Diagnosing Disease Outbreaks – P5: If and Where?

The Goal

We would like to know… if and where there are clusters of disease outbreaks. This critical information will be used to dispatch resources where they are needed.

Utilizing data from a worldwide sample of recent headlines with 'health mentions' extracted from news articles over the first few months of 2016, this brief report attempts to determine "if and where there are (geographic) clusters of disease outbreaks." Before considering the intertwined questions of 'if' and 'where', a concise overview of the dataset is provided.

Sample of Headlines

Originally, the dataset on which the analysis is based consisted solely of 650 headlines typically focused on the spread (either directly or indirectly) of a particular disease in a particular city, usually with no other geographic reference (e.g. a state or country). The following headlines are representative of the sample:

> Lima tries to address Zika Concerns
> More people in Toronto are infected with Hepatitis E every year
> Brisbane is infested with Meningitis
> Cholera outbreak in Dakar

There are instances, however, where either a specific disease isn't mentioned all, or the medical term employed is too general to determine the specific disease, or no mention is made of the geographical location. Each of these situations is illustrated in the headlines shown below:

Mystery Virus Spreads in Recife, Brazil
The CDC in Atlanta is Growing Worried
Zika Virus Sparks 'International Concern'

In these latter cases, the headlines were eliminated from the sample, leaving a total of 611 out of the 650 headlines to consider.

With the remaining sample the major requirement for analysis was to determine both the disease and city mentioned in the headlines. In terms of the diseases mentioned, a list of possible diseases was manually constructed from all the headlines. There were 36 unique diseases mentioned in the sample covering a wide range of infectious diseases (of which many are mosquito borne), sexually transmitted diseases, respiratory diseases, and even animal diseases (e.g. Mad Cow).

With the cities, a list of all possible cities in the world was constructed using a specialized computer function based on the *Geonames* database (www.geonames.org), which also provides other geographic information such as the associated continent, country and latitude-longitude. This list of cities was then used to determine the specific city referenced in the headline. Unlike the diseases, there were almost as many unique city names referenced in the headlines as there were headlines with most city names occurring once or twice. However, if you consider the associated country names, it's clear that the sample is weight towards the United States since it's referenced in almost half the headlines (303 out of 611). For

this reason, most of the analysis that follows revolves around two segments – the US and the 'rest of the world'.

Before we delve into the specific findings, it is important to note that none of the headlines references how the disease was explicitly transmitted which has implications for diseases that can be 'borne' in multiple ways and for determining the resources needed to control the disease.

Findings

Question 1: Is there evidence of disease outbreaks in early 2016?

Figure 1A displays the worldwide frequency distribution of the diseases mentioned in the sample of headlines.

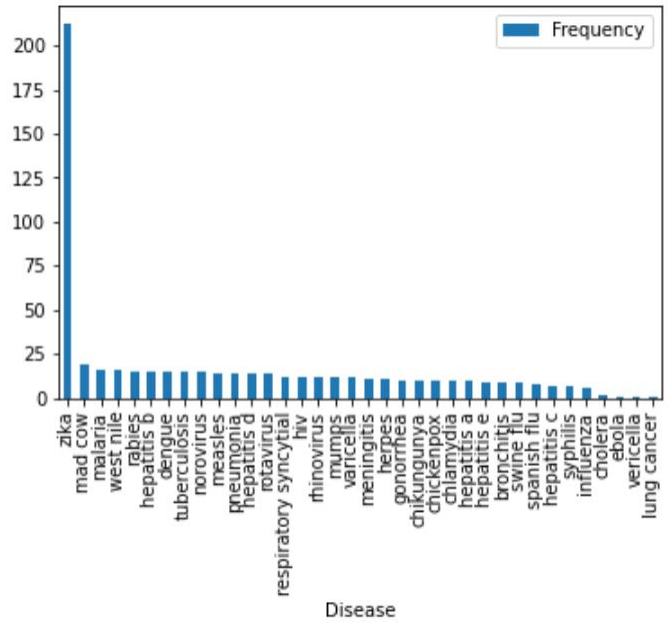


Figure 1A. Diseases mentioned in Worldwide Sample of News Headlines from early 2016
(36 Diseases in 586 out of 611 Headlines)

It's apparent that the Zika virus dominates the headlines in the sample. This pattern holds true for both the US and the 'rest of the world' (see Figures 1B & 1C), although the specifics of the other diseases mentioned varies between the two segments.

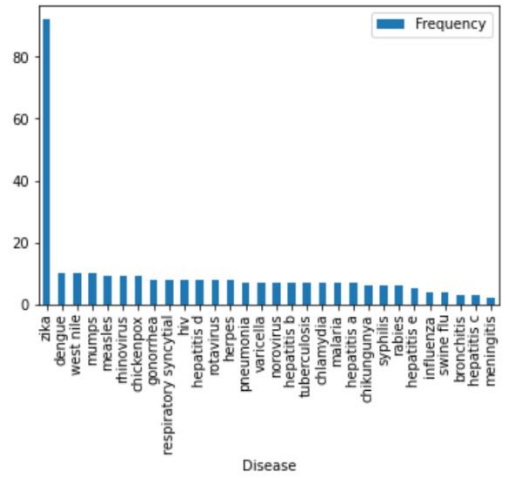


Figure 1B. Diseases in US Headlines
(30 Diseases in 292 Headlines)

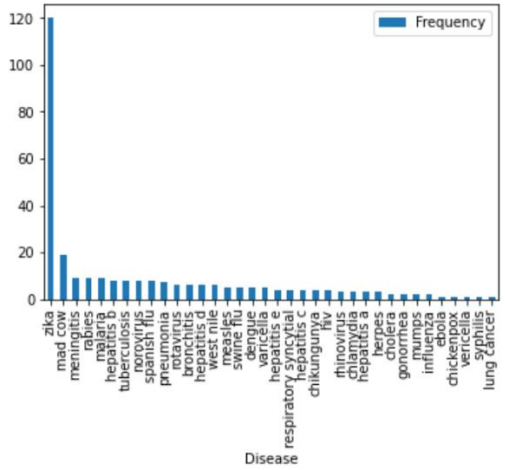


Figure 1C. Diseases in Rest of World Headlines
(36 Diseases in 294 Headlines)

In formal terms, a disease *outbreak* is defined as 'the occurrence of disease cases in excess of normal expectancy.' If the outbreak is 'widespread in a given population' or 'sudden and widespread across many populations globally', then it qualifies as an *epidemic* or *pandemic*, respectively. If the spread is consistently heightened in a given population overtime, then it's *endemic*.

Potential Outbreaks

Zika Virus

Given historical patterns, it's clear from this sample of headlines that Zika demands particular attention.

Zika is mosquito borne by the *Aedes aegypti* species which thrives in warm, wet environments. Zika can also be transmitted from an infected individual to an uninfected individual through sexual intercourse and from an infected pregnant mother to her unborn child.

Zika was discovered in Uganda in 1947. The first human case was identified in Africa in 1954. The first outbreak was in 2007 on the island of Yap in Micronesia. Since 2014 there has been a steady rise in the number cases in Central and South America. It's an open question whether it has reached epidemic proportions at the beginning of 2016, although the number of headlines seen here provides compelling evidence.

Other Infectious Diseases

In terms of the remainder of the diseases referenced in the headlines, there may be other potential candidates that need further inspection.

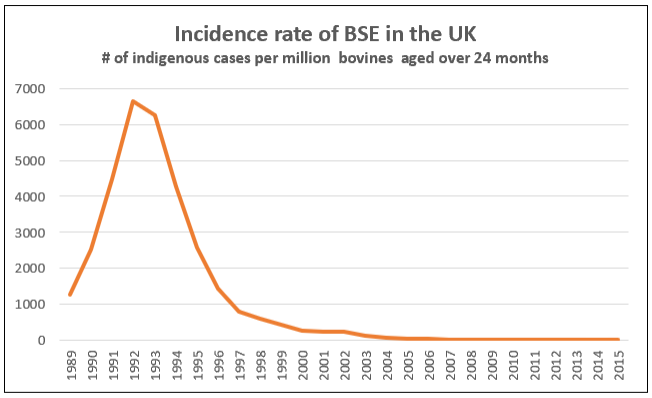
Many of these diseases are best described as endemics. A case in point is malaria which appears to require constant attention in developing countries. Another case in point is the 'measles' which appears to be borderline endemic. In the US measles was officially eradicated in 2000. By 2011 measles cases began to appear regularly in the US in larger numbers with periodic outbreaks in select cities. This appears to be what is reflected in the headlines here. The same is true for the mumps in the US. Before 2005, the

number of mumps cases in the US was 200-300 per year. In 2006 it spiked to ~6500 cases and dropped back to ~200 cases in 2012. Over the past 3 years it has been steadily climbing to ~1200 cases. Like measles, mumps outbreaks tend to be isolated in select areas.

Mad Cow (BSE)

One anomaly among the potential candidates is BSE or 'Mad Cow' disease. There are 19 headlines that reference 'bovine spongiform encephalopathy' (BSE), known more popularly as 'Mad Cow' disease. Without going into details, the headlines come from a variety of cities across Europe with a heavy emphasis (8 out 19) on the spread of the disease in various cities of the UK.

BSE originated in the UK in 1986 and peaked in 1992-1993 with ~6500 cases per million head of cattle over age 24 months (note: there are between 9-10 million bovines in the UK any given year). As Figure 2 indicates, since 2003 the incidence per year has been negligible. In the rest of Europe, there are 6 other countries (Belgium, France, Germany, Portugal, Spain, and Switzerland) that have experienced substantial issues with BSE. However, at their peaks (primarily in 2001-2002) the incidence was between 20-80 per million head of cattle. Like the UK, the incidence in the rest of Europe dropped to fractions per million after 2003 (due to bans on shipments of British beef and changes in feeding practices in the UK). The bottom line is that BSE should continue to be monitored, but these headlines probably don't portend a major 'outbreak'.



Source: https://www.oie.int/en/animal-health-in-the-world/bse-situation-in-the-world-and-annual-incidence-rate/annual-incidence-rate/

Figure 2. Incidence of BSE in the UK by Years

Question 2: Where are the Clusters of Disease Outbreaks?

The above analysis suggests that the Zika Virus is the only disease that provides substantial "evidence of an outbreak in early 2016." For that reason, the focus is solely on Zika in answering the second question, i.e., "where are the geographical clusters of the Zika outbreaks?"

Macro Geographical View

212 of the headlines reference Zika. Of these, the majority of the cities in the Zika headlines come from the North Americas, Asia, and South America in that order with literally a handful from Oceania and Europe (see Figure 3)

N.Americas	112	Asia	51	S.America	39	Oceania	4	Europe	1
US	92	Philippines	14	Brazil	20	Australia	3	Spain	1
Mexico	10	India	10	Columbia	9	Marshall Is.	1		
12 Others	10	Malaysia	8	3 Others	10				
		Thailand	5						
		8 Others	14						

Figure 3. Number of Headlines Referencing Cities by Continent

Of the 92 headlines from the US, 39 originated from cities in Florida, 22 from cities and Texas with the remaining 31 headlines coming from cities in 22 states spread across the country.

Any map or maps based on macro geographical areas like continents, countries or administrative areas like states in the US are not likely to provide a detailed understanding of the spread of the Zika virus. That said, these macro areas eventually must be considered because they: (1) are part of the reporting and organizational structure of WHO; and (2) control the legal and administrative structures and resources need to combat and control the spread.

Geographical Clusters Based on Longitude and Latitude

Given that each headline only references a single city, we can use the city's longitude and latitude to pinpoint the specific location associated with the headline. These values also enable us to determine the distances between the locations of the various headlines, using one of the common linear or spherical distance formulas. In turn, these distances provide a way to determine 'where the clusters of disease outbreaks are.'

Clustering Process

Clustering is a type of *machine learning* that is designed to identify similar instances and assign them to *clusters* or groups of similar instances. There are wide variety of clustering algorithms that can be used to derive clusters from empirical data (like the longitude and latitude of the cities referenced in the headlines). Some of them 'look for instances centered around a particular point, while others look for continuous regions of densely packed instances that can take any shape.' The one employed in this analysis, DBSCAN, is of the latter sort. Again, for those who are interested, the exact details can be found in the notebooks accompanying this report.

In addition to using DBSCAN to cluster the headlines, clusters in both segments were divided into two groups - those where more than 50% of the headlines referenced Zika and those where the percentage was less. Those where the percentage was less than 50% were eliminated. The implication being that a preponderance of evidence (headlines) was required to designate an area as having an outbreak.

Clustering Results

When DBSCAN was applied to the longitudes/latitudes of the cities in the two segments of headlines, 32 clusters were identified in the 'rest of the world' and 14 clusters within the US. Out of the 32 clusters in the 'rest of the world', 17 of the clusters had more than 50% of their headlines referencing Zika (the average was ~85%). Out of the 14 clusters in the US, only 3 referenced Zika more than 50% of the time (the average was 90%).

The following maps display the clusters that qualify as 'outbreaks' in each segment. For each map, different colors represent different clusters.

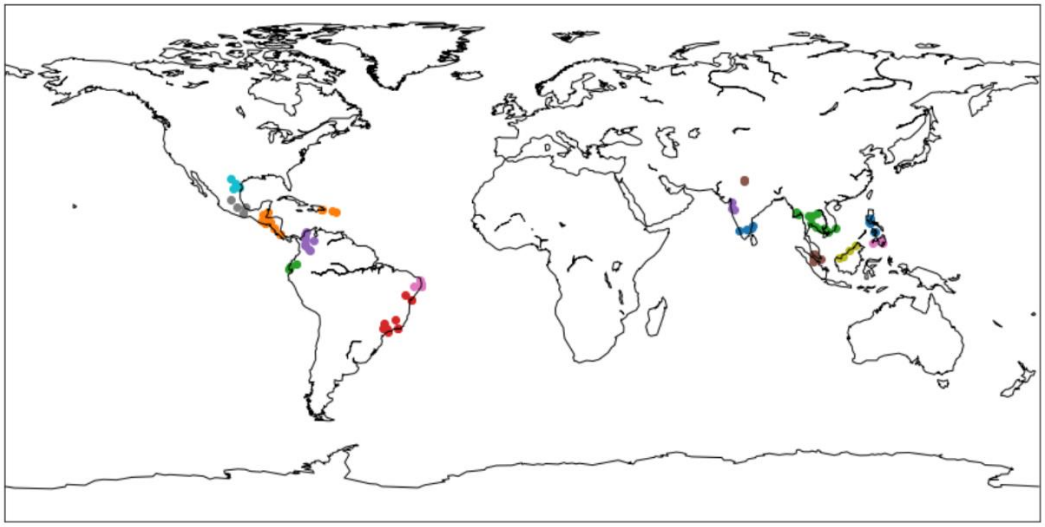


Figure 4A. Location of Clusters in 'Rest of World' with Zika Outbreaks (N=17)

From the standpoint of WHO's reporting structures, the clusters with Zika outbreaks are in three reporting regions: Americas, South East Asia, and Western Pacific. However, visually it appears the Americas region is broken into 2 groupings -- one hovering around Mexico, Central America and the top of South America and the other in Brazil.

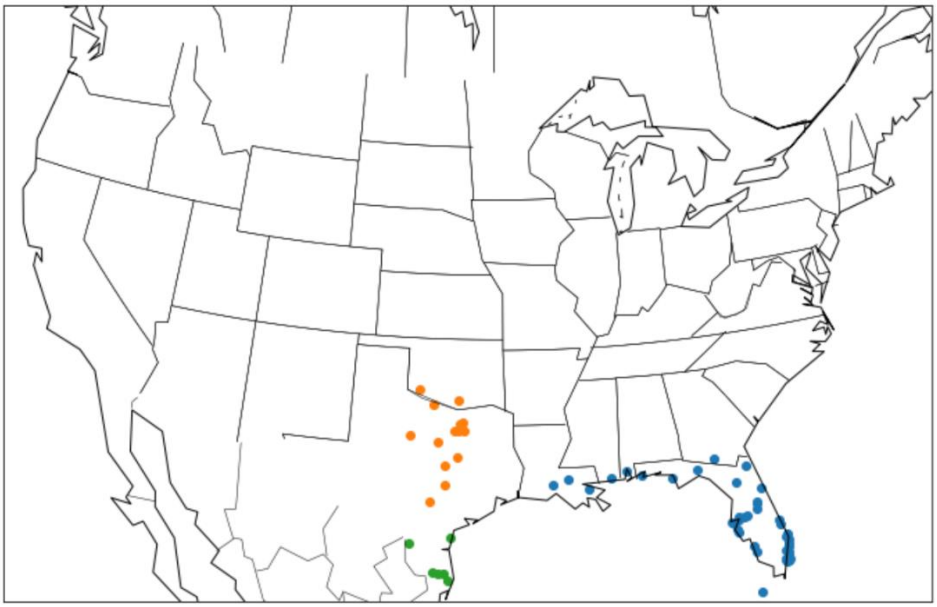


Figure 4B. Location of Clusters in US with Zika Outbreaks (N=3)

In the US segment, there are three clusters of outbreaks located in the Gulf Coast States and up into Texas and Oklahoma.

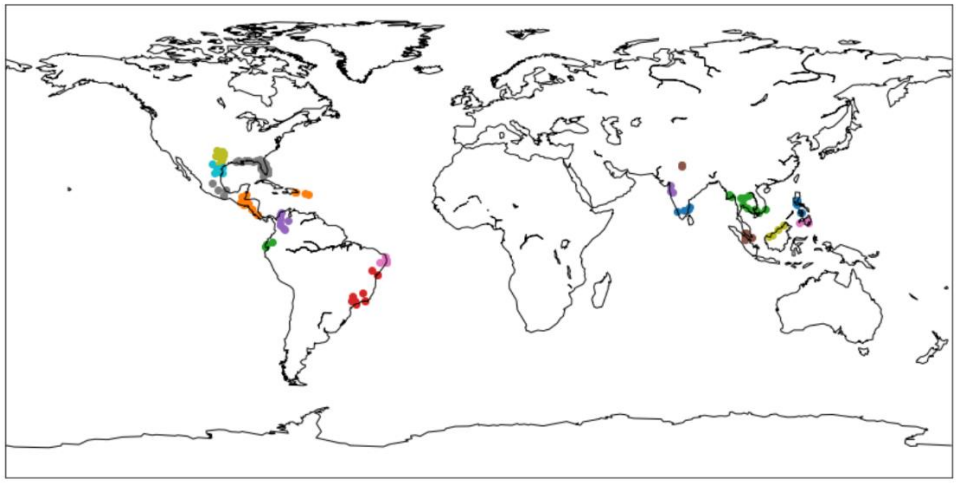


Figure 4C. Worldwide Location of Clusters with Zika Outbreaks (N=20)

When the clusters of outbreaks are combined into a total worldwide view, the 4 groupings in the Americas are highlighted further with one large group of clusters of outbreaks running from Florida through Mexico into Central America and down to Columbia and Ecuador in South America.

Interpretation

The specific location of the clusters with Zika outbreaks is not surprising. These are all areas that have high concentrations of the *Aedes aegypti* mosquitos which can bear Zika (see Figure 5). The fact that it is not more widespread throughout South America and Asia may indicate that this is the beginning of larger epidemic in these areas.

What is surprising is that there are no outbreaks reported in Africa, which also has a large concentration of areas with *Aedes aegypti* mosquitos. Overall, 31 of the headlines in the sample reference Africa, but none of these also reference Zika. There are a variety of potential reasons for this discrepancy like underreporting, misreporting, or widespread immunity in the population. However, there are other sources that have documented a widespread outbreak in Cape Verde in west Africa that began in October, 2015 and was ongoing at the beginning of 2016. This potential oversight indicates that we should revisit the way in which the sample of headlines was collected.

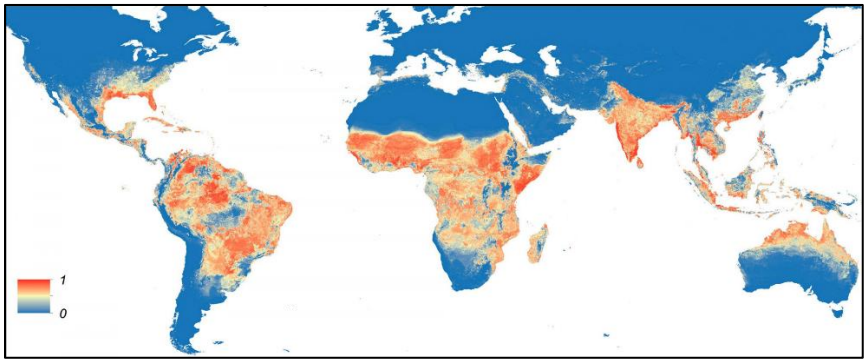


Figure 5. Predicted distribution of *Aedis aegypti* mosquito. Red is higher probability.
Source: ww.keranews.org/post/how-contagious-zika

Traveling to and from Affected Areas

The fact that the clusters of outbreaks have occurred in regions with higher concentrations of Zika bearing mosquitos leads us to conclude that the infections were locally acquired. However, it's also the case that someone could travel to a country with ongoing infections and be bitten by an infected mosquito or acquire the disease through sexual intercourse. In fact, in 2015 the Center for Disease Control (see cdc.gov/zika/reporting/2015-case-counts.html) reported 62 symptomatic Zika disease cases in the US where all 62 were due to travelers returning from affected areas. In contrast they reported 10 cases in the US territories with 1 case resulting from a returning traveler and 9 cases through presumed local mosquito-borne transmission.

Conclusion

To reiterate there is substantial evidence provided by this analysis that there is a wide spread outbreak of Zika that has the potential in 2016 to reach epidemic status in the Americas (from the Gulf States in US down to Brazil), in Southeast Asia (especially India), and the Western Pacific region. There is also reason to further investigate the case of Africa, which has an ongoing outbreak that does it not referenced in this sample.

The Who protocols for dealing with Zika outbreaks are well-established and in this instance should include advisories for: mosquito control, travelers to affected areas, personal protections (clothing, use of repellants, etc.), for individual in affected areas, and for protecting pregnant who plan to travel to or live in affected areas.