*Nigerian Veterinary Journal 2010 Vol 31(1):1-7*

**PERSPECTIVE**

# TRANSFORMING ZOONOTIC DISEASE SURVEILLANCE AND SPATIAL EPIDEMIOLOGY: IMPROVING CONTENT, DETAIL, AND DATA INFRASTRUCTURE

PETERSON, A. TOWNSEND
*Biodiversity Institute, The University of Kansas, Lawrence, Kansas 66045 USA*

Author Email: town@ku.edu; Tel: +1 7858643926

## Disease Geography and Dynamics

Zoonotic diseases represent significant challenges for veterinary and public health initiatives, and the field of spatial epidemiology sets out to characterize risk of transmission of these diseases across landscapes. Frequently, these diseases are quiescent for extended periods, and then "emerge," causing major outbreaks or isolated cases, with serious consequences for human and animal well-being (indeed, humans are affected both by the diseases directly, *and* by the negative effects on their domestic animals). Unfortunately, however, zoonotic diseases also frequently remain poorly known, poorly documented, incompletely diagnosed, and thereby underappreciated as to the significant role that they play in human well-being.

Zoonotic diseases represent interacting systems of elements of biodiversity. Pathogens (bacteria, protozoans, viruses, etc.) circulate in populations of some host or hosts (often mammals or birds), and may be transmitted among hosts either directly, by vectors (mosquitoes, sandflies, ticks, fleas, etc.), or via the environment (e.g., soil in anthrax transmission). In general, then, if any of these elements is lacking, transmission ceases, and the disease is likely not to circulate further in the region (Peterson, 2007). This framework of thinking emphasizes a vital linkage between spatial epidemiology as a field and the area termed biogeography in biodiversity science.

This contribution offers a general perspective on paths toward improving this scenario, and transforming spatial epidemiology into a more synthetic, predictive, and functional science. That is, spatial epidemiology at present is based on a poor data infrastructure and inadequate system of archiving important samples, and then interprets those data within two different, but both largely inappropriate analytical frameworks. Solutions exist, but require some rather radical changes in how the field "does business."

## Present Situation in Disease Geography and Spatial Epidemiology

*Data formats and content*.—All disease occurrences carry some sort of geographic reference, but these references can be quite heterogeneous in form. The present system in most regions is based chiefly on reference to areas (e.g., states, provinces, counties) rather than points, and offers no summary of likely precision or certainty of that locality (e.g., Fang *et al.*, 2006). The first point—that of reference to areas (i.e., some sort of polygon) immediately constrains the results of the analysis to resolutions that are coarser than the area of that polygon—no detail of mapping risk of this disease will be possible at finer resolutions, because the disease occurrence

information offers no information at finer resolutions.

The second point (that of estimating uncertainty of disease occurrences) requires an example: imagine two domestic animals on the same farm, each of which contracts a particular disease. One animal has spent its entire lifetime on that farm, and thus must have contracted the disease in the immediate vicinity, whereas the other might be a work animal that has ranged broadly over the entire region in recent weeks. The difference in implications of these two disease occurrences for mapping disease transmission risk is dramatic—one leaves a quite specific record of where a disease is transmitted, whereas the other only a vague impression.

In the present system of recording disease occurrence, however, none of these details is captured and expressed in summary data records. That is, two case occurrences may fall into very similar or radically different environmental situations, but this information may be masked by the imprecise spatial specification of the locality (e.g., a particular state or a particular county). Worse still, in the example, the two animals living on the same farm would appear identical in most epidemiological data sets, even though one pinpoints the site of exposure much more precisely than the other, and the imprecise point may prove positively misleading (i.e., the animal might have been infected in a very different environmental situation in a different state).

The biggest problem, at the end of the day, is that analyses are constrained to the base resolution of the occurrence data available. With the current system, that base resolution is unknown, which can bias results in unknown ways. When occurrence data are accompanied by precision estimates, they can be filtered to yield those cases that are sufficiently precisely known to be informative to a given analysis. In this way, a given data record can be "recycled," and used productively for other analyses not envisioned when the data point was originally captured.

*Lack of data and specimen infrastructure*.—Data in veterinary and public health applications are stored in an odd, eclectic, and often *ad hoc* system, which does not in any way foster recycling and reuse of data for future applications. Too frequently, indeed, such data are considered personal research resources, and as such never become openly available. In other cases, data sets are maintained in local, regional, provincial, national, or international agencies, in formats that are not necessarily interoperable or interchangeable. Only high-profile diseases (e.g., H5N1 avian influenza) are tracked in databases that see more careful attention and integration (O.I.E., 2009).

This system, if it can be called as such, acts generally to limit analyses to the goals of the study for which a data point was originally collected. When presented with a new challenge, such as a disease emergence event or an intriguing case, quite frequently, existing data are either unavailable or accessing them is overly cumbersome. Too often, these obstacles are sufficiently large as to prevent reuse of existing data, and synthetic analyses must await accumulation of new data. To complicate the situation further, existing data are shared only relatively rarely, and (when shared) are shared in cumbersome, inefficient formats (e.g., O.I.E., 2009) that in no way support and promote creative exploration of the data.

In parallel to the data challenge goes a series of issues regarding diagnostic specimen materials (Peterson, 2010). These specimens include both what can be termed "voucher specimens" (i.e.,

specimens of hosts or vectors to document identifications) and diagnostic tissue specimens, which are collected specifically for testing for presence of pathogens. Such materials, however, have the potential to document distributions across space and across potential hosts of pathogens once an emerging threat is identified, and yet rarely are referenced in veterinary or public health publications, and diagnostic materials in disease surveillance are rarely organized in formal, permanent, curated collections.

*Inappropriate approaches to analysis*.—Even when data are collected and stored appropriately, and become available to a researcher for analysis, the analyses that are standard in spatial epidemiology are frequently neither fully appropriate, nor as powerful as they could be. The weakest such analyses are developed in spatial dimensions only, and as such ignore environmental variation that underlies the spatial pattern (see, e.g., Fang et al. 2006)—that is, although zoonotic diseases frequently show broad spatial trends, the details of their behavior are invariably driven by environmental variation. This environmental variation is not manifested in results based on exclusively spatial analyses, which presents a serious limitation.

Even when environmental factors *are* considered, however, analyses in spatial epidemiology are nonetheless not always developed appropriately. Any spatial prediction exercise must manage two types of error: omission error (predicting areas of known presence as absent) and commission error (predicting areas of actual absence as present). Multivariate statistical approaches are often employed, but with overall optimizations that weight these two error components equally, and minimize overall error. Species' geographic distributions, however, present rather odd challenges—omission error is almost always genuine error (except for sink populations, erroneous geographic references, and/or erroneous taxonomic identifications), but commission "error" is usually only partly error. That is, areas from which a disease is not known are counted as areas of absence, yet may simply be areas of presence from which the disease has not been reported, where humans and associated animals may not be present, where no studies have been developed, etc.—as such, commission error rates will often appear to be quite a bit higher than they really are. As a consequence, in such analyses, omission error must be accorded a significantly greater weight than commission error, or optimization efforts may be seriously biased (Anderson *et al.*, 2003).

Specifically, if an algorithm is allowed simply to minimize overall error (be it omission or commission), it will weight an omission and a commission equally. The omission represents a real case of a known occurrence of a disease being left out of the predicted area, while the commission represents an area from which a disease has not been recorded proving, in reality, to be suitable for the disease. Clearly, the former case is much more serious than the latter, and yet currently accepted modeling approaches do not take these differences into account, perhaps owing to a statistical modeling focus, as opposed to a more biogeographic approach focused on reconstructing full geographic distributions of species.

*Overall picture*.—Taking this rather broad-spectrum view of data and analyses towards mapping disease transmission risk, it becomes clear that current infrastructures and approaches will produce an incomplete picture. Because of structural considerations for data, certain spatial resolutions will prove inaccessible to mapping efforts, and because data and samples are stored and/or shared only

rather ineptly, much information will be off-limits to researchers desiring to develop maps. Finally, because the analyses *per se* are not developed in a biogeographic context, considering real features of species' geographic distributions and how they are characterized, the results will frequently not be useful.

## The "Fixes"

The situation characterized above is complex, and fixing its problems will require a number of serious changes in how disease reporting is achieved in veterinary science and human public health. Many of these fixes represent improvements that will require additional time and attention, but the reward will be greatly improved risk maps, and greater flexibility in developing novel analyses. As such, the reward for these investments will take some time to perceive, as a data infrastructure must be constructed, but will be significant. Many of the solutions detailed below are drawn from the world of biodiversity science, where the same (or parallel) issues have been explored for some years (Soberón & Peterson, 2004).

*Data formats and content*.—A first step that would make a world of difference in veterinary and public health spatial epidemiology work is that of developing a standardized data framework. In biodiversity science, early in the development of biodiversity informatics, the Darwin Core was developed to summarize crucial data fields that express taxonomic identification, place of occurrence, time of occurrence, and some specifics of the record; the Darwin Core has now been approved as an official metadata standard for biodiversity data by the Taxonomic Database Working Group.

Spatial epidemiological data take much the same form—describing the occurrence of a particular pathogen at a particular site at a particular point in time, but would require careful thought as to which additional fields would prove necessary (e.g., relation to host and/or vector, method of determination of taxonomic identification, titres or prevalences, etc.).

For expression of geographic references, only a point-based system will be able to take full advantage of the detail available in some records. Linking these point-based records with measures of uncertainty (usually expressed as a radius around the point) provides additional critical information that can be used to decide the suitability of particular points for inclusion in particular analyses (Wieczorek *et al.*, 2004). This point-radius method is simple, and could easily be adapted for data recording, even by non-specialists. When privacy concerns are an issue, which is more common with human disease, these point-radius georeferences can easily be "dumbed down" and returned back, e.g., to county-level spatial resolution, for public data sharing.

*Effective data and specimen infrastructure*.—Development of an appropriate and effective data infrastructure represents a major and important challenge for spatial epidemiology. A major question has been that of centralizing data sets (or not)—in the centralized case, data are sent to a central repository, where they are stored and served. This data structure has the advantage of simple management (i.e., changes can be made globally to the entire data set) and simple data serving (i.e., one dataset placed for search online), but can result in "divorcing" data sets from the institutions and organizations that produce them and care for them. In the latter case, the currency and integrity of the data may decline over time (e.g., mosquitoes identified as *Anopheles gambiae*, now recognized as a complex of species, so the data records might refer to any of *A. gambiae sensu lato*, *A. arabiensis*, *A.*

*bwambae*, *A. merus*, *A. melas*, *A. quadriannulatus*, or *A. gambiae sensu stricto*). A distributed data architecture, in which data reside at the "home" institution, but that are shared via the Internet to form a single virtual database, represents a potential solution to this challenge, and also functions to preserve institutional "ownership" of the data.

An effective system of specimen documentation and archiving is a further challenge. One important partnership that can constitute an easy and immediate "fix" is that of linking veterinary and public health efforts to the broader biodiversity community. The latter has a well-established system of effectively permanent archiving of biological specimen resources, which would be more than pleased to receive documentary specimens in deposit, as the same specimens can be important to their own research in systematics. Biodiversity institutions also have well-established unique references to individual specimens that can and should be cited in veterinary and public health publications (Peterson, 2010), and these references are in the process of considerable refinement (Clark *et al.*, 2004). Diagnostic (tissue) specimens can be maintained at veterinary or public health institutions (which often have better biosecurity capabilities), or at biodiversity institutions, and can be catalogued and data served to permit effective and efficient access by researchers.

*Improving approaches to analysis*.— Analyses of spatial distributions of biological phenomena should be based on direct measures of the environmental factors that determine them. This approach has been termed "ecological niche modeling," emphasizing the critical link between spatial models and the set of environmental conditions within which a species can maintain populations without immigrational subsidy (modified from the original definition from Joseph Grinnell). Niche modeling emphasizes the realities of unequal weighting of presence *versus* absence information, in a clear biogeographic context.

In niche modeling, a first priority is full characterization of ecological niches of species (or biological phenomena, such as disease transmission cycles), which requires data on occurrences across the entire spatial distribution of the phenomenon. A second requirement is that of characterizing the arena that is appropriate for analysis, taking into account the geographic factors that constrain the distributional potential of species—effectively the area within which analyses should be carried out. This area is that which has likely been "sampled" by the species for possible colonization (i.e., present distributional area + dispersal distance, and taking into account past distributional shifts), and can represent a serious challenge for analyses.

Once data are assembled and environmental arenas of analysis defined, the analyses may begin. In these analyses, it is crucial to ponder (1) what rates of error likely characterize the occurrence data set (e.g., an animal or human infected in one site, but diagnosed and "georeferenced" in another site), which has been quantified as the parameter *E* (Peterson *et al.*, 2008), and (2) what relative weights should be applied to omission and commission errors. The weights assigned to different error components can be used directly in many niche modeling algorithms, several of which have built-in means of prioritizing omission error over commission error (Anderson *et al.*, 2003). Once the raw model is in hand, its interpretation requires further thinking, beginning with establishing thresholds for separating prediction of presence (or at least suitability) from prediction of likely

absence (Peterson *et al.*, 2007)—in general, the appropriate solution will be to select a threshold that includes (100 - E)% of the presence data set on which the model was based. This threshold will take into account error inherent in the presence data, and sets a point of separation between presences and absences that is most appropriate biogeographically for characterizing spatial distributions of biological phenomena.

## CONCLUSIONS

The goal of spatial epidemiology is to offer a predictive view of spatial and environmental dimensions of disease transmission risk. That is, the objective of the field is to process existing information into useful, predictive interpretations of disease risk in terms of space (i.e., identifying "hotspots" of disease transmission) and in terms of environment (i.e., identifying environmental risk factors), and any possible interactions between these two suites of factors. This emerging field, however, has had what would best be termed marginal success in these endeavors—the data infrastructure for the field is inefficient and frequently requires duplication of effort, and the analytical approaches used often fail to reconstruct distributions of biological phenomena realistically.

In this perspective, I offer a series of reflections and suggestions regarding paths forward for this field, with the goal of achieving an infrastructure of data and tools that meet the goals of the field. These "fixes" involve significant investment of time, logistics, and thinking, and almost certainly cannot *all* be followed by a single institution or for a single region. However, the hope is that some portion of this framework of thinking will prove useful to some sectors that look to spatial epidemiology for adequate risk mapping regarding disease transmission.

## ACKNOWLEDGMENTS

I thank my valued colleague Folorunso Oludayo Fasina for the kind invitation to prepare this perspective piece, and R. Ryan Lash for his careful thinking and detailed input into questions of geographic referencing.

## REFERENCES

ANDERSON, R.P., LEW, D. and PETERSON, A.T. (2003): Evaluating predictive models of species' distributions: Criteria for selecting optimal models. *Ecological Modelling*, **162**, 211-232

CLARK, T., MARTIN, S. and LIEFELD, T. (2004): Globally distributed object identification for biological knowledgebases. *Briefings in Bioinformatics*, **5**, 59-70

FANG, L., YAN, L., LIANG, S., VLAS, S.J.D., FENG, D., HAN, X., ZHAO, W., XU, B., BIAN, L., YANG, H., GONG, P., RICHARDUS, J.H. and CAO, W. (2006): Spatial analysis of hemorrhagic fever with renal syndrome in China. *BMC Infectious Diseases*, **6**, 77

O.I.E. (2009): *World Animal Health Information Databases (WAHID), version 1.4*. World Organisation for Animal Health, Paris.

PETERSON, A.T. (2007): Ecological niche modelling and understanding the geography of disease transmission. *Veterinaria Italiana*, **43**, 393-400

PETERSON, A.T. (2010): The critical role of permanent voucher specimens of hosts and vectors in public health

and epidemiology. *Emerging Infectious Diseases*, **16**, 341-342

PETERSON, A.T., PAPEŞ, M. and EATON, M. (2007): Transferability and model evaluation in ecological niche modeling: A comparison of garp and maxent. *Ecography*, **30**, 550-560

PETERSON, A.T., PAPEŞ, M. and SOBERÓN, J. (2008): Rethinking receiver operating characteristic analysis applications in ecological niche modelling. *Ecological Modelling*, **213**, 63-72

SOBERÓN, J. and PETERSON, A.T. (2004): Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London B*, **359**, 689-698

WIECZOREK, J., GUO, Q. and HIJMANS, R. (2004): The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, **18**, 745-767