Forum

The rise of big data in disease ecology

Jean-François Doherty,[1] Xuhong Chai,[1] Laurie E. Cope,[1] Daniela de Angeli Dutra,[1] Marin Milotic,[1] Steven Ni,[1] Eunji Park,[1] and Antoine Filion[1,*]

Big data have become readily available to explore patterns in large-scale disease ecology. However, the rate at which these public databases are exploited remains unknown. We highlight trends in big data usage in disease ecology during the past decade and encourage researchers to integrate big data into their study framework.

Integrating big data into disease ecology

Emerging infectious diseases are of major concern worldwide for their negative impacts on wildlife and human health [1] as well as on livestock and crops [2]. Most of these diseases are zoonoses that have originated from natural populations [3]. Within the past few decades the rise of these zoonotic diseases has been increasingly linked to environmental anthropogenic changes [4]. Accordingly, to better predict the emergence and spread of infectious diseases globally, the 'One Health' approach, which integrates data from human, non-human animal, and environmental health, has been proposed as an ideal solution to tackle large-scale ecological questions related to zoonotic diseases and host–parasite interactions [5]. In this light, macroecological studies provide critical information on how human and natural systems are connected by investigating the patterns and processes at both regional and global scales to better understand disease distribution and emergence [6]. Inherently, these types of studies require the use of large datasets, compiled from multiple sources and often available to the public [6], which we refer to as 'big data' (Box 1; see also Table S1 in the supplemental information online). While this term is commonly used in computer sciences[i], here we define big data in an ecological framework as large datasets assembled from heterogeneous primary sources of data [7], such as those collated from published peer-reviewed studies. Still, data obtained from local surveys and field studies on disease-causing agents in natural host populations, referred to as 'original data' here, remain essential for addressing system-specific questions and understanding the fine-scale processes and mechanisms that are responsible for large-scale disease patterns [8]. Here, we highlight the advantages and disadvantages of using original versus big data to address questions related to the emergence and spread of disease-causing agents in natural and human populations. Following this, we explore the trends in the use of big data with an example on avian malaria research during the past decade. We finish with a discussion on other large datasets that can be used to complement the big data sources explored in Box 1.

Pros and cons of big data

Regardless of their source, data are ultimately used to answer specific scientific questions. This is why big data are especially important to answer spatial and temporal questions at larger scales [7]. Big data are useful to explore general patterns and complex processes of disease ecology at both regional and global scales, whereas original data are typically limited to local or regional host–parasite systems, giving the latter an indisputable advantage of being fully independent for specific studies [8]. Big data can be complementary to original data, in the sense that original data feed into big data and allow us to explore fine-scale processes that are impossible to test globally, and big data allow us to investigate and generate new hypotheses relating to large-scale disease patterns that are useful for testing locally in new systems. However, big data come with obvious disadvantages: there will always be some noise from study biases, such as variations in sampling effort and statistical nonindependence [6]; these are mainly due to heterogeneous study effort associated with the sources of original data that feed into big data, creating knowledge gaps in taxonomic or geographical range, potentially hindering predictions of disease drivers. In addition, country-specific metrics such as the Gross Domestic Product, the level of freedom of the press, spoken languages, and communication structures, can induce additional biases in the way that data are reported [9,10]. This is why big data need to be curated and regularly updated by a devoted team who take the time to ensure the quality of the original data (Box 1). One of the main advantages of working with big data is avoiding all the constraints associated with acquiring original data. Field and laboratory studies are often costly and time-consuming, they usually require study permits, and they are subject to environmental restrictions such as the seasonality of host–parasite systems. While generating original data necessary for big data obviously requires a huge effort behind the scenes, big data per se, available to all, require little effort to obtain online, thus representing an inexpensive solution to tackling large-scale questions on disease ecology. Even though both sources of data have their advantages and disadvantages, research is driven by questions, and these questions ultimately determine what type of data should be used.

MalAvi as big data

Here, we provide a broad overview of the use of MalAvi, a public database on avian haemosporidian parasites, since just after its foundation in 2009 (Box 1). We searched

Box 1. Big data and disease ecology

In the current digital era, 'big data' (i.e., enormous quantities of data that are difficult to collect and process individually) have become readily and freely available for researchers in disease ecology. Large amounts of data are being collected by research facilities and individual researchers around the world. In the field of disease ecology, big data can be used to estimate the global distribution and seasonal dynamics of parasites, and determine host–parasite networks, codiversification patterns, and environmental drivers of disease. To investigate what databases are currently available for disease ecology, we conducted an extensive but nonexhaustive search in Google, Google Scholar, and Web of Science. We used different combinations of keywords, such as 'database', 'public', and 'parasite', or specific names of parasite groups, for example, 'nematode' and 'malaria'. Our search efforts resulted in 26 reliable and currently accessible databases (Table I and see Table S1 in the supplemental information online). Ten of these databases are specific to human-infecting parasites, including databases of human *Plasmodium* spp. and helminths (Table I). As seen here, various types of data, such as parasite occurrence and distribution, genomic and transcriptomic (mostly of human parasites), morphological, and images available from these databases, can be used to elucidate patterns in large-scale disease ecology.

Table I. Summary of 'big data' databases and complementary databases openly available for disease ecology[a]

		Number of databases	Type of data available	Resources
Host group	Aquatic animals	2	Morphology, occurrence, distribution	International database on aquatic animal diseases[viii], Fish parasite ecology software tool[ix]
	Birds	2	Occurrence, distribution, genetic and taxonomic information	MalAvi[x], Global associations between birds and vane-dwelling feather mites[xi]
	Humans	10	Genomic, transcriptomic, epidemiology, occurrence, distribution, cDNA, etc.	ENHanCEd Infectious Diseases (EID2)[xii], WormBase ParaSite[xiii], PhenoPlasm[xiv], ToxoDB[xv], PlasmoDB[xvi], FULL-malaria[xvii], Centralized information system for infectious diseases (CISID)[xviii], TickReport[xix], PlasmoGem[xx], CryptoDB[xxi]
	Mammals	2[b]	Occurrence and distribution	Global Mammal Parasite Database[xxii], LeishDB[xxiii]
Parasite group	*Cryptosporidium* spp.	1	Genomic	CryptoDB[xxi]
	Fungi	1	Occurrence and distribution	Fungal databases[xxiv]
	Non-human haemosporidians	1	Occurrence, distribution, genetic and taxonomic information	MalAvi[x]
	Helminths	3	Occurrence, distribution, and genomic	WormBase ParaSite[xiii], Host–parasite database – Natural History Museum[xxv], Helminth.net v1.0[xxvi]
	Leishmania spp.	1	Occurrence, distribution, and genomic	LeishDB[xxiii]
	Plasmodium spp.	4	Genomic, transcriptomic, epidemiology, cDNA, disease eradication strategies, etc.	PhenoPlasm[xiv], PlasmoDB[xvi], FULL-malaria[xvii], PlasmoGem[xx]
	Tick-borne diseases	1	Occurrence and distribution	TickReport[xix]
	Feather mites	1	Occurrence, distribution, taxonomic information	Global associations between birds and vane-dwelling feather mites[xi]
	Toxoplasma spp.	1	Genomic	ToxoDB[xv]
Multiple hosts and parasite groups		7	Occurrence, distribution, images, genomic	Awesome Parasite[xxvii], World Animal Health Information System (OIE-WAHIS)[xxviii], National Notifiable Diseases Surveillance System (NNDSS)[xxix], United States Geological Survey (USGS)[xxx], VEuPathDB[xxxi], Host–parasite webs (Interaction Web Database)[xxxii], Web-of-life[xxxiii]
Additional complementary sources of data		6	Biodiversity, ecological traits, environmental data	Global Biodiversity Information Facility (GBIF)[xxxiv], Open Traits Network (OTN)[xxxv], United Nation Environmental Protection (UNEP)[xxxvi], Global Lakes and Wetlands Database (GLWD)[xxxvii], Food and Agriculture Organization of the United Nations (FAO)[xxxviii], Water Quality Portal[xxxix]

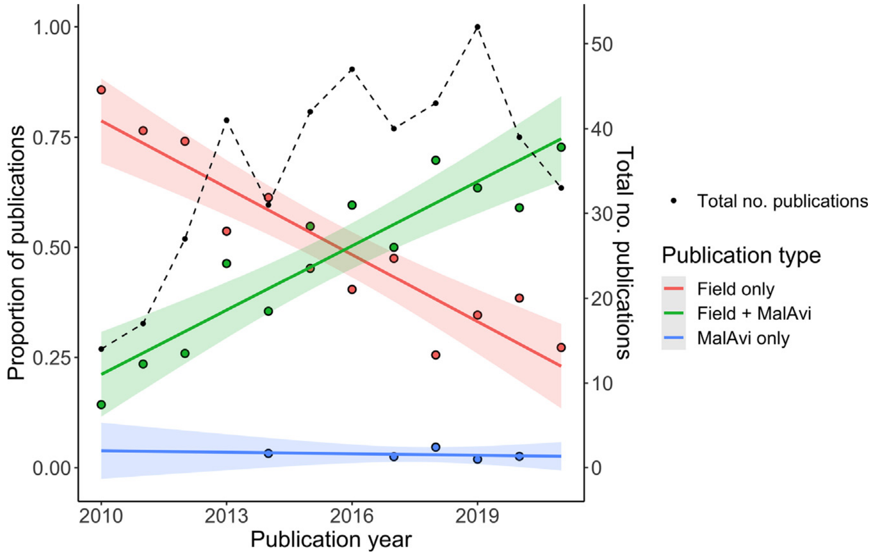
[a] See Table S1 for details.

[b] One of these databases, the Global Mammal Parasite Database[xxii], requires approval from the authors prior to accessing the full dataset.

Web of Science (for methods, see Table S2 in the supplemental information online) to examine the yearly trends of three publication categories, from January 2010 to June 2021 inclusively: the number of publications on avian malaria using fieldwork (original) data, MalAvi data, or both. We found a total of 432 relevant publications, which were plotted in Figure 1 as the proportion of publications by year for each category. First, the proportion of papers that utilised data exclusively from the field decreased linearly from 2010 to 2021. Second, we found no trend in the proportion of studies that used data exclusively from MalAvi over the years. Only five studies, four of which were published in the past 5 years, used MalAvi to answer broad ecological questions on avian malaria (Figure 1). Lastly, the proportion of studies using fieldwork and MalAvi data together increased linearly, thus inversely to the studies using only fieldwork data (Figure 1). Most of these studies, however, used MalAvi only to assign taxonomy to genetic data of haemosporidians, that is, BLAST searches. Some of these papers also used GenBank to perform BLAST searches, but these were counted as artifacts since we searched exclusively for MalAvi as the source of big data (Table S2).

Complementary sources of data

While data from original studies and big data can be used to answer questions related to specific study designs, the past few years have seen a massive increase in the availability of additional tools that can be combined with big data to broaden the framework of disease ecology. For instance, to understand how disease occurrence in the wild is mitigated by anthropogenic impacts on climate and landscape, both WorldClim[ii] climate data and MODIS Terra[iii] vegetation indices for land usage are readily available and provide open access metrics that can help to answer a vast array of scientific questions, thus broadening the scope of disease ecology at any spatial scale.



Trends in Parasitology

Figure 1. Publication trends on avian malaria research. Yearly proportions of publications that used either fieldwork (original) data, MalAvi data, or both (left *y*-axis). Number of publication search results for each year (right *y*-axis). Each coloured circle represents a proportion (for full methods and results see Table S2 in the supplemental information online), and each coloured area represents the 95% confidence bands of the trend line (calculated with linear model smoothing in *ggplot2*). Data were compiled from January 2010 to June 2021 inclusively.

From a physiological perspective, one may want to answer important questions related to the traits of both disease-causing agents and hosts, to better understand niche-related mechanisms and ecosystem resilience towards disease [6]. While trait data on disease-causing agents remain scarce [11], host trait data for plants and animals are readily available in open-access databases (see Table A1 in [12]). For example, the Amniote Life History Database[iv] and PanTHERIA[v] have a variety of species-wide traits for vertebrates. All of these resources offer ecologically relevant variables useful for disease ecologists using big data, both at regional and global scales.

From an evolutionary perspective, an ecophylogenetic approach can be useful to understand coevolutionary patterns, coinfection patterns, and community structure in disease dynamics [13]. While genomic approaches can be costly, numerous resources now exist to simplify logistics and help incorporate this information into a disease ecology framework. For instance, the TimeTree[vi] public database provides evolutionary timelines for a vast array of organisms, thus allowing researchers to build consensus trees based on their own datasets and incorporate them into their study.

These are but a few examples of open-access tools that can be combined with most study designs on disease ecology that incorporate big data (for more examples, see Box 1 and Table S1).

Concluding remarks and future directions

Here, we show that researchers in disease ecology are incorporating big data increasingly to complement their original field data. However, few studies appear to use big data exclusively to answer broad ecological questions; these studies are obviously limited to the number of unique questions that can exploit big data to their fullest advantage. Despite the limitations discussed earlier, new untapped sources of information on many different systems related to disease ecology are emerging. For instance, the use of internet data passively collected by the public at large, for example, iNaturalist[vii], has been proposed as a novel source of big data and has already been used to discover new patterns in disease distribution [9,14]. Moreover, new machine learning tools, for example, deep-learning algorithms, now allow researchers to extract information from text in multiple publications or sets of images from various internet sources, creating even more comprehensive big datasets from sources that were traditionally disregarded because of their complexity [9,15]. After lagging behind other study areas for some time, we posit that disease ecology now has a wide range of tools to bridge the technological gaps with other areas of science and answer key ecological questions within a broader infectious disease framework.

Acknowledgments

This forum article was prepared during a working retreat hosted by Cass Field Station (University of Canterbury). We thank Robert Poulin for critical feedback on an early draft of the manuscript. We also thank the two anonymous reviewers for their comments and suggestions that improved the manuscript. Lastly, we sincerely thank Emerson's Brewery (Dunedin) for their generous liquid sponsorship.

Declaration of interests

The authors declare no conflict of interest.

Supplemental information

Supplemental information associated with this article can be found online at https://doi.org/10.1016/j.pt.2021.09.003.

Resources

[i] www.sas.com/en_us/insights/big-data/what-is-big-data.html

[ii] www.worldclim.org/

[iii] https://earthdata.nasa.gov/earth-observation-data

[iv] https://datarepository.wolframcloud.com/resources/Amniote-Life-History-Database

[v] http://esapubs.org/archive/ecol/E090/184/

[vi] www.timetree.org/

[vii] www.inaturalist.org/

[viii] www.cefas.co.uk/international-database-on-aquatic-animal-diseases/

[ix] https://panic.alwaysdata.net/

[x] http://130.235.244.92/Malavi/

[xi] http://onlinelibrary.wiley.com/doi/10.1002/ecy.1528/suppinfo

[xii] https://eid2.liverpool.ac.uk

[xiii] https://parasite.wormbase.org/index.html

[xiv] http://phenoplasm.org/

[xv] https://toxodb.org/toxo/app

[xvi] https://plasmodb.org/plasmo/app

[xvii] https://fullmal.hgc.jp/

[xviii] http://data.euro.who.int/cisid/

[xix] www.tickreport.com/stats

[xx] https://plasmogem.sanger.ac.uk/

[xxi] https://cryptodb.org/cryptodb/app

[xxii] https://parasites.nunn-lab.org/

[xxiii] https://github.com/fgtorres/LeishDB

[xxiv] https://nt.ars-grin.gov/fungaldatabases/fungushost/fungushost.cfm

[xxv] www.nhm.ac.uk/research-curation/scientific-resources/taxonomy-systematics/host-parasites/database/index.jsp

[xxvi] www.helminth.net/

[xxvii] https://github.com/ecohealthalliance/awesome-parasite

[xxviii] https://wahis.oie.int/#/home

[xxix] www.cdc.gov/nndss/index.html

[xxx] https://my.usgs.gov/confluence/display/biodata/BioData+Taxonomy+Downloads

[xxxi] https://veupathdb.org/veupathdb/app

[xxxii] www.ecologia.ib.usp.br/iwdb/resources.html#host_parasite

[xxxiii] www.web-of-life.es/map.php

[xxxiv] www.gbif.org/

[xxxv] https://opentraits.org/

[xxxvi] www.unep.org

[xxxvii] www.worldwildlife.org

[xxxviii] www.fao.org

[xxxix] www.waterqualitydata.us/portal/

[1] Department of Zoology, University of Otago, Dunedin,
New Zealand

*Correspondence:
afilion90@gmail.com (A. Filion).

https://doi.org/10.1016/j.pt.2021.09.003

References

1. Daszak, P. *et al.* (2000) Wildlife ecology – emerging infectious diseases of wildlife – threats to biodiversity and human health. *Science* 287, 443–449
2. Fisher, M.C. *et al.* (2012) Emerging fungal threats to animal, plant and ecosystem health. *Nature* 484, 186–194
3. Jones, K.E. *et al.* (2008) Global trends in emerging infectious diseases. *Nature* 451, 990–993
4. Jones, B.A. *et al.* (2013) Zoonosis emergence linked to agricultural intensification and environmental change. *Proc. Natl. Acad. Sci. U. S. A.* 110, 8399–8404
5. Zinsstag, J. *et al.* (2011) From 'one medicine' to 'one health' and systemic approaches to health and well-being. *Prevent. Vet. Med.* 101, 148–156
6. Stephens, P.R. *et al.* (2016) The macroecology of infectious diseases: a new perspective on global-scale drivers of pathogen distributions and impacts. *Ecol. Lett.* 19, 1159–1171
7. Farley, S.S. *et al.* (2018) Situating ecology as a big-data science: current advances, challenges, and solutions. *Bioscience* 68, 563–576
8. McGill, B. (2003) Strong and weak tests of macroecological theory. *Oikos* 102, 679–685
9. Poulin, R. *et al.* (2021) iParasitology: mining the internet to test parasitological hypotheses. *Trends Parasitol.* 37, 267–272
10. Smith, K. *et al.* (2014) Global rise in human infectious disease outbreaks. *J. R. Soc. Interface* 11, 20140950
11. Llopis-Belenguer, C. *et al.* (2019) Towards a unified functional trait framework for parasites. *Trends Parasitol.* 35, 972–982
12. Schneider, F.D. *et al.* (2019) Towards an ecological trait-data standard. *Methods Ecol. Evol.* 10, 2006–2019
13. Fountain-Jones, N.M. *et al.* (2018) Towards an eco-phylogenetic framework for infectious disease ecology. *Biol. Rev.* 93, 950–970
14. Doherty, J.-F. *et al.* (2021) The people versus science: can passively crowdsourced internet data shed light on host–parasite interactions? *Parasitology* 1313–1319
15. Thieu, T. *et al.* (2012) Literature mining of host–pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics* 28, 867–875