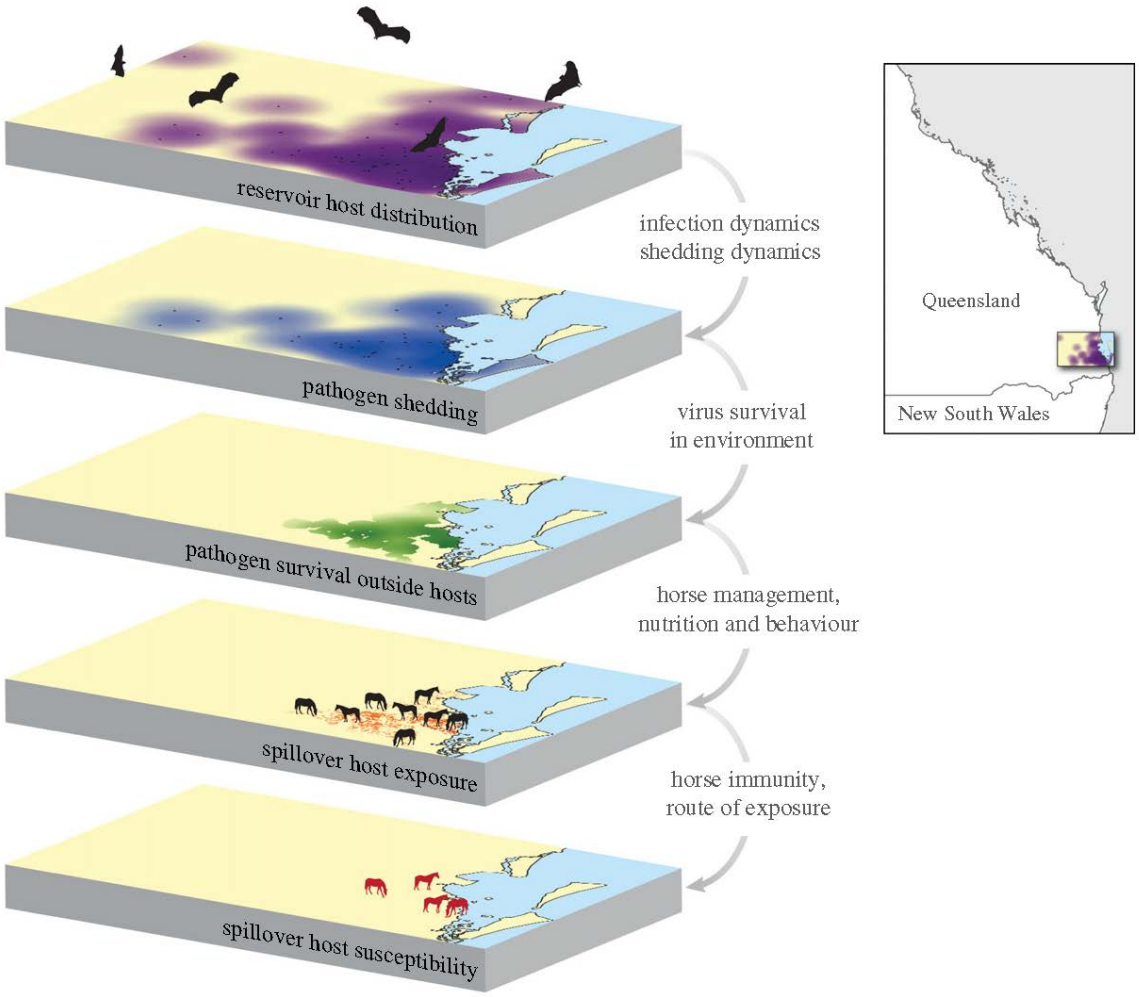


Figure 9.1. Understanding the various filters of risk. Plowright *et al.* investigate the various factors influencing the spatial distribution of Hendra virus cases. Here different aspects of the transmission cycle can act as filters for subsequent viral transmission (from Plowright *et al.* (2015)).

When using occurrence data to generate risk maps, it is always important to bear in mind how this information can be related to case numbers and other burden indicators such as prevalence

and incidence. Relationships between species distribution models and species abundance has been shown in more standard ecological contexts *(Oliver et al., 2012)*. Therefore, there should be some theoretical linkage between suitability models and prevalence/incidence rates, as was investigated for dengue *(Bhatt et al., 2013)*. However, this estimation process is likely to be more complex for leishmaniasis due to a number of different factors *(Bern et al., 2008)*. There is the possibility, as discussed earlier, that predictions may represent false positives. Whilst the evidence consensus layer, resolved at the district level, will allow for some reliable differentiation of false positive predictions, this will not be complete. Nevertheless, such estimates can be beneficial when no other informed measures exist, particularly in areas of poor-reporting or healthcare provisioning *(Shepard et al., 2014)*.

Public health interventions will similarly have an impact on leishmaniasis presence. During the Global Malaria Eradication Programme of the 1950s, significant reductions in leishmaniasis transmission were achieved as a by-product of indoor residual spraying programmes using dichlorodiphenyltrichloroethane (DDT) *(Ostyn et al., 2008)*. Ironically, more significant inroads to leishmaniasis elimination were achieved by these programmes than with malaria. Contemporary spraying, combined with the use of insecticide-treated bednets may lead to significant reduction in disease case load and potentially local elimination *(Davies et al., 1994; Golding et al., 2015; Ritmeijer et al., 2007; Wilson et al., 2014)*. Finding ways of incorporating this as a proxy by adapting evidence consensus methods or more robustly as a modelling covariate should be considered.

A number of human factors similarly will influence the burden of the disease that we observe *(Alvar et al., 2006; Boelaert et al., 2009; Desjeux, 2001)*. Malnutrition and a variety of other physical stresses will impact the severity of infection that results *(Bern et al., 2010)*. Furthermore, there can be significant variation in at-risk groups within a population. In Western

Europe for example there is a large burden of disease linked to a variety of non-environmental risk factors including intra-venous drug users, HIV-positive individuals and poor screening for parasites during blood and organ donation processes *(Alvar et al., 2008; Alvar et al., 1997; Cardo, 2006; de Silva et al., 2015)*. In addition, large influxes of susceptible populations, such as refugees from conflict, can be displaced into leishmaniasis endemic areas, which often results in a subsequent increase in mortality and morbidity *(Alawieh et al., 2014; Jacobson, 2011; Kolaczinski et al., 2004; Rowland et al., 1999)*. In other locations, it seems that general economic development, such as that observed in Western Europe, mitigates much of the risk associated with leishmaniasis, whether it be through improved population health levels or a reduction in sandfly breeding sites in areas of high urban development. Given considerable advances in leishmaniasis vaccine development *(Kumar and Engwerda, 2014)*, if we do see its widespread implementation, this will further limit the number of cases we see in areas where transmission may have occurred.

These issues touch upon a much wider limitation of niche-based modelling approaches. As mentioned in the introduction to this thesis, the gap between Option 4 and 5 diseases is regulated by data constraints *(Hay et al., 2013a; Hay et al., 2013b)*. For leishmaniasis, in certain locations, there is sufficient data to allow for the more sophisticated MBG approaches. For instance, Brazil has an online Ministry of Health data portal *(SINAN)* which reports cases of a variety of diseases, including leishmaniasis and dengue, at a municipality level. As a result, sufficiently detailed information to estimate prevalence and incidence of such diseases is possible to allow for this more data-driven processing *(Karagiannis-Voules et al., 2013)*. However, whilst for diseases such as malaria sufficiently detailed information exists across a broader-scale *(Moyes et al., 2013)*, for leishmaniasis it is only sporadically available. Obviously, improving the nature of the information reported by healthcare systems, or making these more readily available is a long term objective, but is far from trivial to implement. Advocating for more systematic reporting of leishmaniasis will be central to future success

(Alvar et al., 2012), and subsequent to publication of these global maps, the WHO have expressed their interest in collaboration to incorporate these outputs into their own global analyses.

Filovirus disease – The wider epidemiology of these viruses still remains poorly characterised *(Groseth et al., 2007; Olival and Hayman, 2014)*. Whilst there have been significant advances in our understanding of infection in animal reservoirs over the last decade, several unknowns still remain, particularly how reservoir viral dynamics can inform human risk.

Identifying the remaining reservoir species is a critical step. To-date only three species have had Ebola virus RNA isolated from them (*Hysignathus monstrosus*, *Epomops franqueti* and *Myonycteris torquata*) *(Leroy et al., 2005)*, whilst a number, including these three, have been found to be seropositive for Ebola/Ebola-like viruses *(Caillaud et al., 2006; Hayman et al., 2010; Hayman et al., 2012; Pourrut et al., 2007; Pourrut et al., 2009; Rouquet et al., 2005; Wittmann et al., 2007)*. Whether this result reflects true reservoir status, or fortunate sampling of the correct bats at the right time, is always a possibility. Isolation of Marburg virus from *Rousettus aegyptiacus* populations has been reported on a number of occasions *(Amman et al., 2012; Amman et al., 2014; Swanepoel et al., 2007; Towner et al., 2009)*, and similar to Ebola virus, a number of other species have been identified as seropositive *(Pourrut et al., 2009; Swanepoel et al., 2007; Towner et al., 2009)*. Consideration of risk should take into account the possibility for additional reservoir species *(Olival and Hayman, 2014)*, particularly given how their varying life histories might influence the way in which viral infections can or cannot be sustained *(Hayman, 2015)*. Indeed, the nature of roost and population connectivity for many of these species remains unknown, with variable degrees of roost faithfulness, combined with some species, such as *Eidolon helvum*, that are highly interconnected across much of sub-Saharan Africa *(Peel et al., 2013)*. Unique approaches, such as defining which life-history traits

typify those found to be filovirus seropositive could provide an important lead on other species to take note of *(Han, 2015; Han et al., 2015)*. Knowing the true reservoirs of the disease has obvious implications for broadscale refinement of risk regions.

How these species interact, not only as a local population, but with the rest of their own species as well as others, will have significant impacts on subsequent viral dynamics. Two hypotheses have been proposed that describe how viral load could be distributed within bats: infection is ubiquitous across the entire range or restricted to specific locations and passed from region to region *(Plowright et al., 2015)*. Theoretical depictions of how variability in bat population dynamics could maintain viruses or indeed cause them to become extinct have been developed, although this is yet to be demonstrated in wildlife populations *(Hayman, 2015)*. Nevertheless, this does suggest a much more complex situation than appreciated. For Marburg virus survey sites, seasonality in infection has been shown as well as transmission between mothers and their juveniles *(Amman et al., 2012)*. For Ebola virus this level of detail has yet to be determined. The possibility of transmission of these viruses between different bat species is an unknown – if this were the case, future models and sampling efforts would have to be expanded to include these co-occurring species. Whilst much basic science in this arena needs to be performed, the results will have important implications on how we characterise the risk that different populations and species pose.

Transmission from bat hosts to susceptible species also remains an unknown. A hypothesised route of Ebola virus transmission between bats and non-human primates, similar to that identified between bats and pigs for Nipah virus in Malaysia *(Daszak et al., 2013)*, has been suggested. It is thought that fruit bats could transfer viral particles to dropped fruit *via* their saliva *(Gonzalez et al., 2007; Walsh et al., 2009)*. Chimpanzees and Gorillas feeding under the same trees will therefore be at risk and can possibly transmit between each other *(Walsh et al.,*

2007). It is unknown for how long the virus remains viable within the fruit however. Transfer directly between bats and humans has been hypothesised as the reason for index infection in one outbreak through the butchering of a bat carcass after which ten individuals became infected *(Leroy et al., 2009)*. Human outbreaks of Ebola virus seem to be initiated more often by contact with non-human primates, often post-mortem and when the carcasses are being prepared for food *(Mylne et al., 2014)*. For Marburg virus, caves and mines represent sites of most common infection *(Adjemian et al., 2011; Towner et al., 2006)* unsurprising given the contrast between the underground roosting *Rousettus aegyptiacus* compared to the tree-dwelling hypothesised reservoirs of Ebola virus. Whilst the first outbreak of Marburg virus disease was the result of infection from vervet monkeys *(Henderson et al., 1971)*, no subsequent infection in non-bat animals has been reported. From the perspective of refining at-risk areas, understanding these interactions is important. As Chapters 4 and 5 demonstrated, whilst viruses could be circulating over a wider area, specific activities (such as hunting) or locations (such as caves) may be critical for allowing spillover transmission to occur. Engaging with anthropological surveys assessing how different populations engage with these different reservoir and susceptible species will be important, particularly as risk may be very variable between gender, age and ethnicity even within the same location *(Woldehanna and Zimicki, 2015)*.

Finally, the majority of EVD focus has been towards Zaire ebolavirus, therefore it could well be that the distribution of the other three pathogenic species (Taï Forest, Bundibugyo and Sudan ebolaviruses) have different dynamics that have, to date, gone uninvestigated. Currently too little data exists to consider this in a meaningful way, but we must be aware of the differences that may exist in the transmission cycles of these different species.

Crimean-Congo haemorrhagic fever – Similar to leishmaniasis, a variety of vector species, in this instance ticks, are responsible for disease transmission *(Ergonul, 2006)*. Integrating their

distributions more formally into distribution models is important, particularly given the nuances in their respective life histories *(Walker et al., 2014)*. In addition, further evidence making the distinction between accidentally infected ticks (who happen to have taken a blood meal from infected mammals) and actual disease vectors is critical. To date a large number of ticks have been definitively shown to be disease vectors, however an even larger number have mixed evidence for support *(Ergonul and Whitehouse, 2010)*. Filtering non-vector species out is important. In addition, a large number of reservoir animals have been identified whose relative contribution, as well as inclusion into the modelling framework, should be considered *(Ergonul, 2006)*. High resolution surfaces of livestock are available *(Robinson et al., 2014)* – their inclusion was considered in the initial modelling effort carried out in Chapter 7 but were not significant contributors (possibly due to sharing a large number of covariates, particularly in land cover type). Regardless of confounding influence, a more appropriate way of including such layers in the analysis is as post-hoc risk surfaces indicating locations where high livestock densities co-occur with regions of environmental suitability for CCHFV transmission.

The future impact of tick-borne disease is often considered when the influence of climate change on infectious disease is discussed *(Medlock and Leach, 2015)*. One focus for CCHF has been the possibility of migratory birds importing infected ticks from endemic settings into non-endemic countries where environmental conditions may be suitable for sustained disease transmission *(Jameson et al., 2012)*. However, when quantified the risk that this poses is considered as minimal *(Gale et al., 2012)*. Nevertheless, understanding how climate change may impact disease transmission cycles is important *(Estrada-Pena et al., 2014)*, although this must be considered in conjunction with a variety of other changing factors *(Messina et al., 2015a)*. As demonstrated by tick-borne encephalitis virus, socio-economic changes can have just as profound an impact as temperature *(Randolph, 2001)*.

Refining how occupational risk is characterised is also important for CCHF. Farms in endemic settings, particularly given the role that livestock species play in the transmission cycle, remain areas where increased awareness of tick bites should be made, particularly when simple measures such as wearing longer clothes will prevent transmission *(Maltezou and Papa, 2010)*. Another important transmission setting is workers in slaughterhouses, where contact with infected livestock blood can allow transmission *(Ergonul, 2006)*. Once again, maintaining a high hygiene standard can significantly minimise the likelihood of this occurring. Finally, with a more effective vaccine showing early signs of success *(Buttigieg et al., 2014)*, not only can these maps be used to target priority areas for vaccination, but also show their continued relevance through iterative updates that incorporate this control data.

Lassa fever – In comparison to filoviral diseases, we are far more certain on the primary reservoir species of Lassa virus, the Natal multimammate mouse, *Mastomys natalensis (McCormick et al., 1987)*. Similar to many other rodent-borne arenaviruses, there seems a more consistent one host-one virus species relationship *(Palmer et al., 2013)*, although surveys to the far west of Guinea have reported infection in *Mastomys erythroleuceus*, which could suggest a potential range expansion into more arid northern regions of the Sahel *(Sogoba et al., 2012)*.

Mastomys natalensis populations seem to be fairly distinct in space, due in part to the reportedly low dispersal ability of the mouse *(Van Hooft et al., 2008)* – it seems unclear how infection is be passed from region to region, although some have supposed the possibility of transmission from human to mouse *(Johnson, 2013)*. Understanding how the virus could be transmitted between populations will provide an important refinement step. The two main endemic foci of Guinea and Nigeria may potentially have more interconnected rodent populations, or just naturally have had higher viral loads, whereas smaller outbreaks, such as that which occurred in Benin *(ProMED-mail, 2014)*, could be restricted to just one local host population, or a chance

introduction into this region. Other rodent-borne viruses have shown prevalence of infection to be tightly related to population dynamics of the host *(Carver et al., 2015)*. It has been suggested that Lassa virus does not show a similar density dependence response, but irrespective of the population prevalence, seasonal variation in population size may be an important component in influencing Lassa fever transmission *(Fichet-Calvet et al., 2007; Fichet-Calvet and Rogers, 2009)*. Accounting for such spatio-temporal variation in host populations and their corresponding viral load would allow for significant improvements in understanding true disease burden. Understanding the interplay between host population and viral dynamics will be important if refining future risk estimates.

Whilst no real definitive evidence of the means of transmission of the virus has been found, rodent to human transmission likely occurs in the household, where humans are most likely to come into contact with rodent urine or droppings or through consumption of contaminated rodent meat *(McCormick et al., 1987; ter Meulen et al., 1996)*. The theory that such matter becomes aerosolised however has never been definitively proven. It has been suggested that crop yield may act as a key indicator for cases found in humans *(Jones, 2014)* – if this is a causal relationship it is likely through altered interactions between humans and rodents, influencing the rate at which we see spillover events occur, as rodents are drawn into the house due to the large crop stockpiles.

Of equal importance for Lassa fever outbreaks is an understanding of the role of the individual in secondary transmission. Evidence suggests that super-spreaders are responsible for the majority of human-to-human cases *(Lo Iacono et al., 2015)* - appreciating this risk, and understanding what aspects of local infrastructure are most likely to influence this will be critical from a hospital case management and outbreak containment point of view *(Bannister, 2010)*.

Viral haemorrhagic fever outbreak risk – The final chapter brings together the four VHF risk maps in order to better understand the outbreak risk these diseases present across the African continent. Two important aspects of the assessment require the greatest attention should these estimates be revised. Firstly, the spatial resolution of many key indicators is poor, constraining the assessment to the national level. Spillover potential and spread measures can be assessed at a pixel level given the resolution to which we have this information, but assessments of healthcare seeking behaviours and other components of the vulnerability and lack of coping capacity measures prevent evaluation at this level. It is possible to provide this analysis for individual nations, however to maintain a regional level assessment it was necessary to retain the same level of assessment for all countries. Hopefully this output will reinforce calls for improving the way in which many of these socioeconomic covariates are reported, particularly at ever more local levels *(Kraemer et al., 2015a)*.

Secondly, providing increasingly relevant VHF indicators rather than more general assumed proxies for these indicators will be important. As an initial first step, broader measures of health infrastructure, such as physicians per capita, will provide good approximate information. However, considering indicators such as personal protective equipment and emergency response capacities will be key in understanding the nature of risk. These indicators, coupled with improving the understanding of where spillover could occur, outlined above, will provide increasingly realistic risk profiles for Africa. The WHO, amongst others, is trying to improve our understanding of just where shortfalls in preparedness lie *(WHO, 2015)*, but this detailed assessment should be considered across a much broader range of countries than initially sent out to.

9.5. Bringing science and policy together

One of the most important indicators of impact for public health epidemiologists is the exposure of their work to public health policymakers. However, this often remains a stumbling block perhaps due to the work being presented in too complex a nature, or the objectives not being specific enough. Chapter 2 presents a methodology that, I believe, could be relatively easily adapted to meet the needs of any policymaker. Importantly, once sufficiently detailed metrics are available (and often policymakers may be privy to indicators that are not publicly released), a priority ranking can be achieved independently of a more involved expert opinion process – the expert involvement was in generating the input data in the first place. Obviously, the relative ranking and methodology would likely be subject to further scientific scrutiny, but this methodology would allow for a first level assessment to be performed well before any panel convenes. Exploring the possibility of creating a tool that could allow for individual users to prioritise diseases based upon a variety of different grouping and comparator datasets would provide a very useful output for policymakers.

The SDM outputs suffer from many of the issues models face in general when presented to policymakers, which is their ability to be misinterpreted or misused. For the most part the onus is on the author to keep as clear as possible the main outcomes of their work, but there is always the opportunity for misunderstanding. One feature that the disease presence maps suffer from is that they fall halfway between more conventional clinical maps and niche maps – as a result there can be confusion when the maps are interpreted incorrectly as clinical incidence or prevalence estimates. These clinical indicators are related to environmental suitability, but require further processing steps before they can be directly inferred from the maps. As seems to be the case with leishmaniasis, policymakers have taken an interest in helping to further improve the outputs, but this should not always be relied upon.

The final chapter of this thesis counters this by specifically addressing an important question with direct policy relevance – where could the next viral haemorrhagic fever outbreak begin, and which regions of the worlds are most vulnerable should its continued transmission not be averted? The key outputs, consisting of relative rankings, can easily be interpreted and provide a strong rationale for action – importantly, they convert surfaces that may be hard to interpret (*e.g.* environmental suitability of zoonotic transmission of a virus) into a more relatable measures (*e.g.* which countries are more likely to see VHF outbreaks). The key for this output will be in highlighting the limitations of the approach, and using them to advocate for improving the key input parameters for future iterations.

9.6. Expanding to other zoonotic diseases

Zoonotic diseases, as has been shown, are responsible for a large number of conditions of clinical significance and represent an interesting mapping conundrum in that they must account not only for the pathogen, but also the added complication of reservoir populations, and often a variety of disease vectors *(Palmer et al., 2013).* These conditions represent the whole range of mapping options described in Chapter 1. Some, such as Lobomycosis *(Paniz-Mondolfi et al., 2012),* are so rare that more information is required on the basic epidemiology of the disease before additional work can be undertaken. Others, such as Venezuelan haemorrhagic fever *(Tesh, 1994),* can be considered by using SDMs on relevant vector and reservoir species. Where more data is available however, we can use the species distribution modelling framework outlined throughout this thesis in order to better understand the spatial risk these pathogens present. In identifying environmentally suitable regions for disease transmission, important first steps can be taken, particularly when compared to existing global surfaces, from which we can build upon when more data are found both on the condition in humans and its wider epidemiology.

Beyond the public health benefit such a survey would provide, a systematic consideration of zoonotic disease will provide an important dataset for discussing more macro-ecological disease patterns. Global assessments of the determinants of pathogen diversity have implicated the role of not only climatic factors *(Guernier et al., 2004)* but also of alternate host diversity and richness *(Dunn et al., 2010)*. However, these studies have suffered from poor spatial resolution, both in terms of disease distribution information and the use of national level spatial tagging. In addition, by considering all infectious diseases, global trends will likely be confounded by a mixture of pathogen taxonomy and life history bias. A highly spatially resolved risk surface for all zoonotic diseases would enable a more nuanced approach to teasing apart the processes determining global pathogen diversity.

This information can also be leveraged as part of a broader discussion of emerging infectious disease risk. Previous assessments have highlighted zoonotic disease as one important source of novel disease *(Jones et al., 2008)*. Cross-referencing zoonotic disease diversity with disease emergence events (after incorporating appropriate methods of correcting for reporting bias) could provide an important resource. Are the zoonotic diseases that we currently recognise the tip of a much larger potentially pathogenic pool, or are these diseases the modern day manifestation of historically important changes in animal-human interaction, such as shifts in diet or animal husbandry *(Barrett and Armelagos, 2013)*? This will have important policy ramifications, since concluding that existing animal diversity, pathogenic diversity, or novel human-animal interactions are driving disease emergence potential will ultimately influence how best to track these diseases *(Anthony et al., 2013)*.

9.7. Conclusions

Disease maps can play an important role in informing a variety of different epidemiological and public health questions. Over the last decade significant advances have been made in the methods used to investigate disease distributions, yet updates of globally comprehensive disease maps lag behind. This thesis has investigated not only how we can prioritise mapping of infectious diseases, but also demonstrated the role of species distribution models in defining the spatial limits of possible disease transmission. Focussing on zoonotic diseases, this work produced the first evidence based maps of the leishmaniases as well as modelling the distributions of the four contagious viral haemorrhagic fevers that are present in Africa. Finally, this thesis presented a way to process these maps into policy relevant documents. In doing so, this thesis has hopefully demonstrated the potential benefits of disease cartography, as well as increased awareness of their shortcomings and what data are required to overcome them.

Much of the work performed in this thesis is equally applicable to a wide array of diseases, whether other zoonoses or vector-borne diseases. Given sufficient data, any disease that shows spatial variation in presence can be mapped using SDMs. As we begin to study some of these conditions in more detail, with surveys and epidemiological questions provoked by initial spatial assessments, this newly generated input data can begin to feedback into the system. The future of disease cartography aims to improve the input data as well as iterate maps more frequently. These objectives are mutually beneficial – as more data becomes available, the maps will improve and become more detailed. These updated maps will provide more useful information to those working in the field and therefore encourage further data generation. The potential benefits that high resolution maps can provide are great. Methods will continue to develop in complexity allowing for ever more detailed analyses to be performed on existing datasets. The work presented in this thesis therefore provides a good basis for future improvement and refinement.

9.8. References

Adjemian J, Farnon EC, Tschioko F, Wamala JF, Byaruhanga E, et al. 2011. Outbreak of Marburg hemorrhagic fever among miners in Kamwenge and Ibanda districts, Uganda, 2007. *J Infect Dis* **204**: S796-S799. doi:10.1093/infdis/jir312.

Alawieh A, Musharrafieh U, Jaber A, Berry A, Ghosn N, et al. 2014. Revisiting leishmaniasis in the time of war: the Syrian conflict and the Lebanese outbreak. *Int J Infect Dis* **29**: 115-119. doi:10.1016/j.ijid.2014.04.023.

Alvar J, Aparicio P, Aseffa A, Den Boer M, Canavate C, et al. 2008. The relationship between leishmaniasis and AIDS: the second 10 years. *Clin Microbiol Rev* **21**: 334-359. doi:10.1128/Cmr.00061-07.

Alvar J, Canavate C, GutierrezSolar B, Jimenez M, Laguna F, et al. 1997. *Leishmania* and human immunodeficiency virus coinfection: the first 10 years. *Clin Microbiol Rev* **10**: 298-&.

Alvar J, Velez ID, Bern C, Herrero M, Desjeux P, et al. 2012. Leishmaniasis worldwide and global estimates of its incidence. *PLoS One* **7**: e35671. doi:10.1371/journal.pone.0035671.

Alvar J, Yactayo S, Bern C. 2006. Leishmaniasis and poverty. *Trends Parasitol* **22**: 552-557. doi:10.1016/J.Pt.2006.09.004.

Amman BR, Carroll SA, Reed ZD, Sealy TK, Balinandi S, et al. 2012. Seasonal pulses of Marburg virus circulation in juvenile *Rousettus aegyptiacus* bats coincide with periods of increased risk of human infection. *PLoS Pathog* **8**: e1002877. doi:10.1371/journal.ppat.1002877.

Amman BR, Nyakarahuka L, McElroy AK, Dodd KA, Sealy TK, et al. 2014. *Marburgvirus* resurgence in Kitaka mine bat population after extermination attempts, Uganda. *Emerg Infect Dis* **20**: 1761-1764. doi:10.3201/eid2010.140696.

Anthony SJ, Epstein JH, Murray KA, Navarrete-Macias I, Zambrana-Torrelio CM, et al. 2013. A strategy to estimate unknown viral diversity in mammals. *Mbio* **4**: e00598-00513. doi:10.1128/mBio.00598-13.

Bannister B. 2010. Viral haemorrhagic fevers imported into non-endemic countries: risk assessment and management. *Brit Med Bull* **95**: 193-225. doi:10.1093/Bmb/Ldq022.

Banuls AL, Bastien P, Pomares C, Arevalo J, Fisa R, et al. 2011. Clinical pleiomorphism in human leishmaniases, with special mention of asymptomatic infection. *Clin Microbiol Infect* **17**: 1451-1461. doi:10.1111/j.1469-0691.2011.03640.x.

Barrett R, Armelagos GJ. 2013. An unnatural history of emerging infections. Oxford: Oxford University Press.

Bern C, Courtenay O, Alvar J. 2010. Of cattle, sand flies and men: a systematic review of risk factor analyses for South Asian visceral leishmaniasis and implications for elimination. *PLoS Negl Trop Dis* **4**: e599. doi:10.1371/journal.pntd.0000599.

Bern C, Maguire JH, Alvar J. 2008. Complexities of assessing the disease burden attributable to leishmaniasis. *PLoS Negl Trop Dis* **2**: e313. doi:10.1371/journal.pntd.0000313.

Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, et al. 2013. The global distribution and burden of dengue. *Nature* **496**: 504-507. doi:10.1038/Nature12060.

Boelaert M, Meheus F, Sanchez A, Singh SP, Vanlerberghe V, et al. 2009. The poorest of the poor: a poverty appraisal of households affected by visceral leishmaniasis in Bihar, India. *Trop Med Int Health* **14**: 639-644. doi:10.1111/j.1365-3156.2009.02279.x.

Brady OJ, Gething PW, Bhatt S, Messina JP, Brownstein JS, et al. 2012. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLoS Negl Trop Dis* **6**: e1760. doi:10.1371/Journal.Pntd.0001760.

Brady OJ, Golding N, Pigott DM, Kraemer MU, Messina JP, et al. 2014. Global temperature constraints on *Aedes aegypti* and *Ae. albopictus* persistence and competence for dengue virus transmission. *Parasit Vectors* **7**: 338. doi:10.1186/1756-3305-7-338.

Brady OJ, Johansson MA, Guerra CA, Bhatt S, Golding N, et al. 2013. Modelling adult *Aedes aegypti* and *Aedes albopictus* survival at different temperatures in laboratory and field settings. *Parasit Vectors* **6**: 351. doi:10.1186/1756-3305-6-351.

Buttigieg KR, Dowall SD, Findlay-Wilson S, Miloszewska A, Rayner E, et al. 2014. A novel vaccine against Crimean-Congo haemorrhagic fever protects 100% of animals against lethal challenge in a mouse model. *PLoS One* **9**: e91516. doi:10.1371/journal.pone.0091516.

Caillaud D, Levrero F, Cristescu R, Gatti S, Dewas M, et al. 2006. Gorilla susceptibility to Ebola virus: the cost of sociality. *Curr Biol* **16**: R489-491. doi:10.1016/j.cub.2006.06.017.

Cardo LJ. 2006. *Leishmania*: risk to the blood supply. *Transfusion* **46**: 1641-1645. doi:10.1111/j.1537-2995.2006.00941.x.

Carver S, Mills JN, Parmenter CA, Parmenter RR, Richardson KS, et al. 2015. Toward a mechanistic understanding of environmentally forced zoonotic disease emergence: Sin Nombre Hantavirus. *BioScience*: in press.

Cauchemez S, Ledrans M, Poletto C, Quenel P, de Valk H, et al. 2014. Local and regional spread of chikungunya fever in the Americas. *Euro Surveill* **19**: 15-23.

Colacicco-Mayhugh MG, Masuoka PM, Grieco JP. 2010. Ecological niche model of *Phlebotomus alexandri* and *P. papatasi* (Diptera: Psychodidae) in the Middle East. *Int J Health Geogr* **9**: e2. doi:10.1186/1476-072x-9-2.

Cuong HQ, Hien NT, Duong TN, Phong TV, Cam NN, et al. 2011. Quantifying the emergence of dengue in Hanoi, Vietnam: 1998-2009. *PLoS Negl Trop Dis* **5**: e1322. doi:10.1371/journal.pntd.0001322.

Daszak P, Zambrana-Torrelio C, Bogich TL, Fernandez M, Epstein JH, et al. 2013. Interdisciplinary approaches to understanding disease emergence: The past, present, and future drivers of Nipah virus emergence. *Proc Natl Acad Sci USA* **110**: 3681-3688. doi:10.1073/pnas.1201243109.

Davies CR, Llanoscuentas A, Canales J, Leon E, Alvarez E, et al. 1994. The fall and rise of Andean cutaneous leishmaniasis - transient impact of the DDT campaign in Peru. *Trans R Soc Trop Med Hyg* **88**: 389-393. doi:10.1016/0035-9203(94)90395-6.

de Araujo VEM, Pinheiro LC, Almeida MCD, de Menezes FC, Morais MHF, et al. 2013. Relative risk of visceral leishmaniasis in Brazil: a spatial analysis in urban area. *PLoS Negl Trop Dis* **7**: e2540. doi:10.1371/Journal.Pntd.0002540.

de Silva AA, Silva APE, Sesso RDC, Esmeraldo RD, de Oliveira CMC, et al. 2015. Epidemiologic, clinical, diagnostic and therapeutic aspects of visceral leishmaniasis in renal transplant recipients: experience from thirty cases. *BMC Infect Dis* **15**: e96. doi:10.1186/S12879-015-0852-9.

Desjeux P. 2001. The increase in risk factors for leishmaniasis worldwide. *Trans R Soc Trop Med Hyg* **95**: 239-243. doi:10.1016/S0035-9203(01)90223-8.

Dunn RR, Davies TJ, Harris NC, Gavin MC. 2010. Global drivers of human pathogen richness and prevalence. *Proc R Soc Lond B Biol Sci* **277**: 2587-2595. doi:10.1098/rspb.2010.0340.

Ergonul O. 2006. Crimean-Congo haemorrhagic fever. *Lancet Infect Dis* **6**: 203-214. doi:10.1016/S1473-3099(06)70435-2.

Ergonul O, Whitehouse CA, editors (2010) Crimean-Congo hemorrhagic fever. A global perspective: Springer. 328 p.

Estrada-Pena A, Ostfeld RS, Peterson AT, Poulin R, de la Fuente J. 2014. Effects of environmental change on zoonotic disease risk: an ecological primer. *Trends Parasitol* **30**: 205-214. doi:10.1016/J.Pt.2014.02.003.

Fichet-Calvet E, Lecompte E, Koivogui L, Soropogui B, Dore A, et al. 2007. Fluctuation of abundance and Lassa virus prevalence in *Mastomys natalensis* in Guinea, West Africa. *Vector Borne Zoonotic Dis* **7**: 119-128. doi:10.1089/vbz.2006.0520.

Fichet-Calvet E, Rogers DJ. 2009. Risk maps of Lassa fever in West Africa. *PLoS Negl Trop Dis* **3**: e388. doi:10.1371/journal.pntd.0000388.

Foley DH, Wilkerson RC, Dornak LL, Pecor DB, Nyari AS, et al. 2012. SandflyMap: leveraging spatial data on sand fly vector distribution for disease risk assessments. *Geospat Health* **6**: S25-S30.

Gale P, Stephenson B, Brouwer A, Martinez M, de la Torre A, et al. 2012. Impact of climate change on risk of incursion of Crimean-Congo haemorrhagic fever virus in livestock in Europe through migratory birds. *J Appl Microbiol* **112**: 246-257. doi:10.1111/j.1365-2672.2011.05203.x.

GBD 2013 Disease and Injury Incidence and Prevalence Collaborators. 2015. Global, regional, and national incidence, prevalence and YLDs for 301 acute and chronic diseases and injuries for 188 countries, 1990-2013: a systemtatic analysis for the Global Burden of Disease Study 2013. *Lancet*: in press.

GBD 2013 Mortality and Causes of Death Collaborators. 2015. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematics analysis for the Global Burden of Disease Study 2013. *Lancet* **385**: 117-171.

Gilsdorf A, Krause G. 2011. Prioritisation of infectious diseases in public health: feedback on the prioritisation methodology, 15 July 2008 to 15 January 2009. *Euro Surveill* **16**: 15-21.

Global Atlas of Helminth Infections. GAHI: global atlas of helminth infections. Available: http://www.thiswormyworld.org/. Accessed: July 2014

Golding N, Moyes CL, Wilson AL, Cano J, Pigott DM, et al. 2015. Integrating vector control across diseases. *BMC Med*: in press.

Gonzalez C, Rebollar-Tellez EA, Ibanez-Bernal S, Becker-Fauser I, Martinez-Meyer E, et al. 2011. Current knowledge of *Leishmania* vectors in Mexico: how geographic distributions of species relate to transmission areas. *Am J Trop Med Hyg* **85**: 839-846. doi:10.4269/ajtmh.2011.10-0452.

Gonzalez C, Wang O, Strutz SE, Gonzalez-Salazar C, Sanchez-Cordero V, et al. 2010. Climate change and risk of leishmaniasis in North America: predictions from ecological niche models of vector and reservoir species. *PLoS Negl Trop Dis* **4**: e585. doi:10.1371/journal.pntd.0000585.

Gonzalez JP, Pourrut X, Leroy E. 2007. Ebolavirus and other Filoviruses. *Curr Top Microbiol* **315**: 363-387.

Groseth A, Feldmann H, Strong JE. 2007. The ecology of Ebola virus. *Trends Microbiol* **15**: 408-416. doi:10.1016/j.tim.2007.08.001.

Guernier V, Hochberg ME, Guegan JFO. 2004. Ecology drives the worldwide distribution of human diseases. *PLoS Biol* **2**: 740-746. doi:10.1371/journal/pbio.0020141.

Han BA. 2015. Unidentified carriers of filoviruses in the wild. Ecology and Evolution of Infectious Diseases 13th Annual Conference. Athens GA.

Han BA, Schmidt JP, Bowden SE, Drake JM. 2015. Rodent reservoirs of future zoonotic diseases. *Proc Natl Acad Sci USA*: in press.

Harris DJ. 2015. Generating realistic assemblages with a joint species distribution model. *Methods Ecol Evol* **6**: 465-473. doi:10.1111/2041-210X.12332.

Hartemink N, Vanwambeke SO, Heesterbeek H, Rogers D, Morley D, et al. 2011. Integrated mapping of establishment risk for merging vector-borne infections: a case study of canine leishmaniasis in southwest France. *PLoS One* **6**: e20817. doi:10.1371/journal.pone.0020817.

Havelaar AH, van Rosse F, Bucura C, Toetenel MA, Haagsma JA, et al. 2010. Prioritizing emerging zoonoses in The Netherlands. *PLoS One* **5**: e13965. doi:10.1371/journal.pone.0013965.

Hay SI, Battle KE, Pigott DM, Smith DL, Moyes CL, et al. 2013a. Global mapping of infectious disease. *Philos Trans R Soc Lond B Biol Sci* **368**: 20120250. doi:10.1098/Rstb.2012.0250.

Hay SI, George DB, Moyes CL, Brownstein JS. 2013b. Big data opportunities for global infectious disease surveillance. *PLoS Med* **10**: e1001413. doi:10.1371/journal.pmed.1001413.

Hayman DT. 2015. Biannual birth pulses allow filoviruses to persist in bat populations. *Proc R Soc Lond B Biol Sci* **282**: doi:10.1098/rspb.2014.2591.
Hayman DTS, Emmerich P, Yu M, Wang LF, Suu-Ire R, et al. 2010. Long-term survival of an urban fruit bat seropositive for Ebola and Lagos Bat viruses. *PLoS One* **5**: e11978. doi:10.1371/journal.pone.0011978.
Hayman DTS, Yu M, Crameri G, Wang LF, Suu-Ire R, et al. 2012. Ebola virus antibodies in fruit bats, Ghana, West Africa. *Emerg Infect Dis* **18**: 1207-1209. doi:10.3201/eid1807.111654.
Henderson BE, Kissling RE, Williams MC, Kafuko GW, Martin M. 1971. Epidemiological studies in Uganda relating to the Marburg agent. In: Martini GA, Siegert R, editors. Marburg virus disease. Berlin: Springer-Verlag. pp. 166-176.
Jacobson RL. 2011. Leishmaniasis in an era of conflict in the Middle East. *Vector Borne Zoonotic Dis* **11**: 247-258. doi:10.1089/vbz.2010.0068.
Jameson LJ, Morgan PJ, Medlock JM, Watola G, Vaux AGC. 2012. Importation of *Hyalomma marginatum*, vector of Crimean-Congo haemorrhagic fever virus, into the United Kingdom by migratory birds. *Ticks and tick-borne diseases* **3**: 95-99. doi:10.1016/j.ttbdis.2011.12.002.
Johnson N, editor (2013) The role of animals in emerging viral diseases: Academic Press. 364 p.
Jones KE. 2014. Predicting the global emergence and spread of zoonotic infectious diseases. British Society for Parastiology 52nd Annual Spring meeting.
Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, et al. 2008. Global trends in emerging infectious diseases. *Nature* **451**: 990-994. doi:10.1038/Nature06536.
Karagiannis-Voules DA, Scholte RG, Guimaraes LH, Utzinger J, Vounatsou P. 2013. Bayesian geostatistical modeling of leishmaniasis incidence in Brazil. *PLoS Negl Trop Dis* **7**: e2213. doi:10.1371/journal.pntd.0002213.
Kolaczinski J, Brooker S, Reyburn H, Rowland M. 2004. Epidemiology of anthroponotic cutaneous leishmaniasis in Afghan refugee camps in northwest Pakistan. *Trans R Soc Trop Med Hyg* **98**: 373-378. doi:10.1016/j.trstmh.2003.11.003.
Kraemer MUG, Hay SI, Pigott DM, Smith DL, Wint GRW, et al. 2015a. Progress and challenges in infectious disease cartography. *Trends Parasitol*: under review.
Kraemer MUG, Sinka ME, Duda KA, Mylne A, Shearer F, et al. 2015b. The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. *eLife*: accepted manuscript.
Krause G, Prioritisation Working Group. 2008. How can infectious diseases be prioritized in public health? A standardized prioritization scheme for discussion. *EMBO Rep* **9**: S22-S27. doi:10.1038/embor.2008.76.
Kumar R, Engwerda C. 2014. Vaccines to prevent leishmaniasis. *Clin Transl Immunology* **3**: e13. doi:10.1038/cti.2014.4.
Leroy EM, Epelboin A, Mondonge V, Pourrut X, Gonzalez JP, et al. 2009. Human Ebola outbreak resulting from direct exposure to fruit bats in Luebo, Democratic Republic of Congo, 2007. *Vector Borne Zoonotic Dis* **9**: 723-728. doi:10.1089/vbz.2008.0167.
Leroy EM, Kumulungui B, Pourrut X, Rouquet P, Hassanin A, et al. 2005. Fruit bats as reservoirs of Ebola virus. *Nature* **438**: 575-576. doi:10.1038/438575a.
Lo Iacono G, Cunningham AA, Fichet-Calvet E, Garry RF, Grant DS, et al. 2015. Using modelling to disentangle the relative contributions of zoonotic and anthroponotic transmission: the case of Lassa fever. *PLoS Negl Trop Dis* **9**: e3398. doi:10.1371/Journal.Pntd.0003398.
Mahalingam S, Herrero LJ, Playford EG, Spann K, Herring B, et al. 2012. Hendra virus: an emerging paramyxovirus in Australia. *Lancet Infect Dis* **12**: 799-807. doi:10.1016/S1473-3099(12)70158-5.
Maltezou HC, Papa A. 2010. Crimean-Congo hemorrhagic fever: risk for emergence of new endemic foci in Europe? *Travel Med Infect Di* **8**: 139-143. doi:10.1016/j.tmaid.2010.04.008.
McCormick JB, Webb PA, Krebs JW, Johnson KM, Smith ES. 1987. A prospective study of the epidemiology and ecology of Lassa fever. *J Infect Dis* **155**: 437-444.

McIntyre KM, Hawkes I, Waret-Szkuta A, Morand S, Baylis M. 2011. The *h*-index as a quantitative indicator of the relative impact of human diseases. *PLoS One* **6**: e19558. doi:10.1371/journal.pone.0019558.

Medlock JM, Leach SA. 2015. Effect of climate change on vector-borne disease risk in the UK. *Lancet Infect Dis*: S1473-3099(1415)70091-70095. doi:10.1016/S1473-3099(15)70091-5.

Messina JP, Brady OJ, Pigott DM, Golding N, Kraemer MU, et al. 2015a. The many projected futures of dengue. *Nat Rev Microbiol* **13**: 230-239. doi:10.1038/nrmicro3430.

Messina JP, Pigott DM, Duda KA, Brownstein JS, Myers MF, et al. 2015b. A global compendium of human Crimean-Congo haemorrhagic fever virus occurrence. *Sci Data* **2**: 150016. doi:10.1038/sdata.2015.16.

Moyes CL, Temperley WH, Henry AJ, Burgert CR, Hay SI. 2013. Providing open access data online to advance malaria research and control. *Malar J* **12**: e161. doi:10.1186/1475-2875-12-161.

Murray CJL, Ezzati M, Flaxman AD, Lim S, Lozano R, et al. 2012. GBD 2010: design, definitions, and metrics. *Lancet* **380**: 2063-2066. doi:10.1016/S0140-6736(12)61899-6.

Murray CJL, Lopez AD. 2013. Measuring the Global Burden of Disease. *N Engl J Med* **369**: 448-457. doi:10.1056/Nejmra1201534.

Murray CJL, Richards MA, Newton JN, Fenton KA, Anderson HR, et al. 2013. UK health performance: findings of the Global Burden of Disease Study 2010. *Lancet* **381**: 997-1020. doi:10.1016/S0140-6736(13)60355-4.

Mylne A, Brady OJ, Huang Z, Pigott DM, Golding N, et al. 2014. A comprehensive database of the geographic spread of past human Ebola outbreaks. *Sci Data* **1**: e140042. doi:10.1038/sdata.2014.42.

Nunes MR, Faria NR, de Vasconcelos JM, Golding N, Kraemer MU, et al. 2015. Emergence and potential for spread of Chikungunya virus in Brazil. *BMC Med* **13**: 102. doi10.1186/s12916-015-0348-x.

Nunes MRT, Palacios G, Faria NR, Sousa EC, Pantoja JA, et al. 2014. Air travel is associated with intracontinental spread of dengue virus serotypes 1-3 in Brazil. *PLoS Negl Trop Dis* **8**: doi:10.1371/journal.pntd.0002769.

Olival KJ, Hayman DTS. 2014. Filoviruses in bats: current knowledge and future directions. *Viruses* **6**: 1759-1788. doi:10.3390/V6041759.

Oliver TH, Gillings S, Girardello M, Rapacciuolo G, Brereton TM, et al. 2012. Population density but not stability can be predicted from species distribution models. *J Appl Ecol* **49**: 581-590. doi:10.1111/j.1365-2664.2012.02138.x.

Ostyn B, Vanlerberghe V, Picado A, Dinesh DS, Sundar S, et al. 2008. Vector control by insecticide-treated nets in the fight against visceral leishmaniasis in the Indian subcontinent, what is the evidence? *Trop Med Int Health* **13**: 1073-1085. doi:10.1111/j.1365-3156.2008.02110.x.

Palmer SR, Soulsby EJL, Torgerson PR, Brown DWG. 2013. Oxford textbook of zoonoses: biology, clinical practice, and public health control. Oxford: Oxford University Press. 884 p.

Paniz-Mondolfi A, Talhari C, Hoffmann LS, Connor DL, Talhari S, et al. 2012. Lobomycosis: an emerging disease in humans and delphinidae. *Mycoses* **55**: 298-309. doi:10.1111/j.1439-0507.2012.02184.x.

Peel AJ, Sargan DR, Baker KS, Hayman DTS, Barr JA, et al. 2013. Continent-wide panmixia of an African fruit bat facilitates transmission of potentially zoonotic viruses. *Nat Commun* **4**: e2770. doi:10.1038/Ncomms3770.

Peterson AT. 2006. Ecologic niche modeling and spatial patterns of disease transmission. *Emerg Infect Dis* **12**: 1822-1826.

Peterson AT, Bauer JT, Mills JN. 2004a. Ecologic and geographic distribution of filovirus disease. *Emerg Infect Dis* **10**: 40-47. doi:10.3201/eid1001.030125.

Peterson AT, Carroll DS, Mills JN, Johnson KM. 2004b. Potential mammalian filovirus reservoirs. *Emerg Infect Dis* **10**: 2073-2081. doi:10.3201/eid1012.040346.

Peterson AT, Lash RR, Carroll DS, Johnson KM. 2006. Geographic potential for outbreaks of Marburg hemorrhagic fever. *Am J Trop Med Hyg* **75**: 9-15.
Peterson AT, Moses LM, Bausch DG. 2014. Mapping transmission risk of Lassa fever in West Africa: the importance of quality control, sampling bias, and error weighting. *PLoS One* **9**: e100711. doi:10.1371/journal.pone.0100711.
Peterson AT, Soberon J, Pearson RG, Anderson RP, Martinez-Meyer E, et al. 2011. Ecological niches and geographic distributions. Princeton: Princeton University Press. 314 p.
Pigott DM, Atun R, Moyes CL, Hay SI, Gething PW. 2012. Funding for malaria control 2006-2010: a comprehensive global assessment. *Malar J* **11**: 246. doi:10.1186/1475-2875-11-246.
Pigott DM, Golding N, Messina JP, Battle KE, Duda KA, et al. 2014. Global database of leishmaniasis occurrence locations, 1960-2012. *Sci Data* **1**: 140036. doi:10.1038/sdata.2014.36.
Pigott DM, Golding N, Mylne A, Huang Z, Weiss DJ, et al. 2015a. Mapping the zoonotic niche of Marburg virus disease in Africa. *Trans R Soc Trop Med Hyg*: doi:10.1093/trstmh/trv024.
Pigott DM, Howes RE, Wiebe A, Battle KE, Golding N, et al. 2015b. Prioritising infectious disease mapping. *PLoS Negl Trop Dis*: in press.
Plowright RK, Eby P, Hudson PJ, Smith IL, Westcott D, et al. 2015. Ecological dynamics of emerging bat virus spillover. *Proc R Soc Lond B Biol Sci* **282**: e20142124. doi:10.1098/Rspb.2014.2124.
Pollock LJ, Tingley R, Morris WK, Golding N, O'Hara RB, et al. 2014. Understanding co-occurrence by modelling species simultaenously with a Joint Species Distribution Model (JSDM). *Methods Ecol Evol* **5**: 397-406. doi:10.1111/2041-210X.12180.
Pourrut X, Delicat A, Rollin PE, Ksiazek TG, Gonzalez JP, et al. 2007. Spatial and temporal patterns of *Zaire ebolavirus* antibody prevalence in the possible reservoir bat species. *J Infect Dis* **196**: S176-S183. doi:10.1086/520541.
Pourrut X, Souris M, Towner JS, Rollin PE, Nichol ST, et al. 2009. Large serological survey showing cocirculation of Ebola and Marburg viruses in Gabonese bat populations, and a high seroprevalence of both viruses in *Rousettus aegyptiacus*. *BMC Infect Dis* **9**: e159. doi:10.1186/1471-2334-9-159.
ProMED-mail. Lassa fever - Benin (02) 20141126.2992727. Available: www.promedmail.org. Accessed: 24th March 2015
Randolph SE. 2001. The shifting landscape of tick-borne zoonoses: tick-borne encephalitis and Lyme borreliosis in Europe. *Philos Trans R Soc Lond B Biol Sci* **356**: 1045-1056.
Reiter P, Sprenger D. 1987. The used tire trade - a mechanism for the worldwide dispersal of container breeding mosquitos. *J Am Mosquito Contr* **3**: 494-501.
Ritmeijer K, Davies C, van Zorge R, Wang SJ, Schorscher J, et al. 2007. Evaluation of a mass distribution programme for fine-mesh impregnated bednets against visceral leishmaniasis in eastern Sudan. *Trop Med Int Health* **12**: 404-414. doi:10.1111/j.1365-3156.2006.01807.x.
Robinson TP, Wint GRW, Conchedda G, Van Boeckel TP, Ercoli V, et al. 2014. Mapping the global distribution of livestock. *PLoS One* **9**: e96084. doi:10.1371/journal.pone.0096084.
Rouquet P, Froment JM, Bermejo M, Kilbourn A, Karesh W, et al. 2005. Wild animal mortality monitoring and human Ebola outbreaks, Gabon and Republic of Congo, 2001-2003. *Emerg Infect Dis* **11**: 283-290. doi:10.3201/eid1102.040533.
Rowland M, Munir A, Durrani N, Noyes H, Reyburn H. 1999. An outbreak of cutaneous leishmaniasis in an Afghan refugee settlement in north-west Pakistan. *Trans R Soc Trop Med Hyg* **93**: 133-136. doi:10.1016/S0035-9203(99)90285-7.
Samy AM, Campbell LP, Peterson AT. 2014. Leishmaniasis transmission: distribution and coarse-resolution ecology of two vectors and two parasites in Egypt. *Rev Soc Bras Med Trop* **47**: 57-62. doi:10.1590/0037-8682-0189-2013.

Shepard DS, Halasa YA, Tyagi BK, Adhish SV, Nandan D, et al. 2014. Economic and disease burden of dengue illness in India. *Am J Trop Med Hyg* **91**: 1235-1242. doi:10.4269/ajtmh.14-0002.

SINAN. Sistema de Informação de Agravos de Notificação. Available: http://dtr2004.saude.gov.br/sinanweb/index.php. Accessed: March 2013

Sinka ME, Bangs MJ, Manguin S, Chareonviriyaphap T, Patil AP, et al. 2011. The dominant *Anopheles* vectors of human malaria in the Asia-Pacific region: occurrence data, distribution maps and bionomic precis. *Parasit Vectors* **4**: 89. doi:10.1186/1756-3305-4-89.

Sinka ME, Bangs MJ, Manguin S, Coetzee M, Mbogo CM, et al. 2010a. The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic precis. *Parasit Vectors* **3**: 117. doi:10.1186/1756-3305-3-117.

Sinka ME, Rubio-Palis Y, Manguin S, Patil AP, Temperley WH, et al. 2010b. The dominant *Anopheles* vectors of human malaria in the Americas: occurrence data, distribution maps and bionomic precis. *Parasit Vectors* **3**: 72. doi:10.1186/1756-3305-3-72.

Smith DL. 2014. Evaluating outbreak control for mosquito-borne pathogens. American Society of Tropical Medicine and Hygiene 63rd Annual Meeting. New Orleans.

Sogoba N, Feldmann H, Safronetz D. 2012. Lassa fever in West Africa: evidence for an expanded region of endemicity. *Zoonoses Public Hlth* **59**: 43-47. doi:10.1111/j.1863-2378.2012.01469.x.

Stoddard ST, Forshey BM, Morrison AC, Paz-Soldan VA, Vazquez-Prokopec GM, et al. 2013. House-to-house human movement drives dengue virus transmission. *Proc Natl Acad Sci USA* **110**: 994-999. doi:10.1073/pnas.1213349110.

Stoddard ST, Wearing HJ, Reiner RC, Jr., Morrison AC, Astete H, et al. 2014. Long-term and seasonal dynamics of dengue in iquitos, peru. *PLoS Negl Trop Dis* **8**: e3003. doi:10.1371/journal.pntd.0003003.

Swanepoel R, Smit SB, Rollin PE, Formenty P, Leman PA, et al. 2007. Studies of reservoir hosts for Marburg virus. *Emerg Infect Dis* **13**: 1847-1851.

ter Meulen J, Lukashevich I, Sidibe K, Inapogui A, Marx M, et al. 1996. Hunting of peridomestic rodents and consumption of their meat as possible risk factors for rodent-to-human transmission of Lassa virus in the Republic of Guinea. *Am J Trop Med Hyg* **55**: 661-666.

Tesh RB. 1994. The emerging epidemiology of Venezuelan hemorrhagic fever and Oropouche fever in tropical South America. *Disease in Evolution* **740**: 129-137. doi:10.1111/j.1749-6632.1994.tb19863.x.

Towner JS, Amman BR, Sealy TK, Carroll SAR, Comer JA, et al. 2009. Isolation of genetically diverse Marburg viruses from Egyptian fruit bats. *PLoS Pathog* **5**: e1000536. doi:10.1371/journal.ppat.1000536.

Towner JS, Khristova ML, Sealy TK, Vincent MJ, Erickson BR, et al. 2006. *Marburgvirus* genomics and association with a large hemorrhagic fever outbreak in Angola. *J Virol* **80**: 6497-6516. doi:10.1128/JVI.00069-06.

Van Hooft P, Cosson JF, Vibe-Petersen S, Leirs H. 2008. Dispersal in *Mastomys natalensis* mice: use of fine-scale genetic analyses for pest management. *Hereditas* **145**: 262-273. doi:10.1111/j.1601-5223.2008.02089.x.

Velasquez-Tibata J, Graham CH, Munch SB. 2015. Using measurement error models to account for georeferencing error in species distribution models. *Ecography* **38**: 1-12. doi:10.1111/ecog.01205.

Walker AR, Bouattour A, Camicas JL, Estrada-Pena A, Horak IG, et al. 2014. Ticks of domestic animals in Africa: a guide to indentification of species. International Consortium on Ticks and Tick Borne Diseases.

Walsh PD, Bermejo M, Rodriguez-Teijeiro JD. 2009. Disease avoidance and the evolution of primate social connectivity: Ebola, bats, gorillas, and chimpanzees. In: Huffman MA, Chapman CA, editors. Primate parasite ecology: the dynamics and study of host-parasite relationships. Cambridge: Cambridge University Press. pp. 183-197.

Walsh PD, Breuer T, Sanz C, Morgan D, Doran-Sheehy D. 2007. Potential for Ebola transmission between gorilla and chimpanzee social groups. *Am Nat* **169**: 684-689. doi:10.1086/513494.

WHO. 2010. Control of the Leishmaniases. Report of a Meeting of the WHO Expert Committee on the Control of Leishmaniases, Geneva, 22-26 March 2010. Geneva: World Health Organization. 186 p.

WHO. Consolidated Ebola virus disease preparedness checklist. Available: http://www.who.int/csr/resources/publications/ebola/ebola-preparedness-checklist/en/. Accessed: 5th June 2015

Wilson AL, Dhiman RC, Kitron U, Scott TW, van den Berg H, et al. 2014. Benefit of insecticide-treated nets, curtains and screening on vector borne diseases, excluding malaria: a systematic review and meta-analysis. *PLoS Negl Trop Dis* **8**: e3228. doi:10.1371/journal.pntd.0003228.

Wittmann TJ, Biek R, Hassanin A, Rouquet P, Reed P, et al. 2007. Isolates of *Zaire ebolavirus* from wild apes reveal genetic lineage and recombinants. *Proc Natl Acad Sci U S A* **104**: 17123-17127. doi:10.1073/pnas.0704076104.

Woldehanna S, Zimicki S. 2015. An expanded One Health model: Integrating social science and One Health to inform study of the human-animal interface. *Soc Sci Med* **129**: 87-95. doi:10.1016/j.socscimed.2014.10.059.

Wood JLN, Leach M, Waldman L, MacGregor H, Fooks AR, et al. 2012. A framework for the study of zoonotic disease emergence and its drivers: spillover of bat pathogens as a case study. *Philos Trans R Soc Lond B Biol Sci* **367**: 2881-2892. doi:10.1098/rstb.2012.0228.

Yang GH, Wang Y, Zeng YX, Gao GF, Liang XF, et al. 2013. Rapid health transition in China, 1990-2010: findings from the Global Burden of Disease Study 2010. *Lancet* **381**: 1987-2015.

Appendix

This appendix includes the following material:

A.1. Appendix to Chapter 2

A.2. Appendix to Chapter 3

A.3. Appendix to Chapter 4

A.4. Appendix to Chapter 5

A.5. Appendix to Chapter 6

A.6. Appendix to Chapter 7

A.7. Hay *et al.* (2013), referred to in Chapter 1

A.8. Pigott *et al.* (2014) – further details on the occurrence database creation for leishmaniasis

A.9. Mylne *et al.* (2014) – further details on the index case database with additional information on their subsequent geographic spread

A.10. Messina *et al.* (2015) – further details on the occurrence database creation for CCHF

A.1. Appendix to Chapter 2

This appendix consists of four parts:

(i) Table of diseases not recommended for mapping (Option 1) with an explanation for this classification (n=171) and potential alternative mapping opportunities.

(ii) Allocating GBD estimates to all mapping diseases.

(iii) Ranking the disease-specific priorities of the global public health community.

(iv) Final cluster prioritisation ranking.

Supporting Information for "Prioritising infectious disease mapping"

Table: Diseases not recommended for mapping (Option 1) with an explanation for this classification (n=171) and potential alternative mapping opportunities.

Disease	Reason for Option 1 (adapted from *Hay et al. (2013)*)	Other mapping opportunities and existing global outputs
Actinomycosis	Endogenous origin[1]	
Acute febrile respiratory disease, Adenoviral	Distribution tied to humans[2]	Mapping of vaccine coverage
Adenovirus infection	Distribution tied to humans	Mapping of vaccine coverage
Adenoviral haemorrhagic conjunctivitis	Distribution tied to humans	Mapping of vaccine coverage
Amoeba – free living	Source of infection present worldwide	
Amoebic abscess	Source of infection present worldwide	
Amoebic colitis	Source of infection present worldwide	
Animal bite-associated infection	Source of infection present worldwide	
Anisakiasis	Source of infection present worldwide	Reported cases [1]
Aspergillosis	Source of infection present worldwide	
Bacillary angiomatosis	Source of infection present worldwide	
Bacillus cereus food poisoning	Source of infection present worldwide	
Bacterial vaginosis	Source of infection present worldwide	
Bartonellosis – cat borne	Source of infection present worldwide	Reported outbreaks [1]
Bartonellosis – other systemic	Source of infection present worldwide	Reported outbreaks [1]
Blastocystis hominis infection	Source of infection present worldwide	
Botulism	Source of infection present worldwide	Mapping of vaccine coverage; Reported outbreaks (using ProMED) [1]
Brain abscess	Endogenous origin	
Brucellosis	Source of infection present worldwide	National annual incidence levels [2]
Campylobacteriosis	Source of infection present worldwide	
Candidiasis (Yeast)	Endogenous origin	
Chancroid	Distribution tied to humans	
Chlamydia infections, misc.	Distribution tied to humans	
Chlamydophila pneumonia infection	Source of infection present worldwide	
Cholecystitis & cholangitis	Endogenous origin	
Chronic fatigue syndrome	Unknown aetiological agent	
Chronic meningococcemia	Source of infection present worldwide	
Clostridial food poisoning	Source of infection present worldwide	
Clostridial myonecrosis	Source of infection present worldwide	Mapping of vaccine coverage
Clostridium difficile colitis	Endogenous origin	
Common cold	Distribution tied to humans	
Conjunctivitis – viral	Distribution tied to humans	Mapping of vaccine coverage
Cryptococcosis (Yeast)	Source of infection present worldwide	
Cryptosporidiosis	Source of infection present worldwide	
Cutaneous larva migrans	Source of infection present worldwide	
Cyclosporiasis	Source of infection present worldwide	
Cysticercosis	Source of infection present worldwide	National level evidence of infection [3,4]
Cytomegalovirus infection (Human herpesvirus 5)	Distribution tied to humans	Mapping of vaccine coverage; national estimates of seroprevalence in females [5]
Dermatophytosis	Source of infection present worldwide	
Dientamoeba fragilis infection	Source of infection present worldwide	
Diphtheria	Distribution tied to humans	Mapping of vaccine coverage [1]; national reported cases [6]
Diphyllobothriasis	Source of infection present worldwide	Reported cases [1]
Dipylidiasis	Source of infection present worldwide	
Dirofilariasis	Source of infection present worldwide	
Endocarditis – infectious	Endogenous origin	
Enterobiasis	Source of infection present worldwide	
Enteroviral hemorrhagic conjunctivitis	Source of infection present worldwide	
Enterovirus infection	Source of infection present worldwide	
Epidural abscess	Endogenous origin	
Erysipelas or cellulitis	Source of infection present worldwide	
Erysipeloid	Source of infection present worldwide	
Erythrasma	Endogenous origin	
Escherichia coli diarrhea	Source of infection present worldwide	
Fungal infection – invasive	Endogenous origin	
Gastroenteritis – viral	Source of infection present worldwide	
Gianotti-Crosti syndrome	Unknown aetiological agent	

Giardiasis	Source of infection present worldwide	
Gonococcal infection	Distribution tied to humans	
Granuloma inguinale (Donovanosis)	Distribution tied to humans	Reported cases [1]
Hepatitis A	Source of infection present worldwide	Mapping of vaccine coverage; national estimates of seroprevalence [7]
Hepatitis B	Distribution tied to humans	Mapping of vaccine coverage; national estimates of seroprevalence and distribution of genotypes [8]
Hepatitis C	Distribution tied to humans	National estimates of seroprevalence [9] and distribution of genotypes [10]
Hepatitis D	Distribution tied to humans	National estimates of seroprevalence and distribution of genotypes [11-13]
Hepatitis E	Source of infection present worldwide	Endemic status [14] and distribution of genotypes [15]
Hepatitis G	Distribution tied to humans	
Herpes B infection	Source of infection present worldwide	
Herpes simplex encephalitis	Distribution tied to humans	
Herpes simplex infection	Distribution tied to humans	
Herpes zoster	Source of infection present worldwide	Mapping of vaccine coverage
Hymenolepis diminuta infection	Source of infection present worldwide	
Hymenolepis nana infection	Source of infection present worldwide	
Infection of wound, puncture, IV line *etc*.	Distribution tied to humans	
Infectious mononucleosis or EBV infection	Distribution tied to humans	
Influenza	Source of infection present worldwide	Mapping of vaccine coverage
Intestinal spirochetosis	Endogenous origin	
Intra-abdominal abscess	Endogenous origin	
Intracranial venous thrombosis	Endogenous origin	
Isosporiasis	Source of infection present worldwide	
Kawasaki disease	Unknown aetiological agent	
Keratoconjunctivitis, Adenoviral	Source of infection present worldwide	
Kikuchi's disease and Kimura disease	Unknown aetiological agent	
Kingella infection	Endogenous origin	
Laryngotracheobronchitis	Source of infection present worldwide	
Legionellosis	Source of infection present worldwide	
Leptospirosis	Source of infection present worldwide	Mapping of vaccine coverage; reported outbreaks [1]
Listeriosis	Source of infection present worldwide	Reported outbreaks [1]
Liver abscess – bacterial	Endogenous origin	
Lymphocytic choriomeningitis	Source of infection present worldwide	
Lymphogranuloma venereum	Distribution tied to humans	
Malignant otitis externa	Endogenous origin	
Measles	Source of infection present worldwide	Mapping of vaccine coverage; national reported cases [6]
Meningitis – aseptic (viral)	Source of infection present worldwide	Mapping of vaccine coverage
Meningitis – bacterial	Source of infection present worldwide	Epidemic risk models [16]
Microsporidiosis	Source of infection present worldwide	
Moniliformis and Macaracanthorhynchus	Source of infection present worldwide	
Mumps	Source of infection present worldwide	Mapping of vaccine coverage; national reported cases [6]
Mycetoma	Source of infection present worldwide	
Mycobacteriosis – *M. marinum*	Source of infection present worldwide	
Mycobacteriosis – *M. scrofulaceum*	Source of infection present worldwide	
Mycobacteriosis – miscellaneous nontuberculosis	Source of infection present worldwide	
Mycoplasma (miscellaneous) infections	Distribution tied to humans	
Mycoplasma pneumoniae infection	Source of infection present worldwide	
Myiasis	Source of infection present worldwide	
Necrotizing skin/soft tissue infections	Endogenous origin	Mapping of vaccine coverage
Orf	Source of infection present worldwide	
Ornithosis	Source of infection present worldwide	
Osteomyelitis	Endogenous origin	
Otitis media	Endogenous origin	Mapping of vaccine coverage
Parainfluenza virus infection	Source of infection present worldwide	
Parvovirus B19 infection	Source of infection present worldwide	
Pediculosis	Source of infection present worldwide	
Pericarditis – bacterial	Endogenous origin	Mapping of vaccine coverage
Perinephric abscess	Endogenous origin	
Perirectal abscess	Endogenous origin	
Peritonitis – bacterial	Endogenous origin	

Pertussis	Source of infection present worldwide	Mapping of vaccine coverage [1]; national reported cases [6]
Pharyngeal and cervical space infections	Endogenous origin	
Pharyngitis – bacterial	Source of infection present worldwide	
Pityriasis rosea	Unknown aetiological agent	
Plesiomonas infection	Source of infection present worldwide	
Pleurodynia	Source of infection present worldwide	
Pneumocystis pneumonia	Source of infection present worldwide	
Pneumonia – bacterial	Endogenous origin	Mapping of vaccine coverage; national estimates of mortality [17]
Protothecosis and chlorellosis	Source of infection present worldwide	
Pseudocowpox	Source of infection present worldwide	
Pyodermas (impetigo, abscess, *etc*.)	Endogenous origin	
Pyomyositis	Distribution tied to humans	
Q fever	Source of infection present worldwide	Mapping of vaccine coverage; reported outbreaks [1]
Rat bite fever – spirillary	Source of infection present worldwide	Reported cases [1]
Rat bite fever – streptobacillary	Source of infection present worldwide	Reported cases [1]
Respiratory syncytial virus infection	Source of infection present worldwide	Mapping of vaccine coverage
Respiratory viruses – miscellaneous	Source of infection present worldwide	
Reye's syndrome	Unknown aetiological agent	
Rheumatic fever	Source of infection present worldwide	
Rhinoscleroma and ozena	Distribution tied to humans	
Rhodococcus equi infection	Source of infection present worldwide	
Roseola or human herpesvirus 6	Distribution tied to humans	
Rotavirus infection	Source of infection present worldwide	Mapping of vaccine coverage; national estimates of mortality [18]
Rubella	Source of infection present worldwide	Mapping of vaccine coverage; national reported cases [6]
Salmonellosis	Source of infection present worldwide	
Sarcocystosis	Source of infection present worldwide	
SARS	Transmission contained worldwide	
Scabies	Source of infection present worldwide	
Scarlet fever	Source of infection present worldwide	
Septic arthritis	Endogenous origin	
Septicemia – bacterial	Endogenous origin	
Shigellosis	Source of infection present worldwide	
Sinusitis	Distribution tied to humans	Mapping of vaccine coverage
Smallpox	Eradicated globally; present only in laboratory reserves	
Sporotrichosis	Source of infection present worldwide	
Staphylococcal food poisoning	Source of infection present worldwide	
Staphylococcal scaled skin syndrome	Endogenous origin	
Streptococcus suis infection	Source of infection present worldwide	Reported cases [1,19]
Strongyloidiasis	Source of infection present worldwide	
Subdural empyema	Endogenous origin	Mapping of vaccine coverage
Suppurative parotitis	Endogenous origin	
Syphilis	Distribution tied to humans	
Taeniasis	Source of infection present worldwide	National level evidence of infection [3]
Tetanus	Source of infection present worldwide	Mapping of vaccine coverage [1]; national reported cases [6]
Thelaziasis	Source of infection present worldwide	
Toxic shock syndrome	Endogenous origin	Mapping of vaccine coverage
Toxocariasis	Source of infection present worldwide	Mapping of vaccine coverage
Toxoplasmosis	Source of infection present worldwide	Mapping of vaccine coverage; seroprevalence status [20]
Typhoid and enteric fever	Source of infection present worldwide	Mapping of vaccine coverage; national estimates of annual incidence [21,22]
Typhus – endemic/murine (flea-borne)	Source of infection present worldwide	
Urinary tract infection	Endogenous origin	
Varicella	Distribution tied to humans	Mapping of vaccine coverage
Vibrio parahemolyticus infection	Source of infection present worldwide	Reported outbreaks [23]
Whipple's disease	Unknown transmission route	
Yersiniosis	Source of infection present worldwide	
Zygomycosis	Source of infection present worldwide	

1. Endogenous origin refers to infections caused by an agent already present on the body.
2. Distribution tied to humans refers to infections believed to be present wherever humans are.

Allocating GBD estimates to all mapping diseases

Quantified estimates allowing comparison of relative disease burden are an essential component to prioritising the diseases for mapping. The most contemporary and systematic estimates of disease burden are provided by the Global Burden of Disease Study (GBD) 2013, which calculates disability-adjusted life years (DALYs) for the major infectious, non-infectious and injury causes of death and morbidity globally [24,25]. For consistency across the diseases, DALY numbers were exclusively taken from the GBD 2013 estimates.

GBD considered 67 categories or groupings relating to infectious diseases (including specific conditions such as "dengue" as well as larger groupings *e.g.* "Encephalitis", "Food-borne trematodiases", and also aggregated "Diarrheal diseases" categories), whilst the present study considered 176 infectious diseases identified for mapping. The process of appropriately allocating disease burden estimates across this larger number of diseases is described in this Appendix. It is worth specifying that many of the 67 infectious diseases were included in the Option 1 (*i.e.* diseases not currently recommended for mapping) category, due to being globally endemic. The table below summarises the 176 infectious diseases and their assigned DALYs. Further explanation is provided in the footnotes beneath.

Clusters were first classified by taxonomy of causative agent, then by similarities in transmission route [26].

Table outlining clusters, constituent diseases, Global Burden of Disease assignment and DALY score. (B) – bacteria, (N) – nematode, (Pl) – platyhelminth, (V) – virus.

Cluster	Disease name [a]	GBD assignment	DALYs assigned
Direct contact (B)	Anthrax	Other IDs	116,065[1]
Direct contact (B)	Brazilian purpuric fever	Other IDs	116,065[1]
Direct contact (B)	Leprosy	Leprosy	39,707[2]
Direct contact (B)	Mycobacteriosis – *M. ulcerans*	Other IDs	116,065[1]
Direct contact (B)	Trachoma	Trachoma	171,170[2]
Direct contact (B)	Tropical phagedenic ulcer	Not identified	100[3]

Food/Water-borne (B)	*Aeromonas* and marine *Vibrio* infections	Diarrheal diseases	3,308,935[4]
Food/Water-borne (B)	Cholera	Diarrheal diseases	3,308,935[4]
Food/Water-borne (B)	Enteritis necroticans	Diarrheal diseases	3,308,935[4]
Food/Water-borne (B)	Sennetsu neorickettsiosis	Other NTD	35,198[5]
Burkholderiaceae	Glanders	Other IDs	116,065[1]
Burkholderiaceae	Melioidosis	Other IDs	116,065[1]
Treponematoses	Endemic syphilis (bejel)	Other IDs	116,065[1]
Treponematoses	Pinta	Other IDs	116,065[1]
Treponematoses	Yaws	Other IDs	116,065[1]
Flea/Mite-borne (B)	Plague	Other IDs	116,065[1]
Flea/Mite-borne (B)	*Rickettsia felis* infection	Other NTD	35,198[5]
Flea/Mite-borne (B)	Rickettsialpox	Other NTD	35,198[5]
Flea/Mite-borne (B)	Typhus – epidemic	Other NTD	35,198[5]
Flea/Mite-borne (B)	Typhus – scrub (mite-borne)	Other NTD	35,198[5]
Flea/Mite-borne (B)	Relapsing fever	Other NTD	35,198[5]
Other VBD (B)	Bartonellosis – South American	Other IDs	116,065[1]
Tick-borne (B)	Tularemia	Other IDs	116,065[1]
Tick-borne (B)	Anaplasmosis	Other NTD	35,198[5]
Tick-borne (B)	Ehrlichiosis – human monocytic	Other NTD	35,198[5]
Tick-borne (B)	African tick bite fever	Other NTD	35,198[5]
Tick-borne (B)	Astrakhan fever	Other NTD	35,198[5]
Tick-borne (B)	Flinders Island spotted fever	Other NTD	35,198[5]

Tick-borne (B)	Israeli spotted fever	Other NTD	35,198[5]
Tick-borne (B)	Japanese spotted fever	Other NTD	35,198[5]
Tick-borne (B)	North Asian tick typhus	Other NTD	35,198[5]
Tick-borne (B)	Queensland tick typhus	Other NTD	35,198[5]
Tick-borne (B)	*Rickettsia sibirica mongolotimonae* infection	Other NTD	35,198[5]
Tick-borne (B)	Rocky Mountain spotted fever	Other NTD	35,198[5]
Tick-borne (B)	Lyme disease	Other NTD	35,198[5]
Tuberculosis	Tuberculosis	Tuberculosis	49,816,215[2]
Fungal infections	Chromomycosis	Other IDs	116,065[1]
Fungal infections	Entomophthoramycosis	Other IDs	116,065[1]
Fungal infections	Histoplasmosis – African	Other IDs	116,065[1]
Fungal infections	Lobomycosis	Other IDs	116,065[1]
Fungal infections	Penicilliosis	Other IDs	116,065[1]
Fungal infections	Rhinosporidiosis	Other IDs	116,065[1]
Fungal infections	Blastomycosis	Other IDs	116,065[1]
Fungal infections	Coccidioidomycosis	Other IDs	116,065[1]
Fungal infections	Histoplasmosis	Other IDs	116,065[1]
Fungal infections	Paracoccidioidomycosis	Other IDs	116,065[1]
Oomycetes	Pythiosis	Not identified	100[3]
Unknown agents	Brainerd diarrhoea	Diarrheal diseases	3,308,935[4]
Unknown agents	Tropical pulmonary eosinophilia	Not identified	100[3]
Unknown agents	Tropical sprue	Not identified	100[3]
Unknown agents	Viliuisk encephalomyelitis	Encephalitis	300,265[6]
Arthropoda	Pentastomiasis – *Armillifer*	Not identified	100[3]
Arthropoda	Pentastomiasis – *Linguatula*	Not identified	100[3]
Arthropoda	Tungiasis	Not identified	100[3]
Leishmaniasis	Leishmaniasis – cutaneous/mucosal,	Cutaneous leishmaniasis	20,826[2]

	New World		
Leishmaniasis	Leishmaniasis – cutaneous/mucosal, Old World	Cutaneous leishmaniasis	20,826[2]
Leishmaniasis	Leishmaniasis – visceral	Visceral leishmaniasis	4,241,487[2]
Malaria	*Plasmodium knowlesi*	Malaria	13,098,627[2]
Malaria	*Plasmodium malariae*	Malaria	13,098,627[2]
Malaria	*Plasmodium ovale*	Malaria	13,098,627[2]
Malaria	*Plasmodium vivax*	Malaria	13,098,627[2]
Malaria	*Plasmodium falciparum*	Malaria	13,098,627[2]
Babesiosis	Babesiosis	Other NTD	35,198[5]
Trypanosomiasis	African trypanosomiasis	Trypanosomiasis	390,075[2]
Trypanosomiasis	American trypanosomiasis	Chagas	338,489[2]
Filariasis	Filariasis - Bancroftian	Lymphatic filariasis	674,033[2]
Filariasis	Filariasis – *Brugia malayi*	Lymphatic filariasis	674,033[2]
Filariasis	Filariasis – *Brugia timori*	Lymphatic filariasis	674,033[2]
Fly-borne (N)	Mansonelliasis – *M. ozzardi*	Other NTD	35,198[5]
Fly-borne (N)	Mansonelliasis – *M. perstans*	Other NTD	35,198[5]
Fly-borne (N)	Mansonelliasis – *M. streptocerca*	Other NTD	35,198[5]
Fly-borne (N)	Loiasis	Other NTD	35,198[5]
Fly-borne (N)	Onchocerciasis	Onchocerciasis	1,179,826[2]
Food/Water-borne (N)	Angiostrongyliasis	Other NTD	35,198[5]
Food/Water-borne (N)	Angiostrongyliasis – abdominal	Other NTD	35,198[5]
Food/Water-borne (N)	Baylisascariasis	Not identified	100[3]
Food/Water-	Capillariasis – extra	Other NTD	35,198[5]

borne (N)	intestinal		
Food/Water-borne (N)	Capillariasis – intestinal	Other NTD	35,198[5]
Food/Water-borne (N)	Dioctophyme renalis infection	Other NTD	35,198[5]
Food/Water-borne (N)	Gnathostomiasis	Other NTD	35,198[5]
Food/Water-borne (N)	Gongylonemiasis	Other NTD	35,198[5]
Food/Water-borne (N)	Lagochilascariasis	Other NTD	35,198[5]
Food/Water-borne (N)	Mammomonogamiasis	Other NTD	35,198[5]
Food/Water-borne (N)	Oesophagostomiasis	Other NTD	35,198[5]
Food/Water-borne (N)	Trichostrongyliasis	Other NTD	35,198[5]
Food/Water-borne (N)	Dracunculiasis	Other NTD	35,198[5]
Soil-transmitted helminth	Ascariasis	Ascariasis	1,271,708[2]
Soil-transmitted helminth	Hookworm	Hookworm	2,181,665[2]
Soil-transmitted helminth	Trichuriasis	Trichuriasis	576,030[2]
Other parasites	Balantidiasis	Other intestinal infectious diseases	61,179[8]
Other parasites	*Entamoeba polecki* infection	Other intestinal infectious diseases	61,179[8]
Cestodes	*Bertiella* and *Inermicapsifer*	Other NTD	35,198[5]
Cestodes	Coenurosis	Other NTD	35,198[5]
Cestodes	Echinococcosis – American polycystic	Cystic echinococcosis	60,557[2]
Cestodes	Echinococcosis – mulitlocular	Cystic echinococcosis	60,557[2]

Cestodes	Echinococcosis - unilocular	Cystic echinococcosis	60,557[2]
Cestodes	Sparganosis	Other NTD	35,198[5]
Food/Water-borne (Pl)	Dicroceliasis	Food-borne trematodiases	302,902[7]
Food/Water-borne (Pl)	Echinostomiasis	Food-borne trematodiases	302,902[7]
Food/Water-borne (Pl)	Fasciolopsiasis	Food-borne trematodiases	302,902[7]
Food/Water-borne (Pl)	Gastrodiscoidiasis	Food-borne trematodiases	302,902[7]
Food/Water-borne (Pl)	Heterophyid infections	Food-borne trematodiases	302,902[7]
Food/Water-borne (Pl)	Metagonimiasis	Food-borne trematodiases	302,902[7]
Food/Water-borne (Pl)	Metorchiasis	Food-borne trematodiases	302,902[7]
Food/Water-borne (Pl)	Nanophyetiasis	Food-borne trematodiases	302,902[7]
Food/Water-borne (Pl)	Paragonimiasis	Food-borne trematodiases	302,902[7]
Food/Water-borne (Pl)	Clonorchiasis	Food-borne trematodiases	302,902[7]
Food/Water-borne (Pl)	Fascioliasis	Food-borne trematodiases	302,902[7]
Food/Water-borne (Pl)	Opisthorchiasis	Food-borne trematodiases	302,902[7]
Water-borne (Pl)	Cercarial dermatitis	Schistosomiasis	437,549[2]
Water-borne (Pl)	*Schistosoma haematobium*	Schistosomiasis	437,549[2]
Water-borne (Pl)	*Schistosoma intercalatum*	Schistosomiasis	437,549[2]
Water-borne (Pl)	*Schistosoma japonicum*	Schistosomiasis	437,549[2]
Water-borne (Pl)	*Schistosoma mansoni*	Schistosomiasis	437,549[2]
Water-borne (Pl)	*Schistosoma mattheei*	Schistosomiasis	437,549[2]
Water-borne (Pl)	*Schistosoma mekongi*	Schistosomiasis	437,549[2]

Mosquito-borne (V)	California serogroup viruses	Encephalitis	300,265[6]
Mosquito-borne (V)	Rift Valley fever	Other NTD	35,198[5]
Mosquito-borne (V)	Dengue	Dengue	1,142,734[2]
Mosquito-borne (V)	Ilheus and Bussuquara	Encephalitis	300,265[6]
Mosquito-borne (V)	Japanese encephalitis	Encephalitis	300,265[6]
Mosquito-borne (V)	Murray valley encephalitis	Encephalitis	300,265[6]
Mosquito-borne (V)	Rocio	Encephalitis	300,265[6]
Mosquito-borne (V)	Spondweni	Other NTD	35,198[5]
Mosquito-borne (V)	St. Louis encephalitis	Encephalitis	300,265[6]
Mosquito-borne (V)	Wesselsbron	Other NTD	35,198[5]
Mosquito-borne (V)	West Nile fever	Other NTD	35,198[5]
Mosquito-borne (V)	Yellow fever	Yellow fever	30,680[2]
Mosquito-borne (V)	Zika	Other NTD	35,198[5]
Mosquito-borne (V)	Coltiviruses – Old World	Other NTD	35,198[5]
Mosquito-borne (V)	Barmah Forest disease	Other NTD	35,198[5]
Mosquito-borne (V)	Chikungunya	Other NTD	35,198[5]
Mosquito-borne (V)	Eastern equine encephalitis	Encephalitis	300,265[6]
Mosquito-borne (V)	Karelian fever	Other NTD	35,198[5]

Mosquito-borne (V)	Mayaro	Other NTD	35,198[5]
Mosquito-borne (V)	Ockelbo disease	Other NTD	35,198[5]
Mosquito-borne (V)	O'nyong nyong	Other NTD	35,198[5]
Mosquito-borne (V)	Pogosta disease	Other NTD	35,198[5]
Mosquito-borne (V)	Ross River virus	Other IDs	116,065[1]
Mosquito-borne (V)	Sindbis	Other NTD	35,198[5]
Mosquito-borne (V)	Venezuelan equine encephalitis	Other NTD	35,198[5]
Mosquito-borne (V)	Western equine encephalitis	Encephalitis	300,265[6]
Other arbovirus	Bunyaviridae infections – miscellaneous	Other NTD	35,198[5]
Other arbovirus	Group C viral fevers	Other NTD	35,198[5]
Other arbovirus	Oropouche virus	Other NTD	35,198[5]
Other arbovirus	Sandfly fever	Other NTD	35,198[5]
Tick-borne (V)	Crimean-Congo hemorrhagic fever	Other NTD	35,198[5]
Tick-borne (V)	Alkhurma hemorrhagic fever	Other NTD	35,198[5]
Tick-borne (V)	Kyasanur Forest disease	Other NTD	35,198[5]
Tick-borne (V)	Louping ill	Encephalitis	300,265[6]
Tick-borne (V)	Omsk hemorrhagic fever	Other NTD	35,198[5]
Tick-borne (V)	Powassan	Encephalitis	300,265[6]
Tick-borne (V)	Tick-borne encephalitis	Encephalitis	300,265[6]
Tick-borne (V)	Tick-borne encephalitis - Russian spring summer	Encephalitis	300,265[6]
Tick-borne (V)	Thogoto	Encephalitis	300,265[6]
Tick-borne (V)	Colorado tick fever	Other NTD	35,198[5]

Avian contact (V)	Avian influenza	Not identified	100[3]
Mammal contact (V)	Chandipura and Vesicular stomatitis	Other NTD	35,198[5]
Mammal contact (V)	Hendra virus disease	Encephalitis	300,265[6]
Mammal contact (V)	Monkeypox	Other IDs	116,065[1]
Mammal contact (V)	Nipah and Nipah-like virus disease	Encephalitis	300,265[6]
Mammal contact (V)	Rabies	Rabies	1,242,902[2]
Mammal contact (V)	Tanapox virus disease	Not identified	100[3]
Mammal contact (V)	Vaccinia and cowpox	Not identified	100[3]
Filoviridae	Ebola	Other NTD	35,198[5]
Filoviridae	Marburg	Other NTD	35,198[5]
Picornaviridae	Poliomyelitis	Other IDs	116,065[1]
Robovirus	Argentine hemorrhagic fever (Junin virus)	Other NTD	35,198[5]
Robovirus	Bolivian hemorrhagic fever (Machupo virus)	Other NTD	35,198[5]
Robovirus	Brazilian hemorrhagic fever (Sabia virus)	Other NTD	35,198[5]
Robovirus	Lassa fever	Other NTD	35,198[5]
Robovirus	Venezuelan hemorrhagic fever	Other NTD	35,198[5]
Robovirus	Whitewater Arroyo virus infection	Other NTD	35,198[5]
Robovirus	Hantavirus infection – Old World	Other NTD	35,198[5]
Robovirus	Hantavirus pulmonary syndrome	Other IDs	116,065[1]
HIV	HIV	HIV	69,480,661[2]

[a] – for a more thorough description of pathogenic agents included in these terms, please refer to the publication by *Hay et al. (2013)*.

Table footnotes

1. **Other infectious diseases**

A variety of diseases are included in the GBD "Other infectious diseases" category (see appendix of GBD outputs for correspondence between GBD categorisations and ICD-10 codes [24,25,27]). Of the 355 diseases listed in a previous study considering the diversity of infectious diseases [26], 66 were classified as "Other infectious diseases", each assigned 7,660,320/66 = 116,065 DALYs.

2. **Direct correspondences**

Diseases for which direct associations could be made between the nomenclature of the GBD and the list of mapping diseases are included in the table above. Some differences are apparent in the level of the nomenclature at the disease level, with GBD entries corresponding to clusters rather than individual diseases. For instance, the GBD "malaria" entry corresponds to the cluster level in the mapping list, which is made up of the five disease species (*Plasmodium falciparum, P. vivax, P. ovale, P. malariae* and *P. knowlesi*).

3. **Non-corresponding diseases**

For a handful of diseases, reconciliation of the disease with GBD (directly or through an ICD-10 code) was not possible. Typically these diseases were of endogenous origin (and were often classified as non-communicable within GBD) or related to ectoparasites. These diseases were each allocated a nominal 100 DALYs. Of the diseases included in the GBD, the infectious disease with lowest DALYs was yellow fever with 30,680, followed by leprosy with 39,707. The nominal DALY allocation to diseases not identified in the GBD list would therefore not skew any of the analytical outputs but enabled the diseases to be included in the analytical plots.

4. **Diarrhoeal diseases**

A variety of diseases are included in the GBD "Diarrheal diseases" category (see appendix of GBD outputs for correspondence between GBD categorisations and ICD-10 codes [24,25,27]). Of the 355 diseases listed by *Hay et al. (2013)*, 22 were classified as "Diarrheal diseases", each assigned 72,796,573/22 = 3,308,935 DALYs.

5. **Other neglected tropical diseases**

A variety of diseases are included in the GBD "Other neglected tropical diseases" category (see appendix of GBD outputs for correspondence between GBD categorisations and ICD-10 codes [24,25,27]). Of the 355 diseases listed by *Hay et al. (2013)*, 89 were classified as "Other neglected tropical diseases", each assigned 3,132,659/89 = 35,198 DALYs.

6. Encephalitis

A variety of diseases are included in the GBD "Encephalitis" category (see appendix of GBD outputs for correspondence between GBD categorisations and ICD-10 codes [24,25,27]). Of the 355 diseases listed by *Hay et al. (2013)*, 16 were classified as "Encephalitis", each assigned 4,804,232/16 = 300,265 DALYs.

7. Food-borne trematodiases

A variety of diseases are included in the GBD "Food-borne trematodiases" category (see appendix of GBD outputs for correspondence between GBD categorisations and ICD-10 codes [24,25,27]). Of the 355 diseases listed by *Hay et al. (2013)*, 12 were classified as "Food-borne trematodiases", each assigned 3,634,820/12 = 302,902 DALYs.

8. Other intestinal infectious diseases

A variety of diseases are included in the GBD "Other intestinal infectious diseases" category (see appendix of GBD outputs for correspondence between GBD categorisations and ICD-10 codes [24,25,27]). Of the 355 diseases listed by *Hay et al. (2013)*, 7 were classified as "Other intestinal infectious diseases", each assigned 428,255/7 = 61,179 DALYs.

Ranking the disease-specific priorities of the global public health community.

An important factor in determining a prioritization road-map for disease mapping is to ensure that diseases for which maps are in greatest demand are addressed first. This was achieved by a survey of a subset of anticipated end-users of the maps, to see which diseases were being targeted by the major public health stakeholders: publicly-funded health agencies, private companies (*e.g.* vaccine developers), political commitments, non-governmental organisation (NGOs), advocates and practitioners, as well as the scientific research community. Examples from the different categories of stakeholders were looked into to capture a sample across the spectrum of interest groups.

1. **International global health agencies**

Diseases of primary interest to the NGO and private sector of the global health community were surveyed from a varied spectrum of agencies. Agencies whose interests were focused on a single cause were excluded (*e.g.* President's Malaria Initiative and the Polio Global Eradication Initiative); as were those whose scope was too comprehensive and far-reaching to allow determination of specific diseases which they prioritized (*e.g.*UNICEF, WHO and World Bank). The agencies included in the review are listed in the table below and were selected to cover a broad range of fields interested in public health. It was found that the diseases identified by such agencies reached a critical threshold of around 45 diseases – this number remained constant beyond approximately 15 agencies, as a result the specific agencies selected is not significant. Webpages describing each agency's scope and priorities were searched, and all diseases cited as being a focus of the agency's work were recorded. Diseases were allocated a point for each agency which targeted them.

Tag-line descriptions and URLs of the fifteen agencies surveyed for their diseases of focus.

Agency	**Description (paraphrased from self-description)**	**URL**
BMGF: Bill & Melinda Gates Foundation	Funding agency: harness advances in science and technology to save lives in developing countries.	http://www.gatesfoundation.org/What-We-Do
Carter Center	Focus on eradication and elimination of diseases. Health education and simple, low-cost methods. Surveillance and delivery systems.	http://www.cartercenter.org/health/index.html
CDC Neglected Parasite Infections	Focussed public health action focus on diseases selected based on: number of people infected, severity of the illness, preventability and treatability.	http://www.cdc.gov/parasites/npi.html
Clinton Foundation	Forms partnerships to strengthen health systems, increase access to lifesaving services, fight climate change, expand economic opportunities.	http://www.clintonfoundation.org/our-work/clinton-health-access-initiative
CORE Group	Brings together 70+ members to improve and	http://www.coregroup.

	expand community health practices. Supports NGOs and governments.	org/index.php
GAVI	Increasing access to immunisation.	http://www.gaviallian ce.org/support/nvs/
IVI	Private company: vaccine development	http://www.ivi.int/web /www/02_04
London 2012 Declaration/WHO 2020 goals for NTDs	Commitment & collaboration by varied set of partners to address NTD disease burden.	http://unitingtocombat ntds.org/
MSF: Médecins Sans Frontières	Medical consultations: medical activities range from vaccination campaigns to complex surgery	http://www.msf.org.uk /medical-issues-0
PATH	Driving transformative innovation to save lives. Development to delivery.	http://www.path.org/o ur-work/emerging- and-epidemic- diseases.php
PSI: Population Services International	Focus on serious challenges to global health. Communication & distribution efforts to ensure acceptance & proper use of health services & products.	http://www.psi.org/ou r-work/healthy-lives
Sabin	Private company: vaccine development. Development of sustainable, cost-effective vaccines.	http://www.sabin.org/ programs/vaccine- development
Shantha Biotechnics	Private company: vaccine development. Develop cost-effective human health care products.	http://www.shanthabio tech.com/r&d.html
TDR: Special Programme for Research and Training in Tropical Diseases	Hosted by WHO, global programme for scientific collaboration to help facilitate, support and influence efforts to combat diseases of poverty.	http://www.who.int/td r/diseases-topics/en/
USAID	US government agency. Works to end extreme global poverty and enable resilient, democratic societies realise their potential.	http://www.usaid.gov/ what-we-do/global- health

2. Notifiable Diseases

Diseases which were notifiable to any of the following Public Health agencies were given one point. These countries were chosen in order to reflect the main GBD regions (‘High Income’, ‘Latin America and Caribbean’, ‘Sub-Saharan Africa’, ‘North Africa and Mediterranean’, ‘South Asia’, ‘Southeast Asia East Asia and Oceania’ and ‘Central and Eastern Europe and Central Asia’.

USA - http://wwwn.cdc.gov/NNDSS/script/ConditionList.aspx?Type=0&Yr=2014

Brazil - http://www.anvisa.gov.br/hotsite/cruzeiros/documentos/2013/Annex%20II%20-%20%20NOTIFIABLE%20DISEASES%20IN%20BRAZIL.pdf

Zambia – http://www.hsa.org.za/misc/Notifiable%20Diseases.pdf

United Arab Emirates - https://www.haad.ae/HAAD/LinkClick.aspx?fileticket=NR4IhJoy5Bo%3D&tabid=1177

India - http://health.puducherry.gov.in/details_of_notifiable_diseases.htm

Malaysia - http://www.jknselangor.moh.gov.my/documents/pdf/sharingDoc/pdf/leptospirosis/Notifikasi.pdf

Croatia - http://hzjz.hr/wp-content/uploads/2013/11/definicije_zb_12.pdf

Final cluster prioritisation ranking. B = bacteria, Pl = platyhelminth, N = nematode, V = virus, VBD = vector-borne disease.

Ranking	Cluster
1	Malaria
2	HIV
3	Tuberculosis
4	Food/Water-borne (B)
5	Water-borne (Pl)
6	Trypanosomiasis
7	Filariasis
8	Soil-transmitted helminths
9	Leishmaniasis
10	Unknown agent
11	Picornaviridae
12	Food/Water-borne (N)
13	Fly-borne (N)
14	Direct contact (B)
15	Mosquito-borne (V)
16	Mammal contact (V)
17	Tick-borne (V)
18	Tick-borne (B)
19	Food/Water-borne (Pl)
20	Fungus
21	Flea/Mite-borne (B)
22	Robovirus
23	Filoviridae
24	Treponematoses
25	Babesiosis
26	Cestodes
27	Burkholderiaceae
28	Avian contact (V)
29	Other arbovirus
30	Other parasites
31	Other VBD (B)
32	Arthropoda
33	Oomycetes

References

1. Wertheim HFL, Horby P, Woodall JP (2012) Atlas of human infectious diseases (available online at: https://infectionatlas.org/). Oxford: Wiley-Blackwell. 280 p.
2. Pappas G, Papadimitriou P, Akritidis N, Christou L, Tsianos EV (2006) The new global map of human brucellosis. Lancet Infect Dis 6: 91-99.
3. Willingham AL, Engels D (2006) Control of *Taenia solium* cysticercosis/taeniosis. Adv Parasit 61: 509-566.
4. WHO Countries and areas at risk of cysticercosis, 2009. Available: http://gamapserver.who.int/mapLibrary/Files/Maps/Global_cysticercosis_2009.png. Accessed: July 2014
5. Cannon MJ, Schmid DS, Hyde TB (2010) Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. Rev Med Virol 20: 202-213.
6. WHO Immunization surveillance, assessment and monitoring: reported incidence time series. Available: http://apps.who.int/immunization_monitoring/data/data_subject/en/index.html. Accessed: July 2014
7. Jacobsen KH, Wiersma ST (2010) Hepatitis A virus seroprevalence by age and world region, 1990 and 2005. Vaccine 28: 6653-6657.
8. Kurbanov F, Tanaka Y, Mizokami M (2010) Geographical and genetic diversity of the human hepatitis B virus. Hepatol Res 40: 14-30.
9. Lavanchy D (2011) Evolving epidemiology of hepatitis C virus. Clin Microbiol Infect 17: 107-115.
10. Messina JP, Humphreys I, Flaxman A, Brown A, Cooke GS, et al. (2015) The global distribution and prevalence of HCV genotypes. Hepatology 61: 77-87.
11. Hughes SA, Wedemeyer H, Harrison PM (2011) Hepatitis delta virus. Lancet 378: 73-85.
12. Pascarella S, Negro F (2011) Hepatitis D virus: an update. Liver Int 31: 7-21.
13. Wedemeyer H, Manns MP (2010) Epidemiology, pathogenesis and management of hepatitis D: update and challenges ahead. Nat Rev Gastroenterol Hepatol 7: 31-40.
14. Goens SD, Perdue ML (2004) Hepatitis E viruses in humans and animals. Anim Health Res Rev 5: 145-156.
15. Purcell RH, Emerson SU (2008) Hepatitis E: An emerging awareness of an old disease. J Hepatol 48: 494-503.
16. Savory EC, Cuevas LE, Yassin MA, Hart CA, Molesworth AM, et al. (2006) Evaluation of the meningitis epidemics risk model in Africa. Epidemiol Infect 134: 1047-1051.
17. O'Brien KL, Wolfson LJ, Watt JP, Henkle E, Deloria-Knoll M, et al. (2009) Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. Lancet 374: 893-902.
18. Glass RI, Parashar UD, Bresee JS, Turcios R, Fischer TK, et al. (2006) Rotavirus vaccines: current prospects and future challenges. Lancet 368: 323-332.
19. Lun ZR, Wang QP, Chen XG, Li AX, Zhu XQ (2007) *Streptococcus suis*: an emerging zoonotic pathogen. Lancet Infect Dis 7: 201-209.
20. Pappas G, Roussos N, Falagas ME (2009) Toxoplasmosis snapshots: Global status of *Toxoplasma gondii* seroprevalence and implications for pregnancy and congenital toxoplasmosis. Int J Parasitol 39: 1385-1394.
21. Connor BA, Schwartz E (2005) Typhoid and paratyphoid fever in travellers. Lancet Infect Dis 5: 623-628.
22. Crump JA, Luby SP, Mintz ED (2004) The global burden of typhoid fever. Bull World Health Organ 82: 346-353.
23. Nair GB, Ramamurthy T, Bhattacharya SK, Dutta B, Takeda Y, et al. (2007) Global dissemination of *Vibrio parahaemolyticus* serotype O3:K6 and its serovariants. Clin Microbiol Rev 20: 39-48.
24. GBD 2013 Disease and Injury Incidence and Prevalence Collaborators (2015) Global, regional, and national incidence, prevalence and YLDs for 301 acute and chronic

diseases and injuries for 188 countries, 1990-2013: a systemtatic analysis for the Global Burden of Disease Study 2013. Under submission.
25. GBD 2013 Mortality and Causes of Death Collaborators (2015) Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematics analysis for the Global Burden of Disease Study 2013. Lancet 385: 117-171.
26. Hay SI, Battle KE, Pigott DM, Smith DL, Moyes CL, et al. (2013) Global mapping of infectious disease. Philos Trans R Soc Lond B Biol Sci 368: 20120250.
27. WHO International statistical classification of diseases and related health problems 10th revision. Available: http://apps.who.int/classifications/icd10/browse/2010/en. Accessed: June 2014

A.2. Appendix to Chapter 3

This appendix consists of three parts:

(i) Global evidence consensus for Cutaneous Leishmaniasis – due to space constraints, this appendix is available online at http://datadryad.org/resource/doi:10.5061/dryad.05f5h/2. Paper versions are available on request.

(ii) Global evidence consensus for Visceral Leishmaniasis – due to space constraints, this appendix is available online at http://datadryad.org/resource/doi:10.5061/dryad.05f5h/1. Paper versions are available on request.

(iii) Populations living in areas of environmental risk of Leishmaniasis.

These data tables reported the estimated populations living in areas of environmental risk of cutaneous leishmaniasis (CL) or visceral leishmaniasis (VL). For a full methodology and indications of geographic range of these populations, please consult the full publication that this dataset is associated with: Pigott et al. (2014) “Global distribution maps of the leishmaniases”

Country	Estimated Population at Risk of CL
Brazil	148786750
Colombia	44869432
Venezuela	26249248
Peru	20095782
Mexico	11929507
Ecuador	12469102
Guatemala	8991769
Argentina	7970827
Haiti	6559161
Dominican Republic	5910084
Bolivia	5727962
Honduras	5488859
Paraguay	5209033
USA	4438373
El Salvador	4660440
Costa Rica	4243499
Nicaragua	4003127
Panama	3308074
Guyana	651271
Suriname	520165
Guadeloupe	415227
Martinique	368280
Belize	237741
French Guiana	216055
New World Total	333,319,768

Table 1: Population at risk estimates for CL in the New World

Country	Estimated Population at Risk of CL
India	319099420
Pakistan	156427700
China	88989808

Egypt	81115232
Iran	60143656
Nigeria	52041104
Bangladesh	44769740
Italy	35991912
Algeria	30969660
Morocco	30282374
Iraq	29738302
Spain	28491340
Saudi Arabia	21682114
Sudan	21327576
Syria	20784102
Yemen	20909202
Turkey	20876026
Ethiopia	20051208
Afghanistan	15616552
DR Congo	13122353
Uzbekistan	12793034
Kenya	12633773
Ghana	12069712
France	10412137
Tunisia	9711311
Côte d'Ivoire	79260723
Sri Lanka	7250986
Greece	6818007
Jordan	6358596
Israel	6241936
Cameroon	6020805
Portugal	5996409
Libya	5716142
Senegal	5869205
Burkina Faso	5379299
Uganda	5133872
Mali	4558003
Nepal	4574548

Turkmenistan	3957351
Tajikistan	3792547
Malawi	3715664
Togo	3708668
Lebanon	3469485
Kashmir	3325686
Azerbaijan	3336433
West Bank	2842185
Niger	2852365
Chad	2634509
Oman	2438769
South Sudan	1632959
Somalia	1554121
Gaza Strip	1515670
Myanmar	1766342
Albania	1451897
Eritrea	1449614
Kuwait	1360076
Georgia	1182311
Armenia	1038948
Gambia	1085648
Guinea	1157852
Mauritania	933688
Central African Republic	873835
South Africa	739913
Djibouti	651927
Kazakhstan	605987
Namibia	538628
Jammu Kashmir	461560
Guinea-Bissau	409057
Cyprus	383121
Bulgaria	373336
Kyrgyzstan	458086
Malta	340932
Croatia	277165

Western Sahara	276087
Macedonia	123064
Thailand	146535
Bosnia and Herzegovina	49072
Dhekelia and Akrotiri SBA	25129
Timor-Leste	7553
Old World Total	1,378,171,654

Table 2: Population at risk estimates for CL in the Old World

Country	Estimated Population at Risk of VL
Brazil	103739410
Colombia	11152233
Mexico	2917514
Peru	9544072
Venezuela	7910364
Guatemala	3834989
Bolivia	1053440
Haiti	1544787
El Salvador	2955878
Paraguay	2764511
Ecuador	2650906
Honduras	2001578
Nicaragua	1431155
Argentina	1055045
Panama	1108801
Costa Rica	506949
Suriname	304616
Guadeloupe	133458
New World Total	156,609,706

Table 3: Population at risk estimates for VL in the New World

Country	Estimated Population at Risk of VL
India	495733890
China	205894780
Pakistan	84579008
Bangladesh	72436664

Egypt	64553600
Nigeria	47012432
Italy	44722576
Iran	45267152
Ethiopia	36732612
Spain	31302114
Turkey	27892914
Algeria	27377244
Morocco	21228010
Uzbekistan	16710510
Sudan	16259580
Myanmar	14260147
Iraq	16618775
Kenya	14151164
Nepal	13864455
Yemen	12390054
DR Congo	12956326
France	12602770
Syria	10079691
Ghana	8797417
Tunisia	8903542
Saudi Arabia	6791114
Afghanistan	8350620
Greece	8662354
Portugal	7716353.5
Thailand	7357333
Côte d'Ivoire	6163972
Burkina Faso	5888366
Senegal	5951419
Uganda	4603301
Azerbaijan	5478662
Malawi	4266093
Kazakhstan	3494969
Kashmir	4140429
Mali	4540417

Libya	4890205
Jordan	4676653
Zambia	4129548
Israel	4664590
South Sudan	3749817
Togo	4209735
Cameroon	3426504
Tajikistan	3349900
Somalia	2363005
Kyrgyzstan	3003517
Eritrea	2665204
Lebanon	3000633
Sri Lanka	1948560
Turkmenistan	2730315
Bulgaria	2543448
Chad	2679550
West Bank	2750155
Serbia	1975425
Niger	2398483
Guinea	2152579
Albania	2267011
Oman	2081964
Georgia	1799810
Tanzania	1575703
Armenia	1634395
Mozambique	1545311
Gaza Strip	1515670
Angola	1276322
Macedonia	1366606
Gambia	1155756
Central African Republic	1069348
Jammu Kashmir	749959
Guinea-Bissau	678762
Bosnia and Herzegovina	514575
Ukraine	109877

Djibouti	652505
Croatia	524861
Malta	340932
Viet Nam	234543
Mongolia	159656
Western Sahara	275970
Cyprus	237389
Timor-Leste	164646
Montenegro	132288
Bhutan	24241
Old World Total	1,531,128,756

Table 4: Population at risk estimates for VL in the Old World

A.3. Appendix to Chapter 4

This appendix consists of three parts

(i) Figure 5 – figure supplement 1: Covariates used in predicting zoonotic transmission niche of Ebola.

(ii) Figure 5 – figure supplement 2: Marginal effect plots for each covariate used in the Ebola virus distribution model.

(iii) Figure 5 – figure supplement 3: Comparison of predictions for zoonotic niche of Ebola virus excluding the Guinea outbreak.

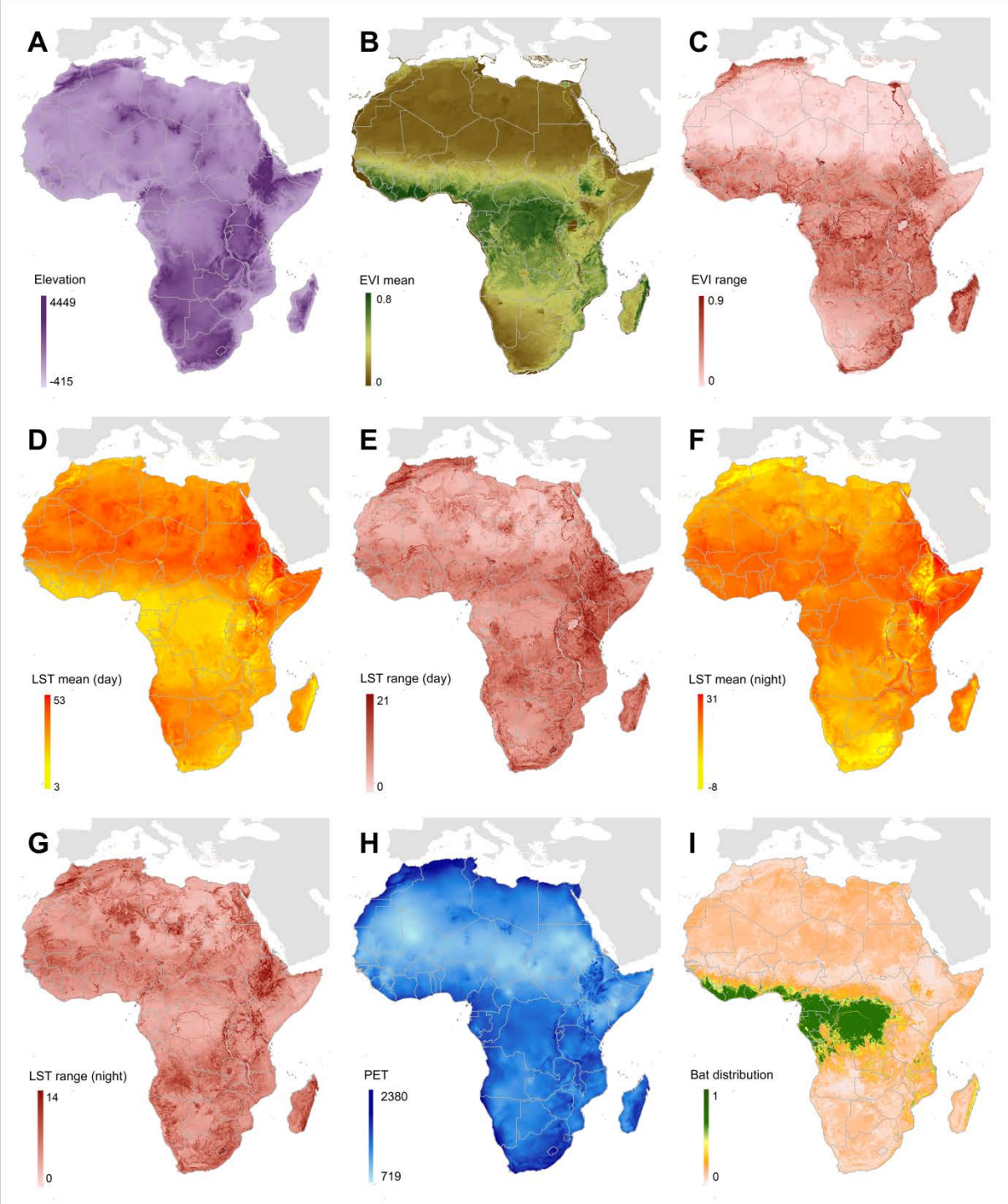


Figure 5—figure supplement 1.Covariates used in predicting zoonotic transmission niche of Ebola.

(**A**) Displays elevation across Africa measured in metres, relative to sea level. (**B** and **C**) Show enhanced vegetation index (EVI) values (mean and spatial range respectively) on a scale from 0 to 1. (**D**–**G**) Display land surface temperature (LST) (mean and spatial range for day and night respectively) measured in degrees Celsius. (**H**) Shows potential evapotranspiration (PET) for Africa, in millimetres per month and (**I**) gives the composite, relative probability of occurrence of the three suspected reservoir bat species. For details of how each of these covariate layers was derived see ‘Materials and methods’.

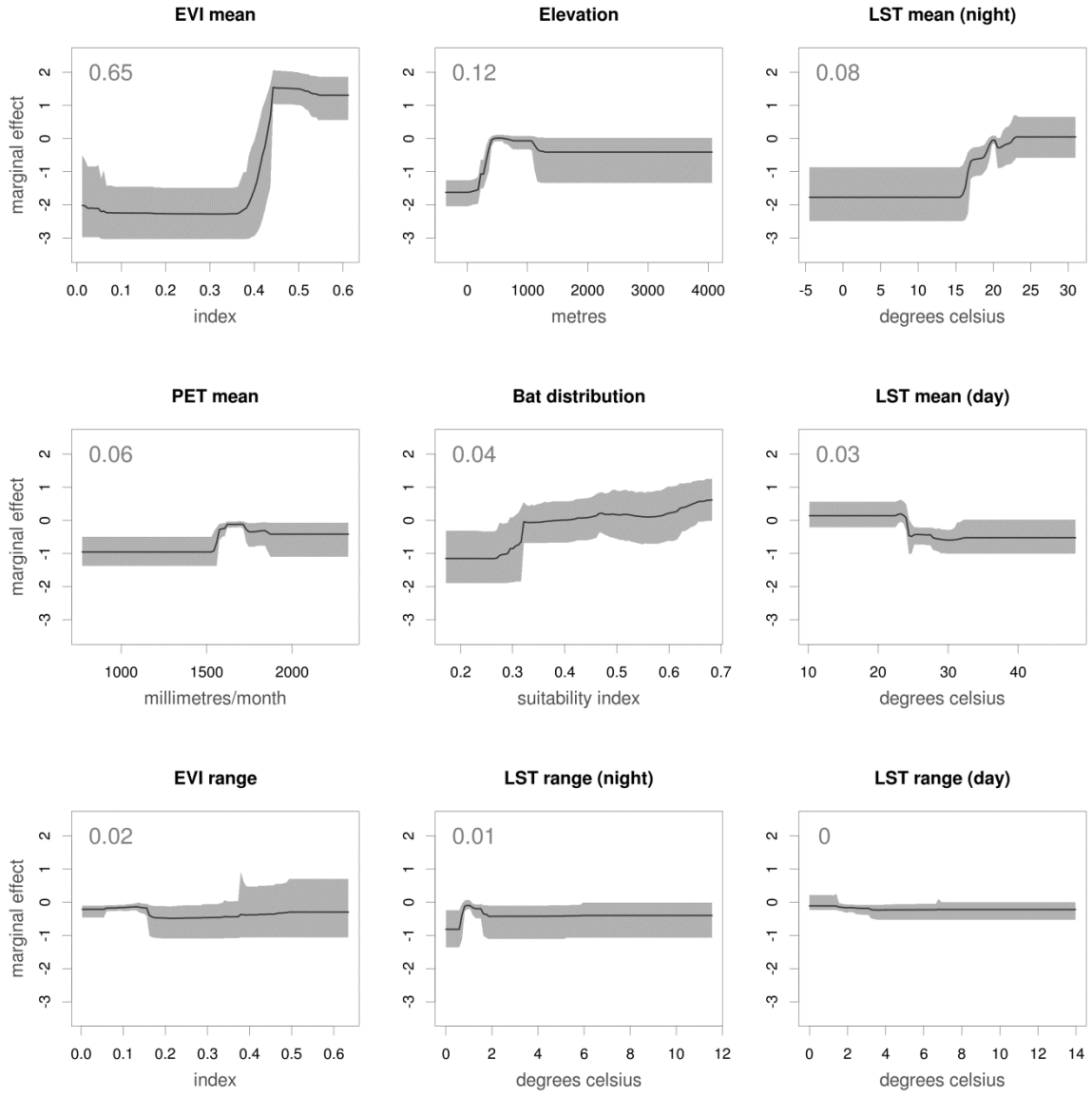


Figure 5—figure supplement 2.Marginal effect plots for each covariate used in the Ebola virus distribution model.

Each panel illustrates the marginal effect (averaging over the effects of other covariates) that changes in each of the covariates has on the predicted relative probability of occurrence of zoonotic Ebola virus transmission. Grey regions and solid lines give the 95% confidence region (a metric of uncertainty) and mean value calculated across all 500 submodels. The mean relative contribution of the covariate to the model (the proportion of iterations in which the covariate was selected by the model-fitting algorithm, indicating sensitivity to the covariates) is given as an inset number. The dependency plots are ordered by mean relative contribution of the covariate. EVI = enhanced vegetation index, LST = land surface temperature and PET = potential evapotranspiration

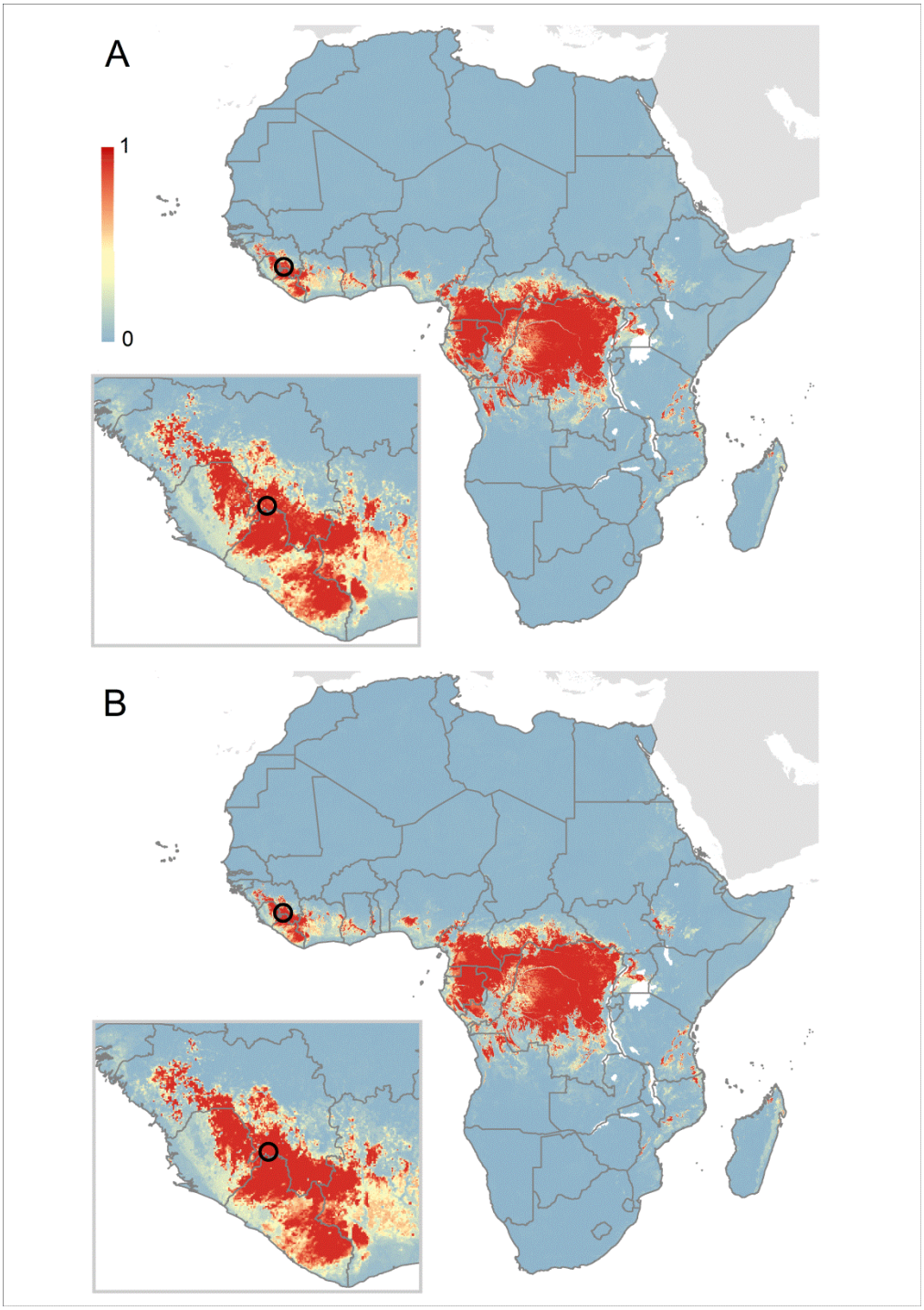


Figure 5—figure supplement 3.Comparison of predictions for zoonotic niche of Ebola virus excluding the Guinea outbreak.

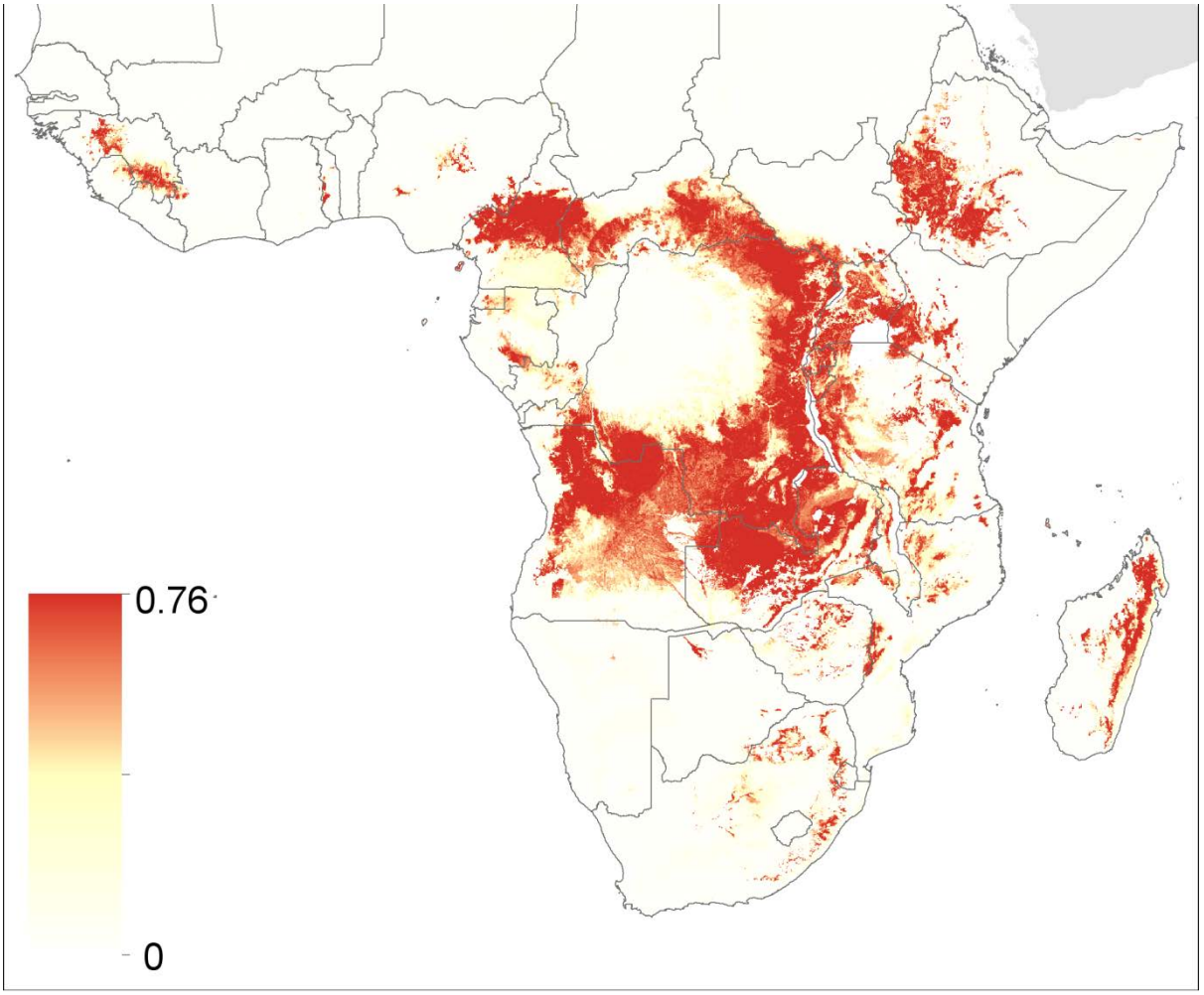
(**A**) Shows the predicted zoonotic niche excluding the index case for the Guinea outbreak from the dataset used to train the model. (**B**) Shows the prediction when including the Guinea data in the model (the model presented in Figure 5). The circle depicts the location of the Guinean index case (#23 in Table 1). As per Figure 5, the scale reflects the relative probability that zoonotic transmission of Ebola virus could occur at these locations; areas closer to 1 (red) are more likely to harbour zoonotic transmission than those closer to 0 (blue).

A.4. Appendix to Chapter 5

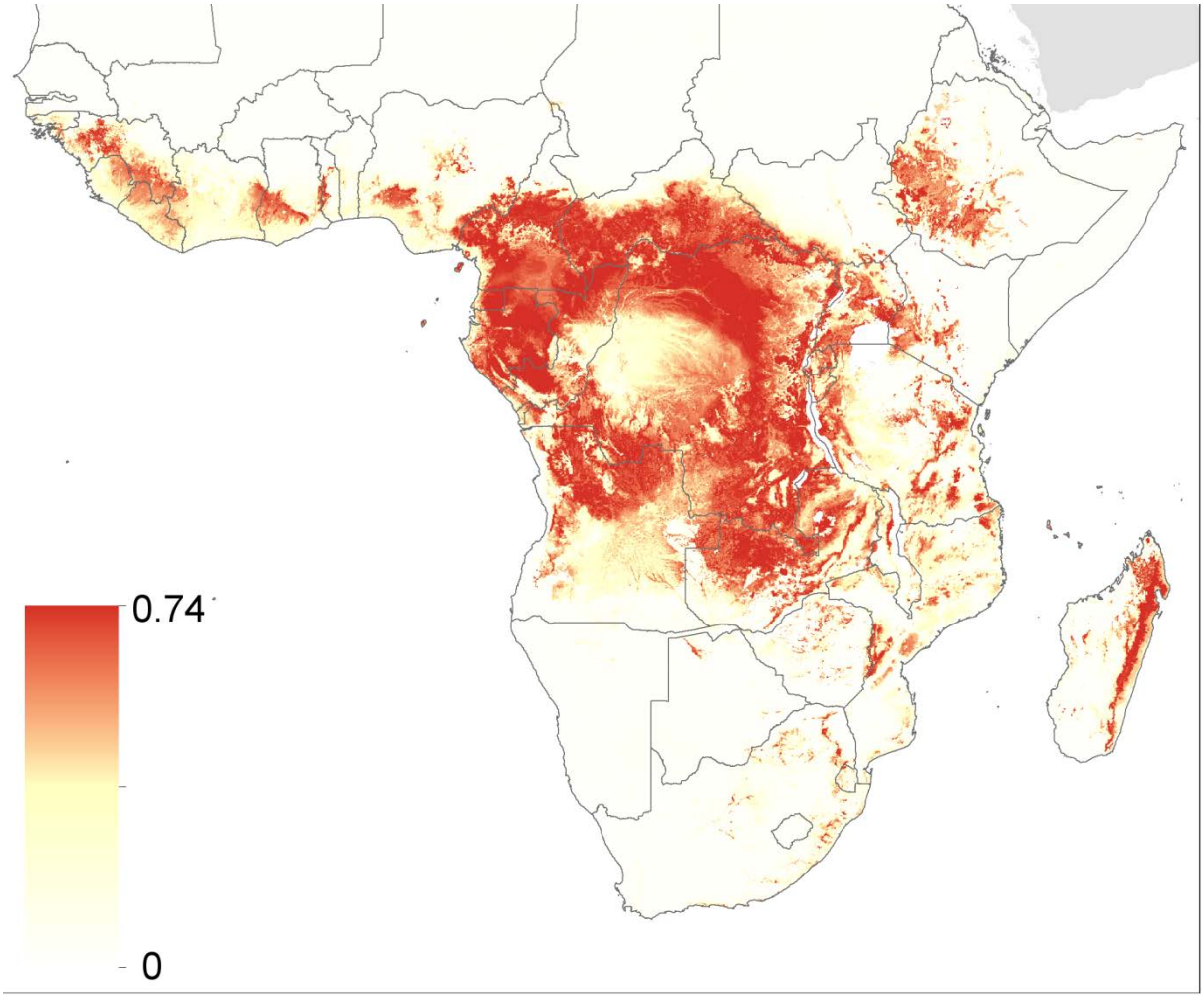
This appendix consists of two parts:

(i) Supplementary Figure 1: Prediction range of model 1: human cases only.

(ii) Supplementary Figure 2: Prediction range of model 2: both human index cases and infections in animals.



Supplementary Figure 1. Prediction range of model 1: human cases only. The difference between the 5 and 95% confidence interval of predicted values was calculated. Areas in red have the greatest range in prediction values whilst areas in white, the smallest. The maximum range of pixel values is 0.76.



Supplementary Figure 2. Prediction range of model 2: both human index cases and infections in animals. The difference between the 5 and 95% confidence interval of predicted values was calculated. Areas in red have the greatest range in prediction values whilst areas in white, the smallest. The maximum range of pixel values is 0.74.

A.5. Appendix to Chapter 6

This appendix consists of three parts:

(i) Supplementary information 1: Plots of all covariates entered into the boosted regression tree (BRT).

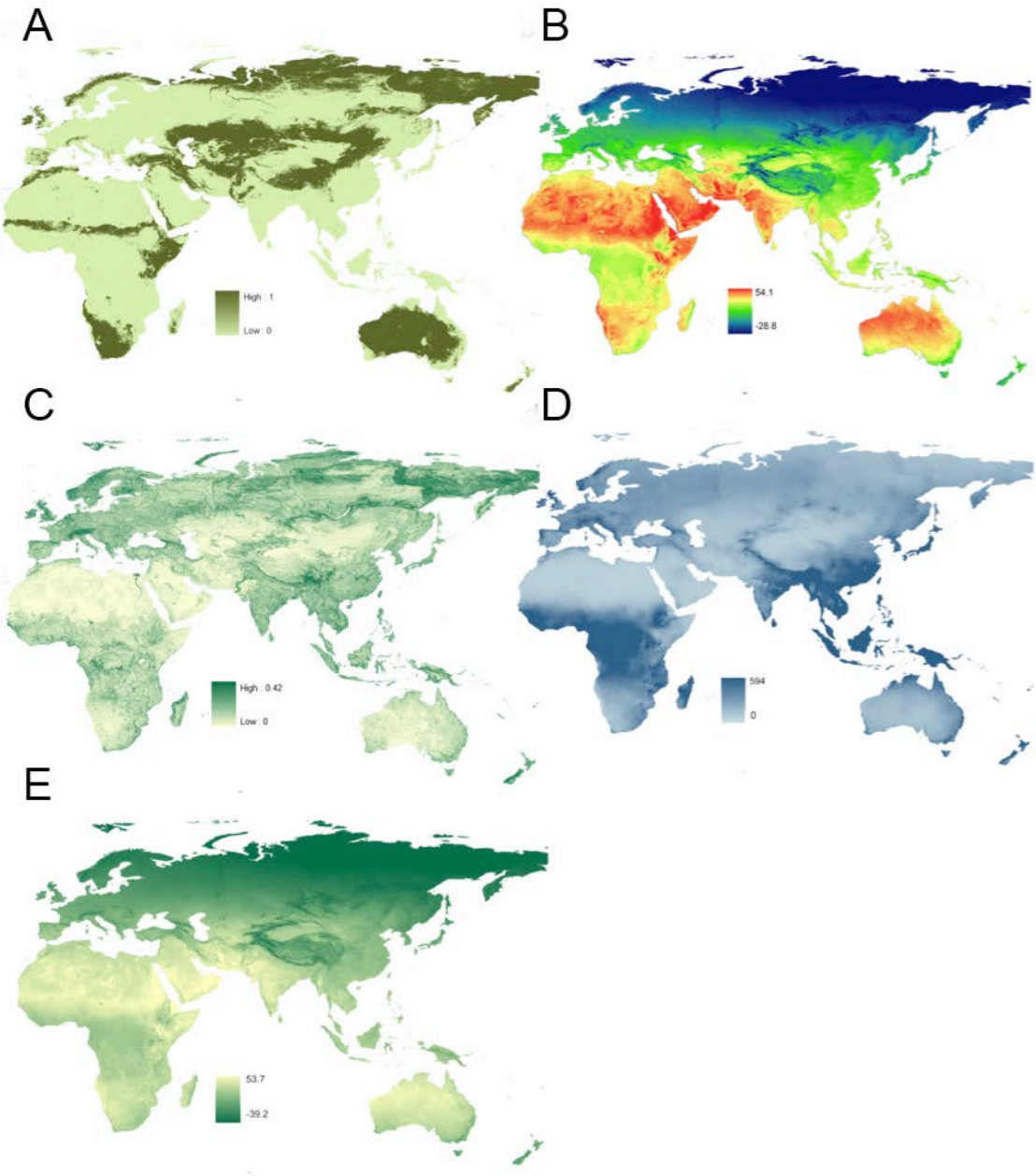
(ii) Supplementary information 2: Effect plots for BRT covariates.

(iii) Supplementary information 3: Evidence consensus scoring for Crimean-Congo haemorrhagic fever – due to size constraints this will be made available upon request

Supplementary Information 1

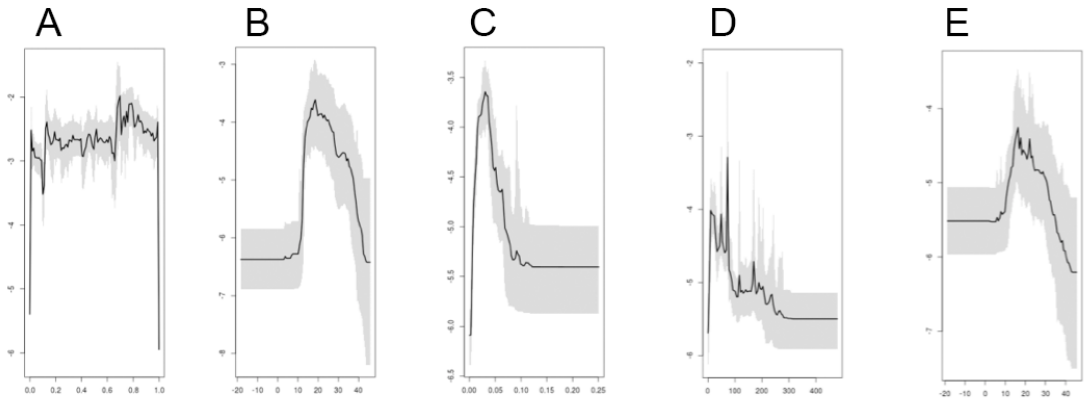
Figure 1. Plots of all covariates entered into the boosted regression tree (BRT).

(a) % grass or shrub land cover; (b) mean annual land surface temperature; (c) standard deviation of mean annual EVI; (d) mean annual precipitation (mm); (e) mean annual EVI



Supplementary Information 2

Figure 1. Effect plots for BRT covariates.



(a) % grass or shrub land cover; (b) mean annual land surface temperature; (c) standard deviation of mean annual EVI; (d) mean annual precipitation (mm); (e) mean annual EVI

A.6. Appendix to Chapter 7

This appendix consists of four parts:

(i) File S1: Evidence consensus.

(ii) File S2: Model outputs.

(iii) File S3: Partial dependency plots.

(iv) File S4: Population at risk.

Supplementary Information for Mylne *et al.* (2015) "Mapping the zoonotic niche of Lassa fever in Africa" – File S1 Evidence consensus

Methodology

Evidence consensus is designed to assess a variety of information sources to evaluate the likely presence or absence of a disease in a specific country.[1] Whilst highlighting countries where conflicting information or data gaps and uncertainty cloud the ability for a definitive presence/absence assessment is important in its own right, these surfaces can also be incorporated into models of disease risk as a way of informing background data selection and masking out areas which are environmentally suitable for transmission of disease, but where the disease is absent for other reasons such as human control or biogeographic features.[2-4]. Four main data sources were used to assess the variety of information associated with Lassa fever across the African continent. They included: health organisation status; peer-reviewed literature evidence of infection in humans; case data looking at outbreak sizes; and animal infection information.

Health organisation status (-3 to +3)

Three bodies that assess national status of endemicity for Lassa fever were included: the World Health Organization (WHO);[5] the Centers for Disease Control (CDC);[6] and the Global Infectious Diseases and Epidemiology Online Network (GIDEON).[7] For each organisation, +1 was scored if a country was indicated as disease endemic or had the disease present and -1 was scored if Lassa fever was reported as absent.

Reports of human infection (+2 to +6)

A search for articles on Lassa fever was performed using PubMed, Web of Science and Scopus. Within these articles, cases of Lassa fever in humans were identified for as many African countries as possible. Each case was evaluated for contemporariness (+3 2007-2014, +2 1999-2006, +1 1998 and earlier) and diagnostic accuracy (+3 for PCR or genetic isolation, +2 for serological based detection, +1 for reported cases without diagnostic support). The highest scoring report was included in the final table.

Case data for outbreaks (-6 to +6)

Outbreaks were identified and assessed by size and contemporariness as outlined below:

Case numbers	Date	Score
20+	2007-2014	+6
20+	1999-2006	+5
20+	1998 and earlier	+4
5-19	2007-2014	+3
5-19	1999-2006	+2

5-19	1998 and earlier	+1

Where no outbreak was identified, we assessed healthcare expenditure and adjacency to countries with reported cases to quantify the likelihood of genuine absence versus underreporting. Information on healthcare expenditure was derived from the WHO World Health Statistics [8] listing of per capita total expenditure on health at the average exchange rate in USD for the year 2011. Healthcare expenditure was categorised into three classes: Healthcare Expenditure (HE) Low (<$150), HE Medium ($150<x<$500), HE High (>$500). Adjacency was considered where countries shared a common border. Adjacency was categorised into four classes: (i) no neighbours reported cases of Lassa fever; (ii) one neighbour reported cases (unless the country only bordered one or two countries); (iii) two neighbours reported cases (unless the country only bordered two countries) or half of the bordering countries had cases; and (iv) three neighbours reported cases or all neighbouring countries had cases. The complication in categorisation was necessary in order to avoid unfairly penalising those countries with fewer shared borders. Healthcare expenditure and adjacency were then considered within the context of each other and scores assigned as follows:

Healthcare class	No neighbours	One neighbour	Two neighbours/half of all neighbours	Three or more neighbours/all neighbours
HE Low	-2	0	+3	+6
HE Medium	-4	0	+2	+5
HE High	-6	0	+1	+4

Animal infection (+1)

In addition to collecting data on human infections, animal infection reports were also included. If there was evidence of infection in the rodent host (the Natal multimammate mouse, *Mastomys natalensis*) +1 was added as a supplementary score.

Final evaluation

All the scores were added together to generate the final evidence consensus value. The denominator (maximum possible score) varied depending on which criteria were included. If no peer-review literature cases could be found the denominator was 9 (3 from Health Reporting Organisation status and 6 from case data/healthcare expenditure status). If peer-reviewed reports were present, the denominator was 15 (3 from Health Reporting Organisation status, 6 from the peer-review literature cases and 6 from case data/healthcare expenditure status). Where a country had evidence of rodent infection, the denominator was increased by 1 (*i.e.* /10 or /16). The final score was then reported as a percentage. The full evidence consensus table is outlined below.

Table S1.1: Evidence consensus

Country	GIDEON	WHO	CDC	Peer Review	Case Data	Animal	Score
Algeria	Not endemic	Absent	Absent -3	-	Neighbours Mali HE Medium ($233) 0	-	-3+0/9 -33%
Angola	Not endemic	Absent	Absent -3	-	No neighbours HE Medium ($178) -4	-	-3-4/9 -78%
Benin	Not endemic	Present	Present +1	2014 Serology [9] +5	15 cases in 2014 [9] +3	-	1+5+3/15 +60%
Botswana	Not endemic	Absent	Absent -3	-	No neighbours HE Medium ($404) -4	-	-3-4/9 -78%
Burkina Faso	Endemic	Absent	Present +1	1975 Serology [10] +3	Neighbours Ivory Coast, Ghana, Benin, and Mali HE Low ($39) +6	-	1+3+6/15 +67%
Burundi	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($21) -2	-	-3-2/9 -56%
Cameroon	Not endemic	Absent	Absent -3	-	Neighbours Nigeria HE Low ($64) 0	Serological evidence in 1989 [11] +1	-3+0+1/10 -20%
Cape Verde	Not endemic	Absent	Absent -3	-	No neighbours HE Medium ($153) -4	-	-3-4/9 -78%
Central	Endemic	Absent	Absent	-	No	-	-1-2/9

African Republic			-1		neighbours HE Low ($19) -2		-33%
Chad	Not endemic	Absent	Absent -3	-	Neighbours Nigeria HE Low ($25) 0	-	-3+0/9 -33%
Comoros	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($31) -2	-	-3-2/9 -56%
Congo	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($85) -2	-	-3-2/9 -56%
Cote d'Ivoire	Endemic	Absent	Present +1	1975 Serology [10] +3	All neighbours have reported Lassa cases HE Low ($84) +6	-	1+3+6/15 +67%
Democratic Republic of the Congo	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($15) -2	-	-3-2/9 -56%
Djibouti	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($119) -2	-	-3-2/9 -56%
Egypt	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($137) -2	-	-3-2/9 -56%
Equatorial Guinea	Not endemic	Absent	Absent -3	-	No neighbours HE High ($1051) -6	-	-3-6/9 -100%
Eritrea	Not endemic	Absent	Absent	-	No neighbours	-	-3-2/9

			-3		HE Low ($12) -2		-56%
Ethiopia	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($14) -2	-	-3-2/9 -56%
Gabon	Endemic	Absent	Absent -1	-	No neighbours HE Medium ($401) -4	-	-1-4/9 -56%
Gambia	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($24) -2	-	-3-2/9 -56%
Ghana	Endemic	Absent	Present +1	2012 PCR [12] +6	Neighbours Ivory Coast and Burkina Faso HE Low ($83) +3	-	1+6+3/15 +67%
Guinea	Endemic	Present	Present +3	2013 PCR [13] +6	22 cases between 1996 and 1999 [14] +2	Multiple infections reported [15, 16] +1	3+6+2+1/16 +75%
Guinea-Bissau	Not endemic	Absent	Absent -3	-	Neighbours Guinea HE Low ($35) +3	-	-3+3/9 0%
Kenya	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($35) -2	-	-3-2/9 -56%
Lesotho	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($35) -2	-	-3-2/9 -56%
Liberia	Endemic	Present	Present +3	2010 PCR [17]	21 cases in 2007 [18-20]	-	3+6+6/15 +100%

				+6	+6		
Libya	Not endemic	Absent	Absent -3	-	No neighbours HE Medium ($211) -4	-	-3-4/9 -78%
Madagascar	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($19) -2	-	-3-2/9 -56%
Malawi	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($30) -2	-	-3-2/9 -56%
Mali	Endemic	Absent	Present +1	1975 Serology [10] +3	Neighbours Guinea, Ivory Coast and Burkina Faso HE Low ($51) +6	PCR isolation in 2013 [21] +1	1+3+6+1/16 +69%
Mauritania	Not endemic	Absent	Absent -3	-	Neighbours Mali HE Low ($51) 0	-	-3+0/9 -33%
Mauritius	Not endemic	Absent	Absent -3	-	No neighbours HE Medium ($450) -4	-	-3-4/9 -78%
Mayotte	Not endemic	Absent	Absent -3	-	No neighbours Assumed HE High -6	-	-3-6/9 -100%
Morocco	Not endemic	Absent	Absent -3	-	No neighbours HE Medium ($195) -4	-	-3-4/9 -78%
Mozambique	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($33)	-	-3-2/9 -56%

					-2		
Namibia	Not endemic	Absent	Absent -3	-	No neighbours HE Medium ($486) -4	-	-3-4/9 -78%
Niger	Not endemic	Absent	Absent -3	-	Neighbours Nigeria, Benin, Burkina Faso and Mali HE Low ($25) +6	-	-3+6/9 +33%
Nigeria	Endemic	Present	Present +3	2013 PCR [22, 23] +6	208 cases in 2014 [24] +6	Serological evidence (2005) [25] and viral isolation (1975) [26] +1	3+6+6+1/16 +100%
Réunion	Not endemic	Absent	Absent -3	-	No neighbours Assumed HE High -6	-	-3-6/9 -100%
Rwanda	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($62) -2	-	-3-2/9 -56%
Sao Tome and Principe	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($108) -2	-	-3-2/9 -56%
Senegal	Endemic	Absent	Absent -1	-	Neighbours Guinea and Mali HE Low ($54) +3	-	-1+3/9 +22%
Seychelles	Not endemic	Absent	Absent -3	-	No neighbours HE Medium	-	-3-4/9 -78%

					($413) -4		
Sierra Leone	Endemic	Present	Present +3	2011 PCR [27] +6	153 cases in 2010 [28] +6	Multiple infections reported [29, 30] +1	3+6+6+1/16 +100%
Somalia	Not endemic	Absent	Absent -3	-	No neighbours Assumed HE Low -2	-	-3-2/9 -56%
South Africa	Not endemic	Absent	Absent -3	-	No neighbours HE High ($670) -6	-	-3-6/9 -100%
South Sudan	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($32) -2	-	-3-2/9 -56%
Sudan	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($119) -2	-	-3-2/9 -56%
Swaziland	Not endemic	Absent	Absent -3	-	No neighbours HE Medium ($270) -4	-	-3-4/9 -78%
Togo	Not endemic	Absent	Present -1	-	All neighbours have reported cases HE Low ($43) +6	-	-1+6/9 +56%
Tunisia	Not endemic	Absent	Absent -3	-	No neighbours HE Medium ($304) -4	-	-3-4/9 -77%

Uganda	Endemic	Absent	Absent -1	-	No neighbours HE Low ($41) -2	-	-1-2/9 -33%
United Republic of Tanzania	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($38) -2	-	-3-2/9 -56%
Western Sahara	Not endemic	Absent	Absent -3	-	No neighbours Assumed HE Low -2	-	-3-2/9 -56%
Zambia	Not endemic	Absent	Absent -3	-	No neighbours HE Low ($87) -2	-	-3-2/9 -56%
Zimbabwe	Not endemic	Absent	Absent -3	-	No neighbours Assumed HE Low -2	-	-3-2/9 -56%

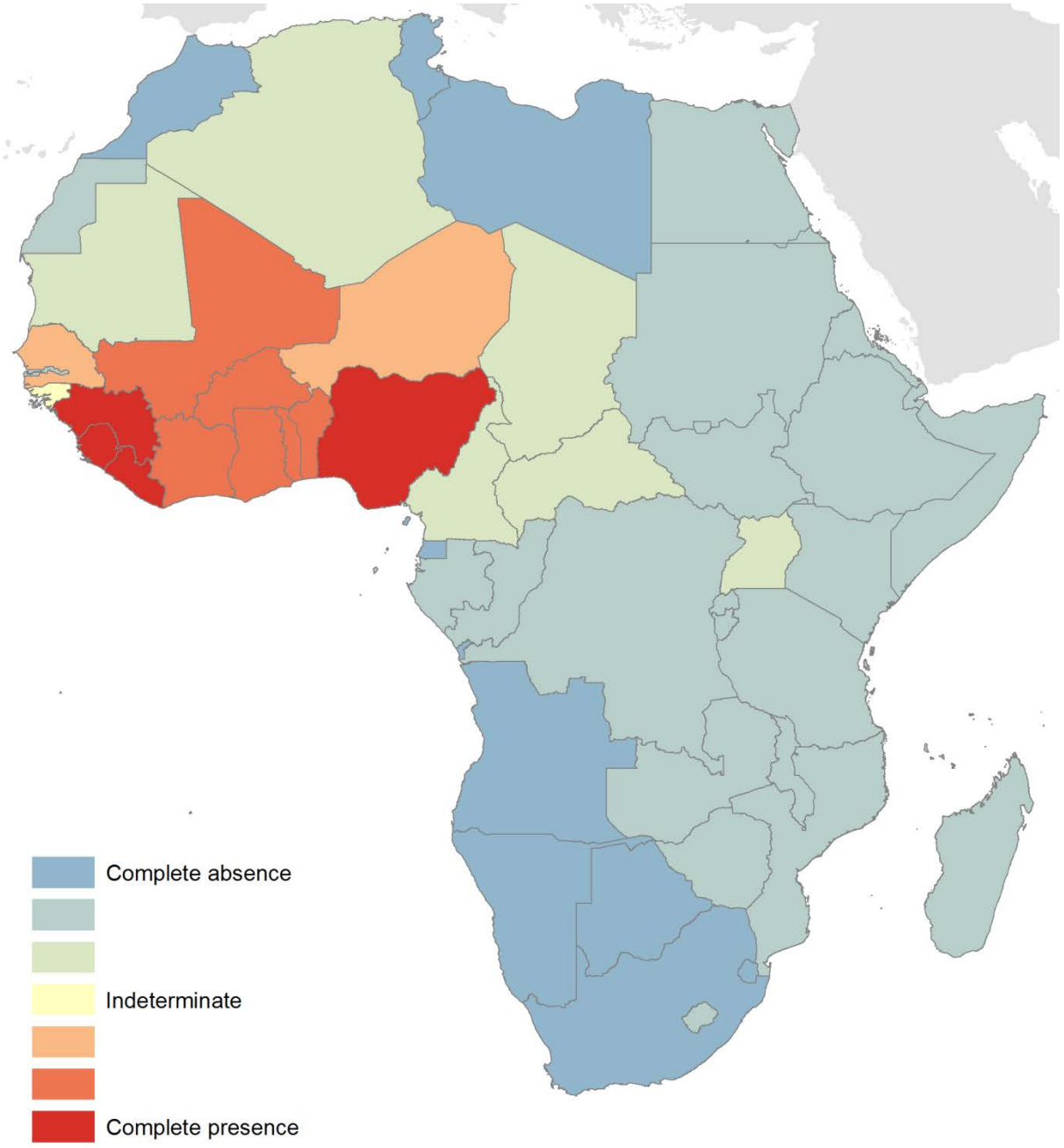


Figure S1.1: Map of Africa showing evidence consensus scores for Lassa fever presence/absence

Areas in red indicate consensus on disease presence whilst areas in blue have consensus on disease absence.

References

1. Brady OJ, Gething PW, Bhatt S, et al. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLoS Negl Trop Dis* 2012;6:e1760.
2. Pigott DM, Bhatt S, Golding N, et al. Global distribution maps of the leishmaniases. *Elife* 2014;3:e02851.
3. Bhatt S, Gething PW, Brady OJ, et al. The global distribution and burden of dengue. *Nature* 2013;496:504-7.
4. Cano J, Rebollo MP, Golding N, et al. The global distribution and transmission limits of lymphatic filariasis: past and present. *Parasit Vectors* 2014;7:466.
5. World Health Organization. *Lassa fever - fact sheet #179*. http://www.who.int/mediacentre/factsheets/fs179/en/ [accessed 25th March 2015].
6. CDC. *Lassa fever*. http://www.cdc.gov/vhf/lassa/ [accessed 25th March 2015].
7. Edberg SC. Global Infectious Diseases and Epidemiology Network (GIDEON): a world wide Web-based program for diagnosis and informatics in infectious diseases. *Clin Infect Dis* 2005;40:123-6.
8. WHO. *World Health Statistics 2014*. Geneva: World Health Organization, 2014.
9. ProMED-mail. *Lassa fever - Benin (02) 20141126.2992727*. www.promedmail.org [accessed 24 March 2015].
10. Frame JD. Surveillance of Lassa fever in missionaries stationed in West Africa. *Bull World Health Organ* 1975;52:593-8.
11. Gonzalez JP, Josse R, Johnson ED, et al. Antibody prevalence against haemorrhagic fever viruses in randomized representative Central African populations. *Res Virol* 1989;140:319-31.
12. Dzotsi EK, Ohene SA, Asiedu-Bekoe F, et al. The first cases of Lassa fever in Ghana. *Ghana Med J* 2012;46:166-70.
13. Klempa B, Koulemou K, Auste B, et al. Seroepidemiological study reveals regional co-occurrence of Lassa- and Hantavirus antibodies in Upper Guinea, West Africa. *Trop Med Int Health* 2013;18:366-71.
14. Bausch DG, Demby AH, Coulibaly M, et al. Lassa fever in Guinea: I. Epidemiology of human disease and clinical observations. *Vector Borne Zoonotic Dis* 2001;1:269-81.
15. Lalis A, Leblois R, Lecompte E, et al. The impact of human conflict on the genetics of *Mastomys natalensis* and Lassa virus in West Africa. *PLoS One* 2012;7:e37068.
16. Fichet-Calvet E, Lecompte E, Koivogui L, et al. Fluctuation of abundance and Lassa virus prevalence in *Mastomys natalensis* in Guinea, West Africa. *Vector Borne Zoonotic Dis* 2007;7:119-28.
17. Amorosa V, MacNeil A, McConnell R, et al. Imported Lassa fever, Pennsylvania, USA, 2010. *Emerg Infect Dis* 2010;16:1598-600.
18. ProMED-mail. *Lassa fever - Liberia: RFI 20070410.1210*. www.promedmail.org [accessed 24 March 2015].

19. ProMED-mail. *Lassa fever - Liberia (02): confirmed 20070413.1235*. www.promedmail.org [accessed 24 March 2015].
20. ProMED-mail. *Lassa fever - Liberia (03) 20070430.1406*. www.promedmail.org [accessed 24 March 2015].
21. Safronetz D, Sogoba N, Lopez JE, et al. Geographic distribution and genetic characterization of Lassa virus in sub-Saharan Mali. *PLoS Negl Trop Dis* 2013;7:e2582.
22. Dongo AE, Kesieme EB, Iyamu CE, et al. Lassa fever presenting as acute abdomen: a case series. *Virol J* 2013;10:123.
23. Ajayi NA, Nwigwe CG, Azuogu BN, et al. Containing a Lassa fever epidemic in a resource-limited setting: outbreak description and lessons learned from Abakaliki, Nigeria (January-March 2012). *Int J Infect Dis* 2013;17:e1011-6.
24. ProMED-mail. *Lassa fever - Nigeria, Liberia 20140328.2363217*. www.promedmail.org [accessed 24 March 2015].
25. Okoror LE, Esumeh FI, Agbonlahor DE, et al. Lassa virus: seroepidemiological survey of rodents caught in Ekpoma and environs. *Trop Doct* 2005;35:16-7.
26. Wulff H, Fabiyi A, Monath TP. Recent isolations of Lassa virus from Nigerian rodents. *Bull World Health Organ* 1975;52:609-13.
27. Branco LM, Boisen ML, Andersen KG, et al. Lassa hemorrhagic fever in a late term pregnancy from northern Sierra Leone with a positive maternal outcome: case report. *Virol J* 2011;8:404.
28. ProMED-mail. *Lassa fever - Sierra Leone (04): (NO) 20101111.4101*. www.promedmail.org [accessed 24 March 2015].
29. Leski TA, Stockelman MG, Moses LM, et al. Sequence variability and geographic distribution of lassa virus, Sierra Leone. *Emerg Infect Dis* 2015;21:609-18.
30. McCormick JB, Webb PA, Krebs JW, et al. A prospective study of the epidemiology and ecology of Lassa fever. *J Infect Dis* 1987;155:437-44.

Supplementary Information for Mylne *et al.* (2015) "Mapping the zoonotic niche of Lassa fever in Africa" – File S2 Model outputs

Figure S2.1: Predicted geographical distribution of the zoonotic niche for Lassa virus using a 1:1 ratio diagnostic weighting schema for human or animal infections diagnosed via PCR/viral isolation and serological tests, respectively (Model 1)

The scale reflects the environmental suitability for zoonotic transmission of Lassa virus. Areas closer to 1 (red) are more suitable than those closer to 0 (blue). The area under the curve statistic, calculated under a stringent cross-validation procedure is 0.79±0.02.

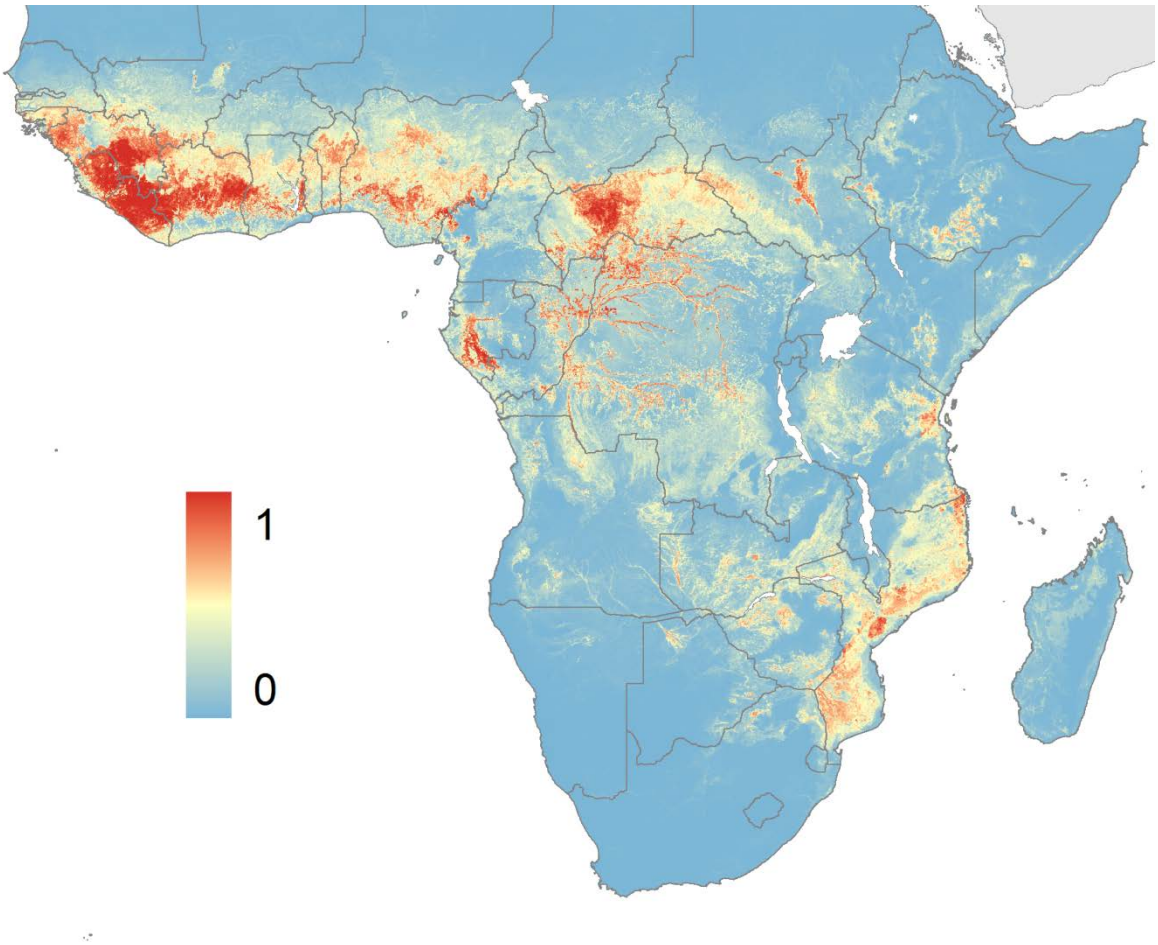


Figure S2.2: Predicted geographical distribution of the zoonotic niche for Lassa virus using a 2:1 ratio diagnostic weighting schema for human or animal infections diagnosed via PCR/viral isolation and serological tests, respectively (Model 2)

The scale reflects the environmental suitability for zoonotic transmission of Lassa virus. Areas closer to 1 (red) are more suitable than those closer to 0 (blue). The area under the curve statistic, calculated under a stringent cross-validation procedure is 0.79±0.02.

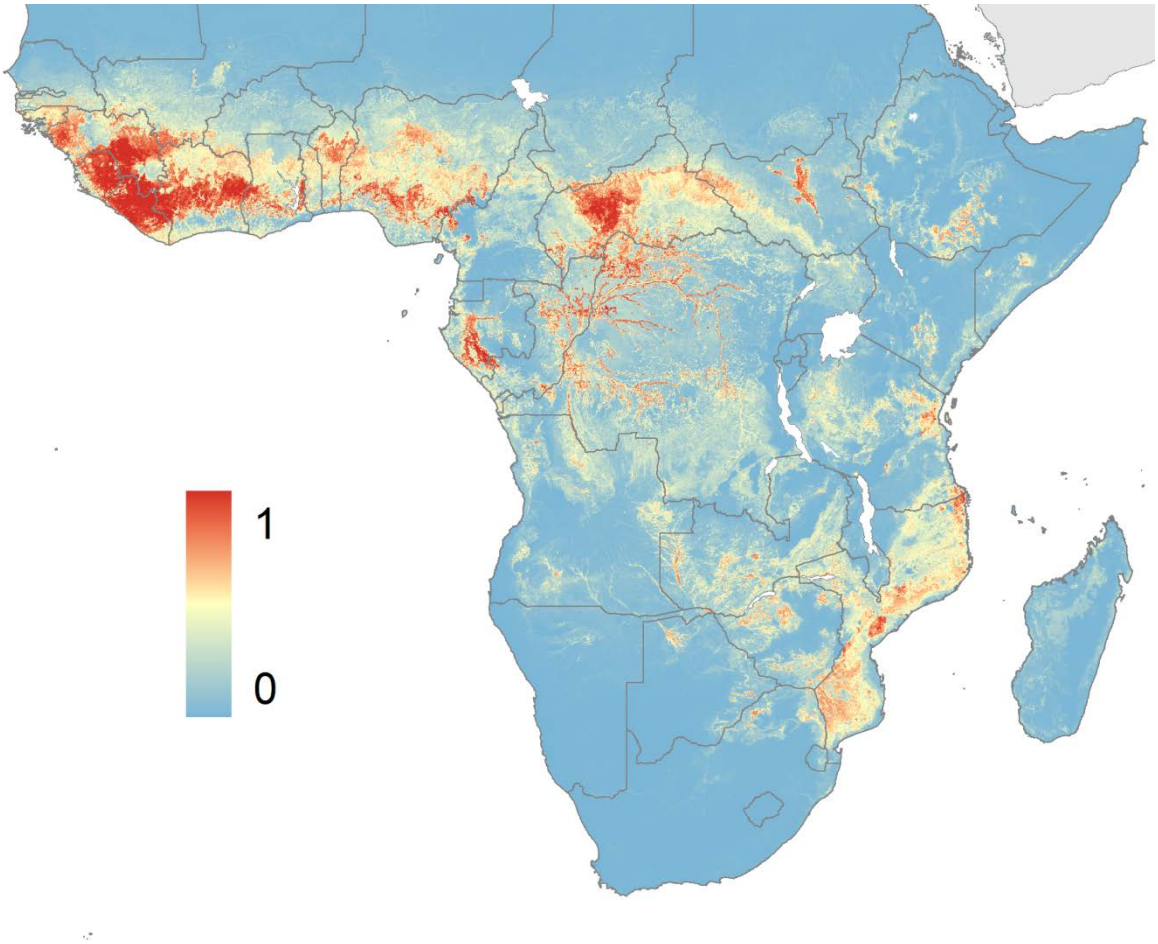


Figure S2.3: Predicted geographical distribution of the zoonotic niche for Lassa virus using a 4:1 ratio diagnostic weighting schema for human or animal infections diagnosed via PCR/viral isolation and serological tests, respectively (Model 3)

The scale reflects the environmental suitability for zoonotic transmission of Lassa virus. Areas closer to 1 (red) are more suitable than those closer to 0 (blue). The area under the curve statistic, calculated under a stringent cross-validation procedure is 0.78±0.02.

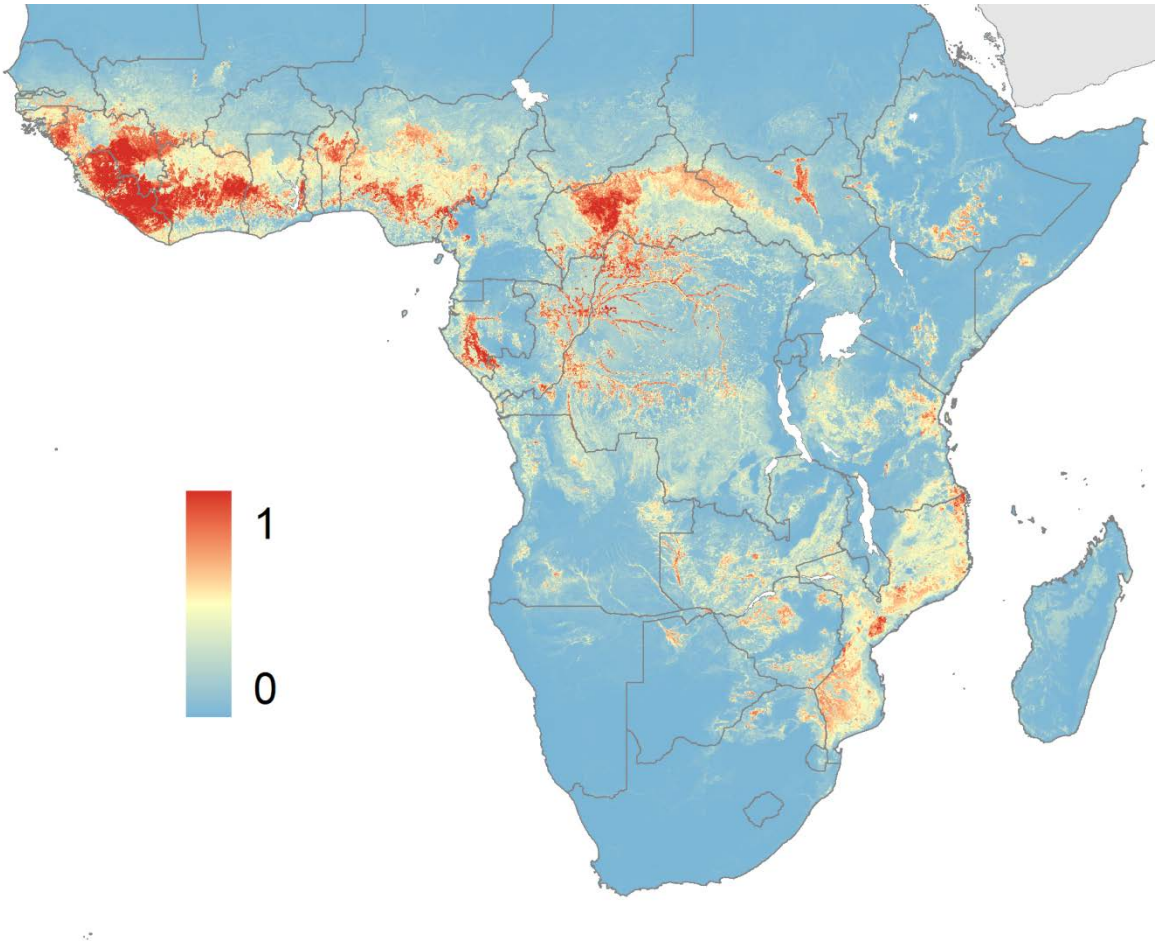


Figure S2.4: Predicted geographical distribution of the zoonotic niche for Lassa virus for human or animal infections diagnosed via PCR/viral isolation tests (Model 4)

The scale reflects the environmental suitability for zoonotic transmission of Lassa virus. Areas closer to 1 (red) are more suitable than those closer to 0 (blue). The area under the curve statistic, calculated under a stringent cross-validation procedure is 0.73±0.04.

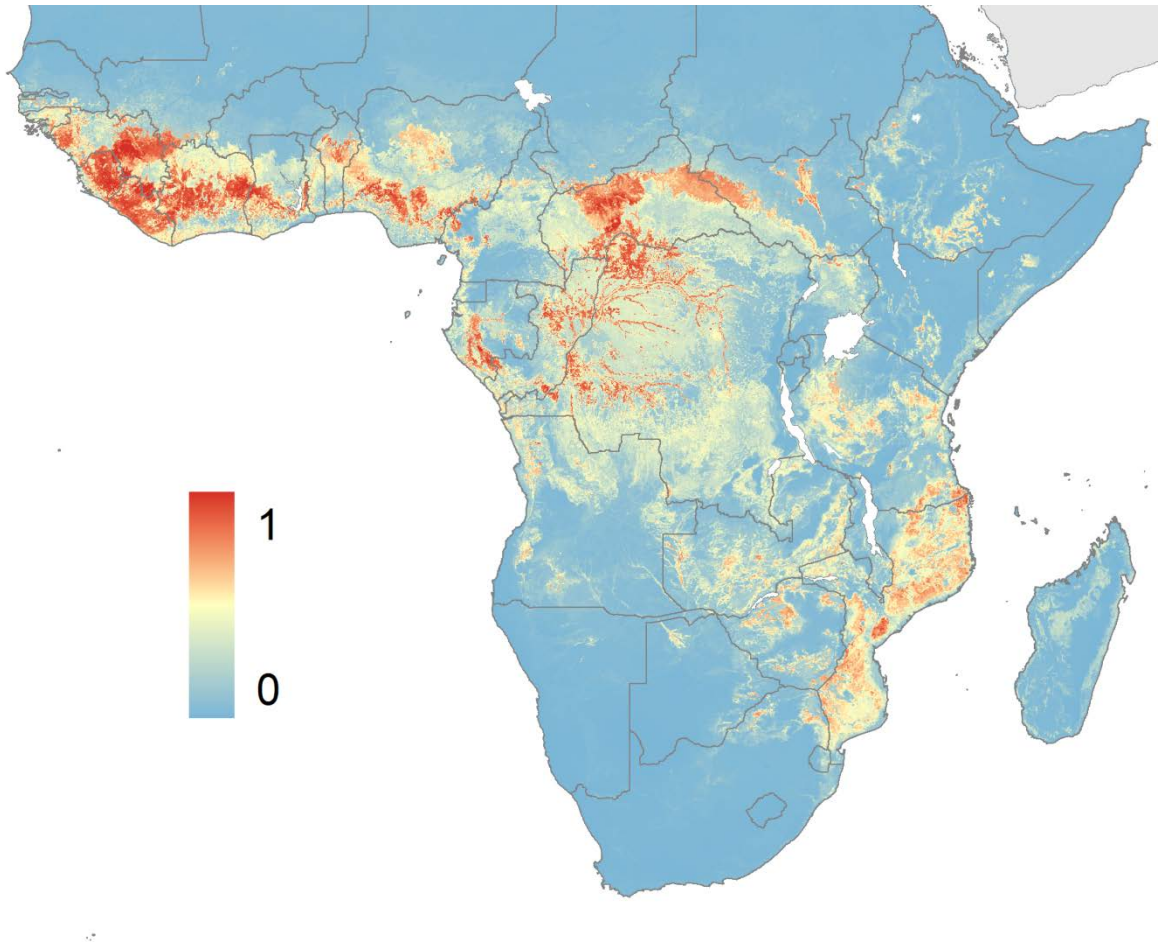


Figure S2.5: Predicted geographical distribution of the zoonotic niche for Lassa virus using a 1:1 ratio diagnostic weighting schema for human infections diagnosed via PCR/viral isolation and serological tests, respectively (Model 5)

The scale reflects the environmental suitability for zoonotic transmission of Lassa virus. Areas closer to 1 (red) are more suitable than those closer to 0 (blue). The area under the curve statistic, calculated under a stringent cross-validation procedure is 0.77±0.02.

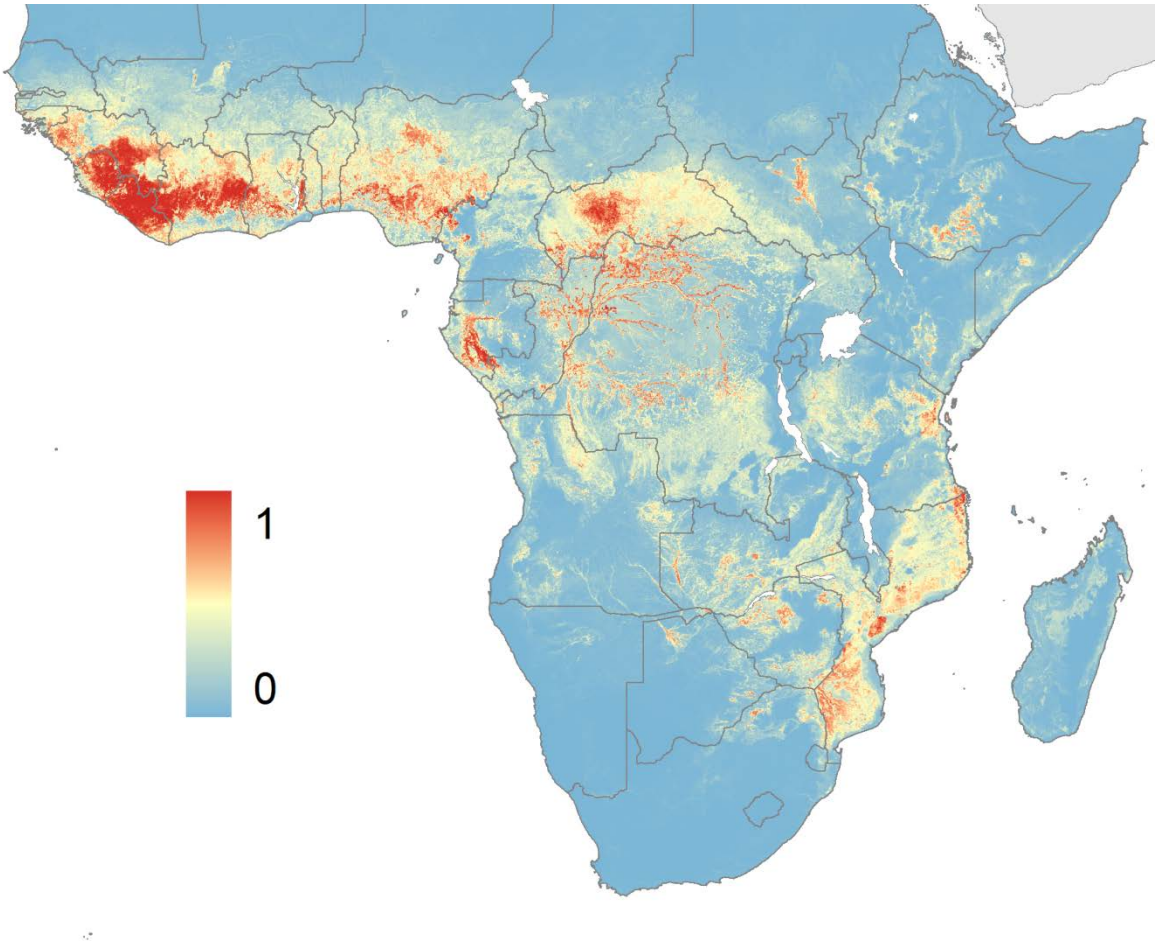


Figure S2.6: Model 1 prediction range

The difference between the 5% and 95% confidence interval of predicted values was calculated. Areas in red have the greatest range in prediction values whilst areas in white, the smallest. The maximum range of pixel values is 0.800275.

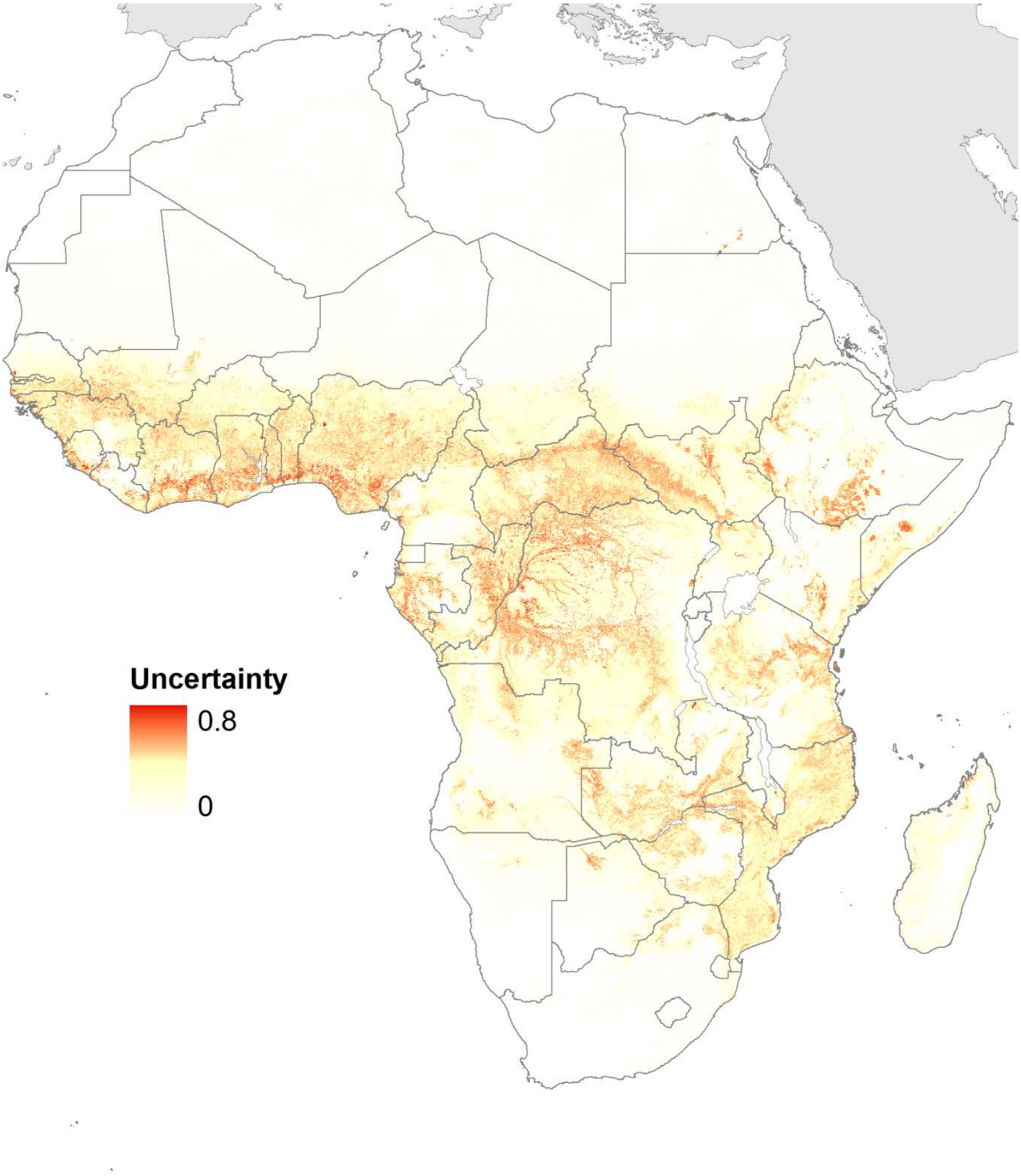


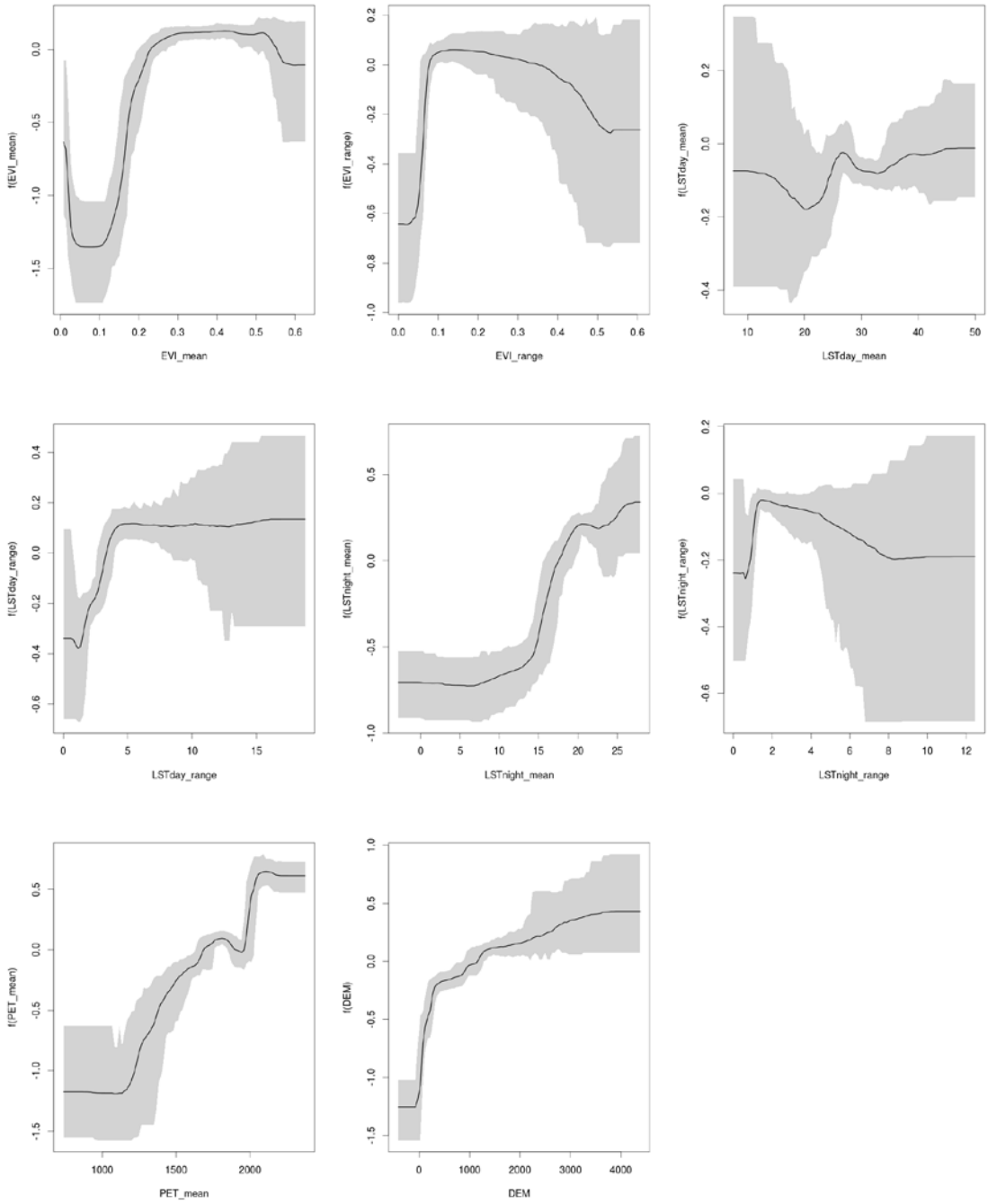
Table S2.1: Summary statistics for model outputs. Relative contributions for each of the top five predictors are reported as a percentage

Statistic	Model 1	Model 2	Model 3	Model 4	Model 5
AUC ± s.d	0.79±0.02	0.79±0.02	0.78±0.02	0.73±0.04	0.77±0.02
1st predictor	Mean EVI: 26.5%	Mean EVI: 25.6%	Mean EVI: 25.0%	Mean EVI: 24.5%	Night-time mean LST: 25.3%
2nd predictor	Night-time mean LST: 19.2%	Night-time mean LST: 17.9%	Night-time mean LST: 17.2%	Night-time mean LST: 18.1%	Day-time mean LST: 19.5%
3rd predictor	Predicted host distribution: 13.6%	Elevation (DEM): 14.3%	Elevation (DEM): 14.7%	Day-time mean LST: 15.8%	Mean EVI: 15.1%
4th predictor	Elevation (DEM): 11.7%	Predicted host distribution: 12.1%	Predicted host distribution: 12.5%	Mean PET: 12.2%	Predicted host distribution: 11.7%
5th predictor	Mean PET: 10.6%	Mean PET: 11.0%	Day-time mean LST: 11.3%	Elevation (DEM): 10.6%	Elevation (DEM): 9.9%

AUC: area under the curve; host: the Natal multimammate mouse, *Mastomys natalensis*; s.d: standard deviation

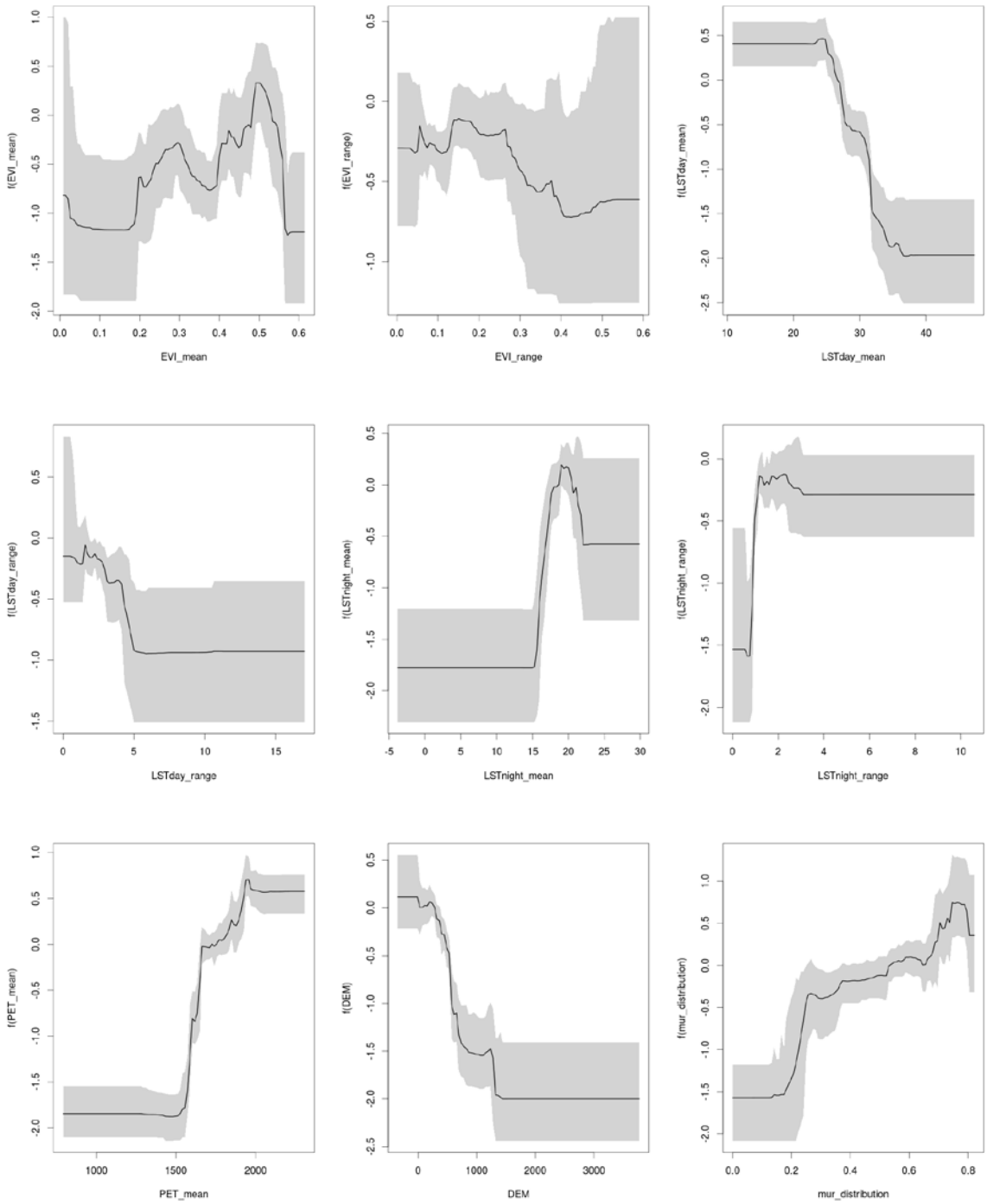
Supplementary Information for Mylne *et al.* (2015) "Mapping the zoonotic niche of Lassa fever in Africa" – File S3 Partial dependency plots

Figure S3.1: Partial dependency plots for the predicted geographical distribution of the Natal multimammate mouse, *Mastomys natalensis*



EVI_mean: enhanced vegetation index mean; EVI_range: enhanced vegetation index range; LSTday_mean: Day time land surface temperature mean; LSTday_range: Day time land surface temperature range; LSTnight_mean: Night time land surface temperature mean; LSTnight_range: Night time land surface temperature range; PET_mean: potential evapotranspiration mean; DEM: digital elevations models

Figure S3.2: Partial dependency plots for the predicted geographical distribution of the zoonotic niche for Lassa virus



EVI_mean: enhanced vegetation index mean; EVI_range: enhanced vegetation index range; LSTday_mean: Day time land surface temperature mean; LSTday_range: Day time land surface temperature range; LSTnight_mean: Night time land surface temperature mean; LSTnight_range: Night time land surface temperature range; PET_mean: potential evapotranspiration mean; DEM: digital elevations models; mur_distribution: predicted geographical distribution of the Natal multimammate mouse, *Mastomys natalensis*

Supplementary Information for Mylne *et al.* (2015) "Mapping the zoonotic niche of Lassa fever in Africa" – File S4 Population at risk

Figure S4.1: Predicted geographical distribution of at-risk populations to Lassa virus

The continuous environmental suitability surface was converted into a binary at-risk (red), not-at-risk (blue) pixel layer. The threshold probability for a 5 km x 5 km at-risk pixel was calculated to be equal to or greater than 0.6459206. Countries with borders outlined by a solid line are those where cases of Lassa fever have previously been reported. Countries with borders outlined by a dash line have not previously reported Lassa fever cases.

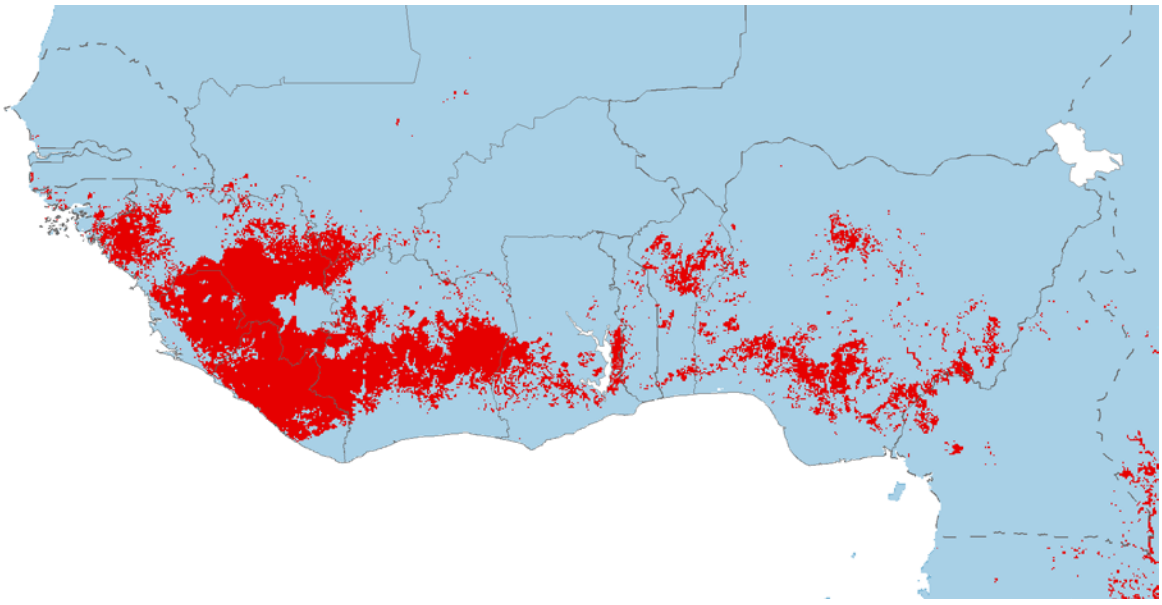


Table S4.1: Location and size of at-risk populations to Lassa virus in countries with past Lassa fever cases

Country	Population size
Benin	964,696
Burkina Faso	22,154
Côte d'Ivoire	6,885,680
Ghana	3,797,054
Guinea	4,930,742
Liberia	2,899,680
Mali	310,155
Nigeria	13,684,312
Sierra Leone	3,423,218
Total	**36,917,691**

Table S4.2: Location and size of at-risk populations to Lassa virus in countries with no past Lassa fever cases

Country	Population size
Cameroon	340,758
Guinea-Bissau	85,662
Niger	1,687
Senegal	23,169
Togo	336,291
Total	**787,567**

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

rstb.royalsocietypublishing.org

Review

Cite this article: Hay SI, Battle KE, Pigott DM, Smith DL, Moyes CL, Bhatt S, Brownstein JS, Collier N, Myers MF, George DB, Gething PW. 2013 Global mapping of infectious disease. Phil Trans R Soc B 368: 20120250. http://dx.doi.org/10.1098/rstb.2012.0250

One contribution of 18 to a Discussion Meeting Issue 'Next-generation molecular and evolutionary epidemiology of infectious disease'.

Subject Areas:
health and disease and epidemiology, bioinformatics, computational biology, ecology

Keywords:
surveillance, biosurveillance, cartography, public health, atlas, crowdsourcing

Author for correspondence:
Simon I. Hay
e-mail: simon.hay@zoo.ox.ac.uk

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2012.0250 or via http://rstb.royalsocietypublishing.org.

Royal Society Publishing
Informing the science of the future

Global mapping of infectious disease

Simon I. Hay[1,2], Katherine E. Battle[1], David M. Pigott[1], David L. Smith[2,3], Catherine L. Moyes[1], Samir Bhatt[1], John S. Brownstein[4], Nigel Collier[5], Monica F. Myers[1], Dylan B. George[2] and Peter W. Gething[1]

[1]Spatial Ecology and Epidemiology Group, Department of Zoology, University of Oxford, Oxford, UK
[2]Fogarty International Center, National Institutes of Health, Bethesda, MD, USA
[3]Department of Epidemiology and Malaria Research Institute, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[4]Department of Pediatrics, Harvard Medical School and Children's Hospital Informatics Program, Boston Children's Hospital, Boston, MA, USA
[5]National Institute of Informatics, Research Organization of Information and Systems, Tokyo, Japan

The primary aim of this review was to evaluate the state of knowledge of the geographical distribution of all infectious diseases of clinical significance to humans. A systematic review was conducted to enumerate cartographic progress, with respect to the data available for mapping and the methods currently applied. The results helped define the minimum information requirements for mapping infectious disease occurrence, and a quantitative framework for assessing the mapping opportunities for all infectious diseases. This revealed that of 355 infectious diseases identified, 174 (49%) have a strong rationale for mapping and of these only 7 (4%) had been comprehensively mapped. A variety of ambitions, such as the quantification of the global burden of infectious disease, international biosurveillance, assessing the likelihood of infectious disease outbreaks and exploring the propensity for infectious disease evolution and emergence, are limited by these omissions. An overview of the factors hindering progress in disease cartography is provided. It is argued that rapid improvement in the landscape of infectious diseases mapping can be made by embracing non-conventional data sources, automation of geo-positioning and mapping procedures enabled by machine learning and information technology, respectively, in addition to harnessing labour of the volunteer 'cognitive surplus' through crowdsourcing.

1. Introduction

The primary goal of this review is to establish the minimum set of information that is needed on the epidemiology of an infectious disease, to make an informed decision on the most appropriate techniques for mapping its global distribution. The assessment is intended to be applicable to all infectious diseases of clinical significance in humans, but makes no attempt to prioritize the case for mapping among the diseases considered.

More than 1400 species of infectious agents have been reported to cause disease in humans [1–3]. These include pathogens for some 347 diseases of sustained clinical importance, for which it is commercially viable to compile information relevant to their diagnosis, epidemiology and therapy, as a decision-support tool for clinicians [4,5]. Logistical constraints required a focus in this review on these clinically important diseases. Among these there are 110 diseases that pose a threat to non-immune travellers [4]. Sixty-two of these clinically significant diseases can be prevented by vaccination; 19 usually as routine childhood immunizations [4,6,7].

There are a variety of reasons for wanting to map the geographical distribution of an infectious disease. Mapping is a primary goal in spatial epidemiology [8–16]. Maps of disease distribution and intensity allow an immediate visualization of the extent and magnitude of the public health

problem. When based on empirical evidence, maps can support carefully weighted assessments by decision makers on the advantages and disadvantages of alternative courses of action [17–19]. These may range from helping plan national scale intervention strategies [20,21] to advice for individuals on whether to vaccinate and/or provide prophylaxis before travel [6,22]. These maps can also document a baseline from which intervention success or failure can be monitored.

In addition, as modes of data gathering evolve and improve (for example, through enhanced electronic surveillance [17] and Internet-based health reporting [23], including HealthMap/ProMED [24,25], BioCaster [26,27] and Argus [28,29]) and techniques develop to exploit these data (for example, semi-automated rapid mapping), these geographical distributions (often referred to in this literature as baseline disease risk assessments) can also provide a 'normal' against which real-time outbreak alerts can be assessed for international biosurveillance [30–32].

Furthermore, as the portfolio of infectious disease distribution maps expands and their fidelity improves, the public health community will be better able to evaluate the factors that predispose a time and place to the origin [33,34], and emergence of infectious disease outbreaks [3,35–42]. Unfortunately, contemporary inferences about the fundamental ecology of infectious diseases (such as decreased species richness [43] and increased range size [44] with latitude and their potential for spread [45,46]) are crude spatially because they rely on data not systematically collected for this purpose and aggregated to the national level [4]. Ultimately, this improved basic understanding will help mitigate the processes that drive the diversity of infectious disease threats with which we contend [47].

There is, therefore, a clear need to perform baseline risk assessments for routine public health, improve biosurveillance and provide better long-term preparedness by improving fundamental epidemiological understanding [31].

An understanding of the public health benefit of the mapping of infectious disease is not new [48–50] and selected old examples for malaria include these references [51–55]. Historical disease cartography usually suffered at least one of the following problems. First, authors very rarely documented the evidence-base that was used to make the map. Second, when mapping was implemented before the advent of geographical information systems, significant errors arose simply as a function of cartographic skill. These errors were magnified enormously when working at global scales. Third, no assessment of the fidelity of the map or how this precision might vary spatially across the map extent was ever given. These limitations constrained significantly the public health utility of the maps and are to a greater or lesser extent resolved in many of the contemporary mapping efforts reviewed here.

Today, there are a range of different geographical distributions or baseline 'risk' maps available [56], which have been derived for a variety of purposes, by a wide community of public health cartographers using a diverse toolbox of mapping methods [8–16]. Moreover, the maps use a variety of disease-related metrics (occurrence, incidence, prevalence), and an even wider array of covariates to inform the predictions [8,57,58]. This complexity means that global comparisons between maps of different diseases are extremely difficult and wider synthesis remains elusive. In part, this review aims to help audit and navigate this diversity and the supplemental information provides an extensive bibliography arising from a systematic review of all diseases of clinical significance [4].

In this review, we also consider the minimum information requirements for disease mapping. When considering cartographic options for diseases of clinical importance, the first question is: do we know the life cycle of the pathogen, its vectors, reservoirs, hosts and routes of transmission? This sounds trivial, but for many pathogens there is still considerable uncertainty around the life history. Second, do we have information about the spatial and temporal patterns of the disease? Third, do we understand the dynamic processes of transmission that determine the patterns we observe in space and time? This level of detail will usually indicate some intimate epidemiological knowledge of covariates (temperature, rainfall, land use patterns, etc.), that can help in understanding the spatial and temporal distribution of a disease. Progression along this gradient of questions reflects increased basic epidemiological understanding and, therefore, an increased ability to map the disease. Fourth, it is important to know what quantity and quality of data are available for mapping. It is self-evident that more high quality contemporary data leads to more robust maps. Many obstacles exist that can make the relevant data scarce, however. For example, health-related data may be closely protected by governments and other institutions or these data may simply be scattered so widely in the formal literature that their systematic assembly is a significant logistical challenge. Fifth, it is also important to know whether previous credible mapping efforts have been conducted. This will help answer questions one through four and, broadly speaking, the longer the history of robust mapping activities, the increased likelihood of reliable mapping outcomes.

The ability to map a disease stems largely from the type of data that are available for mapping [10,15]. The accuracy of maps is then largely determined by the abundance, spatial representativeness and heterogeneity of those data [59]. Point data types used in disease mapping are generally georeferenced occurrence or prevalence records. Occurrence data simply record an observation of a disease at a given location and time, and are characteristic of the data provided routinely by HealthMap/ProMED [24,25], BioCaster [26,27] and Argus [28,29]. The other commonly recorded point data are infection prevalence surveys, which not only locate a disease in time and space, but also measure the infected fraction of the sampled local population and thus, enable the standard quantification of the 'abundance' of a disease. This is often referred to as its endemicity [60]. An accurate global representation of the contemporary endemicity of a disease is a key achievement for infectious disease mapping, because it affords a rich diversity of operationally important public health inferences: for example, clinical burden [61,62] and basic reproductive number estimation [18,63] to inform national elimination feasibility assessment [20,64].

A wide range of approaches have been developed for empirical modelling of species and disease distributions, given data on point observations of occurrence [65], with the objective of identifying the fundamental niche of the target organism [66,67]. Of the plethora available, the boosted regression trees (BRT) method [68,69] is selected by the authors as a default for occurrence mapping. A schematic overview of the occurrence mapping process is provided in figure 1. This selection was based on a number of factors: first, in a review of 16 species modelling methods, BRT was

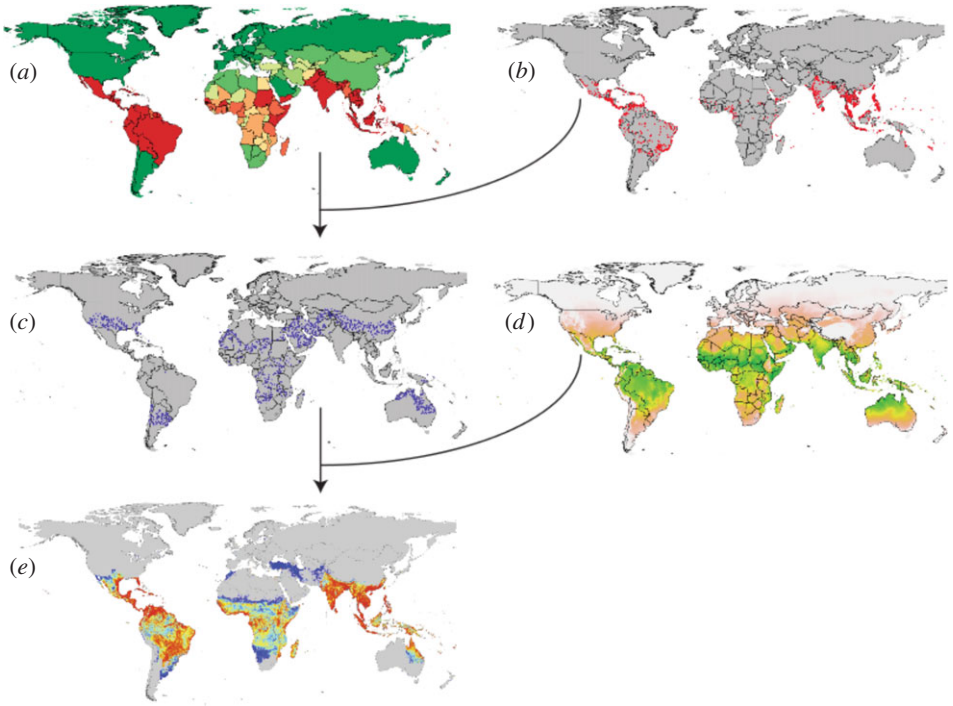


Figure 1. A schematic overview of a niche/occurrence mapping process (for example boosted regression trees (BRT)) that uses pseudo-absence data guided by expert opinion. Consensus based definitive extent layers of infectious disease occurrence at the national level (*a*) are combined with accurately geo-positioned occurrence (presence) locations (*b*) to generate pseudo-absence data (*c*). The presence (*b*) and pseudo-absence data (*c*) are then used in the BRT analyses, alongside a suite of environmental covariates (*d*) to predict the probability of occurrence of the target disease (*e*).

one of the top performing methods evaluated using the area under the receiver operating characteristic curve (AUC) and correlation statistics [16,70]; second, the method is flexible in being able to accommodate different types of predictor variables (e.g. continuous or categorical data); third, it is easy to understand, implement and uses reliable, well documented and freely available R code [71]; and fourth, the resulting maps are simple to interpret and include a ranked list of environmental predictors. The authors also have extensive experience with this technique after a global scale project to map the distribution of the anophelines of public health importance [72–76]. These references provide a detailed statistical explanation and examples of how BRT was applied to species distribution mapping.

Model-based geostatistics (MBG) [77,78] has recently been more widely applied in infectious disease mapping [17,79–83] and is the technique of choice where data allow. There are several reasons for this. First, MBG deals explicitly with the spatial (and with extension temporal) autocorrelation of disease data; this is still widely ignored in occurrence mapping. Second, MBG models can be configured to offer a much more robust parameterization of factors that can affect disease endemicity (such as age of the individuals sampled, the diagnostic technique used, the influence of covariates etc.). Third, by fitting the models using Bayesian inference, outputs can be presented to show the full uncertainty of the prediction in all parts of the predicted maps. The main impediments to its wider use are the lack of bespoke software with which to implement the models and its relatively large computational burden.

We assume that advances with respect to occurrence mapping or MBG techniques may modify our guidance with regard to mapping techniques and elaborate on some of the generic improvements that may be made in infectious disease mapping in §4. Those we have favoured here are proved methods that can be applied now.

In summary, the objective of this review is to formalize the questions outlined in §1, in order to define rules for advocating specific cartographic techniques for a baseline risk assessment for each disease of clinical importance, and then to assess to what level this mapping potential has been realized. A substantial literature review has been conducted to collate the data required to make those cartographic suggestions evidence-based and is provided as electronic supplementary material.

2. Material and methods

(a) Selection of infectious diseases of clinical importance

A total of 347 infectious diseases of clinical importance were selected for review based on the GIDEON database, accessed November 2010. GIDEON is an infectious disease information and diagnostic resource available online through subscription that derives its content from a range of sources including formal peer-reviewed journals and informal sources such as Ministry of Health reports [4,5]. This list was then revised to 355 diseases based on further re-definitions and decoupling of some groups. These diseases were placed into one of 11 classifications based on transmission type: animal contact, blood/body fluid contact, direct contact, endogenous, food/water-borne, respiratory, sexual contact, soil contact, unknown, vector-borne and water contact.

Revisions were as follows: mucosal and cutaneous leishmaniasis were re-classified as cutaneous/mucosal leishmaniasis, Old World and New World; the spotted fevers were also divided into New and Old World to better differentiate between the various species of bacteria and ticks that spread the disease in different parts of the world; malaria was split into *Plasmodium*

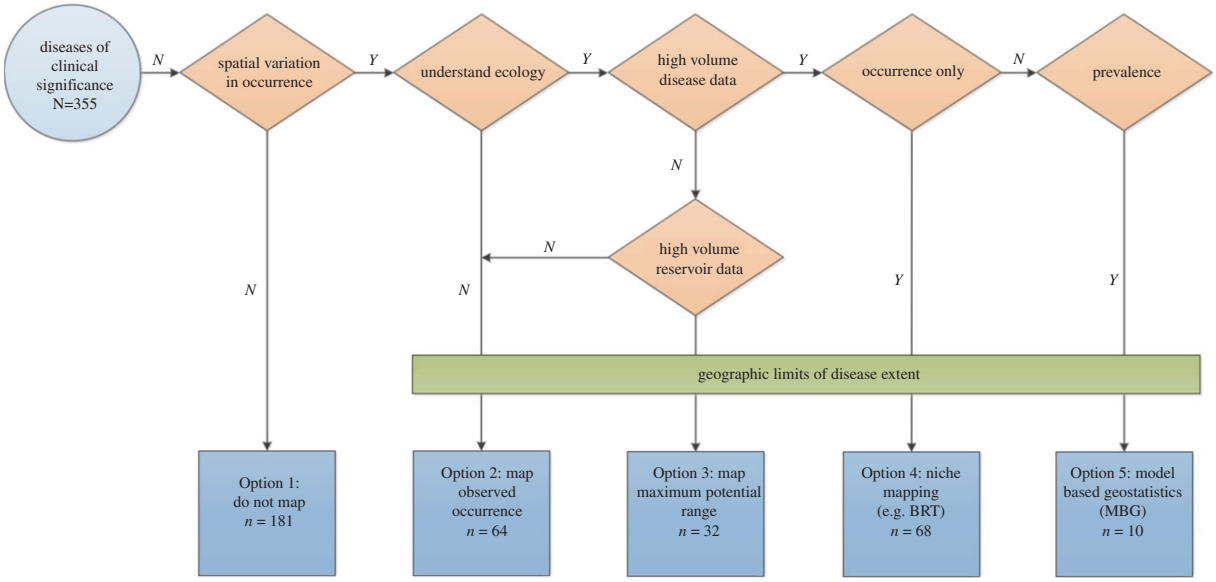


Figure 2. A schematic of the disease classification process. The classification system results in diseases being categorized into one of five options: (1) do not map; (2) map observed occurrence; (3) map maximum potential range of reservoir or vectors; (4) niche/occurrence mapping with BRT and (5) MGB-based endemicity maps.

falciparum, *Plasmodium vivax*, *Plasmodium ovale* and *Plasmodium malariae*, because variation in geographical range and epidemiological patterns of these pathogenic species would be masked if considered together; AIDS was removed and was combined with HIV; conjunctivitis-inclusion was similarly removed, and incorporated into trachoma; the umbrella term 'adenovirus infection' was divided into acute febrile respiratory disease (adenoviral), adenoviral haemorrhagic conjunctivitis, keratoconjunctivitis (adenoviral) and adenovirus infection; similarly, enterovirus infection was divided into enterovirus haemorrhagic conjunctivitis and enterovirus infection; human herpesvirus 6 was renamed Roseola; sandfly fever was added because of its possible impact on travellers; and avian influenza virus serotype H5N1 was added because of its epidemic potential.

(b) Data assembly

(i) Natural history

Data were collected on the natural history of each infectious agent. Information on the genus and species, disease reservoir, vector species (if applicable), mode of transmission, incubation period, vaccine (where relevant) and geographical distribution was obtained using GIDEON. Taxonomic classifications were supplemented by the Tree of Life Project (http://tolweb.org). Further evidence regarding geographical distribution and vaccine development was found in the American Public Health Association's Control of Communicable Disease Manual [7].

(ii) Transmission dynamics

The basic reproduction number (R_0) was used to quantify the transmission potential of the various aetiological agents. The R_0 is defined as the average number of secondary infections produced when a single-infected individual is introduced into a fully susceptible population [84–87]. A literature search was conducted to obtain R_0 values in humans and reservoirs of zoonotic diseases. The search was carried out in PubMed (http://www.pubmed.gov) using the terms '[disease name]' and 'reproduction number' in the 'all fields' search box in September 2011. The search was then repeated replacing 'reproduction number' with 'reproduction ratio', 'reproduction rate', 'reproductive number', 'reproductive ratio' and 'reproductive rate.' That search pattern was reiterated with '[Genus species]' or '[diseases synonym]' replacing '[disease name],' if applicable. This procedure was also performed in ISI-Web of Knowledge (http://isiwebofknowledge.com) in the 'title/keywords/abstract' field. These searches often produced few or no results and the entire search process would be conducted again using Google Scholar (http://scholar.google.co.uk). Data regarding R_0 values and the reservoir species when relevant were abstracted from references obtained, and if multiple R_0 estimates were reported among sources for a single disease, the range of estimates was recorded. The range for all R_0 estimates was assumed to start from 0.

(iii) Thumbnail maps

To visualize the approximate endemic regions of a disease, simple maps were constructed from the distribution data provided by GIDEON. A list of 275 global countries and territories were coded as 1 for endemic and 0 for non-endemic for each listed disease. The database was then imported into ArcGIS 10 (ESRI 2010) and displayed as global maps at the national level.

(iv) Occurrence data availability and quality

To determine the relative amount of information available for the various infectious diseases, a search was done using only the disease name as the text term in PubMed on 4 November 2011 and using the species name in GenBank on 1 March 2012 (for selected diseases). Data on the number of feeds for each disease from the start of data collection were received from HealthMap and ProMED on 23 November 2011 and from BioCaster on 24 February 2012. Because only data from manual searches of PubMed has, to our knowledge, been used in mapping, we base our analyses on PubMed figures only, but provide the potential data from the other sources in the electronic supplementary material. These may improve the prospects for mapping of many of the diseases once the utility of these information sources has been confirmed by experiment.

(c) Decision rules devised to categorize mapping options

Decision rules were created for disease mapping options, shown schematically in figure 2. The Option 1, *do not map*, classification was used for those conditions which are known to occur

worldwide, and hence do not show sustained spatial variation in occurrence. The diseases within this category range from sexually transmitted diseases such as Chlamydia, viral agents such as Epstein–Barr Virus or rhinoviruses causing the common cold and endogenous diseases (infections caused by previously dormant or inapparent pathogens, often from the typical commensal microbial flora of humans—such as urinary tract infections caused by *Escherichia coli* or brain abscesses by *Staphylococcus aureus*). The incidence of these diseases may show enormous spatial variation. These differences are linked often to variation in human or human-related factors, however, and are best mapped using techniques associated with the cartography of non-infectious disease [88]. More traditional surveillance within this cosmopolitan distribution, therefore, may have a public health rationale and this is explored on a case by case basis in the electronic supplementary material. For most of these conditions, it would be useful to apply a simple mask of human population density to give a more realistic picture of where the disease is truly observed globally. Option 2, *map the observed occurrence*, would apply to diseases that have few data available and limited information regarding the disease ecology. A cut-off of fewer than 25 PubMed hits per endemic country was applied to designate a paucity of data for any operationally significant disease. For example, Mayaro virus has 90 search results on PubMed for 11 potentially endemic countries and, therefore, only about eight results per country. There has also not been a definitive reservoir host identified for Mayaro, which would be needed for the following option. Option 3, *map the maximum potential range*, is appropriate for a disease that also has fewer than 25 PubMed results per country, but information is available regarding reservoir or vector species that would place boundaries on the potential disease distribution, as is the case with African tick bite fever with its known vector distribution. Mapping of the disease using ecological niche modelling, Option 4, would implement *BRT technology on observed occurrence data*. Adequate information regarding occurrence of disease (greater than 25 PubMed hits per country) is needed to use this strategy. This information would be usefully supplemented with information on where the disease is not found, obtained through systematic searches or derived by expert opinion maps. If the authors were aware of systematic searches of occurrence data that were significantly richer than the PubMed hits, these were documented and the mapping option re-evaluated accordingly. Option 5, the *implementation of MBG to mapping*, is reserved for diseases that have more than 25 results per country of systematically recorded prevalence data. This strategy uses MBG for the creation of complete endemicity maps with detailed uncertainty metrics. The mapping option to be used is dependent on the amount and nature of the disease data available, implying that diseases currently classified for one option would be eligible for a higher grade in the future as further data become available.

(d) Scoring the quality of existing mapping of the geographical distribution of disease

It was also of critical interest to obtain information regarding the extent to which the diseases had been previously mapped. A search was again conducted in PubMed using the text terms '[disease name/synonym]' and 'map' as well as '[disease name/synonym]' and 'epidemiology,' selecting for reviews in October 2011. If an excess of results were returned (more than 1000), this was further narrowed using the search terms 'distribution' or 'global.' For diseases transmitted by a specific vector, the search was repeated using the text terms '[vector species name]' and 'map.' The same process was repeated for prominent reservoir species. The search was also performed using ISI-Web of Knowledge. Irrelevant references were removed from the search output, and all references regarding the spatial temporal distribution of a disease, vector or reservoir were checked to determine the parameter mapped (for example, occurrence, prevalence, incidence, or risk) and in what geographical region.

In order to allow for both relative and quantitative assessment of each map, we devised a metascore, which evaluated three criteria: data quality, geographical scope and the mapping technique used.

Data quality (out of nine) was scored in three ways. (i) Contemporariness, where three points were awarded if data less than 10 years old was used, two points for the use of data greater than or equal to 10 years to less than 20 years old, and one point for data greater than 20 years old. If no age could be identified, no points were given. For papers reporting a range of dates, the score was based on the most recent, with the exception of databases that provide country-specific estimates that were surveyed across different time periods. In that case, an additional half point (2.5) was given. (ii) Diagnostic accuracy, where three points were awarded for the use of data diagnosed by genotype or PCR, or in the case of vector maps, where advanced modelling techniques had been used on a large number of occurrence points. Two points were given to those studies that had used hospital or national health surveys or confirmed case reports; an additional half a point was gained if serological or immunological data had been used. Vectors maps received two points if simple interpolation techniques had been used on occurrence data. One point was awarded if cited literature had been used. One point was also given for unpublished health organization data collected as part of routine health management information systems (HMIS) or presumptive diagnosis, with a half point given to non-specific numerical data. The use of expert opinion in drawing vector maps was awarded one point. If the data came from an unknown source, or was not listed in the article, no points were awarded. (iii) Geo-positional accuracy, where three points were awarded for the use of data coupled with GPS coordinates, two points if survey coordinates could be derived from supporting maps, or data was provided to administrative level 1; an additional half a point was earned if administrative level 2 was used, or towns and villages were specified. One point was gained if approximate coordinates of unknown provenance or country level data was present. Expert opinion ranges obtained from cited literature received half a point. If no geo-positional data was associated with the map, no points were awarded.

The geographical scope was scored out of 100. The GIDEON endemic country lists for each disease were converted into national populations at risk using the UN population data from 2010 [89]. Each map was assessed for how many countries were included (rounded up to the national level, to match the resolution of GIDEON), and population covered was calculated and expressed as a percentage (out of 100%) of the GIDEON endemic total.

The mapping technique used (mapping option used/theoretically best mapping option) was calculated using the criteria outlined above, each map was evaluated for the mapping option used (for example, if BRT modelling techniques had been used, the map was to Option 4 standard), and was related to the potential mapping option that could be used, based upon the amount and quality of data present for that disease. For instance, if a map of Lassa fever (which is an Option 4 disease owing to there being more than 25 PubMed hits per country) only uses occurrence points (Option 2 standard), a score of 2/4 would be achieved.

The metascore was then calculated as the product of these figures ([Quality]/9 × [Scope] × [Option Used]/[Option Potential]) resulting in a maximum of 100. Scores of greater than or equal to 75 per cent were deemed to have evaluated the global distribution of the specific disease to a satisfactory standard.

Table 1. The number of clinically important infectious diseases and the subset of those with a rationale for mapping by transmission category (see §2).

classification	clinically significant diseases ($n = 355$)	diseases with rationale for mapping ($n = 174$)
animal contact	20	9
blood/body fluid contact	14	5
direct contact	23	7
endogenous[a]	35	0
food/water-borne	82	36
respiratory	39	9
sexual contact	11	2
soil contact	21	14
unknown	11	4
vector-borne	88	80
water contact	11	8

[a]Endogenous infections are those caused by previously inapparent or dormant pathogens arising from the typical commensal microbial flora of humans.

3. Results

The electronic supplementary material provides full details of all the epidemiological and mapping evidence collated and scored and the decision rules applied. The electronic supplementary material includes a summary page on each of the 355 diseases with details of the natural history, transmission, quantity of data available, quality of data from previously published maps and recommendations for future mapping endeavours. The information included on natural history was the ICD-10 code, transmission classification (table 1), type of pathogen (agent), taxonomic details, mode of transmission, reservoir species (host organism that is a source of infection or potential reinfection of humans) and incubation period.

The epidemiological characteristics highlighted include the vaccine availability, and estimates of the basic reproduction number (R_0) in human and reservoir populations, where applicable. A number of diseases (126) were considered to have an R_0 value of less than 1 because they are primarily zoonotic diseases. Citations were provided to support that transmission occurs mainly in animals. The R_0 estimates ranged from point source outbreaks of diarrhoeal diseases or less than 1 for zoonoses to 100 for *P. vivax* malaria and Ross River virus and 1000 for *P. falciparum* malaria. Estimates were not obtained for many of the reservoir species, but for those that were found, the range was from 1.06 for Old World mucocutaneous leishmaniasis in dogs to 28 for West Nile fever virus in birds.

Occurrence details included information on the number of PubMed and GenBank hits, relevant reports from HealthMap, ProMED and BioCaster feeds, and the approximate number of endemic countries. A table of previously published maps was included incorporating information on whether the map is of the disease, vector or host reservoir; geographical scope; data quality score; mapping option used; metascore; citation.

The option for future mapping (figure 2) was determined using the PubMed hits returned and the number of endemic countries per diseases (see the electronic supplementary material). A total of 181/355 were classified as Option 1 (do not map); 64 were classified as Option 2 (map observed occurrence); 32 were classified as Option 3 (map maximum potential range); 68 were classified as Option 4 (map using BRT) and 10 were classified as Option 5 (map using MBG).

There are trends within the diseases that have a strong rationale for mapping. Unsurprisingly, endogenous diseases exhibit little sustained spatial variation in occurrence, whereas those transmission categories that are inherently linked to some feature of the environment, or other factor that varies on a global scale, such as vector-borne disease, water contact and soil contact tend to show greater variation. The remaining transmission types have just under half of the diseases showing differing global patterns of distribution. Similar trends are also apparent when we consider the occurrence of agents of disease—nearly two-thirds of diseases caused by parasites show tendency to vary over a spatial scale, as do 61 per cent of all viruses; on the other hand, there is evidence for spatially variable distributions in only 28 per cent of bacteria. Clearly, these sets of results are inherently linked; of the 61 viral diseases that would benefit from having mapped distributions, 41 are vector-borne and a further eight are soil contact; of those bacterial species that are not endemic worldwide, about two-thirds are vector-borne. Such a trend is not so apparent when considering parasitic diseases and their routes of transmission (many are food/water-borne). This could be due to their requirements for external development, and thus potentially environmentally determined life cycles.

Of the 174 diseases with strong rationale for mapping, only seven had maps that scored higher or equal to 75 per cent on the metascore. These were coltiviruses (Old World), dengue, Lassa fever, Mayaro, monkey pox, *P. falciparum* and *P. vivax*; all vector-borne diseases. Figure 3*a* shows radial plots of all the 174 diseases with a rationale for mapping, as well as separate plots by agent (figure 3*b–e*). The white line represents the highest scoring metascore for each disease; the black space above each individual line equates to the information deficit present.

4. Discussion

We have collated a significant amount of information on 355 diseases of clinical importance and have made evidence-based suggestions on the appropriate cartographic approaches to use in mapping each disease. These have been summarized in the results and are elaborated for each disease in the electronic supplementary material. In the following sections, we review some of the common omissions in existing maps and look to novel data sources, new techniques and information technology developments that may change the future landscape of infectious disease mapping.

This review has provided the opportunity to make some preliminary observations on some of the common omissions in infectious disease mapping that might be considered when embarking on new cartographies. They are as follows.

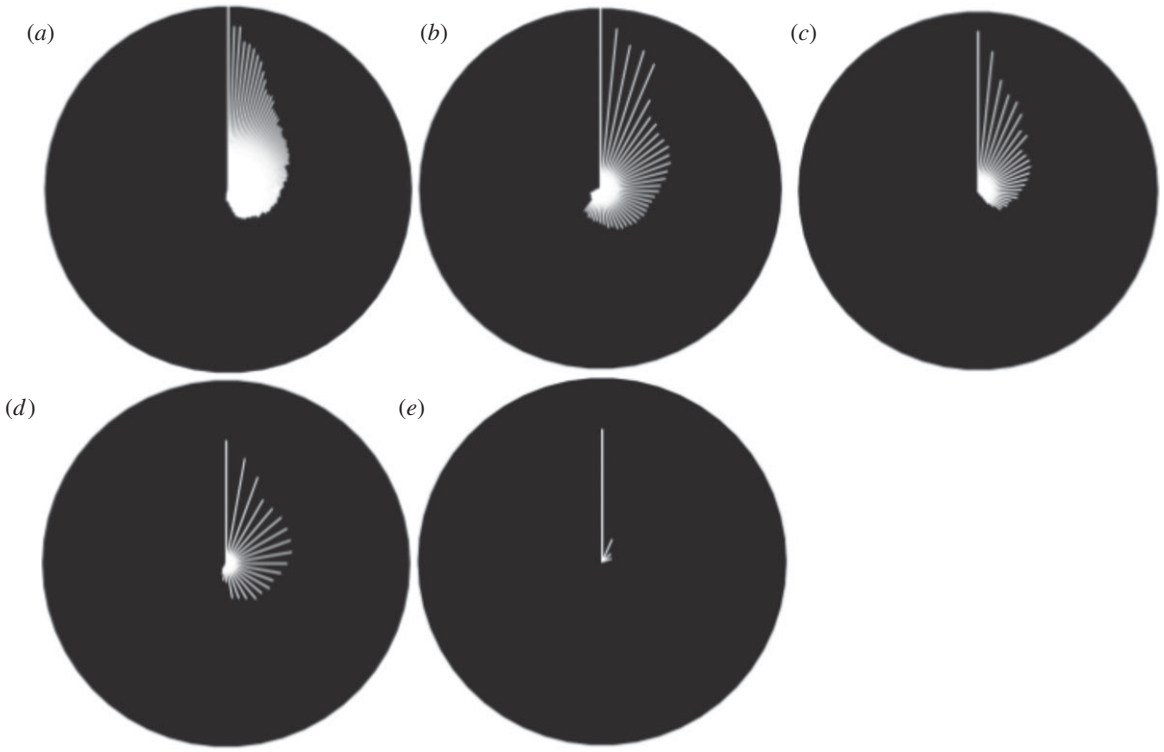


Figure 3. Radial plots for all diseases with a rationale for mapping, ordered clockwise, by metascore (white line). A white line from the centre to the edge of the circle would show a perfect metascore. (*a*) Reflects all diseases ($n = 174$ of 355), (*b*) viral diseases ($n = 62$ of 101), (*c*) parasitic diseases ($n = 61$ of 96), (*d*) bacterial diseases ($n = 36$ of 128), and (*e*) comprises fungal ($n = 9$ of 17), protoctistan ($n = 2$ of 2) and diseases of unknown pathogen ($n = 4$ of 10). Note that there was one algal disease, which did not have a rationale for mapping and is not shown in this diagram.

(a) Other relevant maps

The most consistent omission is the lack of additional information that can provide significant epidemiological insight—often referred to as 'expert opinion'. These definitive extent data can be an ad hoc collection for each disease that may include information on biological and biogeographic limits (often as range maps), as well as, further distribution or occurrence data on intermediate and reservoir hosts. There are several occurrence mapping methods that can use this information, such as weighted forms of BRT that have been trialled extensively with respect to the anophelines [72–75] (figure 1). They do this by overcoming the biogeographic and taxonomic ignorance of all occurrence mapping techniques that assume the globally realized niche approximates the fundamental niche. The careful use of definitive extent data would substantially reduce the degree to which inferences are required.

(b) Formalizing expert opinion

Further investigation is also advised on using the Cooke method to help determine the importance ascribed to the expert opinion [90,91]. Essentially these methods allow a simple way to gauge the accuracy of an expert source by testing their knowledge on a set of subject related questions to which the answers are well known. For a cartographic problem set, this could be very easily formalized by rating answers for a related disease we know the distribution of extremely well. It may be possible to link this with BRT and formalize the weights that are ascribed to other relevant epidemiological information.

(c) Human population distribution

There is a systematic deficit in the use of human population distribution maps [92,93], both as a mapping covariate and for determining the population at risk of infection or the reservoir of infection. Some effort may also be invested in incorporating the latest human population surfaces into the information suite. The diseases for which human population distribution may help refine risk assessments, including both those with a rationale for mapping and those ubiquitous clinically important diseases for which the recommendation was not to map, have been highlighted (see the electronic supplementary material).

(d) Refining of environmental covariates

Most cartographic applications use environmental covariates crudely without any adjustment to the epidemiology of the diseases concerned. Where detailed information and experiments on the environmental responses of a disease have been conducted it has proved valuable to combine this with the covariate. An example would be the way that temperature data have been used not only to map the environmental limits of *P. falciparum* and *P. vivax* globally [94], but have also been transformed into indexes of transmission suitability. These indexes were more strongly selected for by the model than untransformed covariates in endemicity mapping. The diseases to which such advances may be relevant are indicated (see the electronic supplementary material).

(e) Public health interventions

It is still rare for geographically specific intelligence on public health interventions to be used in the mapping of diseases.

Table 2. The cartographically relevant holdings of the National Center for Biotechnology Information PubMed and GenBank systems. The searches were conducted on 4 November 2011 and 1 March 2012, respectively.

system	PubMed	GenBank
start year	1946 [98]	1982 [99]
frequency of updates	daily [98]	Daily [100]
number of species catalogued	>250 000 [100]	>250 000 [100]
approximate number of entries	21 million [101]	340 million [100]
number of clinically relevant diseases for which data are available	168	155
occurrence point sources for mapping	526 564	672 327

Table 3. Geo-positioned occurrence data archived by the HealthMap and BioCaster online disease outbreak reporting systems. HealthMap uses automated text processing to classify and position alerts that are then confirmed by a human analyst [25]. BioCaster has automated text processing to classify and position alerts processed through a multilingual ontology [26]. The totals were assembled using data provided for HealthMap on 23 November 2011 and BioCaster on 24 February 2012.

system	HealthMap	BioCaster
start year	2006	2006
approximate posts per day	300 [24]	100 [29]
number of languages	10 (J. S. Brownstein 2012, personal communication)	11 [102]
number of diseases tagged	245 (J. S. Brownstein 2011, personal communication)	230 (N. Collier 2012, personal communication)
number of clinically relevant diseases for which data are available	84 of 245	99 of 230 (N. Collier 2012, personal communication)
total occurrence points	337 105 (J. S. Brownstein 2011, personal communication)	189 361 (N. Collier 2012, personal communication)
occurrence point sources for mapping	66 284 (J. S. Brownstein 2011, personal communication)	140 038 (N. Collier 2012, personal communication)

Such information could be used in the same way as other 'expert opinion' data sources by BRT. Where human interventions have significantly affected the distribution of a disease, for example vaccine coverage in a population [95–97], this has been identified. We have sought to identify those diseases for which this information may be relevant but have not searched systematically for the availability of relevant public health information.

There are many potential novel data sources that may be used for global infectious disease mapping. The resources described below have never been used systematically to address the paucity in occurrence data across the range of infectious diseases reviewed. Substantial progress will be made from exploiting the geospatial information in the formal literature (e.g. PubMed, www.ncbi.nlm.nih.gov/pubmed) and in genetic and protein sequence databases (e.g. GenBank, www.ncbi.nlm.nih.gov/genbank). The potential information available has been identified for each disease in the electronic supplementary material and is further summarized in table 2.

Significant prospects for the rapid acquisition of occurrence data are also clearly possible from online outbreak alert resources (i.e. HealthMap/ProMED [24,25], BioCaster [26,27] and Argus [28,29] records). The potential information available has been identified for each disease in the electronic supplementary material and is further summarized in table 3 for those systems where data can be freely shared.

Finally, there is a revolution occurring in both the volume and public availability of data about the health and wellbeing of individuals and populations through various forms of social media [103]; most notably Twitter (twitter.com). This is an online social media site that allows users to post 'Tweets'; messages less than or equal to 140 characters which are freely available to all. It took 3 years to reach the first billion Tweets, but by March 2011, it took only a week to reach one billion posts and 140 million Tweets are now posted daily with an increasing number of them automatically geo-positioned. This wealth of accurately geo-positioned information has already begun to be harvested for public health purposes. Twitter feeds surrounding the 2009 H1N1 flu outbreak were analysed and found to predict outbreaks one to two weeks in advance of traditional surveillance [104,105]. Tweets can also be analysed to identify a broader range of health-related terms such as symptoms, syndromes and treatments to illuminate geographical patterns in syndrome surveillance [106].

Our optimism about the future use of social media is tempered by the realization that the main contemporary issue in disease mapping, of dealing with the lack of relevant data, will subside, and that our new challenges will be informatics, developing systems and processes to take on the big data challenges of the future. This is discussed in the following section.

There are also many novel techniques that may be used to improve the prospects of global infectious disease mapping, notably automation through machine learning and harnessing the cognitive surplus. In the defined schema (figure 2), it is more logistically and technically difficult (and thus expensive) to map diseases from Option 1 (do not map) through to Option 5 (map endemicity with MBG). It is also more expensive to deal with conditions for which data retrieval is a significant logistical obstacle. This will be directly proportional to the number of PubMed and other (see earlier) data source hits identified.

The HealthMap and BioCaster systems have pioneered machine learning algorithms that automatically classify relevant reports, identify the infectious disease of interest and determine the geographical location of the outbreak. Scaling these to cope with this potential data deluge is a non-trivial but largely technical problem. Ideally, the results of this process should be audited and verified by subject matter experts but this is non-scalable, time consuming and prohibitively expensive.

As an alternative, developments in social computing have led to increased interest in using large numbers of non-experts as a cheaper and scalable method for data filtering: the so-called crowdsourcing or distributed cognition [107,108]. Currently established ways to crowdsource exist (i) framing filtering tasks as fun online games, incentivizing users to filter data for free [109] and (ii) posting the task online and seeking non-experts using a pay-per-example setting as pioneered by the Amazon Mechanical Turk system [110,111]. The central idea is that, if questions can be structured in a simple and intuitive way, and presented to a large number of individuals, the central tendency of responses is likely to provide an accurate answer. Crowdsourcing is particularly appealing in the context of filtering social media disease reports because of the non-expert nature of key components of the task, such as geo-positioning. Crowdsourcing is not, of course, a panacea for data filtering. The reliability of contributors must be quantitatively assessed and iteratively adjusted for, again with reference to a gold-standard reference set of externally validated results.

In conclusion, this systematic review has shown that we have an astonishingly poor knowledge of the global distribution of the vast majority of infectious diseases of clinical importance. Less than 5 per cent of clinically important infectious diseases have been mapped reliably. This presents clear obstacles to advances in determining the global burden of these conditions, our ability to differentiate outbreaks of concern in international biosurveillance, and our ability to understand the geographical determinants of disease emergence, past, present and future. We have shown that contemporary solutions exist to enable us to use new data and new technology to rapidly improve the cartography of a wide range of clinically important pathogens. Few conceptual barriers exist to making rapid progress and to 'seeing further' into the relatively unknown landscape of infectious disease mapping.

The catalyst for this review was a National Institute for Mathematical and Biological Synthesis (NIMBioS) and the US Department of Defense hosted meeting on infectious disease modelling (23–25 January 2011, Knoxville, TN, USA). NIMBioS also provided K.E.B. with resources to conduct the literature review. S.I.H. is financially supported by a Senior Research Fellowship from the Wellcome Trust (no. 095066) which also supports P.W.G., K.E.B. and D.M.P.; S.I.H., D.L.S. and D.B.G. also acknowledge support from the RAPIDD programme of the Science & Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health (http://www.fic.nih.gov). This work also forms part of the output of the Malaria Atlas Project (MAP, http://www.map.ox.ac.uk), principally financially supported by the Wellcome Trust, UK (http://www.wellcome.ac.uk). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Thanks are extended to Dr Kevin Hanson and Prof. Louis Gross for insightful comments, to the Global Infectious Diseases and Epidemiology Network (GIDEON) for permission to reproduce their data within the 355 maps of the supplementary material. Finally, we are grateful to the editors Dr Oliver Pybus, Prof. Christophe Fraser and Prof. Andrew Rambaut for inviting us to participate in the scientific discussion meeting.

References

1. Cleaveland S, Laurenson MK, Taylor LH. 2001 Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Phil. Trans. R. Soc. Lond. B* **356**, 991–999. (doi:10.1098/rstb.2001.0889)
2. Taylor LH, Latham SM, Woolhouse ME. 2001 Risk factors for human disease emergence. *Phil. Trans. R. Soc. Lond. B* **356**, 983–989. (doi:10.1098/rstb.2001.0888)
3. Woolhouse ME, Gowtage-Sequeria S. 2005 Host range and emerging and reemerging pathogens. *Emerg. Infect. Dis.* **11**, 1842–1847. (doi:10.3201/eid1112.050997)
4. *Global Infectious Diseases and Epidemiology Network* (GIDEON). 2011 *The world's premier global infectious diseases database*. Los Angeles, CA: GIDEON Informatics, Inc. See http://web.gideononline.com/web/epidemiology.
5. Edberg SC. 2005 Global infectious diseases and epidemiology network (GIDEON): a world wide web-based program for diagnosis and informatics in infectious diseases. *Clin. Infect. Dis.* **40**, 123–126. (doi:10.1086/426549)
6. C.D.C. 2009 *CDC Health information for international travel 2010*. Atlanta, GA: Centers for Disease Control and Prevention.
7. Heymann DL. 2008 *Control of communicable diseases manual*, 19th edn. Washington, DC: American Public Health Association.
8. Hay SI. 2000 An overview of remote sensing and geodesy for epidemiology and public health application. *Adv. Parasitol.* **47**, 1–35. (doi:10.1016/S0065-308X(00)47005-3)
9. Rogers DJ, Randolph SE, Snow RW, Hay SI. 2002 Satellite imagery in the study and forecast of malaria. *Nature* **415**, 710–715. (doi:10.1038/415710a)
10. Cromley EK, McLafferty SL. 2002 *GIS and public health*. New York, NY: The Guildford Press.
11. Rogers DJ, Randolph SE. 2003 Studying the global distribution of infectious diseases using GIS and RS. *Nat. Rev. Microbiol.* **1**, 231–237. (doi:10.1038/nrmicro776)
12. Hay SI, Graham AJ, Rogers DJ. (eds) 2006 Global mapping of infectious diseases: methods, examples and emerging applications. In *Advances in parasitology*, vol. 62. London, UK: Academic Press.
13. Hay SI, Snow RW. 2006 The malaria atlas project: developing global maps of malaria risk. *PLoS Med.* **3**, e473. (doi:10.1371/journal.pmed.0030473)
14. Riley S. 2007 Large-scale spatial-transmission models of infectious disease. *Science* **316**, 1298–1301. (doi:10.1126/science.1134695)
15. Pfeiffer DU, Robinson TP, Stevenson M, Stevens KB, Rogers DJ, Clements ACA. 2008 *Spatial analysis in epidemiology*. Oxford, UK: Oxford University Press.
16. Stevens KB, Pfeiffer DU. 2011 Spatial modelling of disease using data- and knowledge-driven approaches. *Spat. Spatio-temporal Epidemiol.* **2**, 125–133. (doi:10.1016/j.sste.2011.07.007)
17. Hay SI *et al*. 2009 A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS*

Med. **6**, e1000048. (doi:10.1371/journal.pmed.1000048)
18. Gething PW, Patil AP, Smith DL, Guerra CA, Elyazar IR, Johnston GL, Tatem AJ, Hay SI. 2011 A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar. J.* **10**, 378. (doi:10.1186/1475-2875-10-378)
19. Gething PW *et al*. 2012 A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. *Public Libr. Sci. Negl. Trop. Dis*. **6**, e1814.
20. Tatem A, Smith D, Gething P, Kabaria C, Snow R, Hay S. 2010 Ranking elimination feasibility among malaria-endemic countries. *Lancet* **376**, 1579–1591. (doi:10.1016/S0140-6736(10)61301-3)
21. Project Global Health Group at Malaria Atlas. 2011 *Atlas of Malaria Eliminating Countries, 2011*. San Francisco, CA: The Global Health Group, Global Health Sciences, University of California.
22. WHO. 2010 *International travel and health: situation as on 1 January 2010*. Geneva, Switzerland: World Health Organization.
23. Fefferman NH, Naumova EN. 2009 Innovation in observation: a vision for early outbreak detection. *Emerg. Health Threats J.* **3**, e6. (doi:10.3134/ehtj.10.006)
24. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. 2008 Surveillance sans frontieres: internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Med.* **5**, e151. (doi:10.1371/journal.pmed.0050151)
25. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. 2008 HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *J. Am. Med. Inform. Assoc.* **15**, 150–157. (doi:10.1197/jamia.M2544)
26. Collier N *et al*. 2008 BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics* **24**, 2940–2941. (doi:10.1093/bioinformatics/btn534)
27. Collier N, Goodwin RM, McCrae J, Doan S, Kawazoe A, Conway M, Kawtrakul A, Takeuchi K, Dien D. 2010 An ontology-driven system for detecting global health events. In *Proc. 23rd Int. Conf. on Computational Linguistics*, *Beijing, China*. Association for Computational Linguistics.
28. Torii M, Yin L, Nguyen T, Mazumdar CT, Liu H, Hartley DM, Nelson NP. 2011 An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics. *Int. J. Med. Inf.* **80**, 56–66. (doi:10.1016/j.ijmedinf.2010.10.015)
29. Hartley DM *et al*. 2010 The landscape of international event-based biosurveillance. *Emerg. Health Threats J.* **3**, e3. (doi:10.3134/ehtj.10.003)
30. Doherr MG, Audige L. 2001 Monitoring and surveillance for rare health-related events: a review from the veterinary perspective. *Phil. Trans. R. Soc. Lond. B* **356**, 1097–1106. (doi:10.1098/rstb.2001.0898)
31. Blazes DL, Russell KL. 2011 Joining forces: civilians and the military must cooperate on global disease control. *Nature* **477**, 395–396. (doi:10.1038/477395a)
32. Khan K *et al*. 2012 Infectious disease surveillance and modelling across geographic frontiers and scientific specialties. *Lancet Infect. Dis.* **12**, 222–230. (doi:10.1016/S1473-3099(11)70313-9)
33. Dobson AP, Carper ER. 1996 Infectious diseases and human population history. *Bioscience* **46**, 115–126. (doi:10.2307/1312814)
34. Wolfe ND, Dunavan CP, Diamond J. 2007 Origins of major human infectious diseases. *Nature* **447**, 279–283. (doi:10.1038/nature05775)
35. Morens DM, Folkers GK, Fauci AS. 2004 The challenge of emerging and re-emerging infectious diseases. *Nature* **430**, 242–249. (doi:10.1038/nature02759)
36. Wolfe ND, Daszak P, Kilpatrick AM, Burke DS. 2005 Bushmeat hunting, deforestation, and prediction of zoonotic disease emergence. *Emerg. Infect. Dis.* **11**, 1822–1827. (doi:10.3201/eid1112.040789)
37. Gayer M, Legros D, Formenty P, Connolly MA. 2007 Conflict and emerging infectious diseases. *Emerg. Infect. Dis.* **13**, 1625–1631. (doi:10.3201/eid1311.061093)
38. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P. 2008 Global trends in emerging infectious diseases. *Nature* **451**, 990–993. (doi:10.1038/nature06536)
39. Randolph SE, Rogers DJ. 2010 The arrival, establishment and spread of exotic diseases: patterns and predictions. *Nat. Rev. Microbiol.* **8**, 361–371. (doi:10.1038/nrmicro2336)
40. Woolhouse M. 2011 How to make predictions about future infectious disease risks. *Phil. Trans. R. Soc. B* **366**, 2045–2054. (doi:10.1098/rstb.2010.0387)
41. Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, McCraw SL, Gurr SJ. 2012 Emerging fungal threats to animal, plant and ecosystem health. *Nature* **484**, 186–194. (doi:10.1038/nature10947)
42. Cliff AD, Smallman-Raynor MR, Haggett P, Stroup DF, Thacker SB. 2009 *Emergence and re-emergence. Infectious diseases. A geographical analysis*. Oxford, UK: Oxford University Press.
43. Guernier V, Hochberg ME, Guegan JF. 2004 Ecology drives the worldwide distribution of human diseases. *PLoS Biol.* **2**, e141. (doi:10.1371/journal.pbio.0020141)
44. Guernier V, Guegan JF. 2009 May Rapoport's rule apply to human associated pathogens? *EcoHealth* **6**, 509–521. (doi:10.1007/s10393-010-0290-5)
45. Smith KF, Guegan J-F. 2010 Changing geograpahic distributions of human pathogens. *Annu. Rev. Ecol. Syst.* **41**, 231–250. (doi:10.1146/annurev-ecolsys-102209-144634)
46. Smith KF, Sax DF, Gaines SD, Guernier V, Guegan JF. 2007 Globalization of human infectious disease. *Ecology* **88**, 1903–1910. (doi:10.1890/06-1052.1)
47. Keesing F *et al*. 2010 Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature* **468**, 647–652. (doi:10.1038/nature09575)
48. Hirsch A. 1883 *Handbook of geographical and historical pathology*. Ann Arbour, MI: New Sydenham Society.
49. Cliff A, Haggett P, Smallman-Raynor M. 2004 *World atlas of epidemic diseases*. London, UK: Arnold Publishers.
50. Koch T. 2011 *Disease maps: epidemics on the ground*. Chicago, IL: University of Chicago Press.
51. Hehir P. 1927 *Malaria in India*. London, UK: Oxford University Press.
52. May JM. 1951 Map of the world distribution of malaria vectors. *Geogr. Rev.* **41**, 638–639. (doi:10.2307/210709)
53. Macdonald G. 1957 *Local features of malaria*. In The epidemiology and control of malaria. pp. 63–99. London, UK: Oxford University Press.
54. Pampana E. 1969 *A textbook of malaria eradication*, 2nd edn. London, UK: Oxford University Press.
55. Lysenko A, Semashko I. 1968 In *Geography of malaria: a medical-geographical study of an ancient desease* (ed. AW Lebedew). Moscow, Russia: Academy of Sciences, USSR.
56. Wertheim HFL, Horby P, Woodall JP. 2012 *Atlas of human infectious diseases*. Oxford, UK: Wiley-Blackwell.
57. Hay SI, Tatem AJ, Graham AJ, Goetz SJ, Rogers DJ. 2006 Global environmental data for mapping infectious disease distribution. *Adv. Parasitol.* **62**, 37–77. (doi:10.1016/S0065-308X(05)62002-7)
58. Scharlemann JPW, Benz D, Hay SI, Purse BV, Tatem AJ, Wint GRW, Rogers DJ. 2008 Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. *PLoS ONE* **3**, e1408. (doi:10.1371/journal.pone.0001408)
59. Patil AP, Gething PW, Piel FB, Hay SI. 2011 Bayesian geostatistics in health cartography: the perspective of malaria. *Trends Parasitol.* **27**, 245–252. (doi:10.1016/j.pt.2011.01.003)
60. Metselaar D, Van Thiel PH. 1959 Classification of malaria. *Trop. Geogr. Med.* **11**, 157–161.
61. Gething PW, Kirui VC, Alegana VA, Okiro EA, Noor AM, Snow RW. 2010 Estimating the number of paediatric fevers associated with malaria infection presenting to Africa's public health sector in 2007. *PLoS Med.* **7**, e1000301. (doi:10.1371/journal.pmed.1000301)
62. Hay SI, Okiro EA, Gething PW, Patil AP, Tatem AJ, Guerra CA, Snow RW. 2010 Estimating the global clinical burden of *Plasmodium falciparum* malaria in 2007. *PLoS Med.* **7**, e1000290. (doi:10.1371/journal.pmed.1000290)
63. Smith DL, Drakeley CJ, Chiyaka C, Hay SI. 2010 A quantitative analysis of transmission efficiency versus intensity for malaria. *Nat. Commun.* **1**, 108. (doi:10.1038/ncomms1107)
64. Smith DL, Cohen JM, Moonen B, Tatem AJ, Sabot OJ, Ali A, Mugheiry SM. 2011 Infectious disease. Solving the Sisyphean problem of malaria in Zanzibar. *Science* **332**, 1384–1385. (doi:10.1126/science.1201398)
65. Rogers DJ. 2006 Models for vectors and vector-borne diseases. *Adv. Parasitol.* **62**, 1–35. (doi:10.1016/S0065-308X(05)62001-5)
66. Hutchinson GE. 1957 Concluding remarks. *Cold Spring Harb. Symp. Quant. Biol.* **22**, 415–427. (doi:10.1101/SQB.1957.022.01.039)
67. Southwood TRE. 1977 Habitat, templet for ecological strategies? Presidential address to British Ecological Society, 5 January 1977. *J. Anim. Ecol.* **46**, 337–365. (doi:10.2307/3817)

68. Elith J, Leathwick JR, Hastie T. 2008 A working guide to boosted regression trees. *J. Anim. Ecol.* **77**, 802–813. (doi:10.1111/j.1365-2656.2008.01390.x)
69. De'ath G. 2007 Boosted trees for ecological modeling and prediction. *Ecology* **88**, 243–251. (doi:10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2)
70. Elith J *et al*. 2006 Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**, 129–151. (doi:10.1111/j.2006.0906-7590.04596.x)
71. R Development Core Team. 2008 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing (http://www.R-project.org).
72. Hay SI *et al*. 2010 Developing global maps of the dominant *Anopheles* vectors of human malaria. *PLoS Med.* **7**, e1000209. (doi:10.1371/journal.pmed.1000209)
73. Sinka ME *et al*. 2011 The dominant *Anopheles* vectors of human malaria in the Asia-Pacific region: occurrence data, distribution maps and bionomic précis. *Parasite Vectors* **4**, 89. (doi:10.1186/1756-3305-4-89)
74. Sinka ME *et al*. 2010 The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. *Parasite Vectors* **3**, 117. (doi:10.1186/1756-3305-3-117)
75. Sinka ME *et al*. 2010 The dominant *Anopheles* vectors of human malaria in the Americas: occurrence data, distribution maps and bionomic précis. *Parasite Vectors* **3**, 72. (doi:10.1186/1756-3305-3-72)
76. Sinka ME *et al*. 2012 A global map of dominant malaria vectors. *Parasite Vectors* **5**, 69. (doi:10.1186/1756-3305-5-69)
77. Diggle PJ, Ribeiro PJ. 2007 In *Model-based geostatistics*. (eds P Bickel, P Diggle, S Fienberg, U Gather, I Olkin, S Zeger). New York, NY: Springer.
78. Diggle PJ, Tawn JA, Moyeed RA. 1998 Model-based geostatistics. *J. Roy. Stat. Soc. C Appl.* **47**, 299–326. (doi:10.1111/1467-9876.00113)
79. Clements ACA, Moyeed R, Brooker S. 2006 Bayesian geostatistical prediction of the intensity of infection with *Schistosoma mansoni* in East Africa. *Parasitology* **133**, 711–719. (doi:10.1017/S0031182006001181)
80. Diggle PJ *et al*. 2007 Spatial modelling and the prediction of *Loa loa* risk: decision making under uncertainty. *Ann. Trop. Med. Parasitol.* **101**, 499–509. (doi:10.1179/136485907X229121)
81. Vounatsou P, Raso G, Tanner M, N'Goran EK, Utzinger J. 2009 Bayesian geostatistical modelling for mapping schistosomiasis transmission. *Parasitology* **136**, 1695–1705. (doi:10.1017/S003118200900599X)
82. Magalhaes RJ, Clements AC, Patil AP, Gething PW, Brooker S. 2011 The applications of model-based geostatistics in helminth epidemiology and control. *Adv. Parasitol.* **74**, 267–296. (doi:10.1016/B978-0-12-385897-9.00005-7)
83. Raso G *et al*. 2012 Mapping malaria risk among children in Cote d'Ivoire using Bayesian geo-statistical models. *Malar. J.* **11**, 160. (doi:10.1186/1475-2875-11-160)
84. Anderson RM, May RM. 1979 Population biology of infectious diseases: part I. *Nature* **280**, 361–367. (doi:10.1038/280361a0)
85. May RM, Anderson RM. 1979 Population biology of infectious diseases: part II. *Nature* **280**, 455–461. (doi:10.1038/280455a0)
86. Anderson RM, May RM. 1991 *Infectious diseases of humans: dynamics and control*. Oxford, UK: Oxford University Press.
87. May RM, Gupta S, McLean AR. 2001 Infectious disease dynamics: what characterizes a successful invader? *Phil. Trans. R. Soc. Lond. B* **356**, 901–910. (doi:10.1098/rstb.2001.0866)
88. Hutt MSR, Burkitt DP. 1986 *The geography of non-infectious disease*. Oxford, UK: Oxford University Press.
89. UNPD. 2010 *World population prospects: the 2010 revision population database*. New York, NY: United Nations Population Division (U.N.P.D.) See http://esa.un.org/unpp/.
90. Aspinall W. 2010 A route to more tractable expert advice. *Nature* **463**, 294–295. (doi:10.1038/463294a)
91. Cooke RM. 1991 *Experts in uncertainty. Opinion and subjective probability in science*. New York, NY: Oxford University Press.
92. Balk D, Deichmann U, Yetman G, Pozzi F, Hay S, Nelson A. 2006 Determining global population distribution: methods, applications and data. *Adv. Parasitol.* **62**, 119–156. (doi:10.1016/S0065-308X(05)62004-0)
93. Linard C, Tatem AJ. 2012 Large-scale spatial population databases in infectious disease research. *Int. J. Health Geogr.* **11**, 7. (doi:10.1186/1476-072X-11-7)
94. Gething PW, Van Boeckel T, Smith DL, Guerra CA, Patil AP, Snow RW, Hay SI. 2011 Modelling the global constraints of temperature on transmission of *Plasmodium falciparum* and *P. vivax*. *Parasite Vectors* **4**, 92. (doi:10.1186/1756-3305-4-92)
95. Hall R, Jolley D. 2011 International measles incidence and immunization coverage. *J. Infect. Dis.* **204**, S158–S163. (doi:10.1093/infdis/jir124)
96. Harrison LH *et al*. 2011 The Global Meningococcal Initiative: recommendations for reducing the global burden of meningococcal disease. *Vaccine* **29**, 3363–3371. (doi:10.1016/j.vaccine.2011.02.058)
97. Minor PD. 2012 The polio-eradication programme and issues of the end game. *J. Gen. Virol.* **93**, 457–474. (doi:10.1099/vir.0.036988-0)
98. MEDLINE. 2011 *Fact Sheet*. Bethesda, MA: U.S. National Library of Medicine (NLM). (http://www.nlm.nih.gov/pubs/factsheets/medline.html)
99. Bilofsky HS, Burks C. 1988 The GenBank genetic sequence data bank. *Nucleic Acids Res.* **16**, 1861–1863. (doi:10.1093/nar/16.5.1861)
100. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW. 2012 GenBank. *Nucleic Acids Res.* **40**, D48–D53. (doi:10.1093/nar/gkr1202)
101. PubMed. 2011 *PubMed help*. Bethesda, MA: National Center for Biotechnology Information. (http://www.ncbi.nlm.nih.gov/books/NBK3827)
102. Lyon A, Nunn M, Grossel G, Burgman M. 2011 Comparison of web-based biosecurity intelligence systems: BioCaster, EpiSPIDER and HealthMap. *Transboundary Emerg. Dis*. (doi:10.1111/j.1865-1682.2011.01258.x)
103. Salathé M *et al*. 2012 Digital epidemiology. *PLoS Comput. Biol.* **8**, e1002616. (doi:10.1371/journal.pcbi.1002616)
104. Signorini A, Segre AM, Polgreen PM. 2011 The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PLoS ONE* **6**, e19467. (doi:10.1371/journal.pone.0019467)
105. Chew C, Eysenbach G. 2010 Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PLoS ONE* **5**, e14118. (doi:10.1371/journal.pone.0014118)
106. Paul MJ, Dredze M. 2011 *A model for mining public health topics from Twitter. Technical report*. Johns Hopkins University, Baltimore, MD.
107. Howe J. 2006 The rise of crowdsourcing. *Wired* **14**, 1–5.
108. Shirky C. 2010 *Cognitive surplus. Creativity and generosity in a connected age*. London, UK: Penguin Books.
109. von Ahn L, Kedia M, Blum M. 2006 Verbosity: a game for collecting common-sense facts. In *ACM Conf. on Human Factors in Computing Systems*. April 24–27, 2006 Montreal Canada. Pittsburgh, PA: Computer Science Department, Carnegie Mellon University.
110. Sheng VS, Provost F, Panagiotis Ipeirotis G. 2008 *Get another label? Improving data quality and data mining using multiple, noisy labelers*. New York, NY: ACM Digital Library.
111. Kittur A, Chi EH, Suh B. 2008 *Crowdsourcing user studies with Mechanical Turk*. New York, NY: ACM Digital Library.

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Infectious diseases

» Parasitology

» Epidemiology

Global database of leishmaniasis occurrence locations, 1960–2012

David M. Pigott[1], Nick Golding[1], Jane P. Messina[1], Katherine E. Battle[1], Kirsten A. Duda[1], Yves Balard[2], Patrick Bastien[2,3], Francine Pratlong[2,3], John S. Brownstein[4,5], Clark C. Freifeld[4], Sumiko R. Mekaru[4], Lawrence C. Madoff[6,7], Dylan B. George[8], Monica F. Myers[1] & Simon I. Hay[1,8]

Received: 29 July 2014
Accepted: 28 August 2014
Published: 30 September 2014

The leishmaniases are neglected tropical diseases of significant public health importance. However, information on their global occurrence is disparate and sparse. This database represents an attempt to collate reported leishmaniasis occurrences from 1960 to 2012. Methodology for the collection of data from the literature, abstraction of case locations and data processing procedures are described here. In addition, strain archives and online data resources were accessed. A total of 12,563 spatially and temporally unique occurrences of both cutaneous and visceral leishmaniasis comprise the database, ranging in geographic scale from villages to states. These data can be used for a variety of mapping and spatial analyses covering multiple resolutions.

Design Type(s)	observation design • epidemiological study • data integration
Measurement Type(s)	epidemiology
Technology Type(s)	data collection method
Factor Type(s)	period
Sample Characteristic(s)	Leishmania • tropical or subtropical location • anthropogenic habitat

[1]Spatial Ecology and Epidemiology Group, Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford OX1 3PS, UK. [2]University Montpellier 1 (UFR Médecine) & CNRS 5290/IRD 224 (UMR 'MiVEGEC'), Laboratoire de Parasitologie-Mycologie, 34295 Montpellier, France. [3]CHRU de Montpellier, Centre National de Référence des Leishmanioses, Departement de Parasitologie—Mycologie, 34295 Montpellier, France. [4]Children's Hospital Informatics Program, Boston Children's Hospital, Boston, Massachusetts, USA. [5]Department of Pediatrics, Harvard Medical School, Boston, Massachusetts 02115, USA. [6]ProMED-mail, International Society for Infectious Diseases, Worcester, Massachusetts 01655, USA. [7]University of Massachusetts Medical School, Worcester, Massachusetts, 01655, USA. [8]Fogarty International Center, National Institutes of Health, Bethesda, Maryland, 20892, USA.
Correspondence and requests for materials should be addressed to D.M.P. (email: david.pigott@zoo.ox.ac.uk).

Background & Summary

The leishmaniases cause a range of clinical symptoms, from cutaneous lesions to visceral, often fatal, complications[1]. Considered as one of the 'neglected tropical diseases,' knowledge of this disease is relatively poor, particularly when compared to conditions with similar burden profiles[2,3]. Indeed, the real burden of this disease remains unknown[4]. Case survey data such as that presented by Alvar *et al.*[5], whilst informative, is prone to biases associated with reporting and national healthcare provisioning. Alternate methods should be employed in conjunction with such data to compensate for these issues and provide different approaches for estimating disease distribution and burden. An example of mapping the leishmaniases is presented in Pigott *et al.*[6] where the authors were able to estimate the current global distribution of these diseases by using the database of 12,563 unique records of occurrence presented here and statistical modelling to produce a pixel-based assessment of the likelihood of disease occurrence.

This database comprises occurrence records from 1960 to 2012, and represents a significant expansion in data compared to existing databases[7]. Database creation and management procedures are outlined below, along with a description of the final database structure. Information for accessing the full database is provided here as well in order to (a) allow for replication of the Pigott *et al.*[6] study, (b) enable the maps to be improved upon as new modelling methods and datasets become available, (c) provide an additional data source for both local-scale and regional mapping studies and (d) provide information for public health organisations on leishmaniasis in specific regions. All data files are available from Dryad (Data Citation 1).

Methods

The methods described here expand upon those outlined in Pigott *et al.*[6] by providing more details on data collection and the protocol for positioning of the data. This paper also provides the methodology for standardisation and validation of the occurrence dataset not previously described.

Data collection

PubMed and Web of Knowledge were searched using the keyword 'leish*' for all articles dating up until December 2012. Abstracts for all articles were imported into a bibliographic referencing tool and assessed for relevance, removing articles that did not contain information relating to disease occurrence. From these searches, 4,845 articles were identified and the corresponding texts obtained in full, where possible.

Leishmaniasis occurrence, defined as a report indicating one or more confirmed cases of leishmaniasis, within a specific administrative unit or 5×5 km pixel in a given calendar year, was divided into two categories representing the two main forms of the disease: cutaneous (including cases of localised cutaneous leishmaniasis, diffuse cutaneous leishmaniasis and mucosal leishmaniasis) and visceral leishmaniasis (including cases of post-kala-azar-dermal leishmaniasis (PKDL)). Only autochthonous symptomatic cases were included; wherever possible, cases that were imported were traced back to the original source of infection, otherwise they were excluded. If reports of PKDL also included information on prior VL infection, the latter was recorded; if not, the PKDL case was registered in the database as a VL occurrence. Similarly, only confirmed cases were included. Diagnoses via PCR or parasite cultures, serological tests or microscopic identification were included, as were articles indicating non-specific 'laboratory diagnosis'.

In addition to this, information was made available from the strain archives of the Centre National de Référence des Leishmanioses (CNR-L) at the Montpellier University Hospital Centre, France. In total, information from 3,465 strains isolated from humans was provided, collected between 1954 and 2013 (in the final database this was restricted to 1960–2012 to be consistent with the literature searches). These strains have been collected by various groups from around the world, cryopreserved in the International Biological Resources Centre of *Leishmania*, Montpellier, France and have subsequently undergone isoenzymatic identification[8,9]. In addition, information from GenBank was extracted by searching for *Leishmania* spp. known to cause disease in humans[10].

Informal online information sources, including online news articles and ProMED-mail reports[11], were provided by HealthMap (http://healthmap.org)[12]. The automated system systematically searches various news aggregators, open mailing lists, electronic disease surveillance networks and public health outbreak report feeds. The HealthMap classification system parses out disease and disease-related keywords from the body of these articles as well as classifying the report by one of five classifications—breaking (i.e., information relating to an ongoing outbreak), context (an article supplying background information on the disease, or relating to policy or research articles), warning (articles that suggest an outbreak or unusual case load is likely in the near future), not-disease related (false-positives resulting from the disease term classification system), and old news (referencing previous outbreaks). Duplicate articles, such as identical reports issued by different news services, are identified and processed by the system by aggregating contemporary articles together and checking for similarities. Later human review of the automated assignments ensures data quality and identifies opportunities to improve the automated process. Articles included in our dataset were those categorised as 'breaking' with the disease tag 'leishmaniasis' and totalled some 109 reports.

Geo-positioning of data

Each article sourced from the literature search was read and the geographic coordinates of the cases described in each article manually extracted. In many cases, either due to multiple places sharing the same name or to differences in spelling of place names (particularly when translating between different languages), additional contextual information from the article was required for accurate positioning. Each case was reported to the highest degree of spatial resolution available based upon the information provided. This ranged from point locations (indicative of a precise location, such as a village), to areas, termed polygon locations, which correspond approximately to districts (specifically an Admin 2 unit as classified by the Food and Agriculture Organization's Global Administrative Unit Layers (GAUL) coding[13]), and areas corresponding to states or provinces (the first national subdivision, GAUL Admin 1). For cities or towns, the coordinates of the centre were recorded, unless a specific part of the city (or an explicit latitude and longitude) was described. In the case of district and provincial level data, the approximate centroid of the polygon (obtained using Google Maps; https://maps.google.co.uk) was recorded, to allow for consistent identification in subsequent analyses.

Montpellier's CNR-L strain archive has geographic tags associated with it (provided by the original survey teams), and those that included at least a sub-national identifier (ranging from villages to provinces) were geo-positioned using Google Maps by the same rules as with the literature-sourced occurrences. GenBank archived material that had associated spatial tags was similarly geopositioned.

HealthMap automatically geopositions the data it parses using a custom-built gazetteer[12]. An algorithm searches for matches between the online article and language-specific terms linked to locations in the geographic reference database, and therefore can identify potential location tags. These matches are then evaluated by a ruleset that attempts to determine the relevance of each location tag (i.e., to distinguish publication relevant tags from outbreak and disease relevant locations) by assessing the number and position of the location keywords within the text. Certain common misleading phrases, such as 'Brazil nut' and 'guinea pig' are also accounted for in this phase.

Occurrence standardisation

As the occurrence database was derived from a wide range of disparate sources, we attempted to standardise the occurrence reports both spatially and temporally. Firstly, using the centroid coordinates, all polygon-level data was assigned to a specific admin unit (as defined by GAUL). Point-level data was similarly aligned to a 5×5 km pixel gridded surface of the world. For occurrences that spanned several years, these were disaggregated so that a location that reported cases from 2000 to 2003 (for example) would represent four separate occurrences, one for each year. Occurrence records which were duplicates in both space (either per pixel or per polygon) and time (per calendar year) were removed from the database thus removing the potential for duplicated reports of the same cases. As such, a unique record in the database is defined as the occurrence of one or more cases in a given location in a specific year. Therefore, irrespective of the number of cases or reports that occur within a given polygon or pixel in a given year, they are summarised as one unique occurrence, i.e., an area with 500 cases in a given year is equivalent to an area with 5 cases. By focussing on using unique occurrences, we avoided issues of bias associated with oversampling caused by higher reporting in one location over another, and thus gather a more accurate picture of the global distribution of these diseases, which is particularly important when using presence-absence based modelling techniques.

Data Records

The database associated with this article (available via Dryad (Data Citation 1) contains the following fields:

1. OCCURRENCE_ID: a unique identifier assigned to each unique occurrence of disease.
2. SOURCE_TYPE: indicating whether sourced from literature, Montpellier CNR-L strain archives, GenBank or HealthMap.
3. LOCATION_TYPE: indicative of point or polygon level data.
4. ADMIN_LEVEL: the administrative level associated with the occurrence. Values are 1 (state or province), 2 (district) and -999 (point).
5. X: the longitudinal coordinate of the point or polygon centroid in decimal degrees (WGS1984 Datum).
6. Y: the latitudinal coordinate of the point or polygon centroid in decimal degrees (WGS1984 Datum).
7. YEAR: the year of occurrence.
8. COUNTRY: the name of the country within which the occurrence lies.
9. DISEASE: indicating whether cutaneous or visceral form of the disease.

Technical Validation

All occurrences were compared to standardised 5×5 km pixel grids of the world and checked that they fell on land. Those not on land were automatically repositioned to the nearest land pixel. In addition, all occurrences were referenced with the evidence consensus map produced by Pigott *et al.*[6] This map is a quantitative measure of the consensus of various sources of evidence for the presence or absence of

leishmaniasis within that province, scored on a scale from +100 to −100. Importantly, the dataset used to generate the evidence consensus scores included a wider range of data sources (e.g., health organisation statuses and national-level case data) and included data that did not meet our geographic precision inclusion criteria for the occurrence dataset as described below. Occurrences that fell in regions where there was a consensus on disease absence (between −25 and −100) were manually double-checked by the authors to ensure that the latitude and longitude were correct and that the corresponding articles reported a genuine occurrence of leishmaniasis. Finally, any polygon greater in area than one squared degree was removed from the database, as the generalisation of modelling covariates at this scale would be misleading and present a possible source of bias in the model.

The final validated database consisted of 12,563 unique occurrences (6,426 for cutaneous leishmaniasis and 6,137 for visceral leishmaniasis) with source locations and occurrence types listed in Table 1. Figs 1 and 2 show the cumulative distributions of unique occurrence records of cutaneous and visceral leishmaniasis over time, subdivided by continent.

Usage Notes

This dataset can be used to investigate the spatial epidemiology of leishmaniasis at a range of scales, from sub-national studies to regional assessments.

In addition, the framework outlined here can easily be extended to a variety of other diseases, as has already been demonstrated with dengue[14], and given the potential to automate some of these methods further[15], the entire process should become increasingly easier and timely. An initial analysis of this dataset was within a niche modelling framework assessing the global distribution of the leishmaniases[6]. The code used to carry out technical validation of the dataset and to generate the predictive risk maps presented in Pigott *et al.*[6] is freely available as an R software package seegSDM from GitHub (https://github.com/SEEG-Oxford/seegSDM) and is accompanied by a tutorial for its use.

This dataset could be used to assess regional and national variation in disease reporting rates by comparing the density and distribution of occurrence data with existing estimates of case numbers. In areas where there is a lack of regular reporting, occurrences may indicate a potential cryptic burden of disease. Since much of this data is derived independent of centralised healthcare provisioning, it can act as a secondary indicator of disease presence. In addition, the data can be considered in the context of burden estimation, particularly in helping to delineate the spatial limits of leishmaniasis risk.

Despite its comparative ubiquity with respect to other forms of epidemiological data, such as disease prevalence, one of the main limitations of disease occurrence data is the potential for sampling bias[16,17], where disease occurrence is more likely to be reported from some areas (e.g., those with strong healthcare systems) than others. Consequently, analyses of these data must account for this

	Point data	Admin 1 data	Admin 2 data	Total
Origin and resolution of occurrence data				
Cutaneous Leishmaniasis				
Literature	3,680	879	1,220	5,779
CNR-L (Montpellier)	531	47	31	609
HealthMap	31	—	—	31
GenBank	6	—	1	7
Total	4,248	926	1,252	6,426
Visceral Leishmaniasis				
Literature	3,050	1,500	1,068	5,618
CNR-L (Montpellier)	429	24	29	482
HealthMap	32	1	—	33
GenBank	3	—	1	4
Total	3,514	1,525	1,098	6,137

Table 1. Origin and spatial resolution of leishmaniasis occurrence data (reproduced from Pigott *et al.*[6]).

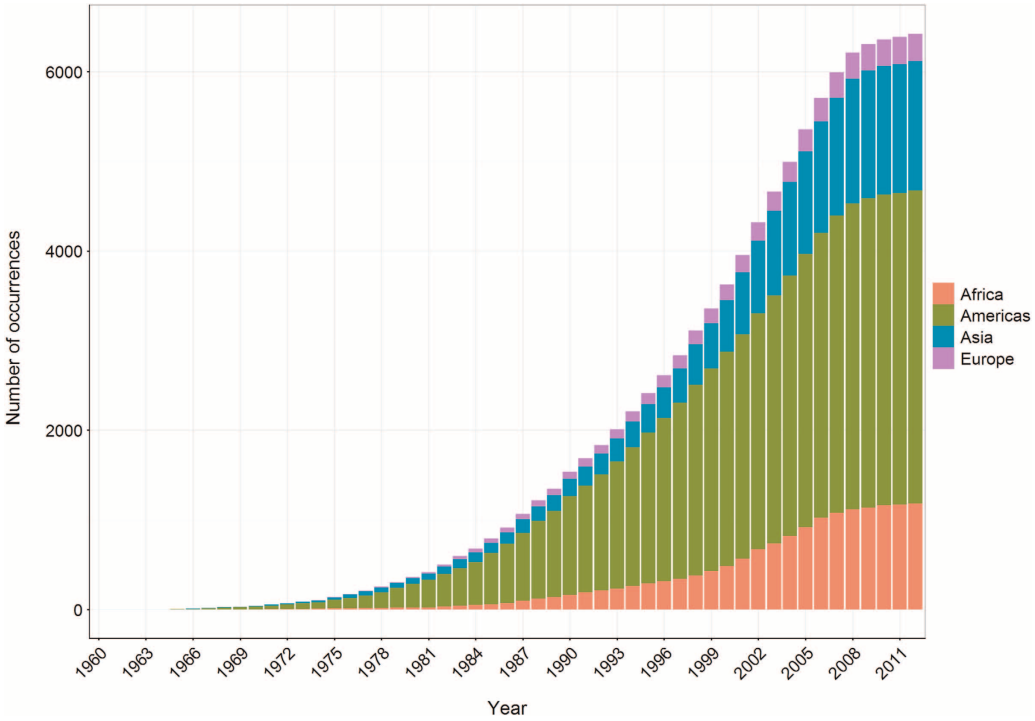


Figure 1. The cumulative number of unique cutaneous leishmaniasis occurrence records per year, from 1960 to 2012, coloured by region (red = Africa, green = Americas, blue = Asia, purple = Europe).

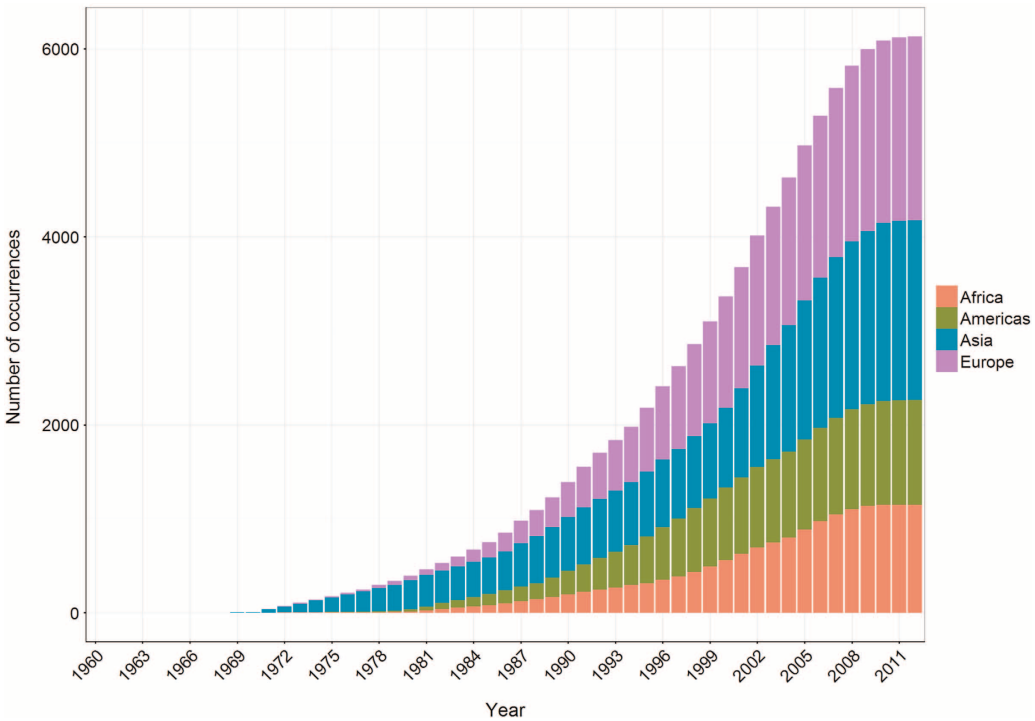


Figure 2. The cumulative number of unique visceral leishmaniasis occurrence records per year, from 1960 to 2012, coloured by region (red = Africa, green = Americas, blue = Asia, purple = Europe).

bias in order to avoid biased conclusions. The importance of this bias is very much dependent on the intended use of the data and a variety of techniques have been developed to assess and accommodate this[17–19]. Pigott *et al.*[6] mitigated this issue in their predictive mapping study by including an 'evidence consensus' map which accounted for reporting rates at a sub-national level. This problem is not only restricted to data from the published scientific literature; data from HealthMap and other internet-based resources are also likely to be influenced by variation in internet usage patterns and online engagement[20].

References

1. Banuls, A. L. *et al.* Clinical pleiomorphism in human leishmaniases, with special mention of asymptomatic infection. *Clin. Microbiol. Infect.* **17,** 1451–1461 (2011).
2. Lozano, R. *et al.* Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380,** 2095–2128 (2012).
3. Murray, C. J. L. *et al.* Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380,** 2197–2223 (2012).
4. Bern, C., Maguire, J. H. & Alvar, J. Complexities of assessing the disease burden attributable to leishmaniasis. *PLoS Negl. Trop. Dis.* **2,** e313 (2008).
5. Alvar, J. *et al.* Leishmaniasis worldwide and global estimates of its incidence. *PLoS One* **7,** e35671 (2012).
6. Pigott, D. M. *et al.* Global distribution maps of the leishmaniases. *eLife* **3,** e02851 (2014).
7. Hendrickx, D., Dujardin, J. C., Pickering, J. & Alvar, J. The leishmaniasis e-compendium: a geo-referenced bibliographic tool. *Trends Parasitol.* **26,** 515–516 (2010).
8. Pratlong, F. *et al.* Geographical distribution and epidemiological features of Old World cutaneous leishmaniasis foci, based on the isoenzyme analysis of 1048 strains. *Trop. Med. Int. Health* **14,** 1071–1085 (2009).
9. Pratlong, F. *et al.* Geographical distribution and epidemiological features of Old World *Leishmania infantum* and *Leishmania donovani* foci, based on the isoenzyme analysis of 2277 strains. *Parasitology* **140,** 423–434 (2013).
10. WHO. *Control of the Leishmaniases. Report of a Meeting of the WHO Expert Committee on the Control of Leishmaniases, Geneva*, 22-26 March 2010. (World Health Organization, 2010).
11. Madoff, L. C. ProMED-mail: An early warning system for emerging diseases. *Clin. Infect. Dis.* **39,** 227–232 (2004).
12. Freifeld, C. C., Mandl, K. D., Ras, B. Y. & Brownstein, J. S. HealthMap: global infectious disease monitoring through automated classification and visualization of internet media reports. *J. Am. Med. Inform. Assn.* **15,** 150–157 (2008).
13. Food and Agriculture Organization of the United Nations. *The Global Administrative Unit Layers (GAUL): Technical Aspects* (Food and Agriculture Organization of the United Nations, EC-FAO Food Security Programme (ESTG), 2008).
14. Messina, J. P. *et al.* A global compendium of human dengue virus occurrence. *Sci. Data* **1,** 140004 (2014).
15. Hay, S. I., George, D. B., Moyes, C. L. & Brownstein, J. S. Big data opportunities for global infectious disease surveillance. *PLoS Med.* **10,** e1001413 (2013).
16. Syfert, M. M., Smith, M. J. & Coomes, D. A. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLoS One* **8,** e55158 (2013).
17. Boria, R. A., Olson, L. E., Goodman, S. M. & Anderson, R. P. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecol. Model.* **275,** 73–77 (2014).
18. Elith, J. *et al.* Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29,** 129–151 (2006).
19. Lobo, J. M. & Tognelli, M. F. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. *J. Nat. Conserv.* **19,** 1–7 (2011).
20. Graham, M., Hogan, B., Straumann, R. K. & Medhat, A. Uneven geographies of user-generated information: patterns of increasing informational poverty. *Ann. Assoc. Am. Geogr.* **104,** 746–764 (2014).

Data Citation

1. Pigott, D. M. *et al. Dryad* http://doi.org/10.5061/dryad.05f5h (2014).

Acknowledgements

D.M.P. is funded by a Sir Richard Southwood Graduate Scholarship from the Department of Zoology at the University of Oxford. S.I.H. is funded by a Senior Research Fellowship from the Wellcome trust (095066) which also supports K.A.D. and K.E.B., N.G. is funded by a grant from the Bill & Melinda gates Foundation (#OPP1053338). J.P.M. is funded by, and S.I.H. acknowledges the support of, the International Research Consortium on Dengue Risk Assessment Management and Surveillance (IDAMS, European Commission 7th Framework Programme (#21803) http://www.idams.eu). Y.B., F.P. and P.B. acknowledge financial support by the French Institut de Veille Sanitaire (InVS). J.S.B., S.R.M. and C.F. acknowledge the funding from NIH National Library of Medicine (R01LM010812). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. S.I.H. had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Author Contributions

D.M.P. drafted the manuscript with editing and approval from all authors. D.M.P., K.E.B., K.A.D. and M.F.M. performed the literature survey for occurrences. Y.B., P.B. and F.P. are responsible for curating and providing the Montpellier strain archive. J.S.B., C.F., S.R.M. and L.C.M. are responsible for the various HealthMap datasets. D.M.P., N.G. and J.P.M. performed database standardisation and technical validation. D.B.G. and S.I.H. conceived the database design and advised on standardisation and validation procedures.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Pigott, D. M. *et al.* Global database of leishmaniasis occurrence locations, 1960–2012. *Sci. Data* 1:140036 doi: 10.1038/sdata.2014.36 (2014).

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Viral infection

» Ecological epidemiology

» Ebola virus

A comprehensive database of the geographic spread of past human Ebola outbreaks

Adrian Mylne[1,*], Oliver J. Brady[1,*], Zhi Huang[1], David M. Pigott[1], Nick Golding[1], Moritz U.G. Kraemer[1] & Simon I. Hay[1,2]

Received: 15 September 2014

Accepted: 06 October 2014

Published: 23 October 2014

Ebola is a zoonotic filovirus that has the potential to cause outbreaks of variable magnitude in human populations. This database collates our existing knowledge of all known human outbreaks of Ebola for the first time by extracting details of their suspected zoonotic origin and subsequent human-to-human spread from a range of published and non-published sources. In total, 22 unique Ebola outbreaks were identified, composed of 117 unique geographic transmission clusters. Details of the index case and geographic spread of secondary and imported cases were recorded as well as summaries of patient numbers and case fatality rates. A brief text summary describing suspected routes and means of spread for each outbreak was also included. While we cannot yet include the ongoing Guinea and DRC outbreaks until they are over, these data and compiled maps can be used to gain an improved understanding of the initial spread of past Ebola outbreaks and help evaluate surveillance and control guidelines for limiting the spread of future epidemics.

Design Type(s)	observation design • epidemiological study • data integration
Measurement Type(s)	Viral Epidemiology
Technology Type(s)	data collection method
Factor Type(s)	emergence and spread locations
Sample Characteristic(s)	Ebolavirus • South Sudan • Sudan • Congo, the Democratic Republic of the • Cote d'Ivoire • Gabon • Uganda • Congo • anthropogenic habitat

[1]Spatial Ecology & Epidemiology Group, Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK. [2]Fogarty International Center, National Institutes of Health, Bethesda, MD 20892-2220, USA. *Joint first authors. Correspondence and requests for materials should be addressed to S.I.H. (email: simon.hay@zoo.ox.ac.uk)

Background & Summary

The genus *Ebolavirus* belongs to the family *Filoviridae* and contains five known species to date that vary in their distribution, reservoir hosts and their pathogenicity to humans. With the exception of *Reston ebolavirus* which has only shown pathogenicity among primates and porcids, all of these have shown some capacity to spill over from their natural reservoirs and cause human cases[1,2]. While only a single human case of *Tai Forest ebolavirus* has been documented[3], the remaining three species (*Zaire ebolavirus*, *Sudan ebolavirus* and *Bundibugyo ebolavirus*) are all known to permit human to human transmission after the initial suspected zoonotic transfer resulting in outbreaks of different sizes, geographic extents and case fatality rates (Fig. 1)[4,5].

Which reservoir species are responsible for maintaining Ebolavirus transmission between outbreaks is not well understood[6], but several candidate species have been identified. In Gabon three species of bats (*Hypsignathus monstrosus*, *Epomops franqueti* and *Myonycteris torquata*) were found to be infected with Ebola virus[7] and some human outbreaks have been directly linked to bat exposure[8]. While it is increasingly clear that gorillas (*Gorilla gorilla*) and chimpanzees (*Pan troglodytes*) are dead end hosts for the virus, infection in populations of these species is frequently found and they still present a risk of animal to human transmission[3,9–14]. For many outbreaks it has been difficult to definitively identify the source of human Ebola index cases, but activities that bring humans into close contact with the blood of mammals through activities such as hunting and the bushmeat trade are common to many of the index cases[8,14–16].

Following the initial suspected zoonotic transfer, secondary transmission can result from close contact between infectious individuals or corpses and other humans, usually through exposure to infectious bodily fluids[2]. Due to the close degree of contact required for secondary transmission, certain specific community activities are commonly associated with hotspots of secondary transmission such as family home care, traditional burial practices that involve washing the corpse, or healthcare settings where sufficient protective measures are not in place[15,17,18]. These focal transmission events combined with an incubation period of 5–9 days mean transmission can often be observed in waves of cases[2,18] in both space and time.

Spread of cases over longer distances is often associated with treatment seeking that draws people from rural villages that typify the index case locations to big urban centres with central medical facilities (Supplementary Figures 1–22). While this mostly involves domestic land travel[19], some instances of international importation by air travel have been documented[18]. Following travel of an infectious individual, either secondary clusters of Ebola cases will occur, or transmission will be interrupted by control methods such as quarantine and patient contact tracing[20,21]. Due to the variable rate of progression of symptoms of Ebola virus disease (EVD) (onset of Ebola Haemorrhagic fever can range from 2–21 days[2]), case fatality rates can vary significantly depending on a number of factors associated

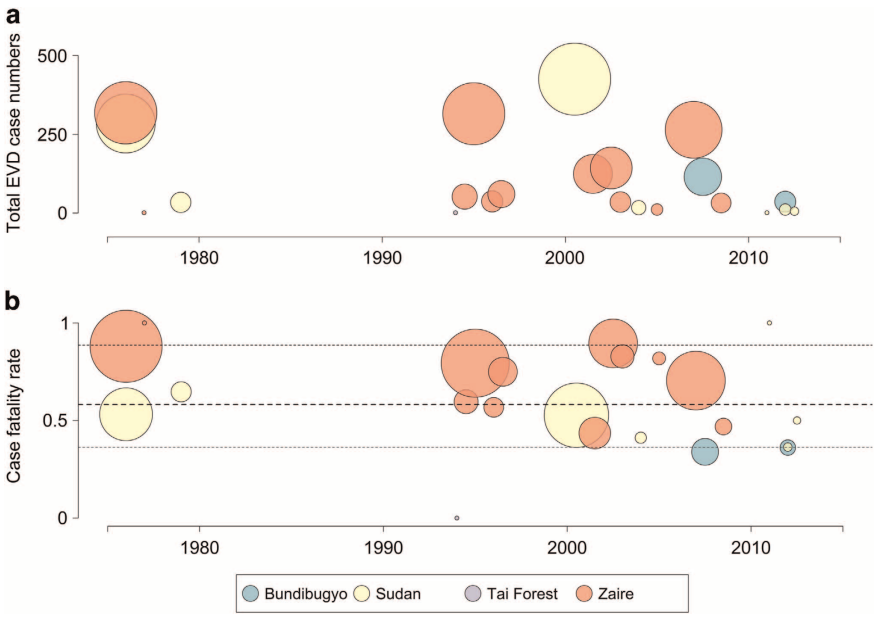


Figure 1. The size (**a**) and case fatality rate (**b**) of the 22 previous Ebola outbreaks (suspected and confirmed cases). Circle area is proportional to the total number of reported cases (**a**) or deaths (**b**) for each outbreak. Circle colour represents different species of Ebolavirus. Black dotted lines in (**b**) show the median and upper and lower 75% quantiles of outbreak case fatality rate.

with Ebola virus pathogenesis and the quality and timing of symptomatic care. The complex interaction between surveillance, control, treatment seeking, patient-contact rates and their combined effects on the dynamics of transmission dictate the spread, magnitude and case fatality rate of an Ebola outbreak.

This database collates existing knowledge on the geographic spread of past Ebola outbreaks in a standardised format that allows the dynamics of different outbreaks to be compared. Procedures for data abstraction are outlined and each outbreak is summarised with a map and brief text description. These data will be useful for conducting spatial analyses of Ebola outbreak spread. We include every outbreak preceding the atypical 2013 Guinea epidemic which has spread further and faster than any previous epidemic. Once the current Guinea and Democratic Republic of the Congo (DRC) outbreaks are over, this database will be updated to include the same standardised data fields for these contemporary outbreaks. Periodic updates to include any additional Ebola outbreaks will also ensure this resource has on-going relevance in Ebola spread analyses. In particular, a comparison between the Guinea 2013 outbreak and historical outbreaks will allow an evaluation of surveillance and control guidelines in terms of their appropriateness for mitigating the spread of future Ebola outbreaks of variable magnitude. In the meantime, it is hoped that these data will support research into EVD epidemiology which can be brought to bear on the current outbreak.

Methods

Data collection

Tables listing proven outbreaks of Ebola virus, sourced from the scientific literature[5] and from health reporting organisations[22], were used to coordinate initial searches of the formal scientific literature using Web of Science and PubMed for each specific outbreak. Relevant papers were abstracted and, where possible, outbreak-specific epidemiological surveys were sourced. The citations in these references were obtained in order to reconstruct the outbreak in detail and extract a range of epidemiological data relating to geographic spread, case and fatality numbers.

In this analysis we excluded ongoing outbreaks meaning that the current Guinea and DRC spread data are not yet included. When this outbreak is over and data has been assimilated we will update this database to incorporate the Guinea and DRC outbreaks using the same procedures described below, which will allow a comparison with historical outbreaks of Ebola.

Definitions of index, secondary and imported cases

Index cases were defined as any human infection resulting from interaction with non-human sources of the disease. Among the sources examined, index cases were identified based on reported interaction with suspected zoonotic reservoirs or hosts, such as non-human mammals during hunting trips[8,14–16,23]. In cases where a mode of suspected zoonotic transmission could not be established, the first reported case was assumed to be the index case. Any cases arising from existing human infections are considered as secondary infections. Cases reported after the index cases were assumed to be secondary cases unless they were accompanied by specific details of likely exposure to a zoonotic reservoir or non-human host. If a case was reported in an area where no further cases occurred and no continued transmission was documented, these were termed imported cases.

Procedure for geo-positioning

For each index, secondary or imported case cluster that could be linked to a unique geographic location we performed a range of procedures similar to methods employed elsewhere[24,25] to assign geographic coordinates. For index and secondary cases these locations were representative of the site of suspected zoonotic transfer or human-to-human transmission respectively. Index cases were geopositioned as the location where exposure to the suspected zoonotic reservoir was likely to be highest. For example hunters who butchered carcases on hunting expeditions were geopositioned as a polygon covering the area of the hunting trip not a point at the hunters' homes. In contrast, if an individual purchased bushmeat from a local market for preparation and consumption at home, the home of the individual was georeferenced as the index case, as it was considered the location of highest exposure to the suspected zoonotic reservoir. For imported cases these locations were representative of all locations that infected patients travelled to, but did not cause onward transmission. For the purpose of this analysis we excluded international imported cases that did not cause autochthonous secondary transmission as, in most cases, they represented diagnosed foreign workers or expatriates who were evacuated for specialised treatment[3,10]. In such circumstances appropriate preventative measures are employed meaning that risk of onward transmission is negligible.

Each occurrence was reported to the highest degree of spatial resolution available based upon the information provided, as long as they could be categorised into one of: index, secondary or imported cases. This ranged from point locations (indicative of a precise location, such as a village), to areas, termed polygon locations, which correspond approximately to administrative regions or custom digitised areas based on site descriptors within the primary articles. Administrative regions were defined as classified by the Food and Agriculture Organization's Global Administrative Unit Layers (GAUL) coding[26]. These classify national boundaries as admin0 units, states or provinces as admin1 units and districts as admin2 units. By classifying Ebola occurrences as polygons we were able to represent the geographic uncertainty around the exact location of Ebola transmission which could have occurred anywhere within the defined

region. For towns, the coordinates of the centre were recorded, unless a specific part of the town (or an explicit latitude and longitude) was described. Coordinates for point locations were extracted using Google Earth (version 7.1.2). If the area concerned could not be assigned to a finer resolution than 5 km × 5 km, it was entered as polygon rather than point data. If specified regions could not be linked to an admin2 or admin1 unit, custom polygons were digitised using site descriptions in the text articles. For imprecise descriptors e.g., '15 km from the town' with no direction specified or 'cases occurred on a north-south road between village X and Y', circular polygons were digitised based on radius distances given or extreme points that defined the geographic limits of transmission. These circular polygons could be trimmed if their area included admin1 or admin2 administrative regions which reported no EVD patients. Some articles referred to 'healthcare districts' that did not correspond to admin1 or admin2 units, but were definable based on maps presented in the primary literature that were digitised or were available on the map sharing website IKI (www.ikimap.com)[27]. For index cases that referred to suspected zoonotic transfer in specific forests or game reserves, polygons were drawn based on the specified park or forest geographic boundary as shown in Google Earth. Two exceptional cases were present. In the first, for outbreak 13, the index case transmission site description merely mentioned a case being reported as 'near the town of Mbandza'. In this instance a circular polygon was defined with radius of half the distance to the next specified location of transmission (7.5 km)[15]. In the second, for outbreak 9, two locations described in the primary literature could not be located, but were described as 'near to the town of Booue'[18]. As a result the same procedure was undertaken and a radius of 30 km was defined around the village of Booue. All digitising was performed using ArcGIS 10.1[28].

Key outbreak metrics recorded

The total number of cases for each outbreak was obtained from the most recent primary source (Table 1 (available online only)). Cases included both clinically suspected and laboratory confirmed cases at the point of care, or diagnosed retrospectively. The number of people who died with a suspected or confirmed diagnosis of EVD was also recorded. These data were spatially disaggregated as much as possible from the information given in the text to give measures of spatial variation in case fatality rate within an outbreak.

Outbreak start and end time were defined by reports of the first (index) and last cases respectively. For each secondary and imported case that occurred in a novel location, the source of importation was recorded where reported. When possible this was reconciled with the date of the first secondary or imported case in each cluster to determine geographic spread order. We only included confirmed sources of origin, as determined by detailed patient histories, not suspected sources and no differentiation of spread order to two new clusters from the same source was made unless both new clusters documented the date of their first secondary case. If two secondary case clusters had a confirmed source but no differentiation of order we assigned them the same spread order. Similarly if a secondary case cluster could have come from two specified sources both were included as sources of spread. Due to a lack of epidemiological investigation at the time, or through information being lost in the reporting chain, the spread order of many secondary and imported case clusters was partially complete or missing altogether (Table 2).

For each outbreak the above details were combined with additional information from the original articles on method and timing of spread to construct the text descriptions accompanying the spread maps in Supplementary Figures 1–22 (selected examples in Figs 2,3,4). Spread is defined as movement of an infected individual from the location of infection to a transmission-free area. For example an individual who was likely infected from the suspected zoonotic reservoir on a hunting expedition in the forest then caused secondary infections back in their local village, would qualify as having spread the virus from the forest to the village. By contrast if the first reported case occurred in a village where subsequent secondary cases then occurred, no spread would have been recorded as index and secondary transmission occur in the same location.

	Spread order (%)	Source (%)	Onset timing (%)	End timing (%)	Case data (%)
Complete	51	69	32	6	27
Partial	44	2	41	0	21
Unknown	5	29	27	94	52

Table 2. Completeness of epidemiological details among the 117 Ebola virus transmission occurrences. Spread order is considered complete if the order of each occurrence in their respective outbreaks is known, partial if their order could not be distinguished from every occurrence in each outbreak and unknown if their order was unknown in the outbreak. Source was considered complete if spread could be linked to one source occurrence, partial if spread is known to have come from one of a number of source occurrences and unknown if source was unidentifiable (index cases excluded). Both onset and end timing were considered complete if day, date and year was known, partial if just month and year was known. Case data was considered complete if the number of cases and deaths was known, partial if either cases or deaths was known.

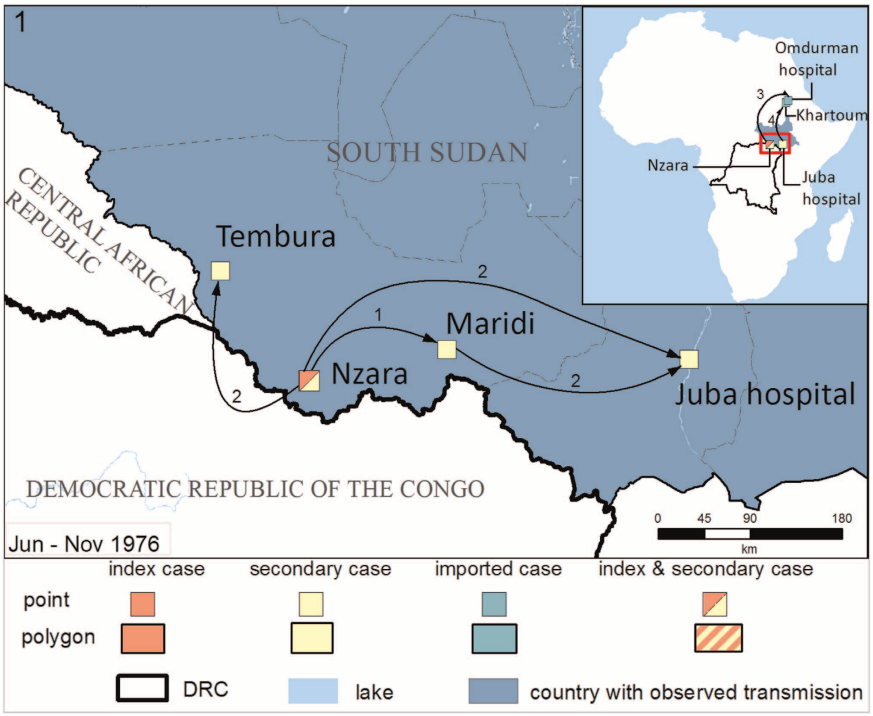


Figure 2. Map of the South Sudan (1976) outbreak. The first reported cases of Sudan Ebola virus were in three workers at a cotton factory in Nzara, in close proximity to three game reserves. The method of acquisition was unknown. The first secondary cases arose in Nzara infecting a total of 67 people who were primarily family members of the factory workers. Further secondary transmission clusters emerged in Maridi following spread from Nzara due to seeking treatment, after which further cases occurred in Juba due to patients who were referred. Additional cases from Maridi were also referred directly to Juba making the source of infection in Juba difficult to identify. Secondary transmission also emerged in Tembura due to a patient seeking family care, although the source of this infection is unknown. Imported cases from Juba to Khartoum and from Nzara to Omdurman were also reported following a patient seeking treatment and a referral for diagnosis respectively (see inset). The principal mode of transmission in this outbreak was initially familial, although in Maridi secondary transmission arose through nosocomial transmission. Seeking of treatment was the principal cause of geographic spread. The first index case became ill on the 27 June 1976 before the first secondary cases in July and subsequent secondary transmission clusters from August to October. Cases peaked in September (138 cases, 65 deaths). The final case was reported on 25 November 1976. Imported cases in Omdurman and Khartoum were reported in August and September, respectively. Overall, 284 cases were reported with 151 deaths giving a CFR of 53.2%. This figure varied in different locations: Nzara (67,31,46%), Maridi (213,116,55%), Tembura (3,3,100%), Juba (1,1,100%). Arrows indicate order of spread. Where spread order is known, numbers are indicative of the order of spread. Arrows sharing the same number indicate that it was not possible to distinguish which spread happened first.

Selected examples are presented in Figs. 2,3,4. In the first, initial focal infection jumped from village to village, as well as longer distance dispersal primarily through treatment seeking of infected patients (Fig. 2). In the second (Fig. 3) more limited outbreak, cases occurred in two main clusters, with one containing working camps surrounding the site of suspected zoonotic transfer and the other villages and a local hospital in the vicinity of Makokou. In the final example outbreak (Fig. 4) secondary transmission spread radially out from the village of Balimba following treatment seeking of a faith healer. Similar to outbreak 1, infected patients travelled to local and national healthcare centres for treatment, in some cases causing secondary transmission at great distances from the index case.

In the text descriptions, methods, sources and order of spread were described in a standardised manner, where information allowed. A brief summary of the evolution of the outbreak over time and the breakdown of total number of cases, deaths and case fatality rate over the course of the entire outbreak is also given.

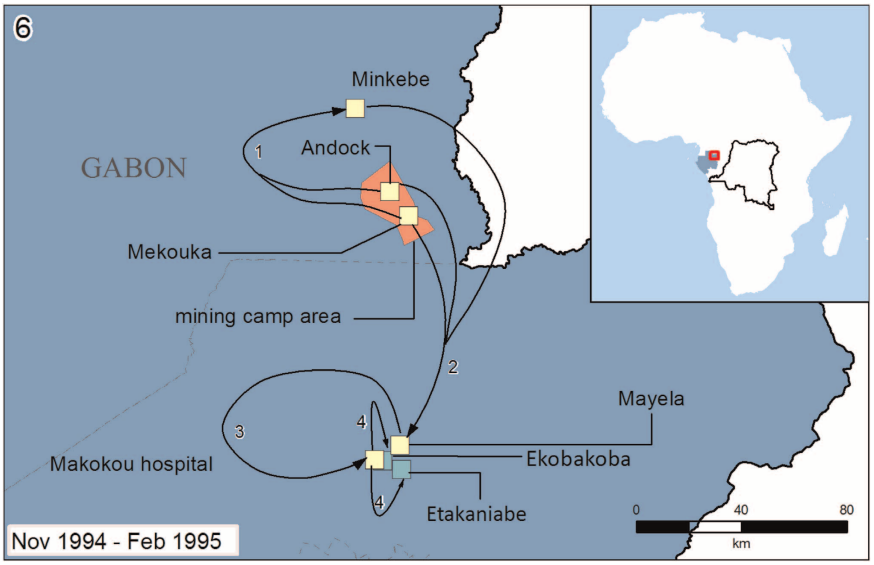


Figure 3. Map of the Gabon (1994) outbreak. The first reported cases of Zaire Ebola virus were in miners from the Mekouka and Andock encampments, suspected to have contracted the infection in the surrounding area. The method of acquisition was unknown. The first secondary cases arose within these two encampments and then spread to the Minkebe camp. Further secondary transmission clusters emerged in Mayela then Makokou general hospital after 32 patients from the forest encampments sought treatment. Cases were also reported in Ekataniabe and Ekobakoba who had recent travel histories to Makokou general hospital. The principal modes of transmission were among workers at first, followed by nosocomial in Makokou general hospital and familial in Mayela (connected by a single traditional healer). The initial case was reported on 13 November 1994 before secondary transmission clusters occurred from the end of January to February 1995. Cases and deaths peaked in December (26 cases, 14 deaths (53.8% CFR)). The final case was reported on 9 February 1995 in Ekobakoba. Overall, 49 cases were reported with 30 deaths, giving a CFR of 61.2%. For map key, see Fig. 2.

A wider-scale map showing the locations of all the outbreaks since 1976 is also available in Supplementary Fig. 23 and a table detailing which type of data were obtained from which source for each outbreak is given in Table 1 (available online only)[3,8,10,13–18,20,29–46].

Data Records

The data from this analysis are summarised in two types of data format (Data Citation 1). First a data table details unique geographic locations of Ebola occurrence, including information on type of transmission, location, spread, timing and case number. These geographic locations are grouped into individual outbreaks ($n = 22$) and summary statistics on timings and case and death numbers are given for each outbreak. Second, geographic information files are provided that match the information presented in the data table to explicit geographic areas. These are available in a variety of formats that can be read by geographic information system (GIS) applications. Information from these two file types were used to make the outbreak summary maps and text in Supplementary Figures 1–22 (selected examples in Figs. 2,3,4).

Data table of unique Ebola virus transmission locations

The table includes the following fields, detailed below. The value 'NA' was entered if information was unknown, unreported or indeterminable. The term 'occurrence' refers to Ebola transmission locations that are either unique in the geographic location or in their type of transmission (index, secondary or imported). Each row in the table represents a unique Ebola occurrence. A group of occurrences make up contained 'outbreaks' and fields with the prefix 'OB' summarise various metrics related to the entire outbreak that each occurrence belongs to.

UNIQ_ID: A unique identification number for each occurrence at which index, secondary or imported cases of EVD have occurred at unique geographic locations ($n = 117$).

NAME: Text description of the point or polygon that defines the location of the occurrence.

COUNTRY: The country where the majority of cases occurred in each outbreak.

VIRUS: The Ebola virus species of each outbreak.

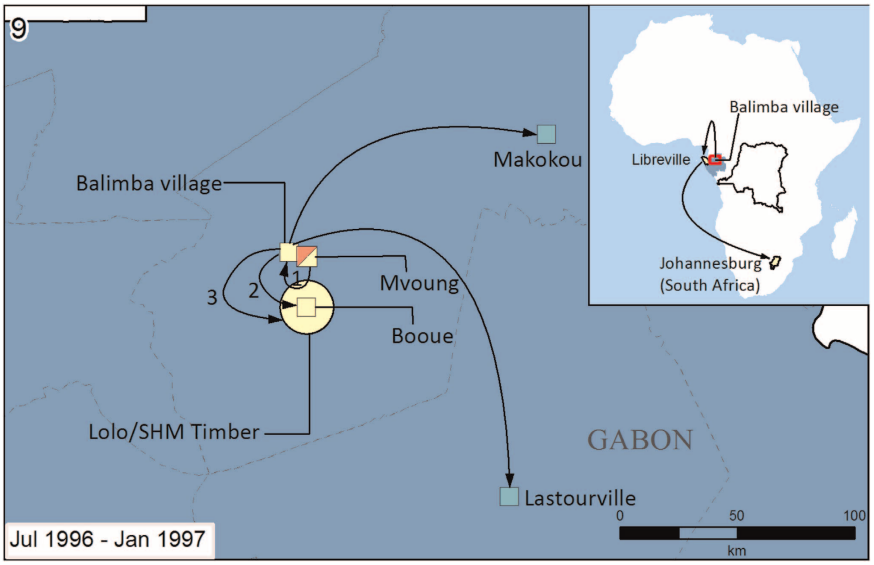


Figure 4. Map of the Gabon (1996b) outbreak. The index case of Zaire Ebola virus likely came from one of three infected hunters in a logging camp near Mvoung. The timing of infection makes it difficult to distinguish index cases from secondary cases during the early stages of this outbreak, but it is likely that the first secondary cases emerged amongst the hunters who then sought treatment from a traditional healer in Balimba. After falling ill, the traditional healer from Balimba sought treatment in Booue, where the disease then radially spread through the communities in the surrounding areas. A further secondary transmission cluster emerged in Libreville (see inset) after patients from Balimba sought treatment there. In Libreville one doctor became infected and flew to Johannesburg, South Africa for treatment before receiving a diagnosis of Ebola. Limited further nosocomial transmission (1 case) occurred upon his arrival in Johannesburg. Imported cases in Makokou General Hospital and Lastourville were also reported after patients from Balimba sought treatment. No clear principal mode of transmission was observed for the early stages of the outbreak, but in Libreville secondary transmission mainly arose through nosocomial transmission. The index case was reported on the 13 July 1976 before the first secondary cases in September and subsequent secondary transmission clusters from September to January. Cases peaked in September and deaths peaked in October. The final case was reported on 18 January 1997. Overall 60 cases were reported with 45 deaths, giving a CFR of 75%. For map key, see Fig. 2.

CASE_TYPE: The type of transmission represented by the Ebola occurrence. Can be either 'index', 'secondary' or 'import'.

DATA_TYPE: Whether the occurrence represents a point or larger polygon location.

LAT: The latitude of the centre point of the point or polygon of the occurrence.

LONG: The longitude of the centre point of the point or polygon of the occurrence.

LOC_NTS: Additional notes describing the site location of the occurrence.

SPR_ORDER: The order of spread between occurrences over the course of the outbreak, as determined by the date of onset of the first case in a given occurrence. Index cases are represented with the value '1'. Two or more occurrences share the same spread order if it is unknown which of the two areas Ebola virus transmission spread to first.

SOURCE_1: The unique identification number of the occurrence where the first EVD patient came from.

SOURCE_2: The unique identification number of the occurrence where the first EVD patient came from. An occurrence may have more than one source if infected patients came from more than one source but it is unknown which triggered secondary transmission.

SOURCE_3: The unique identification number of the occurrence where the first EVD patient came from. An occurrence may have more than one source if infected patients came from more than one source but it is unknown which triggered secondary transmission.

STR_DAY: Day of first reported case in the occurrence.

STR_MNTH: Month of first reported case in the occurrence.

STR_YEAR: Year of first reported case in the occurrence.

END_DAY: Day of last reported case in the occurrence.

END_MNTH: Month of last reported case in the occurrence.

END_YEAR: Year of last reported case in the occurrence.

REP_CASE: The total number of cases (suspected or confirmed) reported over the course of the outbreak, but only within the occurrence.

REP_DEATH: The total number of deaths (suspected or confirmed) reported over the course of the outbreak, but only within the occurrence.

OB_ID: A unique identification number for each outbreak ($n = 22$).

OB_STR_DAY: Day of first reported case of the outbreak.

OB_STR_MNTH: Month of first reported case of the outbreak.

OB_STR_YEAR: Year of first reported case of the outbreak.

OB_END_DAY: Day of last reported case of the outbreak.

OB_END_MNTH: Month of last reported case of the outbreak.

OB_END_YEAR: Year of last reported case of the outbreak.

OB_CASE: The total number of cases (suspected or confirmed) reported over the course of the outbreak in all areas.

OB_DEATH: The total number of deaths (suspected or confirmed) reported over the course of the outbreak in all areas.

Geographic information files

The geographic information files include index, secondary and imported cases linked to geographic locations. All fields match those in the data table. Unknown, unreported or indeterminable data are represented by the character combination 'NA'.

Technical Validation

For each outbreak, all relevant text articles were used to confirm or reach a consensus on the likely result. Where differing results were found (e.g., total case count), figures from primary research articles took preference over review, summary articles or periodical epidemiological reports.

The extracted geographic and epidemiological summary information was cross-checked by at least two different researchers to ensure accuracy.

All point and polygon locations were checked in ArcGIS 10.1[28] against national and subnational boundaries to ensure their location matched the text descriptions. A gridded raster file from the Global Lakes and Wetlands Database[47] giving the locations of rivers and lakes was also included to check points and polygons fell on land rather than water. Any points that fell in water were moved to the nearest land pixel.

Usage Notes

The data presented here can be used in combination with spatial and temporal meteorological and socioeconomic information to generate hypotheses about the factors that may be important in the emergence of index cases and the spread of secondary and imported cases of EVD.

Pigott *et al.* combined data on Ebola index cases with a suite of environmental information in a species distribution model to map the zoonotic niche of Ebola transmission across Africa[48]. Matching more varied and finer scale local information at the sites of the index cases presented here could help develop our understanding of the complex process of Ebola virus emergence and the risk posed by certain human activities and land use patterns.

Using the secondary and imported case data, it would be possible to model and investigate causes of spread of human Ebola outbreaks. Understanding spread of the pathogen in past outbreaks may aid control of the current outbreak. A comparison between the spread of historical outbreaks and the spread of the current ongoing outbreak once it is over will be useful for informing Ebola surveillance and control guidelines to minimise the size and burden of these sporadic zoonoses.

Finally, this dataset can be used to investigate how the rate, extent and the environment in which Ebola outbreaks spread relates to important outbreak measures such as the total number of cases and case fatality rate. Understanding what distinguishes brief, geographically limited and low mortality Ebola outbreaks from those that impose a much higher public health burden will be important for informing future surveillance, control and treatment efforts[49]. This dataset provides the most comprehensive collection of standardised data on Ebola outbreak spread currently available and will be an important resource for these uses. These records will also be updated periodically with data from ongoing and future Ebola outbreaks to ensure the ongoing relevance of this database.

References

1. Barrette, R. W. *et al.* Discovery of swine as a host for the Reston ebolavirus. *Science* **325,** 204–206 (2009).
2. Feldmann, H. & Geisbert, T. W. Ebola haemorrhagic fever. *The Lancet* **377,** 849–862 (2011).
3. Le Guenno, B. *et al.* Isolation and partial characterisation of a new strain of Ebola virus. *The Lancet* **345,** 1271–1274 (1995).
4. King, A. M., Adams, M. J., Lefkowitz, E. J. & Carstens, E. B. *Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses* Vol. 9. (Elsevier, 2012).
5. Kuhn, J. & Calisher, C. H. *Filoviruses: A Compendium of 40 Years of Epidemiological, Clinical, and Laboratory Studies* 1st edn, Vol. 20 (Springer, 2008).
6. Peterson, A. T., Carroll, D. S., Mills, J. N. & Johnson, K. M. Potential mammalian filovirus reservoirs. *Emerg. Infect. Dis.* **10,** 2073–2081 (2004).
7. Leroy, E. M. *et al.* Fruit bats as reservoirs of Ebola virus. *Nature* **438,** 575–576 (2005).

8. Leroy, E. M. *et al.* Human Ebola outbreak resulting from direct exposure to fruit bats in Luebo, Democratic Republic of Congo, 2007. *Vector Borne Zoonot. Dis.* **9,** 723–728 (2009).
9. Bermejo, M. *et al.* Ebola outbreak killed 5000 gorillas. *Science* **314,** 1564–1564 (2006).
10. Formenty, P. *et al.* Ebola virus outbreak among wild chimpanzees living in a rain forest of Cote d'Ivoire. *J. Infect. Dis.* **179,** S120–S126 (1999).
11. Rouquet, P. *et al.* Wild animal mortality monitoring and human Ebola outbreaks, Gabon and Republic of Congo, 2001–2003. *Emerg. Infect. Dis.* **11,** 283–290 (2005).
12. Groseth, A., Feldmann, H. & Strong, J. E. The ecology of Ebola virus. *Trends Microbiol.* **15,** 408–416 (2007).
13. Milleliri, J., Tévi-Benissan, C., Baize, S., Leroy, E. & Georges-Courbot, M. Les épidémies de fièvre hémorragique due au virus Ebola au Gabon (1994-2002). *Bull. Soc. Pathol. Exot. Filiales* **97,** 199–205 (2004).
14. Nkoghe Mba, D. *et al.* Plusieurs épidémies de fièvre hémorragique due au virus Ebola au Gabon, d'octobre 2001 à avril 2002. *Bull. Soc. Pathol. Exot. Filiales* **98,** 224–229 (2005).
15. Boumandouki, P. *et al.* Prise en charge des malades et des défunts lors de l'épidémie de fièvre hémorragique due au virus Ebola d'octobre à décembre 2003 au Congo. *Bull. Soc. Pathol. Exot. Filiales* **98,** 218–223 (2005).
16. Nkoghe, D., Kone, M. L., Yada, A. & Leroy, E. A limited outbreak of Ebola haemorrhagic fever in Etoumbi, Republic of Congo, 2005. *Trans. R. Soc. Trop. Med. Hyg.* **105,** 466–472 (2011).
17. Baron, R. C., McCormick, J. B. & Zubeir, O. A. Ebola virus disease in southern Sudan - hospital dissemination and intrafamilial spread. *Bull. World Health Organ.* **61,** 997–1003 (1983).
18. Georges, A.-J. *et al.* Ebola hemorrhagic fever outbreaks in Gabon, 1994–1997: epidemiologic and health control issues. *J. Infect. Dis.* **179,** S65–S75 (1999).
19. Francesconi, P. *et al.* Ebola hemorrhagic fever transmission and risk factors of contacts, Uganda. *Emerg. Infect. Dis.* **9,** 1430 (2003).
20. Shoemaker, T. *et al.* Reemerging Sudan ebola virus disease in Uganda, 2011. *Emerg. Infect. Dis* **18,** 1480 (2012).
21. WHO. *Interim Manual - Ebola and Marburg Virus Disease Epidemics: Preparedness, Alert, Control, and Evaluation*, http://www.who.int/csr/disease/ebola/manual_EVD/en/ (2014).
22. Centers for Disease Control. *Known Cases and Outbreaks of Ebola Hemorrhagic Fever, in Chronological Order*, http://www.cdc.gov/vhf/ebola/resources/outbreak-table.html (2014).
23. WHO. Outbreak(s) of Ebola haemorrhagic fever, Congo and Gabon, October 2001–July 2002. *Wkly. Epidemiol. Rec.* **78,** 24 (2003).
24. Messina, J. P. *et al.* A global compendium of human dengue virus occurrence. *Sci. Data* **1,** 140004 (2014).
25. Pigott, D. M. *et al.* Global database of leishmaniasis occurrence locations, 1960–2012. *Sci. Data* **1,** 140036 (2014).
26. Food and Agriculture Organization of the United Nations. *The Global Administrative Unit Layers (GAUL): Technical Aspects*. Food and Agriculture Organization of the United Nations, EC-FAO Food Security Programme (ESTG), (2008).
27. IKI maps, http://www.ikimap.com/.
28. ArcGIS Desktop (Environmental Systems Research Institute, Redlands, CA, 2012).
29. WHO/International Study Team. Ebola haemorrhagic fever in Sudan, 1976. *Bull. World Health Organ.* **56,** 247–270 (1978).
30. International Commission. Ebola haemorrhagic fever in Zaire, 1976. *Bull. World Health Organ.* **56,** 271–293 (1978).
31. Heymann, D. *et al.* Ebola hemorrhagic fever: Tandala, Zaire, 1977–1978. *J. Infect. Dis.* **142,** 372–376 (1980).
32. Lamunu, M. *et al.* Containing a haemorrhagic fever epidemic: the Ebola experience in Uganda (October 2000–January 2001). *Int. J. Infect. Dis.* **8,** 27–37 (2004).
33. World Health Organization. Outbreak of Ebola haemorrhagic fever, Uganda, August 2000–January 2001. *Wkly. Epidemiol. Rec.* **76,** 41–48 (2001).
34. Okware, S. *et al.* An outbreak of Ebola in Uganda. *Trop. Med. Int. Health* **7,** 1068–1075 (2002).
35. World Health Organization. Outbreak(s) of Ebola haemorrhagic fever in the Republic of the Congo, January-April 2003. *Wkly. Epidemiol. Rec.* **78,** 285–289 (2003).
36. World Health Organization. Outbreak of Ebola haemorrhagic fever in Yambio, south Sudan, April-June 2004. *Wkly. Epidemiol. Rec.* **80,** 370–375 (2005).
37. Onyango, C. O. *et al.* Laboratory diagnosis of Ebola hemorrhagic fever during an outbreak in Yambio, Sudan, 2004. *J. Infect. Dis.* **196,** S193–S198 (2007).
38. Grard, G. *et al.* Emergence of divergent Zaire ebola virus strains in Democratic Republic of the Congo in 2007 and 2008. *J. Infect. Dis.* **204,** S776–S784 (2011).
39. Wamala, J. F. *et al.* Ebola hemorrhagic fever associated with novel virus strain, Uganda, 2007–2008. *Emerg. Infect. Dis.* **16,** 1087–1092 (2010).
40. MacNeil, A. *et al.* Proportion of deaths and clinical features in Bundibugyo Ebola virus infection, Uganda. *Emerg. Infect. Dis.* **16,** 1969 (2010).
41. Towner, J. S. *et al.* Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda. *PLoS Pathog.* **4,** e1000212 (2008).
42. World Health Organisation. *End of Ebola Outbreak in the Democratic Republic of the Congo*, http://www.who.int/csr/don/2009_02_17/en/ (2009).
43. Albarino, C. *et al.* Genomic analysis of filoviruses associated with four viral hemorrhagic fever outbreaks in Uganda and the Democratic Republic of the Congo in 2012. *Virology* **442,** 97–100 (2013).
44. World Health Organization. *DR Congo: Ebola (Situation as of 01 October 2012)*, http://www.afro.who.int/pt/grupos-organicos-e-programas/ddc/alerta-e-resposta-epidemias-e-pandemias/outbreak-news/3698-dr-congo-ebola-situation-as-of-01-october-2012.html (2012).
45. World Health Organisation. *Ebola in Uganda*, http://www.who.int/csr/don/2012_07_29/en/ (2012).
46. World Health Organisation. *Uganda: Ebola (situation as of 27 August 2012)*, http://www.afro.who.int/en/clusters-a-programmes/dpc/epidemic-a-pandemic-alert-and-response/outbreak-news/3674-uganda--ebola-situation-as-of-27-august-2012-.html (2012).
47. WWF. Global Lakes and Wetlands Database, http://www.worldwildlife.org/pages/global-lakes-and-wetlands-database (2014).
48. Pigott, D. M. *et al.* Mapping the zoonotic niche of Ebola virus disease in Africa. *eLife* **3,** e04395 (2014).
49. Brady, O., Hay, S. & Horby, P. Scale up supply of experimental Ebola drugs. *Nature* **512,** 233 (2014).

Data Citation

1. Mylne, A. *et al. Figshare* http://dx.doi.org/10.6084/m9.figshare.1168886 (2014).

Acknowledgements

O.J.B. is funded by a BBSRC studentship. S.I.H. is funded by a Senior Research Fellowship from the Wellcome Trust (#095066) which also supports A.M. and a grant from the Bill & Melinda Gates Foundation (#OPP1093011). S.I.H. would also like to acknowledge funding support from the RAPIDD

program of the Science & Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health. Z.H. is funded by the Bill & Melinda Gates Foundation (#OPP1106023). D.M.P. is funded by a Sir Richard Southwood Graduate Scholarship from the Department of Zoology at the University of Oxford. N.G. is funded by a grant from the Bill & Melinda Gates Foundation (#OPP1053338). M.U.G.K. is funded by the German Academic Exchange Service (DAAD) through a graduate scholarship. Funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

O.J.B. and A.M. wrote the manuscript with editing and approval from all authors. O.J.B., A.M. and D.M.P. compiled the data records. Z.H. digitized the geographic data and produced the outbreak maps. N.G., D.M.P. and M.U.G.K. helped with data cross-checking. S.I.H. conceived database design.

Additional information

Table 1 is only available in the online version of this paper.

Supplementary information accompanies this paper at http://www.nature.com/sdata

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Mylne, A. *et al.* A comprehensive database of the geographic spread of past human Ebola outbreaks. *Sci. Data* 1:140042 doi: 10.1038/sdata.2014.42 (2014).

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Viral infection

» Literature mining

» Epidemiology

» Viral epidemiology

A global compendium of human Crimean-Congo haemorrhagic fever virus occurrence

Jane P. Messina[1], David M. Pigott[1], Kirsten A. Duda[1], John S. Brownstein[2], Monica F. Myers[1], Dylan B. George[1] & Simon I. Hay[1,3]

Received: 19 December 2014
Accepted: 13 February 2015
Published: 14 April 2015

In order to map global disease risk, a geographic database of human Crimean-Congo haemorrhagic fever virus (CCHFV) occurrence was produced by surveying peer-reviewed literature and case reports, as well as informal online sources. Here we present this database, comprising occurrence data linked to geographic point or polygon locations dating from 1953 to 2013. We fully describe all data collection, geo-positioning, database management and quality-control procedures. This is the most comprehensive database of confirmed CCHF occurrence in humans to-date, containing 1,721 geo-positioned occurrences in total.

Design Type(s)	observation design • epidemiological study • data integration objective
Measurement Type(s)	Viral Epidemiology
Technology Type(s)	data collection method
Factor Type(s)	
Sample Characteristic(s)	Crimean-Congo hemorrhagic fever virus • anthropogenic habitat

[1]Spatial Ecology and Epidemiology Group, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. [2]Department of Pediatrics, Harvard Medical School and Children's Hospital Informatics Program, Boston Children's Hospital, Boston, Massachusetts 02115, USA. [3]Fogarty International Center, National Institutes of Health, Bethesda, Maryland 20892, USA. Correspondence and requests for materials should be addressed to J.P.M. (email: jane.messina@zoo.ox.ac.uk).

Background & Summary

The quality of reporting of Crimean-Congo haemorrhagic fever virus (CCHFV) infection is inconsistent by country and by region and is often biased by difficulties in diagnosis, limited resources for diagnostic testing and the variable reporting capacities of national health systems. Furthermore, active surveillance of human CCHFV infection is rare, so it is difficult to gauge the limits and intensity of transmission in a consistent manner across the world. As such, this database focuses on known occurrences rather than incidence of human CCHFV infection, with each occurrence identified by its unique geographical location on the globe and the year in which it occurred between 1953 and 2013. Sources of information include published literature, case reports and informal online sources, described in detail in our methods section.

Because the purpose of this database is to enable the identification of all locations of known occurrences of human CCHFV infection globally, many different types of information sources were used for compiling the database. These included individual case reports, reports of many cases in a particular location over a given period of time (up to several years), and active surveillance studies. As such, no information about the number of cases represented by each occurrence record was recorded, as the denominator would not be consistent over space or time, and as such attempts to model incidence or prevalence of CCHF would be a misuse of the database. Rather, with this database it is possible to model the probability of occurrence of CCHFV transmission to humans with a high degree of spatial resolution, for example as Bhatt *et al.*[1] did for dengue. In that study, the global probability of occurrence of human dengue infection was derived over long-term average conditions at a 5 km × 5 km resolution using a similar occurrence database as we describe here for CCHF, as well as a suite of environmental covariates. It would be unnecessary to repeat this type of labour-intensive data collection for these diseases or others for which this work has already been done[2,3]. In drawing upon information from studies in multiple locations, this type of database is intended to allow inferences to be made about the likelihood of disease transmission in locations with little or no information and thereby inform future research and surveillance efforts. Because the locations of disease occurrence are recorded at the finest level of detail possible (i.e., points when the exact location was known or else polygons for administrative units when this was the best information available), this database allows these environmental characteristics to be determined as accurately as possible in a consistent manner at larger spatial scales.

It should be noted that we did not discriminate studies or reports based upon the clinical outcome of the human CCHF infections they reported; this information was not consistently reported across all of the varied sources and we aimed to be comprehensive in our inclusion of all known locations of confirmed human CCHFV infection. We did, however, exclude cases where a healthy individual tested positive for CCHFV antibodies due to potential cross-reactivity with other viruses and the inability to determine the site and time of infection.

We describe all data collection processes in full, as well as the geo-positioning, quality-control and database-management procedures. The database's construction from many different sources of information makes it the best currently available standardised data available on global CCHFV infection in humans, comprising 1,721 geo-positioned occurrences in total worldwide. The database is not necessarily confined to use for global analyses, as it contains enough detailed information for certain parts of the world to carry out modelling at a regional or even sub-national scale. Regions and countries are specified in addition to the specific location of every record in order to facilitate sub-setting of the database for the user's needs.

Methods

Data collection

An occurrence database comprising point (e.g., town or city) or polygon (e.g., county or province) locations of confirmed CCHF infection presence was compiled from peer-reviewed literature, Genbank records, and HealthMap alerts. A literature search was conducted on PubMed and Web of Science using the terms 'CCHF' or 'Crimean Congo Hemorrhagic Fever' or 'Crimean Hemorrhagic Fever' or 'Congo Hemorrhagic Fever'. The same terms were used in our Genbank search. An occurrence was defined as one or more laboratory or clinically confirmed infection(s) with CCHFV occurring at a unique location (the same polygon, or 5 km × 5 km pixel for points) within one calendar year. All occurrence data underwent manual review and quality control to ensure reliability and precision of geo-positioning. Original data sources can be provided in PDF format by requesting the files from the corresponding author.

Informal online data sources were collated automatically by the web-based system HealthMap (http://healthmap.org) as described elsewhere[4]. Briefly, HealthMap is an online infectious disease outbreak-monitoring system that captures data from a range of electronic sources in nine different languages. The system performs hourly scans of online news aggregators, listservs, electronic disease surveillance networks and public health outbreak report feeds. It captures four fields: headline (the headline, title or subject line), date (publication date), description (a brief summary) and information text (the main content of the article or report). The information text is passed to HealthMap's classification engine, which parses out one or more disease names and outbreak locations using dictionaries of disease and location patterns. The system then uses a separate algorithm to assign relevance scores that classify alerts as (i) breaking (information about a new outbreak or new information about an on-going outbreak),

(ii) context (content about research, policy or background on a particular disease), (iii) warning (articles that warn about the potential for an outbreak), (iv) not disease-related (articles that are captured by the system because they contain words that match disease names in the dictionary but are not in fact about an infectious disease) or (v) old news (an article that mentions a historical outbreak). Finally, HealthMap handles duplicates by aggregating together highly similar alerts such as those released by a news wire service and published in multiple periodicals. The requirements for including a CCHF occurrence record from the HealthMap data set in our database were that the article or report contained one of thirty-six key words or phrases in twenty different languages, including 'CCHF', 'CCHFV', and 'Crimean-Congo' in English. The article must also have been classified by the system as 'breaking'. The HealthMap data set included in the current database was last updated on 26th May 2012.

Geo-positioning of data

All available location information was extracted from each peer-reviewed article and PROMED case report. The site name was used together with all contextual information provided about the site position to determine its latitudinal and longitudinal coordinates using Google Maps (https://www.maps.google.co.uk/). Place names are often duplicated within a country, so the contextual information was used to ensure the correct site was selected. When the site name was not found, the contextual information was used to scan sites in the approximate area to check for names that had been transliterated in Google Maps in a different way to the published article (e.g., Imichli and Imishly). If the study site could be geo-positioned to a specific city, town or village, its centre was recorded and termed a 'point location' (i.e., associated with a specific latitude and longitude). Point occurrences also included explicit co-ordinates supplied in the article. In reports where more accurate details about the specific location within the city, town or village were available (such as a specific suburb), this was used to define the point occurrence. If the study site could only be identified at an administrative area level (e.g., province or district), the latitude and longitude of a point within the area, along with its name and administrative level, was recorded from Google Maps and then later overlaid in a geographic information system (GIS) to identify the appropriate polygon for the administrative unit. All administrative units were as recognised by the FAO Global Administrative Unit Layer (GAUL) system[3]. These occurrences referring to an area were termed 'polygon locations'. Reports of autochthonous (locally transmitted) cases or outbreaks were entered as an occurrence within the country in which transmission occurred. If imported cases were reported with information about the site of infection, they were geo-positioned to the country where transmission occurred. If imported cases were reported with no information about the site of contagion, they were not entered into the database. All formal occurrence records underwent spatial and temporal standardisation as described below, in order to ensure consistent definition of an occurrence before undergoing technical validation.

Geo-positions for the HealthMap data were generated using a custom-built gazetteer, or geographic dictionary, of over 4,000 relevant phrases and place names and their corresponding geographic coordinates. The system uses a look-up tree algorithm that searches for matches between sequences of words in alert info text and sequences of words in the gazetteer. When a match is found, a set of rules are applied which attempt to determine the relevance of the place name to the outbreak that is being reported based on the position of the phrase in the report text. The gazetteer includes place names at a range of spatial resolutions (e.g., neighbourhoods, cities, provinces and countries) and uses certain phrases to trigger exclusion of a place name (e.g., Brazil nut). All HealthMap records were added to the unstandardised database and then underwent spatial and temporal standardisation, as well as technical validation, along with the records from the literature and ProMED reports.

Occurrence database management: locational and temporal standardisation

As the database was compiled from many different sources and by multiple persons, it was first necessary to standardise the data entries such that identical locations which may have been geo-positioned slightly differently were given the same unique identifier. To do this, polygon records were all assigned a unique administrative code by overlaying the recorded geographic coordinates in the GIS with corresponding administrative unit shapefiles[5]. Point records were given the same unique identifier if they lay in the same 5 km × 5 km pixel within a global grid. Finally, any record associated with a polygon measuring larger than 1° × 1° at the equator was removed from the database, although the authors are happy to provide the records for these polygons upon request.

It was next necessary to temporally standardise the database, as the collected CCHF occurrence data came in a variety of temporal forms. For example, some sources reported multiple cases in a single location throughout a year with no finer-scale temporal information. However, in other sources (particularly online sources), multiple cases in the same location throughout the year were presented as a new report each time subsequent transmission occurred. Furthermore, many sources described endemic transmission occurring across multiple years. As a result, we chose to define a single occurrence at a given unique location (as identified above) as one or more confirmed human cases of CCHFV infection occurring within one calendar year in that location, as this was the finest temporal resolution available across all records. This involved a procedure which: (i) disaggregated any records which were in the same location but spanning multiple years into individual records for each respective year; and then (ii) aggregated all records with the same unique location identifier and occurring within the same year to

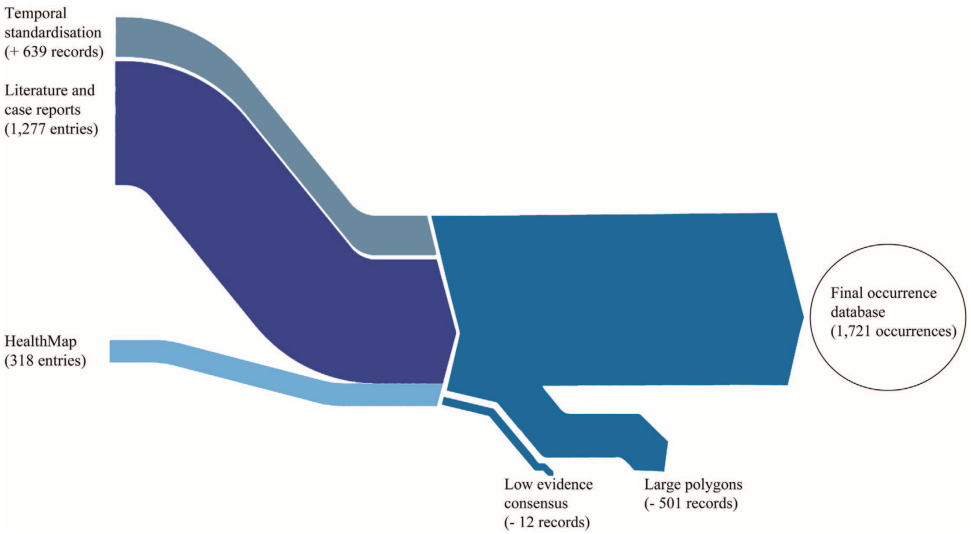


Figure 1. Occurrence data processing summary, beginning with the raw inputs and showing the proportion of data gained or lost through the stages of temporal standardisation and quality control before reaching the final occurrence database.

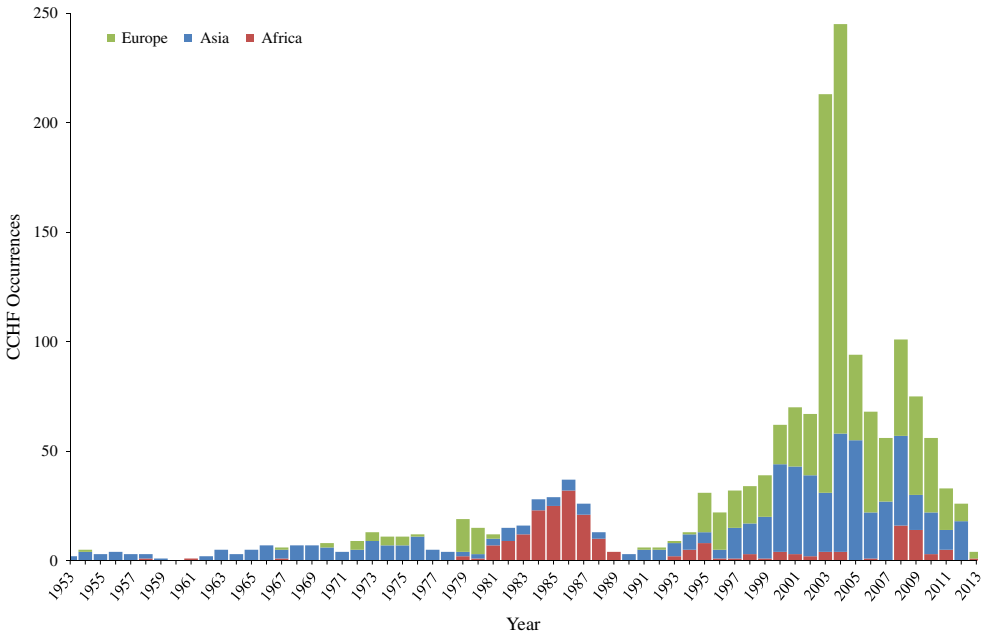


Figure 2. The numbers of occurrence locations per year separated by region.

form a single occurrence record. It should be emphasized that because the studies and case reports from which data was compiled represent various types of both active and passive surveillance, the resulting database alone cannot reveal the actual frequency of CCHF occurrences, and that each unique location where any CCHFV infections occurred were only assigned one occurrence per year, regardless of the number of infections reported. The database next and finally underwent technical validation, as described below.

Data Records

The database is publicly available online (Data Citation 1) as a comma-delimited file for ease of use and the ability to import it into a variety of software programs. Each of the 1,721 rows represents a single occurrence record (one or more CCHF cases in the same unique location within a single calendar year). A summary of the data management procedure, beginning with the raw inputs and showing the proportion of data lost through the stages of quality control before reaching the final occurrence database,

Location Type	Africa	Asia	Europe	Total
point	184	740	473	1397
polygon	43	150	131	324
Total	227	890	604	1721

Table 1. Occurrences broken down by region and location type.

is provided in Fig. 1. Only three records were entered at the country level, and were all in countries with an area greater than 1 degree squared at the equator, so no country-level records existed in the final database as can be seen below. Fig. 2 displays the numbers of occurrence locations per year separated by region. The fields contained in the database are as follows:

1. **OCCURRENCE_ID:** Unique identifier for each occurrence in the database after temporal and locational standardisation.
2. **LOCATION_TYPE:** Whether the record represents a point or a polygon location.
3. **ADMIN_LEVEL:** The administrative level which the record represents when the location type is a polygon. Values are 1 (state or province), 2 (district) and −999 when the location type is a point.
4. **GAUL_AD1:** The first-level GAUL code which identifies the Admin-1 level occurrences as well as the Admin-1 polygon within which any smaller polygons and points lie.
5. **GAUL_AD2:** The second-level GAUL code which identifies the Admin-2 level occurrences as well as the Admin-2 polygon within which any points lie. Values of −999 are assigned when the polygon was Admin-1 level.
6. **UNIQUE_LOCATION:** A unique identifier created for all locations (both points and polygons) based upon the unique point locations (in same 5 km × 5 km pixels) and the GAUL codes.
7. **X:** The longitudinal coordinate of the point or polygon centroid (WGS1984 Datum).
8. **Y:** The latitudinal coordinate of the point or polygon centroid (WGS1984 Datum).
9. **YEAR:** The year of the occurrence.
10. **COUNTRY:** The name of the country within which the occurrence lies.
11. **REGION:** The region within which the occurrence lies—values are Asia, Europe, and Africa (includes the Arabian peninsula).

Technical Validation

The following procedures were carried out on the final database to ensure the accuracy and validity of the occurrence records.

1. A raster distinguishing land from water was created at a 5 km × 5 km resolution and was used to ensure all disease occurrences were positioned on a valid land pixel.
2. We cross-validated all of the unique occurrence locations against CCHFV transmission extent based upon an evidence-based consensus score, derived in a similar manner as Brady *et al.*[6] previously carried out for dengue. In brief, this classification was determined according to a qualitative evidence base that assessed consensus among a wide variety of evidence types on presence or absence of human CCHF cases at a national and sometimes sub-national level. This consensus ranged from complete agreement on absence (score of −100) to complete agreement on presence (100). We chose to exclude points in areas with scores of less than −25. This conservative criterion was intended to preserve points in areas of uncertainty on CCHF status.
3. A random sub-sample of HealthMap occurrence points were manually checked to identify common geo-positioning problems which were rectified in the final database.

The result is a database consisting of 1,721 geo-positioned occurrences in total worldwide, broken down by region and location type in Table 1.

In addition to the main comma-delimited file, three Supplementary Files are included as part of the file set which can be found online (Data Citation 1). These include two text documents: (i) a bibliography of the published literature used for inputting occurrence records, and (ii) a list of accession ID numbers for those records obtained from GenBank. The third Supplementary File is an additional comma-delimited file containing the HealthMap records entered into the database (before standardization and technical validation). This database contains information on the year of each report, as well as the latitude and longitude of the disease occurrence and the source from which this information was extracted.

References

1. Bhatt, S. *et al.* The global distribution and burden of dengue. *Nature* **496,** 504–507 (2013).
2. Pigott, D. M. *et al.* Global database of Leishmaniases occurrence locations, 1960-2012. *Sci. Data* **1,** 140036 (2014).
3. Mylne, A. *et al.* A comprehensive database of the geographic spread of past human Ebola outbreaks. *Sci. Data* **1,** 140042 (2014).

4. Freifeld, C. C., Mandl, K. D., Reis, B. Y. & Brownstein, J. S. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *J. Am. Med. Inform. Assoc.* **15,** 150–157 (2008).
5. Global administrative unit layers. FAO Statistics Division, http://data.fao.org/maps (2010).
6. Brady, O. J. *et al.* Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLoS Negl. Trop. Dis.* **6,** e1760 (2012).

Data Citation

1. Messina, J. P. *et al. Figshare* http://dx.doi.org/10.6084/m9.figshare.1270687 (2014).

Acknowledgements

S.I.H. is funded by a Senior Research Fellowship from the Wellcome Trust (095066), which also funds K. A.D. J.P.M. is funded by the International Research Consortium on Dengue Risk Assessment Management and Surveillance (IDAMS, 21803, http://www.idams.eu). D.M.P. is funded by a Sir Richard Southwood Graduate Scholarship from the Department of Zoology at the University of Oxford. JSB acknowledges funding from NIH National Library of Medicine (R01LM010812) and from the Bill & Melinda Gates Foundation (#OPP1093011). S.I.H. also acknowledges funding support from the RAPIDD program of the Science & Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health.

Author Contributions

J.P.M. drafted the manuscript with editing and approval from all authors. J.P.M., D.M.P., K.A.D., M.F.M., and D.B.G. compiled the data records. J.P.M. performed database standardisation and technical validation. J.S.B. is responsible for the HealthMap component of the database. S.I.H. conceived the database design and advised on standardisation and validation procedures.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Messina, J. P. *et al.* A global compendium of human Crimean-Congo haemorrhagic fever virus occurrence. *Sci. Data* 2:150016 doi: 10.1038/sdata.2015.16 (2015).