World Organisation for Animal Health
Organisation mondiale de la santé animale
Organización Mundial de Sanidad Animal

https://doi.org/10.20506/rst.42.3352

Assessing the quality of data for drivers of disease emergence

V. Horigan* [(1)], L. Kelly [(1)], A. Papa [(2)], M.P.G. Koopmans [(3)], R.S. Sikkema [(3)], L.G.H. Koren [(4)] & E.L. Snary [(1)]

(1) Animal and Plant Health Agency, Woodham Lane, Addlestone, Surrey KT15 3NB, United Kingdom

(2) Aristotle University of Thessaloniki, University Campus, Thessaloniki 54006, Greece

(3) Department of Viroscience, Erasmus MC, PO Box 2040, 3000 CA, Rotterdam, the Netherlands

(4) Statistics Netherlands (CBS), PO Box 24500, 2490 HA, The Hague, the Netherlands

*Corresponding author: verity.horigan@apha.gov.uk

Summary

Drivers are factors that have the potential to directly or indirectly influence the likelihood of infectious diseases emerging or re-emerging. It is likely that an emerging infectious disease (EID) rarely occurs as the result of only one driver; rather, a network of sub-drivers (factors that can influence a driver) are likely to provide conditions that allow a pathogen to (re-)emerge and become established. Data on sub-drivers have therefore been used by modellers to identify hotspots where EIDs may next occur, or to estimate which sub-drivers have the greatest influence on the likelihood of their occurrence. To minimise error and bias when modelling how sub-drivers interact, and thus aid in predicting the likelihood of infectious disease emergence, researchers need good-quality data to describe these sub-drivers. This study assesses the quality of the available data on sub-drivers of West Nile virus against various criteria as a case study. The data were found to be of varying quality with regard to fulfilling the criteria.

The characteristic with the lowest score was completeness, i.e. where sufficient data are available to fulfil all the requirements for the model. This is an important characteristic as an incomplete data set could lead to erroneous conclusions being drawn from modelling studies. Thus, the availability of good-quality data is essential to reduce uncertainty when estimating the likelihood of where EID outbreaks may occur and identifying the points on the risk pathway where preventive measures may be taken.

Keywords

Data quality – Drivers – Emerging infectious disease – Modelling – Predicting disease outbreaks – Prevention – West Nile virus.

Introduction

Global epidemics, or pandemics, have never been more pertinent than in the present day, with coronavirus disease 2019 (COVID-19) resulting in an unprecedented global social, economic and public health crisis. COVID-19 fits into the category of emerging infectious diseases (EIDs), broadly defined as diseases that are increasing in their incidence, geographic or host species range, or impact [1]. Identifying potential drivers of EIDs can help to better understand how, when and where they happen, providing evidence that could help to prevent the next pandemic.

Drivers are factors that have the potential to directly or indirectly influence the likelihood of (re-)emergence of infectious diseases [2]. Drivers predominantly fall into three broad categories: biological, ecological (including demographic), and behavioural [3, 4, 5, 6, 7, 8]. Understanding of the specific effects of many of these drivers is often supported by limited hard data because ecological and biological systems are highly complex and multi-layered [3, 9]. Very often, drivers are only assumed to be responsible for the occurrence of an EID, based on expert elicitation [10], but substantiating data to prove this assumption are lacking or have not been subjected to robust analyses to disentangle association and causation. Thus, there will always be a degree of uncertainty. Additionally, unknown drivers, or a combination of drivers that provide the correct conditions for disease emergence, are rarely accounted for.

It is unlikely that EIDs occur because of only one driver. Rather, it is probably a combination of sub-drivers (factors that can influence a driver) that provide conditions that increase the likelihood of a pathogen (re-)emerging and becoming established [11]. The driving factors may not be successive and there may be extensive time periods between events. For instance, a pathogen may be released into a naive population several times before it becomes able to transmit from animal to animal or person to person, if at all [12]. Thus, more often than not, sub-drivers are required both for the release and emergence of a pathogen and for its subsequent spread or onward transmission for an EID to occur. The likelihood of EIDs occurring could potentially be anticipated through surveillance of their sub-drivers, using both global and regional data, if they are of sufficient quality.

The conditions leading to a new EID have been described as a microbial 'perfect storm', arising from a rare and complex combination of sub-drivers [13]. Contrary to this, however, is the view that these circumstances have arisen not from a 'large number of usually unpredictable factors' but from long-standing and well-understood human actions and inactions [14]. The use of the term 'perfect storm' creates an impression that is reactive rather than proactive, with concepts of randomness and volatility that may undermine the ability to anticipate and prevent pandemics before they emerge. A 'perfect storm' frame of mind emphasises the power of chance over the efficacy of prevention. However, historical EID outbreaks have shown that long-term investments in disease tracking and surveillance, scientific research and public health infrastructure are key to containing at least some emerging threats [14]. It is likely that a more realistic scenario is a hybrid situation, in which surveillance of drivers can indicate changes in conditions that increase the likelihood of EIDs occurring, and the 'perfect storm' occurs when certain factors converge and a 'tipping point' is reached. Following this rationale, levels of concern can be triggered by the timely supply of good-quality data from the surveillance of recognised drivers.

A further challenge is that a study on the use of evidence when predicting the likelihood of EIDs found that the unconscious bias of knowledge of recognised sub-drivers can cause researchers to miss other more local sub-drivers, e.g. the local building of roads and housing in rural agricultural areas. Such an activity is a behavioural sub-driver in this example of local building, causing changes in land use that could affect both the abundance and habitat range of wildlife and vectors. Thus, there is a need for perceived risks to be considered in the local context [15]. Consequently, whilst there may be known hotspots for EID release, effort is also required to identify hidden drivers and challenge preconceptions arising from unconscious bias towards traditional generic drivers.

Partly in view of the complexities described above, current responses to EIDs are largely reactive instead of proactive. The ability to predict the likelihood of EIDs occurring with low uncertainty would allow for prevention, early detection, and swift and effective reaction to mitigate impact [16]. An understanding of the biological, ecological (including demographic) and behavioural factors that contribute to the emergence of infectious diseases is needed for this process. This requires a fusion of data from a broad array of sources [16]. Outbreaks of EIDs are often preceded by changes affecting the factors that influence pathogen presence and spread. Modelling the ways in which EID outbreaks develop, with a focus on these drivers, can help to suggest preventive actions by specifically tackling influential sub-drivers [17]. In addition, such modelling can help to pinpoint where early warning surveillance could be targeted.

Data on sub-drivers have therefore been used by modellers to identify hotspots where EIDs may next occur, or which drivers have the greatest influence on the likelihood of their occurrence [18]. To minimise error and bias when modelling how drivers may interact, it is essential to have good-quality data describing these drivers. A lack of data or poor-quality data can limit modelling outputs as the reliability of these models depends on the robustness of the data inputs. This study assesses the quality of the data available on sub-drivers of West Nile virus (WNV) as a case study, and as an application of these data in estimating the likelihood of potential future outbreaks of WNV through modelling and forecasting.

Examples of data for drivers of emerging infectious diseases

When considering data, it is appropriate to investigate all sources, including the innovative and novel as well as the more traditional. New technologies have enabled rapid access to many more types of information. Integrating data elements from the micro level (genes) to the macro level (social, political, climate, and global mobility patterns) could provide better information systems to anticipate and prepare for epidemics [7].

Traditional data, such as those on passenger travel or trade in live animals and animal products, are useful to monitor routes via which the global spread of diseases can occur, while more novel techniques, such as data mined from Internet search engines or social media, can prove useful to monitor and prevent the more localised spread of an EID. Data scientists at Google were the first to use data gathered online to track infectious diseases with Google Flu Trends infection forecasting. They did this by combing through queries on the search engine to look for small increases in 'flu-related terms', such as symptoms or vaccine availability [19]. The use of such sources of data needs proper scrutiny as it may be impacted by evolving events. For instance, when Google Flu Trends was used during the H1N1 pandemic in 2009, people's behaviour changed in the wake of media reports. Uninfected people also began searching online for flu symptoms, and an increasing number of false results were observed [19]. Similar challenges may happen for EIDs, when Internet searches spike out of general fear or curiosity [19].

A further advance is the use of citizen science derived from active public involvement in scientific research. Data sourced in this way are prone to variable quality. For example, data collection protocols may not be followed or correctly implemented [20]. The use of citizen science surveillance can, however, be beneficial, particularly where resources are limited or where large-scale surveillance, both in time and space, would not otherwise be financially feasible. One such example is Mosquito Alert (http://www.mosquitoalert.com) [21].

Data derived from Earth observation (EO) are of great interest for work on EIDs as they can offer continuous spatial and temporal coverage. Earth observation comprises data on the Earth's physical, chemical and biological systems via remote sensing technologies, usually involving satellites carrying imaging devices. Earth observation data can therefore be used to monitor and assess the status of, and changes in, the natural and human-made environment [22], and can provide a timely source of data across countries, regions and cities on natural resources and ecosystems. These data can be combined with other geo-referenced socio-demographic, economic and public administration data to ensure that indicators and analysis are more relevant and more effectively targeted. This has the added benefit that there will often be historical data, which can be used to explore the relationship between EIDs and the environment by comparing different time periods and deriving trends [23].

Another area where EO data are useful is in monitoring climatic changes as a driver of EIDs [22]. The Climate Change Initiative generates continuous global data records for key aspects of the climate, known as Essential Climate Variables [22]. Earth observation satellites also have the advantage of providing accurate measurements of areas that are difficult to reach, such as polar regions, which are important for the development of climate hazard early warning systems and measurements of climate change.

Although there is a plethora of data sources, a proven system that integrates several sources of data is necessary for EID early warning surveillance. However, combining the various data sets highlights the difficulty in quantifying information at different scales of space and time [19]. One of the key aspects before deciding to start combining data is an assessment of their quality.

Quality of data for drivers

Assessing the quality of data is a challenge and depends on defining what constitutes quality and what is most important to the application(s) for which the data are being used [24]. The key to the use of data characteristics is an understanding of what is important in any specific context – for example, in the case of modelling the drivers of EIDs. These requirements will, in turn, help to define the criteria best used for assessing data quality.

There have been many published sets of data characteristics – see, for example, Pipino *et al.* [25], Cai and Zhu [26], and the United States Centers for Disease Control and Prevention [27] – and their assessment [28, 29, 30, 31]. The data characteristics used here, and their definitions, were selected by the present authors from these citations as the most relevant for reducing uncertainty when modelling EID outbreaks (Table I). Thus, these definitions are a synthesis of those cited above to best fit the requirements for the current authors' purposes.

Data sets for specific sub-drivers are likely to be variable across these characteristics, and it is unlikely that a single data set will satisfy all of them. However, each has a role to play. For example, Internet search engines, such as Google traffic, can provide data with a very small time-lag between occurrence and collection (timeliness), whereas EO data can provide very high coverage of data points at varying levels of resolution (granularity). Some data may not fully satisfy certain quality requirements, e.g. timeliness or accuracy, but there may be an acceptable trade-off, particularly when the benefit of including such data outweighs the uncertainty introduced with an imperfect data set.

Case study – West Nile virus

A quality assessment framework, using the characteristics specified in Table I, was applied to WNV as a zoonotic case study to observe whether sufficient quality data were available to estimate the likelihood of a future outbreak in a country where WNV has not yet been observed. This work was done as part of a European Union-funded project, aimed at improving outbreak reaction and early warning by mining data from multiple sources (Versatile Emerging Infectious Disease Observatory). West Nile virus (family Flaviviridae; genus *Flavivirus*) was chosen because it is currently present in parts of Europe and is thought to be one of the EIDs that may expand towards western and northern Europe as a consequence of climate change. West Nile virus is maintained by a complex transmission cycle involving multiple species of mosquitoes and birds [32]. The enzootic cycle is driven by continuous virus transmission to susceptible bird species through adult mosquito blood-meal feeding, which results in virus amplification. There is genetic diversity among the viral strains identified in vectors and birds in Europe, including several different phylogenetic clades [33], suggesting that there have been repeated introductions of WNV into Europe and the virus has since overwintered in mosquitoes [34], and possibly other hosts (see Reiter [35] and Sambri *et al.* [36] for reviews).

The epidemiological situation of WNV in Europe is heterogeneous; some European countries report outbreaks in humans and animals every year while others have never reported any

Table I

Characteristics and scores (with definitions by the current authors) against which data quality can be measured

	Characteristic	Description	1 = High uncertainty	2 = Medium uncertainty	3 = Low uncertainty
1	**Accuracy and precision (relevance)**	The data capability to measure the sub-driver; the degree to which the data correctly describe the sub-driver; to what degree the data will not cause ambiguity	No precise data are available; data set is a proxy data set for the sub-driver; data available may cause ambiguity; data have not been verified	Data are not proxy data but there may be some ambiguity introduced as the measurement scale may be too broad for modelling purposes	Data are relevant to the purposes for which they are to be used and relate directly to the requirements of the model. There is little uncertainty of the accuracy of the data in describing the sub-driver
2	**Reliability and consistency**	The degree to which the data in the data set are consistent with data in another data set; the absence of difference when comparing two or more representations of a sub-driver	The data values for a specific sub-driver differ by a substantial margin when compared to those of another data set, increasing uncertainty in the model	The data values for a specific sub-driver differ by a small margin when compared to those of another data set, which may introduce some uncertainty in the model	The data values for a specific sub-driver show little or no variation, reducing uncertainty over the reliability of the data
3	**Timeliness**	The degree to which the data represent the sub-driver for the specified time frame, i.e. whether the data are regularly updated and the time delay from data generation to use	Data are old or not available quickly; data are considered out of date and will not capture relevant changes to the sub-drivers that are happening at the present time	Some time-lag between the actual data delivery and the specific time frame addressed in the model. Introduces some uncertainty	Data are collected from within a relevant time period to be able to draw accurate conclusions; data are available quickly to support modelling needs. Data updates occur frequently and allow any changes to be rapidly captured
4	**Completeness and comprehensiveness**	The extent to which the required data are available compared to the amount of data needed; the proportion of available data against the potential of 100% completeness; whether the deficiency of data will impact the use or accuracy of the model	Very incomplete data are available when compared to the requirements of the model, potentially leading to high levels of uncertainty	Some data are available, but more would be needed to achieve the full requirements of the model and reduce uncertainty to a low level	A complete data set is available so that all requirements for the model are fulfilled; data are accompanied by appropriate metadata including limitations in use
5	**Availability and accessibility**	The degree to which data exist and whether access is limited; the extent to which data are available or easily and quickly retrievable; whether a data access interface is provided; the difficulty level for users to obtain the data	Data are not directly available for use and will incur a cost or require an application for use to a private company or competent authority. Applying for the data may take time and the lack of accessibility can prevent access to data, which can make model results highly uncertain	A data access interface is not available, making data mining a difficult process that can introduce error and increase uncertainty	Data are freely available and accessible; there is unrestricted access to data with a comprehensive data access interface allowing easy access to the user
6	**Granularity**	The level of resolution of the data when compared to the requirements of the model	Data are aggregated and/or summarised and thereby potentially missing important insights at a local level or time span. Model results will be generalised, which will introduce high uncertainty	Data are available but in greater intervals/spatial areas than required by the model such that the lack of resolution may introduce some uncertainty into the model's results	The level of detail (or resolution) of the data describing the sub-driver is suitable for modelling requirements so that accurate conclusions can be drawn at the required scale of detail

autochthonous cases. Since 2019, WNV has been reported in humans or horses as far north as Brandenburg, Berlin and Saxony-Anhalt in Germany [37]. The identification of the first seven clinical cases of autochthonous human and wild bird WNV infections in 2020 in the Netherlands confirms the virus's potential to expand its geographical distribution northward [38, 39]. The most recent sequences that clustered with the Dutch WNV sequences originated from Germany, suggesting this may be the origin of the virus found in the Netherlands [39].

A review of the literature on emerging vector-borne zoonotic diseases (VBZDs) in general found that the most common potential driver of disease emergence referred to by authors was a change in land use. But, for many diseases, the driver was unknown, illustrating the complexity and multifactorial nature of VBZD emergence and spread. Land use change and international trade and commerce were more frequently cited as important for emergence than climate and weather. Ecological impacts caused by anthropogenic landscape changes, local contact opportunities, and humans and vectors moving between or to new climatically suitable regions, due to globalisation, were implicated more often than climate change alone as drivers of VBZDs in the literature reviewed [40].

A literature search specifically targeted at the importance of driver interactions for the emergence of WNV is shown in Figure 1 [40–59]. (It is acknowledged that these references are not exhaustive.) Mosquito activity varies both spatially and temporally, according to climatic conditions, and, since the vertebrate hosts are birds, migratory birds have the potential to carry the virus over long distances. The coalescence of bird migration, landing periods, peak vector abundance and suitable temperatures for the extrinsic incubation period [35] are all drivers needed to complete the enzootic cycle of WNV before its initial spread into the human population [60].

The current authors assessed the quality of the available data sources to describe the drivers shown in Figure 1 in a suitable model, to estimate the likelihood of future outbreaks. The results are shown in Table II. Each data source was scored by the authors against the six criteria shown in Table I.

Results

Table II illustrates that the data describing the various sub-drivers identified as contributing to WNV outbreaks are of varying quality. The characteristic with the lowest score over all the data sets was 'completeness'. This is an important characteristic as an incomplete data set can increase uncertainty and potentially lead to erroneous conclusions from modelling.

Genetic sequencing EO/Google Earth for land use and meteorological data sets all scored highly for most of the characteristics assessed. These data sets are freely available and have good granularity, although timeliness can be an issue for some data. For example, genetic sequencing data may only be released after publication in the researcher's original article. The use of such data when modelling the likelihood of a WNV outbreak can reduce uncertainty and increase

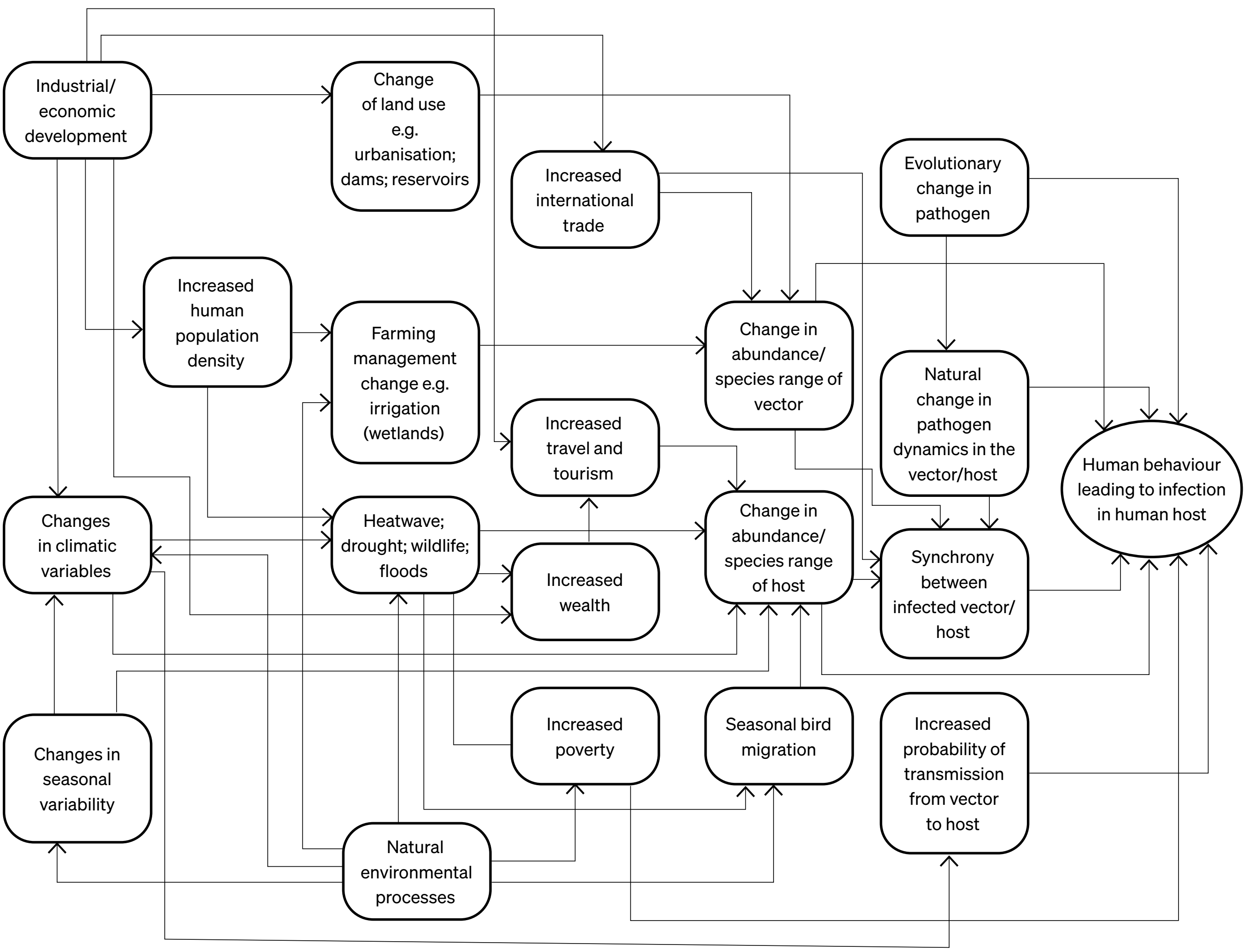

Figure 1

Interaction of sub-drivers resulting in the release of West Nile virus (sub-drivers are not exhaustive) [40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59]

Table II

Examples of data sources for sub-drivers and the quality of their data to predict the emergence of West Nile virus in humans in an area with no previous detection of the virus in humans or birds

See Table I for characteristic/scoring definitions

Sub-driver	Data requirements	Data sources (with example sources)	Accuracy and precision	Reliability and consistency	Timeliness	Completeness	Availability and accessibility	Granularity	Total
Seasonal variables/ changes in seasonal variability/climatic variables	Gridded climate observations	Meteorological data (https://psl.noaa.gov)	3	3	3	3	3	2	**17**
Natural environmental processes	Geological events	Geological survey data (https://www.bgs.ac.uk)	3	2	2	2	2	3	**14**
Industrial and economic development	Change of land use	Urban light intensity (https://urbannext.net)	1	1	3	1	2	2	**10**
		International Monetary Fund (https://www.imf.org)	1	2	1	1	3	1	**9**
		World Bank Open Data (https://data.worldbank.org)	1	2	1	1	3	1	**9**
	Farming management: tree density; proximity to wetlands; deforestation; urbanisation	Earth observation (https://earthobservatory.nasa.gov/images/147612/counting-trees-in-africas-drylands)	3	3	2	3	3	3	**17**
		Google Earth (https://earth.google.com)	3	3	3	2	2	3	**16**
		Lakes and wetlands mapping (https://www.medwet.org/codde/8_EarthObservation/EarthObservation-Manual.pdf)	3	3	2	2	2	3	**15**
		Normalised difference vegetation index (https://eos.com)	3	3	2	2	1	2	**13**
Increasing human population density	Human population density	Nomenclature of territorial units for statistics (https://ec.europa.eu/eurostat/web/nuts/background)	2	2	2	2	3	2	**13**
		Gridded population of the world (https://sedac.ciesin.columbia.edu/data/collection/gpw-v4)	2	2	2	2	2	2	**12**
		Earth observation (https://neo.gsfc.nasa.gov/view.php?datasetId=SEDAC_POP)	2	2	2	2	2	2	**12**
		Urban light intensity (https://urbannext.net)	1	1	3	1	2	2	**10**
		Census data (https://census.gov.uk)	3	3	1	3	1	3	**14**
Increased poverty	Poverty rates	Census data (https://census.gov.uk)	2	2	1	2	1	3	**11**

confidence in the outputs. Six data sources were available to map bird migration routes to describe the sub-driver of seasonal migration and expansion of range of establishment. While these data sets all scored the same, with medium uncertainty for all characteristics, using a combination of data from all six data sources may reduce the uncertainty surrounding these inputs.

Traditional data sources, such as census data and trade data, which are important when modelling changes in human

Table II (cont.)

Sub-driver	Data requirements	Data sources (with example sources)	Accuracy and precision	Reliability and consistency	Timeliness	Completeness	Availability and accessibility	Granularity	Total
Increased travel and tourism	Flight data	International Air Transport Association (https://www.iata.org)	3	3	3	3	1	3	**16**
	Maritime data	Maritime and shipping statistics (https://www.statista.com)	3	3	2	3	1	2	**14**
	Movements and numbers of people by destination	Mobile phone data of localised movements (https://www.nature.com/articles/s41598-021-81873-6)	1	1	2	1	1	2	**8**
Increased international trade	International trade	Food and Agriculture Organization of the United Nations (https://www.fao.org)	2	2	2	2	3	1	**12**
Seasonal bird migration/short-distance migration; expansion in range of establishment	Bird migration routes	Bird migration network (https://ebird.org)	2	2	2	2	2	2	**12**
		Bird ringing (https://euring.org)	2	2	2	2	2	2	**12**
		Bird ringing (https://www.bto.org)	2	2	2	2	2	2	**12**
		Bird tracking (https://www.bto.org/understanding-birds/articles/bird-tracking-%E2%80%94-masterclass)	2	2	2	2	2	2	**12**
		Google Earth (https://earth.google.com)	2	2	3	2	2	3	**14**
Change in abundance/species range of host	Bird numbers and range	Public gardens birdwatch (https://www.birdcount.org)	2	2	2	2	2	3	**13**
		Surveillance (https://pecbms.info)	2	2	2	2	3	2	**13**
		Waterbird population surveillance (https://wpe.wetlands.org)	2	2	2	2	2	2	**12**
		European Bird Census Council (https://www.ebcc.info)	2	2	2	2	3	2	**13**
		Citizen science (http://datazone.birdlife.org/info/citizenscience)	2	2	2	2	2	2	**12**
Change in abundance/species range of vector	Vector numbers	Surveillance (https://www.cdc.gov/westnile/resourcepages/mosqSurvSoft.html)	2	2	2	1	3	2	**12**
		Surveillance (https://www.ecdc.europa.eu/en/disease-vectors/surveillance-and-disease-data/mosquito-maps)	2	2	2	1	3	2	**12**
		Citizen science (http://www.mosquitoalert.com/en)	2	2	2	2	3	2	**13**
Synchrony between infected birds and high mosquito abundance	Bird numbers and range/vector abundance	Surveillance (https://www.frontiersin.org/articles/10.3389/fpubh.2017.00236/full)	1	1	2	1	1	1	**7**
	Rate of transmission	Scientific research (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4520649)	2	2	2	1	3	2	**12**

Table II (cont.)

Sub-driver	Data requirements	Data sources (with example sources)	Accuracy and precision	Reliability and consistency	Timeliness	Completeness	Availability and accessibility	Granularity	Total
Change in pathogen dynamics in the host/vector	Evolutionary change in pathogen in vector/s	Genetic sequencing data (https://parasitesandvectors.biomedcentral.com/articles/10.1186/s13071-014-0542-2)	2	1	2	1	2	3	**11**
		Genetic sequencing data (https://www.ncbi.nlm.nih.gov/genbank)	3	3	2	3	3	3	**17**
	Evolutionary change in pathogen in birds	Genetic sequencing data (https://parasitesandvectors.biomedcentral.com/articles/10.1186/s13071-020-04399-2)	3	2	2	1	2	3	**13**
Human behaviour	Use of repellents	Scientific survey (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6034598)	2	2	2	1	1	1	**9**
	Length of time spent outdoors	Scientific survey (https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2133.2010.10165.x)	2	2	2	1	1	1	**9**
	Human protective measures	Scientific survey (https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-015-1918-8)	2	2	2	1	1	1	**9**
Total			**85**	**84**	**82**	**72**	**83**	**85**	**491**

population density and the potential global movement of pathogens, respectively, tended to suffer from lack of timeliness. These data are usually released annually or, in the case of census data, collected only once every few years (for example, five or ten). A country's demography may not generally change within a few weeks or months, but this could be an important consideration during times of conflict or famine, resulting in unusually large movements of people. Additionally, while drivers of emergence of an EID are processes that can take many years, drivers of spread may need more timely data, since spread can happen over much shorter time spans.

Social media data sources tended to score lower for 'accuracy and precision' and 'reliability and consistency'. On the plus side, Web-based disease surveillance methods are adaptable, low cost, and can be operated in real time. Particular progress has been made in their use during the COVID pandemic, such as in the United Kingdom Zoe Health Study (https://health-study.joinzoe.com). Sub-drivers that required data from scientific studies, such as those relating to human behaviour, did not generally score highly for 'availability and accessibility'. Such studies are very often made public via publication in journals, but full data are sometimes not disclosed, supplied as supplementary material or only available on request to the author. Access to data in publications can also be restricted if the journal is not open access but requires subscription.

While the data sources were assessed in regard to how well they fulfilled the criteria described in Table I, the present authors acknowledge that there may also be differences in the 'weight' of the criteria. For example, human census data usually do not change dramatically, so the timeliness criterion could be given less weight in this instance. Similarly, since suitable temperatures for the extrinsic incubation period of WNV can vary, accuracy and granularity would have more 'weight' for the meteorological data source [61].

Discussion

This paper assessed the quality of the available data for sub-drivers of WNV as a case study for a zoonotic EID. The objective was to gain perspective on their value in predicting potential future outbreaks through modelling and forecasting. The ability to minimise uncertainty when modelling the likelihood of an EID occurring would enable preventive measures to be taken. At the very least, it would permit a rapid response, which could limit the spread and consequences of an EID outbreak. Modelling EIDs can help to simulate the effects of preventive actions aimed at reducing the effects of specific sub-drivers (or combinations of sub-drivers) that result in pathogen emergence.

There are currently no existing standards for mining and analysing data from the Internet and results or decisions based on Internet sources have previously been classified as low quality [62]. However, novel methods of data collection, such as citizen science and EO, have encouraged data quality-assessment schemes to validate their use as data sources [20, 63, 64].

Barriers preventing or delaying data sharing can be a problem for modellers in regard to data availability or accessibility. It is increasingly being recognised that, in order to prevent and control EID outbreaks, interdisciplinary collaborations in all aspects of healthcare for people, animals and the environment are necessary. Most barriers can be identified in the initial stages of sampling and sequencing. These are stages of primary importance for outbreak control and public health response. Early identification of changes in pathogen genomes is of the utmost importance for signalling (re-)emerging infectious diseases and developing diagnostic tests and vaccines. This has clearly been shown in the recent outbreak of severe acute respiratory syndrome coronavirus 2, more commonly known as SARS-CoV-2. The Nagoya Protocol of 2010 was developed to facilitate access to genetic resources and ensure fair and equitable sharing of benefits arising from their use [65].

Table II illustrates the data available to define the known sub-drivers of a zoonotic outbreak of WNV, as documented in the literature. However, assessing only these known drivers leads to the concern that unknown drivers are not being accounted for, particularly at the local level. If potential unknown drivers also have an influence, then the uncertainty surrounding forecasting will increase and detailed planning may be inappropriate. Drivers may also need to occur in a certain synchrony for an EID outbreak to occur.

The sub-drivers documented in Table II are based on knowledge from the literature, which has cited these drivers as being important for WNV outbreaks. Association between occurrence of disease and drivers does not, however, equal causation. That is, there may be an association between the two but this association may be influenced by other sub-drivers. The process then becomes similar to a self-fulfilling prophecy, whereby outputs of predictive models can only tell us about the relationship between drivers that have been selected beforehand. For example, it is still unknown how WNV first arrived in New York City [66] but it has since spread across North, South and Central America and the Caribbean. This is because predicting a WNV outbreak is particularly challenging, due to the complexity of the drivers involved.

It may be possible to use the available data in a different way to generate more knowledge. For example, a recent study by Priyadarsini *et al.* [67] used a theory-building approach and tool to rank anthropogenic drivers by their impacts on other factors and the interdependence among them. The results revealed that changes in any individual factor in the study could directly or indirectly help to cause repeated epidemics. Expanding human populations, globalisation and civil unrest were the top factors. More research could be carried out into the inter-relationships between drivers and how they influence one another. In regard to both 'unknown knowns and unknowns', machine learning could be explored to reveal previously undetected relationships/patterns in data. It may also be possible to introduce a so-called 'chaos driver' as a scenario, to analyse how this affects the pathogen's potential for emergence and spread. Machine learning could also reduce the elements of unconscious bias that may exist in traditional modelling.

By modelling the sub-drivers of any particular EID, it may be possible to assess the likelihood of an outbreak occurring or at least enable preventive measures that could limit the consequences of an outbreak. For example, in the case of African swine fever (ASF), while it is not possible to follow illegal movements of pig meat, it is possible to identify poor biosecurity as a sub-driver for onward transmission of the virus. Knowing the locations of pig farms and encouraging enhanced biosecurity at these premises to prevent further outbreaks of ASF can therefore limit the potential spread of this EID. Thus, the availability of good-quality data is essential to assess the likelihood of where EIDs may occur and to identify points in the risk pathway where preventive measures may be taken. Performing an assessment of the data quality for drivers of EIDs can help to identify areas where additional or different data sources may reduce uncertainty in model outcomes when assessing the likelihood of an outbreak.

Acknowledgements

This work was supported by the Versatile Emerging Infectious Disease Observatory, funded by the European Union Horizon 2020 Research and Innovation Programme under Grant Agreement No. 874735, with additional funding from the Department for Environment, Food and Rural Affairs, United Kingdom.

Évaluer la qualité des données relatives aux facteurs d'émergence de maladies

V. Horigan, L. Kelly, A. Papa, M.P.G. Koopmans, R.S. Sikkema, L.G.H. Koren & E.L. Snary

Résumé

Les facteurs d'émergence sont des éléments ayant le potentiel direct ou indirect d'influencer la probabilité d'émergence ou de réémergence d'une maladie infectieuse. Il est probablement rare qu'une maladie infectieuse émergente apparaisse en raison d'un seul facteur ; c'est plutôt un faisceau de sous-facteurs (éléments pouvant avoir une influence sur un même facteur) qui contribue à ce que les conditions soient réunies pour qu'un agent pathogène puisse (ré)émerger et s'établir. Les concepteurs de modèles ont donc utilisé les données relatives aux sous-facteurs pour identifier les zones sensibles où les prochaines maladies infectieuses émergentes pourraient survenir, ou pour faire une estimation des sous-facteurs ayant la plus grande influence sur la probabilité de leur occurrence. Les chercheurs ont besoin de données de qualité pour décrire ces sous-facteurs, afin de minimiser le risque d'erreur et de biais lors de la modélisation de l'interaction entre les différents sous-facteurs, et de contribuer ainsi à mieux prédire la probabilité d'apparition d'une maladie infectieuse émergente. Les auteurs présentent une étude de cas qui a consisté à évaluer la qualité des données disponibles relatives aux sous-facteurs d'émergence du virus de la fièvre de West Nile au regard de différents critères. Il est apparu que la qualité des données était variable au regard des critères examinés.

Le paramètre dont le score était le plus bas est celui de la complétude – le fait que suffisamment de données soient disponibles pour répondre à toutes les exigences du modèle. Il s'agit pourtant d'un paramètre important car des données incomplètes peuvent inciter à tirer des conclusions erronées des études de modélisation. La disponibilité de données de bonne qualité est essentielle pour réduire l'incertitude lors de l'estimation de la probabilité d'apparition de maladies infectieuses émergentes dans des zones déterminées, ainsi que pour identifier les points critiques de concrétisation du risque où des mesures préventives pourraient être mises en place.

Mots-clés

Facteurs d'émergence – Maladie infectieuse émergente – Modélisation – Prédiction des maladies – Prévention – Qualité des données – Virus de la fièvre de West Nile.

Evaluación de la calidad de los datos sobre los inductores de la aparición de enfermedades

V. Horigan, L. Kelly, A. Papa, M.P.G. Koopmans, R.S. Sikkema, L.G.H. Koren & E.L. Snary

Resumen

Los inductores o factores de inducción [*drivers*] son aquellos que, directa o indirectamente, pueden influir en la probabilidad de que surjan o resurjan enfermedades infecciosas. Todo indica que rara vez una enfermedad infecciosa emergente aparece por efecto de un solo factor de inducción, sino que es probable que haya más bien una combinación de «subfactores de influencia» [*sub-drivers*] (factores que pueden influir en un inductor) que cree condiciones propicias para que un patógeno (re)surja y logre asentarse. Los creadores de modelos, por consiguiente, se han servido de datos sobre estos subfactores de influencia para localizar aquellas zonas donde con mayor probabilidad puedan aparecer próximamente enfermedades infecciosas emergentes o para determinar cuáles son los subfactores que más influyen en la probabilidad de que ello ocurra. Para reducir al mínimo los errores y sesgos al modelizar la interacción entre los subfactores y ayudar así a calcular la probabilidad de que surja una enfermedad infecciosa emergente, los investigadores necesitan datos de buena calidad para caracterizar estos subfactores. En el análisis expuesto por los autores se utilizó el virus del Nilo Occidental como ejemplo de estudio para evaluar, con arreglo a diversos criterios, la

calidad de los datos existentes sobre los subfactores que inciden en la aparición de este virus. Lo que se constató, en relación con el grado de cumplimiento de los criterios, es que esos datos eran de calidad variable.
La característica o parámetro que deparó la puntuación más baja fue la completud, es decir, la existencia de datos suficientes para aportar al modelo toda la información requerida para que este funcione bien. Se trata de una característica importante, pues un conjunto incompleto de datos podría llevar a extraer conclusiones erróneas de los estudios de modelización. Por ello, para reducir la incertidumbre a la hora de calcular la probabilidad de que en cierto lugar surjan brotes de enfermedades infecciosas emergentes y de determinar, dentro de la cadena de materialización del riesgo, aquellos eslabones en los que cabe adoptar medidas preventivas, es indispensable disponer de datos de buena calidad.

Palabras clave

Calidad de los datos – Enfermedades infecciosas emergentes – Factores – Modelización – Predicción de la aparición de enfermedades – Prevención – Virus del Nilo Occidental.

References

[1] Morens D.M. & Fauci A.S. (2012). – Emerging infectious diseases in 2012: 20 years after the Institute of Medicine report. *mBio*, **3** (6), e00494-12. https://doi.org/10.1128/mBio.00494-12

[2] Wentholt M.T.A., Cardoen S., Imberechts H., Van Huffel X., Ooms B.W. & Frewer L.J. (2012). – Defining European preparedness and research needs regarding emerging infectious animal diseases: results from a Delphi expert consultation. *Prev. Vet. Med.*, **103** (2–3), 81–92. https://doi.org/10.1016/j.prevetmed.2011.09.021

[3] Woolhouse M.E.J. & Gowtage-Sequeria S. (2005). – Host range and emerging and reemerging pathogens. *Emerg. Infect. Dis.*, **11** (12), 1842–1847. https://doi.org/10.3201/eid1112.050997

[4] Jones K.E., Patel N.G., Levy M.A., Storeygard A., Balk D., Gittleman J.L. & Daszak P. (2008). – Global trends in emerging infectious diseases. *Nature*, **451** (7181), 990–993. https://doi.org/10.1038/nature06536

[5] Suk J.E. & Semenza J.C. (2011). – Future infectious disease threats to Europe. *Am. J. Public Health*, **101** (11), 2068–2079. https://doi.org/10.2105/AJPH.2011.300181

[6] Olson S.H., Benedum C.M., Mekaru S.R., Preston N.D., Mazet J.A.K., Joly D.O. & Brownstein J.S. (2015). – Drivers of emerging infectious disease events as a framework for digital detection. *Emerg. Infect. Dis.*, **21** (8), 1285–1292. https://doi.org/10.3201/eid2108.141156

[7] World Health Organization (WHO) (2016). – Meeting report: WHO informal consultation: anticipating emerging infectious disease epidemics. WHO, Geneva, Switzerland, 72 pp. Available at: https://apps.who.int/iris/handle/10665/252646 (accessed on 29 April 2022).

[8] Semenza J.C., Lindgren E., Balkanyi L., Espinosa L., Almqvist M.S., Penttinen P. & Rocklöv J. (2016). – Determinants and drivers of infectious disease threat events in Europe. *Emerg. Infect. Dis.*, **22** (4), 581–589. https://doi.org/10.3201/eid2204

[9] Richardson J., Lockhart C. […] & Poppy G. (2016). – Drivers for emerging issues in animal and plant health. *EFSA J.*, **14** (Suppl. 1), e00512. https://doi.org/10.2903/j.efsa.2016.s0512

[10] Loh E.H., Zambrana-Torrelio C., Olival K.J., Bogich T.L., Johnson C.K., Mazet J.A.K., Karesh W. & Daszak P. (2015). – Targeting transmission pathways for emerging zoonotic disease surveillance and control. *Vector Borne Zoonotic Dis.*, **15** (7), 432–437. https://doi.org/10.1089/vbz.2013.1563

[11] Rosenthal S.R., Ostfeld R.S., McGarvey S.T., Lurie M.N. & Smith K.F. (2015). – Redefining disease emergence to improve prioritization and macro-ecological analyses. *One Health*, **1**, 17–23. https://doi.org/10.1016/j.onehlt.2015.08.001

[12] Wolfe N.D., Dunavan C.P. & Diamond J. (2007). – Origins of major human infectious diseases. *Nature*, **447** (7142), 279–283. https://doi.org/10.1038/nature05775

[13] Institute of Medicine, Board on Global Health & Committee on Emerging Microbial Threats to Health in the 21st Century (2003). – Microbial threats to health: emergence, detection, and response (M.S. Smolinski, M.A. Hamburg & J. Lederberg, eds). National Academies Press, Washington, DC, United States of America, 397 pp. https://doi.org/10.17226/10636

[14] Brandt A.M. & Botelho A. (2020). – Not a perfect storm – Covid-19 and the importance of language. *N. Engl. J. Med.*, **382** (16), 1493–1495. https://doi.org/10.1056/NEJMp2005032

[15] Ahmed S., Dávila J.D., Allen A., Haklay M.M., Tacoli C. & Fèvre E.M. (2019). – Does urbanization make emergence of zoonosis more likely? Evidence, myths and gaps. *Environ. Urban.*, **31** (2), 443–460. https://doi.org/10.1177/0956247819866124

[16] Burgos S. & Otte J. (2010). – Managing the risk of emerging diseases: from rhetoric to action. HPAI Research Brief No. 22. Department for International Development, London, United Kingdom, 4 pp. Available at: https://www.gov.uk/research-for-development-outputs/managing-the-risk-of-emerging-diseases-from-rhetoric-to-action (accessed on 29 April 2022).

[17] Sproul T., Zilberman D., Roland-Holst D. & Otte J. (2009). – A case for multilateral investment in avian flu prevention. HPAI Research Brief No. 17. Department for International Development, London, United Kingdom, 4 pp. Available at: https://www.gov.uk/research-for-development-outputs/a-case-for-multilateral-investment-in-avian-flu-prevention (accessed on 29 April 2022).

[18] Pigott D.M., Deshpande A. […] & Hay S.I. (2017). – Local, national, and regional viral haemorrhagic fever pandemic potential in Africa: a multistage analysis. *Lancet*, **390** (10113), 2662–2672. https://doi.org/10.1016/S0140-6736(17)32092-5

[19] Eisenstein M. (2018). – Infection forecasts powered by big data. *Nature*, **555** (7695), S2–S4. https://doi.org/10.1038/d41586-018-02473-5

[20] Balázs B., Mooney P., Nováková E., Bastin L. & Jokar Arsanjani J. (2021). – Data quality in citizen science.

In The science of citizen science (K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson & K. Wagenknecht, eds). Springer, Cham, Switzerland, 139–157. https://doi.org/10.1007/978-3-030-58278-4_8

[21] Lawson B., Petrovan S.O. & Cunningham A.A. (2015). – Citizen science and wildlife disease surveillance. *EcoHealth*, **12** (4), 693–702. https://doi.org/10.1007/s10393-015-1054-z

[22] European Space Agency (ESA) Space Solutions (2020). – Newcomers Earth observation guide. Version 1.9. ESA Space Solutions, Didcot, United Kingdom. Available at: https://business.esa.int/newcomers-earth-observation-guide (accessed on 29 April 2022).

[23] Day J., Ashfield S., Brown D., Gale P., Heeley L., Snary E., Taylor R., Way L. & Jones G. (2021). – Copernicus User Uptake (CUU): applying Earth observation (EO) to horizon scanning for emerging infectious diseases (EIDs). Joint Nature Conservation Committee (JNCC) Report No. 676. JNCC, Peterborough, United Kingdom, 45 pp. Available at: https://hub.jncc.gov.uk/assets/9efd4ce0-b7a9-4ad2-b7ed-f0e7646927b3 (accessed on 23 January 2023).

[24] European Statistical System (ESS) (2019). – Quality assurance framework of the European Statistical System. Version 2.0. ESS, Luxembourg City, Luxembourg, 57 pp. Available at: https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V2.0-final.pdf (accessed on 29 April 2022).

[25] Pipino L.L., Lee Y.W. & Yang R.Y. (2002). – Data quality assessment. *Commun. ACM*, **45** (4), 211–218. https://doi.org/10.1145/505248.506010

[26] Cai L. & Zhu Y. (2015). – The challenges of data quality and data quality assessment in the big data era. *Data Sci. J.*, **14**, 2. https://doi.org/10.5334/dsj-2015-002

[27] Centers for Disease Control and Prevention (CDC) (2021). – The six dimensions of EHDI data quality assessment. CDC, Atlanta, United States of America, 4 pp. Available at: https://www.cdc.gov/ncbddd/hearingloss/documents/dataqualityworksheet.pdf (accessed on 29 April 2022).

[28] Menéndez S., Steiner A., Witschi U., Danuser J., Weber U. & Regula G. (2008). – Data quality of animal health records on Swiss dairy farms. *Vet. Rec.*, **163** (8), 241–246. https://doi.org/10.1136/vr.163.8.241

[29] Chen H., Yu P., Hailey D. & Wang N. (2014). – Methods for assessing the quality of data in public health information systems: a critical review. *Stud. Health Technol. Inform.*, **204**, 13–18. https://doi.org/10.3233/978-1-61499-427-5-13

[30] Solomon M., Addise M., Tassew B., Balcha B. & Abebe A. (2021). – Data quality assessment and associated factors in the health management information system among health centers of southern Ethiopia. *PLoS One*, **16** (10), e0255949. https://doi.org/10.1371/journal.pone.0255949

[31] Schmidt C.O., Struckmann S., Enzenbach C., Reineke A., Stausberg J., Damerow S., Huebner M., Schmidt B., Sauerbrei W. & Richter A. (2021). – Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. *BMC Med. Res. Methodol.*, **21** (1), 63. https://doi.org/10.1186/s12874-021-01252-7

[32] Hadfield J., Brito A.F., Swetnam D.M., Vogels C.B.F., Tokarz R.E., Andersen K.G., Smith R.C., Bedford T. & Grubaugh N.D. (2019). – Twenty years of West Nile virus spread and evolution in the Americas visualized by Nextstrain. *PLoS Pathog.*, **15** (10), e1008042. https://doi.org/10.1371/journal.ppat.1008042

[33] Charrel R.N., Brault A.C., Gallian P., Lemasson J.-J., Murgue B., Murri S., Pastorino B., Zeller H., de Chesse R., de Micco P. & de Lamballerie X. (2003). – Evolutionary relationship between Old World West Nile virus strains. Evidence for viral gene flow between Africa, the Middle East, and Europe. *Virology*, **315** (2), 381–388. https://doi.org/10.1016/s0042-6822(03)00536-1

[34] Rudolf I., Betášová L., Blažejová H., Venclíková K., Straková P., Šebesta O., Mendel J., Bakonyi T., Schaffner F., Nowotny N. & Hubálek Z. (2017). – West Nile virus in overwintering mosquitoes, Central Europe. *Parasit. Vectors*, **10** (1), 452. https://doi.org/10.1186/s13071-017-2399-7

[35] Reiter P. (2010). – West Nile virus in Europe: understanding the present to gauge the future. *Euro Surveill.*, **15** (10), 19508. https://doi.org/10.2807/ese.15.10.19508-en

[36] Sambri V., Capobianchi M. […] & Landini M.P. (2013). – West Nile virus in Europe: emergence, epidemiology, diagnosis, treatment, and prevention. *Clin. Microbiol. Infect.*, **19** (8), 699–704. https://doi.org/10.1111/1469-0691.12211

[37] Ziegler U., Santos P.D. […] & Lühken R. (2020). – West Nile virus epidemic in Germany triggered by epizootic emergence, 2019. *Viruses*, **12** (4), 448. https://doi.org/10.3390/v12040448

[38] Vlaskamp D.R.M., Thijsen S.F.T., Reimerink J., Hilkens P., Bouvy W.H., Bantjes S.E., Vlaminckx B.J.M., Zaaijer H., van den Kerkhof H.H.T.C., Raven S.F.H. & Reusken C.B.E.M. (2020). – First autochthonous human West Nile virus infections in the Netherlands, July to August 2020. *Euro Surveill.*, **25** (46), 2001904. https://doi.org/10.2807/1560-7917.ES.2020.25.46.2001904

[39] Sikkema R.S., Schrama M. & van der Jeugd H. (2020). – Detection of West Nile virus in a common whitethroat (*Curruca communis*) and *Culex* mosquitoes in the Netherlands, 2020. *Euro Surveill.*, **25** (40), 2001704. https://doi.org/10.2807/1560-7917.ES.2020.25.40.2001704

[40] Swei A., Couper L.I., Coffey L.L., Kapan D. & Bennett S. (2020). – Patterns, drivers, and challenges of vector-borne disease emergence. *Vector Borne Zoonotic Dis.*, **20** (3), 159–170. https://doi.org/10.1089/vbz.2018.2432

[41] Paz S. & Semenza J.C. (2013). – Environmental drivers of West Nile fever epidemiology in Europe and Western Asia: a review. *Int. J. Environ. Res. Public Health*, **10** (8), 3543–3562. https://doi.org/10.3390/ijerph10083543

[42] Tran A., Sudre B., Paz S., Rossi M., Desbrosse A., Chevalier V. & Semenza J.C. (2014). – Environmental predictors of West Nile fever risk in Europe. *Int. J. Health Geogr.*, **13**, 26. https://doi.org/10.1186/1476-072X-13-26

[43] Marcantonio M., Rizzoli A., Metz M., Rosà R., Marini G., Chadwick E. & Neteler M. (2015). – Identifying the environmental conditions favouring West Nile virus outbreaks in Europe. *PLoS One*, **10** (3), e0121158. https://doi.org/10.1371/journal.pone.0121158

[44] Stilianakis N.I., Syrris V., Petroliagkis T., Pärt P., Gewehr S., Kalaitzopoulou S., Mourelatos S., Baka A., Pervanidou D., Vontas J. & Hadjichristodoulou C. (2016). – Identification of climatic factors affecting the epidemiology of human West Nile virus infections in northern Greece. *PLoS One*, **11** (9), e0161510. https://doi.org/10.1371/journal.pone.0161510

[45] Rosile P.A. & Bisesi M. (2017). – Novel indices of meteorological drivers of West Nile virus in Ohio *Culex* species mosquitoes from 2002–2006. *J. Environ. Health*, **79** (8), 16–23. Available at: https://www.jstor.org/stable/26329729 (accessed on 24 January 2023).

[46] Paull S.H., Horton D.E., Ashfaq M., Rastogi D., Kramer L.D., Diffenbaugh N.S. & Kilpatrick A.M. (2017). – Drought and immunity determine the intensity of West Nile virus epidemics and climate change impacts. *Proc. R. Soc. B: Biol. Sci.*, **284** (1848), 20162078. https://doi.org/10.1098/rspb.2016.2078

[47] Bakonyi T. & Haussig J.M. (2020). – West Nile virus keeps on moving up in Europe. *Euro Surveill.*, **25** (46), 2001938. https://doi.org/10.2807/1560-7917.ES.2020.25.46.2001938

[48] Candeloro L., Ippoliti C. […] & Conte A. (2020). – Predicting WNV circulation in Italy using Earth observation data and extreme gradient boosting model. *Remote Sens.*, **12** (18), 3064. https://doi.org/10.3390/rs12183064

[49] Fornasiero D., Mazzucato M., Barbujani M., Montarsi F., Capelli G. & Mulatti P. (2020). – Inter-annual variability of the effects of intrinsic and extrinsic drivers affecting West Nile virus vector *Culex pipiens* population dynamics in northeastern Italy. *Parasit. Vectors*, **13** (1), 271. https://doi.org/10.1186/s13071-020-04143-w

[50] Gayle A.A. (2020). – Artificial intelligence predicts and explains West Nile virus risks across Europe: extraordinary outbreaks determined by climate and local factors. *medRxiv*, **2020.07.24**, 20146829. https://doi.org/10.1101/2020.07.24.20146829

[51] Karki S., Brown W.M., Uelmen J., Ruiz M.O. & Smith R.L. (2020). – The drivers of West Nile virus human illness in the Chicago, Illinois, USA area: fine scale dynamic effects of weather, mosquito infection, social, and biological conditions. *PLoS One*, **15** (5), e0227160. https://doi.org/10.1371/journal.pone.0227160

[52] Shocket M.S., Verwillow A.B., Numazu M.G., Slamani H., Cohen J.M., El Moustaid F., Rohr J., Johnson L.R. & Mordecai E.A. (2020). – Transmission of West Nile and five other temperate mosquito-borne viruses peaks at temperatures between 23°C and 26°C. *elife*, **9**, e58511. https://doi.org/10.7554/eLife.58511

[53] Smith K.H., Tyre A.J., Hamik J., Hayes M.J., Zhou Y. & Dai L. (2020). – Using climate to explain and predict West Nile virus risk in Nebraska. *Geohealth*, **4** (9), e2020GH000244. https://doi.org/10.1029/2020GH000244

[54] Chowdhury P. & Khan S.A. (2021). – Global emergence of West Nile virus: threat and preparedness in special perspective to India. *Indian J. Med. Res.*, **154** (1), 36–50. https://doi.org/10.4103/ijmr.IJMR_642_19

[55] García-Carrasco J.-M., Muñoz A.-R., Olivero J., Segura M. & Real R. (2021). – Predicting the spatio-temporal spread of West Nile virus in Europe. *PLoS Negl. Trop. Dis.*, **15** (1), e0009022. https://doi.org/10.1371/journal.pntd.0009022

[56] Lorenz C., Salomão de Azevedo T. & Chiaravalloti-Neto F. (2022). – Impact of climate change on West Nile virus distribution in South America. *Trans. R. Soc. Trop. Med. Hyg.*, **116** (11), 1043–1053. https://doi.org/10.1093/trstmh/trac044

[57] Watts M.J., Monteys V.S.I., Mortyn P.G. & Kotsila P. (2021). – The rise of West Nile virus in southern and southeastern Europe: a spatial−temporal analysis investigating the combined effects of climate, land use and economic changes. *One Health*, **13**, 100315. https://doi.org/10.1016/j.onehlt.2021.100315

[58] Farooq Z., Rocklöv J., Wallin J., Abiri N., Odhiambo Sewe M., Sjödin H. & Semenza J.C. (2022). – Artificial intelligence to predict West Nile virus outbreaks with eco-climatic drivers. *Lancet Reg. Health Eur.*, **17**, 100370. https://doi.org/10.1016/j.lanepe.2022.100370

[59] Lourenço J., Barros S.C. […] & Obolski U. (2022). – West Nile virus transmission potential in Portugal. *Commun. Biol.*, **5** (1), 6. https://doi.org/10.1038/s42003-021-02969-3

[60] Esser H.J., Liefting Y., Ibáñez-Justicia A., van der Jeugd H., van Turnhout C.A.M., Stroo A., Reusken C.B.E.M., Koopmans M.P.G. & de Boer W.F. (2020). – Spatial risk analysis for the introduction and circulation of six arboviruses in the Netherlands. *Parasit. Vectors*, **13** (1), 464. https://doi.org/10.1186/s13071-020-04339-0

[61] Papa A. (2013). – West Nile virus infections in humans: focus on Greece. *J. Clin. Virol.*, **58** (2), 351–353. https://www.sciencedirect.com/science/article/abs/pii/S1386653213000814?via%3Dihub

[62] Osaghae E.N., Okokpujie K., Ndujiuba C., Okesola O. & Okokpujie I.P. (2018). – Epidemic alert system: a Web-based grassroots model. *Int. J. Electr. Comput. Eng.*, **8** (5), 3809–3828. https://doi.org/10.11591/ijece.v8i5.pp3809-3828

[63] Hunter J., Alabri A. & van Ingen C. (2013). – Assessing the quality and trustworthiness of citizen science data. *Concurr. Comput. Pract. Exp.*, **25** (4), 454–466. https://doi.org/10.1002/cpe.2923

[64] Bugbee K., Le Roux J., Sisco A., Kaulfus A., Staton P., Woods C., Dixon V., Lynnes C. & Ramachandran R. (2021). – Improving discovery and use of NASA's Earth observation data through metadata quality assessments. *Data Sci. J.*, **20** (1), 17. https://doi.org/10.5334/dsj-2021-017

[65] Convention on Biological Diversity (2011). – Nagoya Protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the Convention on Biological Diversity: text and annex. Secretariat of the Convention on Biological Diversity, Montreal, Canada, 25 pp. Available at: https://www.cbd.int/abs/doc/protocol/nagoya-protocol-en.pdf (accessed on 18 February 2023).

[66] Johnston B.L. & Conly J.M. (2000). – West Nile virus − where did it come from and where might it go? *Can. J. Infect. Dis.*, **11** (4), 175–178. https://doi.org/10.1155/2000/856598

[67] Priyadarsini S.L., Suresh M. & Huisingh D. (2020). – What can we learn from previous pandemics to reduce the frequency of emerging infectious diseases like COVID-19? *Glob. Transit.*, **2**, 202–220. https://doi.org/10.1016/j.glt.2020.09.003