Supplementary Information

Our bioinformatics approach allowed us to develop a thirty-three year dataset (1980-July 2013) of 12,102 outbreaks of 215 human infectious diseases, comprising more than 44 million total cases occurring in 219 nations (Table 1). Data extracted from outbreak records and used in analyses have been deposited at http://ramachandran-data.brown.edu/datarepo, which is hosted by SR at Brown University.

Overview of The Global Infectious Disease Epidemiology Online Network (GIDEON)

GIDEON (http://www.gideononline.com) is a subscription-based diagnostic and reference website and application providing extensive geographic, ecological and epidemiological information for 347 recognized human infectious diseases across 231 nations. The data are collated through a system of computer macros and dedicated source lists developed over the past 19 years (http://www.gideononline.com/about/gideon). GIDEON is updated weekly; searches of Medline are conducted against a listing of all GIDEON key words, and titles/abstracts of interest are reviewed. All available national Health Ministry publications are scanned for relevance before they are collated and entered into GIDEON, as are standard publications of the World Health Organization (WHO) and the Centers for Disease Control (CDC). A full list of the references and resources used by GIDEON is available at http://www.gideononline.com/features/resources. GIDEON data on national endemic disease distributions (present-day presence/absence of diseases locally acquired within a nation's borders) has been used in a number of studies [e.g. 6-8], including one by this study's lead author [6]. That paper focused on 289 infectious diseases in GIDEON that had extensive ecological and epidemiological data (described below). All of the diseases represented in our database were analyzed by Smith et al. 2007 [6].

Outbreak data in GIDEON

For each disease outbreak, GIDEON provides a record that includes nation and year of occurrence. In addition, the record often lists the number of cases. An outbreak is defined by GIDEON (adopted from the World Health Organization) as the occurrence of cases of disease in excess of what would normally be expected in a *defined* community, geographical area or season. The completeness of the GIDEON outbreak records have been confirmed against those based on PubMed, ISI Web of Knowledge, WHO and CDC reports, validating GIDEON as a source of records of known outbreaks [15]. The GIDEON outbreak data is in textual form that is not easily synthesized for downstream use (e.g., GIS-based or statistical analyses of changes in outbreak diversity or cases over time for a particular disease). We know of only one previous study that overcame GIDEON’s prose; this was done by manually extracting information for a fraction of outbreak records available (1,428 water-associated disease outbreaks 1991-2008) [15]. This approach has some disadvantages: by-hand extraction requires a large investment of person-hours for data extraction, introduces human error into resulting data files, and likely forces a study to focus analysis on only a small amount of the data available through GIDEON. We purchased a 1-year membership (June 2012 – July 2013) to the GIDEON database and developed a bioinformatics pipeline that automates the parsing and encoding of outbreak records from the website, implemented using regular expressions.

Ecological attributes of pathogens

Each outbreak was encoded with the epidemiological and ecological information for the disease's causal pathogen previously compiled from GIDEON by Smith et al. (2007) [6]. Pathogens are classified taxonomically in one of five basic groups: bacteria, fungi, parasites (generally the helminthes which include cestodes, nematodes, trematodes, and acanthocephalans), protozoa (including protistans and algae), and viruses. Pathogens are also classified by host type. Human specific pathogens are those that are currently entirely restricted to human hosts (i.e., contagious only between persons). Examples include measles, smallpox, and syphilis. Zoonotic pathogens develop, mature, and reproduce entirely in non-human hosts, but nonetheless have the potential to spill over and infect human populations. Humans are a dead-end host for pathogens in this group that is, the pathogen does not then spread human-to-human. Examples of zoonotic pathogens include rabies, plague, and hantavirus. Smith et al. (2007) [6] created a third category of host type called multi-host pathogens that can use both human and non-human hosts to complete their lifecycle. These pathogens are often consider zoonotics and are in this study. Finally, pathogens are classified by vector transmitted or non-vector transmitted, where vector hosts are defined as organisms that facilitate transmission of pathogens between hosts. Subsequently, outbreak records were categorized by three classifications: 1) the taxonomy of the causal pathogen (bacteria, fungi, parasites, protozoa, or viruses), 2) pathogen transmission mode (vector transmitted or non-vector transmitted), and 3) host type (human specific or zoonotic) (see also Table 1).

The format of outbreak records in GIDEON

Our ability to automate parsing of outbreak information depended on the format of outbreak record text presented by GIDEON. GIDEON contains 31,788 outbreak records (as of July 29, 2013). Many outbreak records date back to the early 1900s; for this study we focused on those occurring since 1980. (http://web.gideononline.com/web/whatisnew/contents.php?page=outbreaks) as that is the earliest time at which other covariates are reliably available. Two examples of the format of outbreak records in GIDEON are:

> [Adenovirus, China]: 2002 - An outbreak (176 cases, 1 fatal) of Adenovirus 11 B2 infection was reported among students at a middle school in Beijing.
>
> [Dengue, India]: 1988 - An outbreak was reported in Kolkata.

Each outbreak record in GIDEON also lists superscript numeric citations, which are hyperlinks to outbreak documentation (not shown in the examples above).

Each outbreak record is denoted by the disease, nation of occurrence (see below) and year. The record year may be the year the outbreak occurred or the publication year of the cited outbreak documentation (e.g., a scientific publication) (see below). Many outbreak records report a total number of cases; this number is sometimes broken down further among four case types: 1) an estimated total number of cases (referred to hereafter and in the main text as "total"; this total estimate is observed in 74.01% of all records parsed), 2) confirmed (observed in 2.97% of all records parsed), 3) hospitalized (0.97% of parsed records), and/or 4) fatal (17.10% of parsed records). Our analyses are based on total case data (see below).

Bioinformatics pipeline

The pipeline is implemented in a Python script, and takes as input a text file of outbreak records for a particular disease. The input file must contain a disease name as a header in all capital letters, and all outbreak records for a particular nation must be grouped together with the nation's name and the text "Notable outbreaks:" followed by a new line preceding the records. The input file is formatted in the same way that GIDEON presents the information to members visiting the website, but citation hyperlinks are removed before the input file is passed to our script.

To parse and encode the outbreak records, our script identified the following data from each outbreak record: nation where the outbreak occurred, year, and cases. The script noted if any of these data items were missing from the outbreak record. These data were identified using regular expressions: patterns in the input outbreak record text that preceded or succeeded nation names where outbreaks occurred, the year of the outbreak, and case information.

Parsing nation name

Our parsing script first identifies all nations that reported outbreaks for a particular disease; GIDEON reports outbreaks in territories and in nation states, which we collectively refer to as "nations" in this study. The user provides a text file containing the names of all nations included in GIDEON, and the script searches the input file for an exact string match of each nation name in uppercase letters; the script expects nation names to be both preceded and followed by a new line. All nations with outbreak records for the disease of interest are stored in a list as parsing continues.

Parsing outbreak year

Each outbreak record is recorded in a single line in the text file and follows a uniform format:

Five spaces + year of outbreak as four digits [+ "to" + end year as four digits] [+ "(publication year)"] + hyphen + record text.

Items shown in [] above are optional and appear in only some records.

The parsing script leverages this format to identify the year of the outbreak. For outbreak records that list a range of years, we only store the first year listed. For 1913 outbreak records (15%), only the publication year of the reference from which the outbreak information was compiled was available. We randomly selected 100 of these ‘publication year’ records and searched the literature to see if we could determine the actual outbreak year. We were able to do so for 68/100 records. For seven of the 68 publication year records, the publication year was the same as the actual outbreak year. Eleven publication year records were one year off the actual outbreak year, 24 where two years off, 10 were three years off, and 16 were four or more years off. In total, 76% of the 68 publication year records we surveyed were reported within three years of the outbreak year. All analyses were conducted both with and without publication year records. The significance of 12 results reported in Tables S1 and S2 changed when publication year records were removed from analyses and so we report only these results (excluding publication year records) in the paper.

Parsing total number of cases

We use regular expressions to identify the total number of cases for a single outbreak record. Most, but not all, case data is listed within parentheses. For each record, the script isolates the first parenthetical clause in the record text. If no parenthetical clauses are found in the record text, all case data is at first considered missing.

First, a set of regular expressions searches for other case types in the following order: fatal, hospitalized, and confirmed. Data on these case types are reported for a small percentage of outbreaks; only 5.63% of outbreak records which reported cases did not report a total number of cases. The challenge when parsing the total number of cases from an outbreak record is that numerous adjectives may precede or follow the total estimate (e.g., probable, suspected). The other case types have fewer terms preceding and following data (Table S3). We therefore first parse these other case types from the parenthetical clause, so that the total number of cases is extracted easily.

We parse the total number of cases by removing matched text for other case types from the outbreak record, to avoid mismatching the total number of cases. For example, consider an outbreak record with the following parenthetical clause containing case data:

(45 cases, 3 fatal cases)

Fatal cases are located first. Once the number of fatal cases is matched and removed, the string being parsed by the script would become:

(45 cases, ______)

The parenthetical clause includes neither hospitalized nor confirmed case data, so these case types would be logged as missing. The scan for total cases would find a positive match in the record and parse “45 cases”.

If ranges in case data were reported, the total number of cases was parsed as the upper end of the range. Some records listed multiple locations with case data for each location, but no total number of cases. For example, “2002 – An outbreak occurred in Pittsburgh (7 cases), Philadelphia (50 cases), and Atlanta (10 cases).” Here, we summed case data reported and recorded that sum as the total cases. That is, the record above was parsed as reporting 67 total cases.

In general, following a number, the phrases ‘estimated cases’, ‘probable cases’, ‘approximate’, and ‘suspected’ implied a total number of cases and included any other cases listed of different case types (i.e., in ‘45 suspected cases, 3 confirmed’, we assumed that the 45 suspected cases includes the 3 confirmed cases). In a small percent of records, however, the text suggested that these numbers preceding terms such as ‘suspected/probable’ only included the unconfirmed cases rather than the total. In these records, we summed the unconfirmed and confirmed cases to find the true total. These records were identified by the following criteria: the confirmed cases were separated from the unconfirmed by the word “and” (rather than the more common comma separator, implying the two numbers are distinct) or the total number was smaller than either the confirmed, fatal, or hospitalized cases. Figure S3 provides a detailed diagram of our script’s application to two GIDEON dengue outbreak records.

The user provides as input to the script all outbreak records for a disease of interest from GIDEON in a text file (e.g., black text in Figure S3). Parsing identifies information on the nation of origin of the outbreak (orange), outbreak year (maroon), publication year (red, detailed above), multinational locations (purple, detailed below), and total cases (green).

Flagging records for post-processing

Figure S3 highlights an important feature of our pipeline for processing outbreak record text from GIDEON: boolean flags. Boolean flags take on the value 1 for records where *1)* outbreaks listing multiple nations should be further processed by hand to ensure accuracy, and *2)* case information pertains to animals. For all flagged records, our parsed output stores and reports the original outbreak record in its entirety, along with the parsing script results. While the use of regular expressions will sometimes parse information incorrectly (see below), our flagging system allows us to overcome this error by flagging records for post-processing which improved the accuracy of the final dataset extracted from GIDEON.

1) Outbreaks listing multiple nations. In Figure S3, the first outbreak’s location is flagged due to the parsed phrases "Air Force" and "American". In this record, post-processing will allow the user to identify that the nation of this outbreak is actually the Philippines not the United States. In this study, we are interested in the nation where each outbreak occurred: therefore, we update the parsed record as occurring in the Philippines and remove it from the list of outbreaks occurring in the USA. We also flagged and then removed records of outbreaks that occurred on cruise ships, military ships, buses, or on airplanes in transit between nations. These were identified using the following key words/prefixes: "military", "tourist", “travel”, “introduc”, “return”, “neighbor”, “arriv”, “cruise”, “visit”, "missionar", “flight”, “import”, “export”, "Navy", “navy”. Outbreaks that occurred on military / cruise ships while docked in a specified nation were kept (n<50).

2) Outbreak records detailing cases in animals. Records mentioning animals were flagged to determine if outbreak cases occurred in animals or in both animals and humans. Records mentioning animals were identified with the following key words/prefixes: "animal", "mammal", “ dog”, “wolf ”, “wolve ”, deer ”, "goat", “monkey”, “macaque”, “ lion”, “ zoo”, " pig", "bird", "fish", “antelope”, “sheep”, "cattle", "livestock", "poultry", “zebra”, " cow", "horse", " ox", “equine, “ovine”, “canine”, “murine”, " farm". The texts of these records either explicitly stated that the outbreak occurred in animals (e.g., “An outbreak (4 fatal) occurred in a herd of buffalo”) or listed an animal-only strain of the disease (e.g., “An outbreak (17 cases) of bovine anthrax occurred”). We identified and excluded 503 records of outbreaks occurring only in animals. Outbreaks that affected both humans and animals were kept in the dataset for analysis; for these outbreak records, our parsed case data only reflects the total number of humans affected by the disease outbreak.

Validation and quality assurance

We validated our pipeline against outbreak data extracted by hand from GIDEON (both within our research group and from Yang et al. 2012 [15]. Our parsing script can produce spreadsheets in minutes, far faster than manual extraction. Despite the accuracy and increased speed of our script as a parser (compared to data compiled manually), we recognize that some errors and inconsistencies could go undetected in the extraction and encoding of the outbreak data.

To ensure the highest possible accuracy of the data, we also examined each of the extracted outbreak records manually, making corrections where necessary. During this step we were able to identify and remove duplicate records. Some outbreak records included information for multiple diseases. When these records listed specific case data for each disease, we split them into separate outbreak records. For example, “an outbreak (53 cases) of Adenovirus (23 cases) and Enterovirus (30 cases) was reported.” Here we create two outbreak records, one for each disease. Some outbreak records described multiple outbreaks. We created new outbreak records from these when unique locations and multiple outbreaks were made clear. Take, for example, the following outbreak record: “Outbreaks (1,000 cases or more, 24 cases, 57 fatal) of meningococcal infection were reported in Bauchi State (24 cases, 7 fatal), Kano State (40 fatal), Kebbi State (7 fatal) and Katsina State (1,000 cases, 4 fatal).” Here, four new outbreak records were created, one each for Bauchi State, Kano State, Kebbi State, and Katsina State, each with the associated case numbers. The original outbreak record text was maintained for each of these new records and the original record was deleted. For records similar to this, but for which it was clear that many outbreaks occurred, but only some are reported, the record is kept as is and no additional records are created. An example is, “Outbreaks (1,667 cases during January to December) were reported, including Frankfurt (45 cases) and Berlin (52 cases).” Here we treat the record as a single outbreak occurring in Germany with a total of 1667 cases, but for which all specific locations are not known. In this way we are prioritizing total case data over individual outbreak records.

Empirical analyses

Shannon’s diversity index (SDI) is a quantitative metric used by ecologists to measure the diversity of species in a community. SDI provides more information about species community composition than the simple measure of richness (i.e. the total number of species present). SDI (denoted by *H*) simultaneously accounts for how many different species there are in a community, and how evenly individual organisms are distributed among the species.

$$H = -\sum_{i=1}^{R} p_i \ln p_i$$

SDI increases both when the number of species (R) increases and when evenness increases. For a given number of species, the SDI is maximized when all species are equally abundant. If all individuals in a community are of a single species and the other species are very rare (even if there are many of them), SDI will approach 0. When there is only one species in the community, SDI equals 0. In this study we equate unique diseases as ‘species’ and each outbreak of a disease as an ‘individual’. Here, H is disease diversity, R is the total number of unique diseases, and p_i is the proportion of outbreaks caused by the *i*th disease in the overall dataset. In this way we can calculate SDI for infectious disease outbreaks annually for each nation in our dataset. This allows us to quantify, for the first time, global trends in disease diversity. For example, a nation with a very low SDI (close to 0) will be one where a small number of diseases cause the vast majority of outbreaks. Other diseases may occur in this nation but these only rarely cause an outbreak.

Controlling for reporting biases in models

Research has shown that a nation's likelihood of experiencing, identifying and reporting an outbreak or harboring subsets of infectious disease within its' borders is determined in large part by its surveillance capabilities, communication infrastructure, geography, and the availability of hosts for the causal pathogen [3,5-9,12-19,21]. While there are many variables that might serve as proxies for these national attributes we chose to control for six well documented in the literature: latitude, GDP, press freedom, Internet usage, human population size and density [3,5-9,12-19,21]. We describe each of these indicators in more detail in the following paragraphs. Our confounding variables were selected based on precedence in the peer-reviewed literature [3,5-9,12-19,21] and availability of data for each variable in all nations and years represented by the outbreak records.

Previous studies have documented a significant increase in infectious disease occurrence (richness) with decreasing latitude [7,8]. To account for this, we determined the latitudinal centroid of each nation using a spatial data file (geodatabase polygon feature class) representing countries of the world (World Countries, ESRI Data & Maps Edition 10, 6/30/2010). The ArcGIS (version 10.1) "Feature To Point" geoprocessing tool converted the polygons to a point feature class in which the position of the output point was located at the center of gravity (centroid) of each country. Each country's name was retained with this output. The Add XY geoprocessing tool was then used to append the tabular component of the point feature class with the longitude and latitude of each point location. A spatial join was performed with a polygon feature class representing continents in order to append the tabular component with the name of each point's corresponding continent. The tabular component was exported to .csv format.

The relationship between poverty and infectious disease is well documented. The global distribution of per-capita gross domestic product (GDP) shows a striking correlation with a number of infectious diseases though the direction of causality is not straightforward and most likely runs both ways [21]. As nations becoming increasingly prosperous their ability to invest resources into detecting and preventing infectious disease outbreaks should also increase. We control for this using annual per capita GDP values for each of the nations in our data set, compiled from the from the World Bank: http://data.worldbank.org/data-catalog/world-development-indicators.

Press freedom and Internet usage have been shown to influence the likelihood of a nation reporting an outbreak [12-14]. A free press contributes to governmental transparency, providing citizens with information and promoting justice. We compiled indices from Freedom House: http://www.freedomhouse.org/report-types/freedom-press. The index is a numerical value between 0 and 100, where 0 is the freest. The variable Press Freedom is treated as a categorical variable with three levels: Free, Partially Free (PF) and Not Free (NF), and we used the level “Free” as the reference level. Internet is used to collect and disseminate information, promote governmental transparency, and is often the vital communication component needed for public health surveillance. We compiled the number of Internet users per 100 people from the World Bank (beginning in 1990): http://data.worldbank.org/data-catalog/world-development-indicators.

The availability of hosts is the most fundamental determinant of pathogen presence and ultimately infectious disease [6]. We use human population size and density to control for host availability and compiled these data from the World Bank: http://data.worldbank.org/data-catalog/world-development-indicators. Density was calculated as population size of a nation in a given year divided by that nation's surface area. For zoonotic pathogens, animal population size and density also contribute to host availability [6], however it is impossible to compile this data, nationally and annually for each of the zoonotic diseases in the dataset.

Statistical methods

Regression models were fit to study the relationship between the confounding variables and four dependent variables: the total number of outbreaks, disease richness, disease diversity and per capita cases caused by outbreaks, all recorded by nation and year. Let Y_{it} indicate the dependent variable in nation i and year t; then the linear regression model we used to investigate temporal trends of disease diversity and per capita cases takes the form

$$Y_{it} = X\beta + \gamma t \qquad (1)$$

where X is the design matrix, the matrix of all independent variables, with the first column of 1s (hence with 7 columns total) and the rest of the confounding variables described earlier, and β is the vector of the coefficients associated with the design matrix. The coefficient γ is the linear temporal trend of the dependent variable and is of primary interest of this research. For models with total number of outbreaks and disease richness as the dependent variables, we employed quasi-Poisson models that take the same format as (1) after the log link function.

Our statistical analyses are based on the outbreak data in our database between 1980 and 2009, as the confounding variables beyond 2009 and before 1980 were not available (or not applicable, as in the case of Internet usage) at the time of analysis. Therefore, our regression results are based on 4,342 nation-year records when analyzing temporal trends in the total number of outbreaks, richness and diversity, and 5,685 outbreak-nation-year records when analyzing temporal trends in per capita cases. In all analyses, we log-transformed total population, population density and GDP.

The p-values reported test the hypothesis that the null model excluding temporal information (i.e., $H_0: t = 0$) fits the data better than does the full model with temporal information (i.e. $H_A: t \neq 0$). Analyses were also conducted on each sub-category of outbreaks defined by pathogen taxonomy, transmission mode and host type (see section on Ecological attributes above). These results can be found in Table S1.

The purpose of including the confounding variables in our models is not to test whether their regression coefficients are statistically different from zero, but to control for their effects on the temporal trends in the dependent variables. In order to employ the full range of the available outbreak data, we conducted analyses from 1990 to 2009 where Internet usage was included. These results can be found in Table S2.

We also conducted a sensitivity analysis where we fitted a nation-specific intercept term in the models to allow for additional variation at the nation-level that is not captured in the model otherwise. The regression results with nation-specific intercept terms remain very close to those shown in Table S1 and S2 and therefore are not reported here, but are available from the authors on request.

Supplementary Figure Captions

Figure S1. Cumulative human infectious disease outbreaks reported by each nation since 1980. Total number of outbreak records reported for each nation shaded so darkest nations experienced the greatest cumulative number of outbreaks.

Figure S2 Rank-abundance curves for outbreaks caused by zoonotic and human specific diseases. Abundance was measured as the log number of outbreaks caused by each causal zoonotic (black circles) and human specific (grey squares) disease during three time periods: (a) 1980-1989, (b) 1990-1999 and (c) 2000-2009.

Figure S3. Parsing of two Dengue outbreaks. Our parsing script identifies and organizes information on outbreak nation, year, and total number of cases. The parsed information is highlighted with boxes.