Preprint version of proceedings paper presented at the annual meeting of the Society of Veterinary and Preventive Medicine (www.svepm.org.uk) in Inverness on 29-31 March 2017

MIGHTY MODELS FROM LITTLE DATA GROW: ESTIMATING ANIMAL DISEASE PREVALENCE

R.R.L. SIMONS*, V. HORIGAN, M. DE NARDI, G. RU, A.E. PENA AND A. ADKIN

SUMMARY

Global datasets relating to prevalence of animal pathogens are a useful input for risk assessments. However, missing data, and the resulting uncertainty, could potentially bias model outputs. This paper reviews the challenges associated with using freely available datasets and explores methods of estimating data, where necessary. By filling these data gaps, the final models can achieve more robust predictions with reduced uncertainty.

INTRODUCTION

Globalisation, with its associated increase in frequency of movement of animals, humans and trade products around the world, has been acknowledged as a risk factor for the spread of pathogens between countries and is a motivation behind the implementation of many pathogen incursion risk assessments (Simons et al., 2016). Thus, risk assessments are being conducted by both international organisations (ECDC, 2016), and individual countries, specifically those wishing to understand potential threats to their livestock populations and to avoid negative implications for trade of both live animals and animal products (Roberts et al., 2016).

One of the aims of the Animal Health and Welfare ERA-NET consortium (ANIHWA) funded project SPARE ('Spatial risk assessment framework for assessing exotic disease incursion and spread through Europe') is to develop a generic risk assessment framework for the entry of exotic animal pathogens into the European Union (EU) (SPARE, 2016). The first step in a risk assessment is termed the release or entry assessment. In general terms, this is defined as the ‘evaluation of the probability of introduction of an agent from its origin until the point of entry into a country or area’ (USDA, 2016). This necessitates the derivation of estimates for the animal disease situation in each ‘origin’ country of the world requiring data on animal disease prevalence (e.g. the number of recorded outbreaks per year), and demographic livestock data (e.g. the total number of animals and farms per species). In order to ensure a correct interpretation of these data, it is important that these data are comparable between countries. Furthermore, they should ideally arise from robust animal health surveillance which ensures that confidence in the health status of animals resident in the country of origin and those moving between countries is maintained and trade barriers are justified (Hoinville et al., 2013).

* Robin Richard Lacey Simons, Animal and Plant Health Agency (APHA), New Haw, Addlestone, Surrey, KT15 3NB, UK. Email: Robin.Simons@apha.gsi.gov.uk

One such data source is the World Animal Health Information Database (WAHIS) and its predecessor Handistatus II (OIE 2016a; OIE 2016b), which are maintained by the Office International des Epizooties (OIE). The OIE operates a global surveillance system which provides valuable data sources for animal diseases by collating information which all member countries (MCs) are obliged to report. The OIE is recognised as a reference organisation by the World Trade Organization and in 2016 has a total of 180 MCs. Each MC is expected to report the animal diseases that it detects within its territory. Therefore, the data available from the OIE database could be considered to be one of the most globally comprehensive. Nevertheless, all global datasets dealing with such data are inherently subject to differences in the quality of individual reporting and the monitoring systems for disease on which reports are based. From the point of view of utilising the OIE data in a quantitative risk assessment, one of the most important issues is 'missing data', i.e. MCs where the data are only reported as presence of a disease, without a numerical estimate of scale, or for which no information is reported at all. Absence of reported disease could be considered to imply disease free status of a country, when in reality this may not be the case.

To assist in obtaining objective estimates for risk assessments, it is therefore expedient to develop reliable methods that can utilise the data that are available to estimate values for MCs where data are missing. Previous research has used a number of different methods to deal with missing data in human health disease incidence datasets, including grouping countries based on predictive associations (McDonald et al., 2015). Presented here is the development of a method to obtain objective estimates for MCs with missing data. This method is illustrated with a case study for number of outbreaks of classical rabies in MCs, based on data on classical rabies outbreaks from the OIE databases (i.e. those outbreaks recorded as rabies). Results are presented to compare the effectiveness of the different grouping methods across multiple species by conducting statistical analysis of the results for MCs with available data.

MATERIALS AND METHODS

Overview

For the purposes of the analysis presented here, and for future integration within a larger risk assessment model, it was decided that the method employed to fill the data gaps should, as far as possible, be automated with generic rules implemented when certain data are absent. An imputation method, sometimes termed 'farcasting' by analogy with forecasting, was used whereby empirical models were fit to existing data and predictions were generated for the missing values from the fitted model (McDonald et al., 2015). To achieve this, countries were grouped together according to a predefined set of rules, all the historical data from the countries in each group were then combined and a probability distribution fit to these grouped data.

Historical input data

For the analysis presented here, the main inputs were the number of reported outbreaks and animal demographic data (e.g. number of animals in a country). Data for the number of reported outbreaks by country, k, species, s, and year, y, $N_{ob}(k,s,y)$, were obtained from all 180 MCs in the OIE databases over the years 1996-2014. The following species were considered for this analysis: pigs, cattle, sheep, goats, cats, dogs. For the period 1996-2005 these data were obtained from Handistatus II (OIE, 2016b). For the period 2005-2014 these

data were obtained from the country annual reports, located in the OIE reporting history section of the WAHIS interface (OIE, 2016a). Where no reports were available for a country, the data was considered to be missing. Species breakdown of the number of outbreaks were not readily available from the WAHIS data and so were estimated based on the average proportion of cases attributed to each species from both the WAHIS and Handistatus data, e.g. if there were 100 outbreaks of rabies reported in the data and the ratio of reported cases from the available data was 40% pigs and 60% cattle, then it was assumed that 40 outbreaks were attributed to pigs and 60 to cattle.

There were considerable differences in the animal demographics between countries which could influence the number of recorded outbreaks. Therefore, to obtain a statistic that was comparable between countries the number of outbreaks was weighted by an animal demographic metric; the number of 'animal establishments' by country and species, $N_{est}(k,s)$. For pigs, cattle, sheep and goats this was the number of animal establishments as defined in the WAHIS database for 2014; this is essentially equivalent to farms. In the absence of reliable data, it was assumed that it was acceptable to treat cats and dogs as single entities, i.e. the number of establishments was equal to the number of animals. The average number of animals per farm was considered as an alternative statistic, but as it was not directly reported in the OIE database, it was decided that the additional level of uncertainty in deriving this estimate was too great; for example it was not clear for some countries whether the farms were made up of lots of farms of average size, a few very large farms, or lots of small 'back yard' farms. While the reporting of the data for number of establishments was sporadic and also appeared to be subject to uncertainty (e.g. the same numbers of livestock/farms in a country being reported for numerous consecutive years, when one would expect some variation as animal production systems are dynamic and vulnerable to external factors), the value was directly reported in the OIE data and so considered to have less uncertainty than the number of animals per farm. However, due to the high level of uncertainty, only one point estimate of number of animal establishments was used, rather than an attempt to estimate values for each year. If these data were not recorded in the 2014 annual report, then the most recent historical observations were used. Data on the numbers of cats and dogs per country were estimated based on a collection of previous studies/literature reviews (OIE 2016a; WSPA 2008; FEDIAF 2014).

The number of outbreaks, $N_{ob}(k,s,y)$ was divided by the number of establishments, $N_{est}(k,s)$, to obtain the 'establishment prevalence' of disease by country, k, species, s and year, y, $P_{est}(k,s,y)$ (Eq. 1);

$$P_{est}(k,s,y) = N_{ob}(k,s,y) / N_{est}(k,s,y) \quad (1)$$

Groupings of countries

It was essential that the country grouping method was relevant to the subject matter and provided robust predictions. A number of different measures, G, for grouping countries were investigated, with six different groupings considered in this analysis. Two groupings were the official United Nations (UN) geographical regions and sub regions, *UNregion*, *UNsubRegion*, respectively (UN, 2016). A further grouping, *AltUNsubregion*, developed in a previous analysis was a modified version of the *UNsubRegion* accounting for some specific epidemiological considerations (Adkin et al., 2004). Two further groupings used k-means cluster analysis to group countries into 5 and 10 groups based on Gross domestic product (GDP) per person, *Gdp5* and *Gdp10* respectively, under the hypothesis that a country's economic activity might be proportionate to its disease prevalence and/or livestock

demographics (Jacobsen & Koopman, 2005). Estimates using the whole world as the sixth group were used for comparison.

Fitting a probability distribution

For each grouping method, a similar calculation was performed: for each subgroup, g, all non-zero entries of the establishment prevalence, $P_{est}(k,s,y)$, relating to MCs, k, in the subgroup g were extracted. It was decided to fit a gamma distribution to these data as the shape of the distribution is very flexible depending on the choice of parameters, so it was considered likely to provide a reasonable fit for a large number of datasets, and it is restricted to positive values, as is the case with the data. The gamma distribution was fit using the *fitdistr*[†] function in R, which uses maximum likelihood methods to determine the best fitting values for the gamma distribution (Eq. 2);

$$P_{fit}(g,s) = fitdistr(P_{est}(k,s,y) > 0,' gamma'), where\ k \in g. \quad (2)$$

Predictive accuracy

For all countries which had OIE reported outbreaks of rabies in the species considered in this analysis, the accuracy of the model predictions was evaluated by two methods. The first method was a comparison of the overall fit using a Kolmogorov-Smirnov (KS) test in R[‡] to compare the raw data of rabies outbreaks over all years for country k and species s, against the fitted distribution from the group, $K_s(k,s)$ (Eq. 3);

$$K_s(k,s) = ks.test\big(P_{est}(k,s,y) > 0,' pgamma', P_{fit}(g,s)\big), where\ k \in g. \quad (3)$$

A statistically significant result for the KS test implied confidence that the two sets of data (raw and fitted) came from the same underlying distribution and thus the fitted distribution could be considered a reasonable estimate for the observed data. The different grouping scenarios were compared by evaluating the proportion of MCs where this was not the case at the 1% level, i.e. had a KS test p value > 0.01.

The second method assessed the grouping fit on a country level, by comparing the observed median number of outbreaks per country from the OIE data, $N_{ob}(k,s)$, against the median, 25th and 75th percentiles of the predicted number of outbreaks per country using the grouping fit, $N_{obFit}(k,s)$. The predicted number of outbreaks were obtained by multiplying the observed number of establishments, $N_{est}(k,s,y)$, by the 50th, 25th and 75th percentiles of the fitted distribution, $P_{fit}(g,s)$. If the median of the raw OIE data fell between the 25th and 75th percentiles of the fitted distribution then it was considered that the fitted distribution was a reasonable estimate for that country. Additionally the squared difference between the observed and predicted country medians were calculated for different grouping methods, G (Eq. 4);

$$S_{diff}(k,s,G) = (N_{ob}(k,s) - N_{obFit}(k,s,G))^2. \quad (4)$$

Statistics were generated for the total squared difference over all countries

$$S_{diffTot}(s,G) = \sum_k S_{diff}(k,s,G) \quad (5)$$

[†] https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html
[‡] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/ks.test.html

and a comparison between two grouping methods, G_1 and G_2, for the number of countries where the squared difference was smaller for G_1 than for G_2, $N_{diff}(s,G_1)$,

$$N_{diff}(s, G_1) = \sum_k (S_{diff}(k, s, G_1) < S_{diff}(k, s, G_2)). \quad (6)$$

RESULTS

Approximately 60% of observations in the rabies dataset were a non-zero value. This varied between species; while 44% of countries reported at least one value for rabies in dogs, <1% of countries reported at least one value for deer (Table 1). This value for deer was thought to be largely due to the low country level prevalence of rabies in deer; due to the lack of data, results for deer and buffalo were considered highly uncertain and are not considered further. Table 1 also shows that while Southern Africa tends to have a lot of data, Australasia has no reported outbreaks. This highlights an important consideration in the choice of the most appropriate grouping method; a method that provides a very good distribution fit for one region, but has no data for the other regions, may not be as good as one that has reasonable data for all groups but has a less good distibution fit.

Table 1. Percentage of countries with at least one reported outbreak of classical rabies by UN sub region and species

UN Sub Region	**Pig**	**Deer**	**Cattle**	**Sheep**	**Goat**	**Cat**	**Dog**	**Buffalo**
South America	29%	0%	64%	29%	21%	29%	29%	14%
Western Africa	0%	0%	38%	6%	6%	6%	75%	0%
Central America	38%	0%	100%	38%	38%	50%	63%	25%
Eastern Africa	11%	0%	58%	26%	26%	32%	58%	0%
Northern Africa	0%	0%	43%	43%	43%	43%	57%	0%
Middle Africa	0%	0%	11%	0%	0%	0%	33%	0%
Southern Africa	40%	0%	100%	60%	80%	100%	100%	0%
Northern America	25%	0%	0%	25%	25%	0%	25%	25%
Caribbean	9%	0%	13%	9%	17%	9%	13%	0%
Eastern Asia	0%	0%	29%	14%	14%	14%	29%	0%
Southern Asia	22%	0%	44%	11%	22%	11%	56%	33%
South-Eastern Asia	0%	0%	45%	0%	0%	36%	45%	0%
Southern Europe	23%	0%	31%	23%	23%	31%	38%	0%
Australia & New Zealand	0%	0%	0%	0%	0%	0%	0%	0%
Melanesia	0%	0%	0%	0%	0%	0%	0%	0%
Micronesia	0%	0%	0%	0%	0%	0%	0%	0%
Polynesia	0%	0%	0%	0%	0%	0%	0%	0%
Central Asia	20%	0%	100%	80%	0%	80%	100%	0%
Western Asia	6%	0%	33%	33%	28%	33%	50%	6%
Eastern Europe	70%	20%	100%	90%	60%	100%	100%	0%
Northern Europe	7%	0%	21%	21%	14%	21%	21%	0%
Western Europe	0%	0%	38%	13%	0%	50%	50%	0%
Overall	**13.0%**	**0.9%**	**39.5%**	**22.4%**	**19.3%**	**28.7%**	**44.4%**	**4.5%**

Figure 1 shows the results of the overall fit for the different grouping measures for rabies. There was a maximum of 13.8% increase in the number of countries with a KS test p value > 0.01, between treating the whole dataset as one epidemiological unit and the best grouping measure. The *AltUNsubregion* was on average the best across all species, although results varied by species, with the *Gdp* groupings performing better for goats.

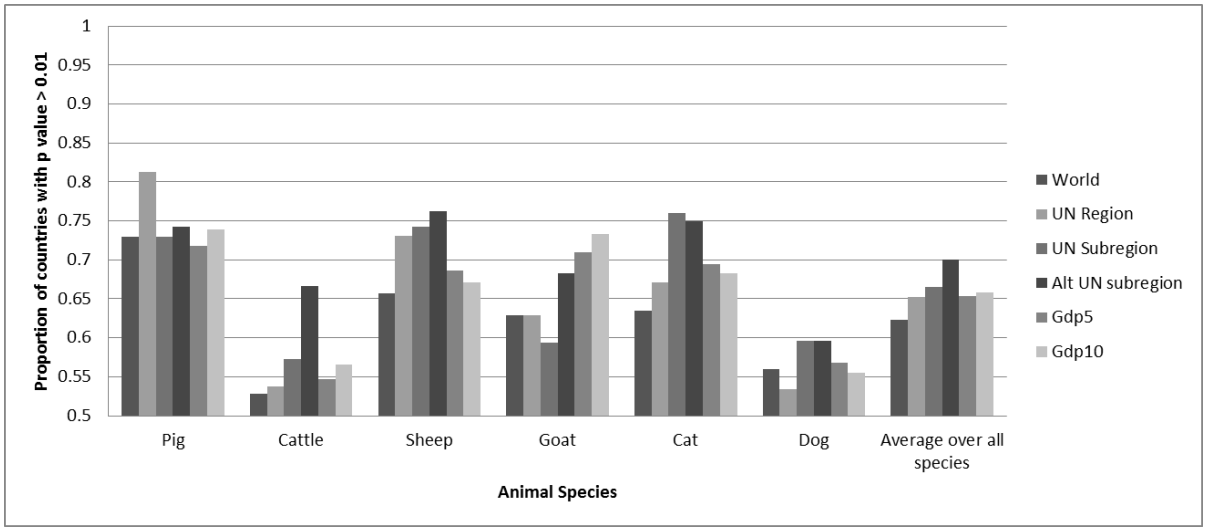

Fig. 1 Percentage of countries with a Kolmogorov-Smirnov *p* value >0.01, by animal species and grouping method, for rabies. A higher proportion means that more countries have a distribution of historical outbreaks that could be considered to come from the same distribution as the one fitted to the grouped data.

Fig. 1 shows results for individual countries with the most observed data, using cattle as an example. It can be seen that for individual countries, the median of the observed data often falls between the 25th and 75th percentiles of the fitted distribution, which suggests both methods provide a reasonable fit.

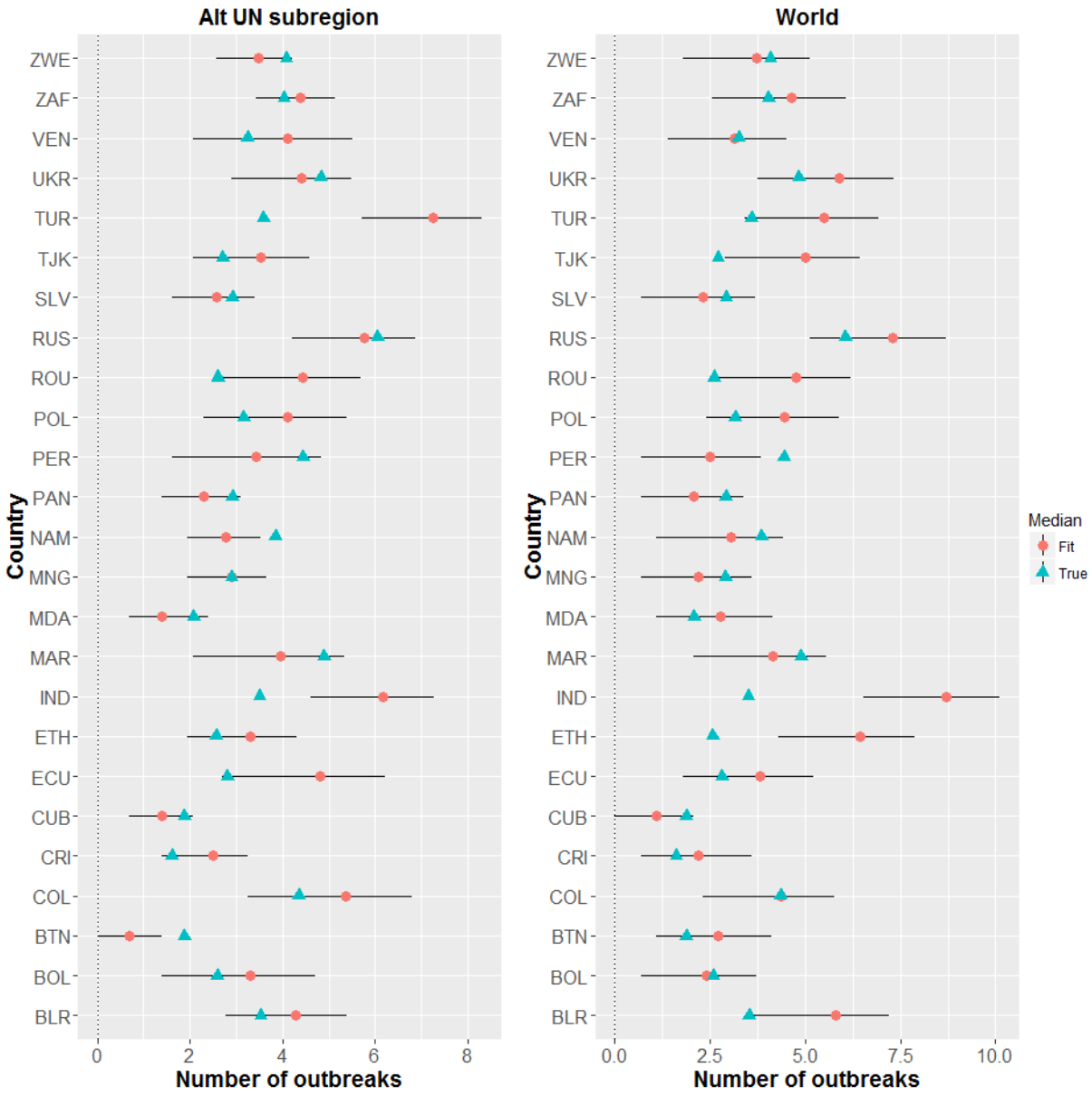

Fig. 1 Comparison of observed and predicted number of rabies outbreaks in cattle for *AltUNsubregion* (predicted values estimated by fitting different distributions to groups of countries based on geographical region and specific epidemiological considerations (Adkin et al., 2004)) and *World* (predicted values estimated using the same distribution for every country), where there were at least 15 observed historical outbreaks. Triangle= median of country specific historical outbreaks, circle= median of fitted distribution, Line= 25^{th} – 75^{th} percentiles of country specific historical outbreaks. Country names are the official UN IS03 codes (UN, 2016).

Further analysis on all the countries showed that the difference between the predicted and observed medians was smaller for the *AltUNsubregion* (compared to the whole world) for 58 countries, and larger for only 29 countries. In addition, the squared sum of the differences between the medians was larger for the whole world than for *AltUNsubregion*. A similar pattern was observed for the other species, suggesting that the *AltUNSubregion* grouping provided the better fit (Table 2).

Table 2: Comparisons between two grouping methods, *AltUNSubRegion (ALT)* and *World* (*Wrld*). The number of countries where the squared difference between observed and predicted median number of rabies outbreaks by species, *s*, and country *k* was smaller for *ALT* than for *Wrld, is depicted by* $N_{diff}(s,ALT)$ and the opposite, depicted by $N_{diff}(s,Wrld)$. The total sum of the squared difference over all countries for *ALT* is depicted by $S_{diffTot}(s,ALT)$ and for *Wrld*, depicted by $S_{diffTot}(s,Wrld)$.

	Pig	**Cattle**	**Sheep**	**Goat**	**Cat**	**Dog**
$N_{diff}(s,ALT)$	23	58	30	25	47	67
$N_{diff}(s,Wrld)$	8	29	13	14	18	34
$S_{diffTot}(s,ALT)$	41	168	70	80	116	287
$S_{diffTot}(s,Wrld)$	148	319	137	164	244	420

DISCUSSION

Risk assessments for animal and public health require good quality data in terms of trade movements, animal health status and production systems (Rodgers et al. 2011). An essential prerequisite for estimating animal health status or disease burden is the availability of comprehensive national-level data on the prevalence of the disease of interest (McDonald et al. 2015). This is not always possible, however, and the treatment of missing values may lead to biased results. An alternative method, as presented here, is to estimate pathogen prevalence for countries with missing national-level data (McDonald et al. 2015). As risk assessments are data driven, with the value of the results being dictated by the quality of the data inputs, filling data gaps will reduce the occurrence of biased results and allow for more usable outputs giving more power to the risk assessment conclusions. Intrinsically, it is essential that the results of these methods provide transparent and robust estimates for the missing data.

To assess the risk of exotic disease incursion and spread through Europe, the OIE database of global animal disease prevalence was used and different imputation methods, based on country groupings, were compared to investigate whether model output accuracy could be improved. In the case study example of rabies, the results presented suggest that the methodology employed provided reasonable estimates for most countries in the world but was not always accurate on a country level. Results suggested that grouping countries before fitting a distribution was more accurate than just using a distribution fit to the whole world, but no grouping method was accurate for every country. As such, it is believed that the methodology is appropriate when estimating missing data, but it would be inadvisable to use the method in order to replicate an observed dataset.

An interesting result of these analyses is that they have the additional benefit of providing a statistic by which to determine which grouping method and/or proxy variables are best to consider for a given situation; the results suggest that this varies between both animal host species and pathogens of interest. For the case study of rabies, a comparison of the average score across species suggested that the *AltUNsubregion* would be the best grouping measure to use. However, it is acknowledged that there are limitations to the analyses presented here. The metrics of comparison used here do not comprehensively describe everything about the goodness of the fit; other metrics such as the average KS score may give different results.

There are also other factors such as the proportion of MCs that fall into groupings where there are no observed data, and thus no estimate can be determined: it should be considered whether a grouping that gives a higher KS score is still the best option if an alternative grouping with a slightly lower KS score can provide estimates for more individual countries. This methodology could be further expanded to consider these issues, as well as test for other proxy variables or combinations of variables to evaluate if they would provide a better fit. Another area where this methodology could be expanded would be to evaluate the fits for distributions other than the gamma (e.g. lognormal or Weibull).

The analysis presented here is designed to fill specific data gaps in global datasets, but is still reliant on good quality input data. While the OIE database is one of the most comprehensive in the world, there are still data gaps that lead to uncertainty about the model results, particularly with regards to the number of animal establishments. The assumption to treat cats and dogs as single entities could underestimate the risk, as it does not account for households that have multiple pets or institutions such as dog homes.

In conclusion, while there is a growing number of available global datasets providing information that could be used for risk assessments, currently there are assumptions which need to be made regarding the data gaps before datasets can be used. Using global datasets as they stand, with missing data treated as a zero, may inadvertently penalise those countries which do report disease outbreaks as opposed to those countries which are affected by a pathogen but do not report outbreak data. The methodology for estimating animal disease prevalence presented here allows for an objective, transparent approach to fill the data gaps and provide a more comprehensive base for data input to risk assessments.

ACKNOWLEDGEMENTS

This work had funding agreed through the Animal Health and Welfare ERA-NET consortium (https://www.anihwa.eu/) under SPARE ('Spatial risk assessment framework for assessing exotic disease incursion and spread through Europe'). Funders are acknowledged as the Department for the Environment, Food and Rural Affairs (Defra) - UK, Ministry of Health - Italy, Spanish National Institute of Agriculture and Food Research and Technology – Spain, and Federal Food Safety and Veterinary Office (FSVO) – Switzerland. The authors would also like to thank Rachel Jinks (APHA), Tony Fooks (APHA), Maria Crescio (IZSTO) and Silvia Bertolini (IZSTO) for their valuable inputs to this paper.

REFERENCES

Adkin, A., Coburn, H., England, T., Hall, S., Hartnett, E., Marooney, C., Wooldridge, M., Watson, E., Cooper, J. and Cox, T.M.S. (2004). Risk assessment for the import of contaminated meat and meat products into Great Britain and the subsequent exposure of GB livestock. http://collections.europarchive.org/tna/20050301192907/http:/www.defra.gov.uk/animalh/illegali/pdf/risk-assessment04.pdf Accessed 29.12.2016

ECDC (2016). Zika Virus disease epidemic - Rapid Risk assessment. http://ecdc.europa.eu/en/publications/Publications/01-08-2016-RRA-eighth-update-Zika%20virus-Americas,%20Caribbean,%20Oceania.pdf Accessed 29.12.2016

FEDIAF (2014). The European pet food industry facts and figures 2014. http://www.fediaf.org/facts-figures/ Accessed 29.12.2016

Hoinville, L.J., Alban, L., Drewe, J.A., Gibbens, J.C., Gustafson, L., Hasler, B., Saegerman, C., Salman, M. and Stark, K.D.C. (2013). Proposed terms and concepts for describing and evaluating animal-health surveillance systems. Prev. Vet. Med. 112, 1-12

Jacobsen, K.H. and Koopman, J.S. (2005). The effects of socioeconomic development on worldwide hepatitis A virus seroprevalence patterns. Int. J. Epidemiol. 34, 600-609

McDonald, S.A., Devleesschauwer, B., Speybroeck, N., Hens, N., Praet, N., Torgerson, P.R., Havelaar, A.H., Wu, F., Tremblay, M., Amene, E.W. and Dopfer, D. (2015). Data-driven methods for imputing national-level incidence in global burden of disease studies. B. World Health Organ. 93, 228-236

OIE (2016a). WAHIS Interface http://www.oie.int/wahis_2/public/wahid.php/Wahidhome/Home Accessed 29.12.2016

OIE (2016b). Handistatus II. http://web.oie.int/hs2/report.asp?lang=en Accessed 29.12.2016

Rodgers C.J., Roque A. and Marcos Lopez, M. (2011). Dataquest: Inventory of data sources relevant for the identification of emerging diseases in the European aquaculture population. EFSA Supporting Publications, 8(1), 1-108

Roberts H., Moir R., Matt C., Spray M. and Boden L.P.B. (2016). Risk assessment for Bluetongue Virus (BTV-8): risk assessment of entry into the United Kingdom https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/499882/qra-BTV8-UK-160212.pdf Accessed 29.12.2016

Simons R.R.L, Horigan V., Gale P., Kosmider R.D., Breed A.C. and Snary, E.L. (2016). A Generic Quantitative Risk Assessment Framework for the Entry of Bat-Borne Zoonotic Viruses into the European Union. PLoS ONE 11, e0165383

SPARE (2016). Spatial Assessment of Risk for Europe for evaluating the incursion and spread of exotic animal disease through Europe. http://www.spare-europe.eu/ Accessed 29.12.2016

UN (2016). Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings. http://unstats.un.org/unsd/methods/m49/m49regin.htm Accessed 29.12.2016

USDA (2016). National Agricultural Library. http://agclass.nal.usda.gov/mtwdk.exe?s=1&n=1&y=0&l=60&k=glossary&t=2&w=release+assessment Accessed 29.12.2016

WSPA (2008). Global Companion Animal Ownership and Trade: Project Summary, June 2008. World Society for the Protection of Animals (WSPA). http://s3.amazonaws.com/zanran_storage/www.wspa.org.uk/ContentPages/48536804.pdf Accessed 29.12.2016