Revealing Influences of Socioeconomic Factors over Disease Outbreaks

S Mahmudul Hasan
1305043.sh@ugrad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Alabi Mehzabin Anisha
mehzabin@Knights.ucf.edu
University of Central Florida
Orlando, Florida, United States

Rudaiba Adnin
1505032.ra@ugrad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Ishrat Jahan Eliza
1605089@ugrad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Ishika Tarin
1805092@ugrad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Sadia Afroz
1505030.sa@ugrad.cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

A. B. M. Alim Al Islam
alim_razi@cse.buet.ac.bd
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

ABSTRACT

The recent Covid-19 pandemic elucidates the need for a better disease outbreak analysis and surveillance system, which can harness state-of-the-art data mining and machine learning techniques to produce better forecasting. In this regard, understanding the correlation between disease outbreaks and socioeconomic factors should pave the way for such systems by providing useful indicators, which are yet to be explored in the literature to the best of our knowledge. Therefore, in this study, we accumulated data on 72 infectious diseases and their outbreaks all over the globe over a period of 23 years as well as corresponding different socioeconomic data. We, then, performed point-biserial and spearman correlation analysis over the collected data. Our analysis of the obtained correlations demonstrates that various disease outbreak attributes are positively and negatively correlated with different socioeconomic indicators. For example, indicators such as lifetime risk of maternal death, adolescent fertility rate, etc., are positively correlated, while indicators such as life expectancy at birth, measles immunization, etc., are negatively correlated, with disease outbreaks that affect the digestive organ system. In this paper, we find and summarize the correlations between 126 outbreak attributes derived from the characteristics of the 72 diseases in consideration and 192 socioeconomic factors which is a novel contribution to the field of disease outbreak analysis and prediction.

CCS CONCEPTS

• **Mathematics of computing → Exploratory data analysis**.

KEYWORDS

Socioeconomic factors, Disease outbreaks, Statistical analysis, correlation

ACM Reference Format:
S Mahmudul Hasan, Alabi Mehzabin Anisha, Rudaiba Adnin, Ishrat Jahan Eliza, Ishika Tarin, Sadia Afroz, and A. B. M. Alim Al Islam. 2022. Revealing Influences of Socioeconomic Factors over Disease Outbreaks. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '22), June 29-July 1, 2022, Seattle, WA, USA*. ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3530190.3534804

1 INTRODUCTION

Disease outbreaks are not a recent phenomenon, they have been around for a very long time [55]. The most recent example, the Covid-19 pandemic, showed us how severely they can hamper the day-to-day lives of citizens and as a result render the world to a standstill. However, as the number of recurring disease outbreaks increases, we accumulate more data regarding their dynamics. Recent innovations in machine learning and data mining have equipped the healthcare sector with new knowledge deriving from the data accumulated over the past decades including the analysis and surveillance of different disease outbreaks.

Several studies were conducted to predict as well as analyze the causes of disease outbreaks [2, 7, 14, 45, 72, 95, 106, 110]. However, a large scale study to find out the roles that socioeconomic factors play over disease outbreaks is yet to be done. Such a study can help us understand how different characteristics of disease outbreaks are related to different socioeconomic factors. Besides, it can pave the way for the development of better disease outbreak surveillance systems that will incorporate socioeconomic data for its prediction

COMPASS '22, June 29-July 1, 2022, Seattle, WA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9347-8/22/06…$15.00
https://doi.org/10.1145/3530190.3534804

and analysis. Finally, such a study can guide further incisive studies to identify the roles socioeconomic factors play in triggering disease outbreaks as well as how they are affected during such outbreaks. Although some similar studies were conducted, they are either limited in terms of the number of diseases they work with or the locale of the outbreaks they took under consideration, or both [7, 8, 12, 22, 76, 95].

In this paper, we accumulated data regarding the occurrences of 72 infectious diseases and their outbreaks and the socioeconomic factors of the regions of the outbreaks. Then, we used point-biserial and spearman correlation to find the nature and strength of the relationships between various socioeconomic factors and attributes of outbreaks. Here, disease outbreak data is the name, place, and year of the outbreak. We accumulated these data from WHO: Disease Outbreak News [104]. Socioeconomic data include attributes such as access to electricity, literacy rate, internet usage, etc., which we collected from the World Bank Open Data [107]. Both of these data are available on a global scale. Disease outbreak attributes are the attributes of the corresponding diseases, which we gleaned from fact sheets provided by WHO and CDC, and various other sources such as Malacards, Mayoclinic, etc.

Thus, we encountered the following set of research questions while conducting this study.

- **RQ1:** How can we accumulate relevant data to find the relationships between disease outbreak characteristics and socioeconomic indicators from numerous online sources (e.g., WHO, CDC, etc.)?
- **RQ2:** How can we quantify the strength and nature of the relationships between disease outbreak characteristics and socioeconomic indicators? Based on the quantification, what are the major trends over those relationships? What are the primary socioeconomic indicators that influence a disease outbreak?

Our study takes a large number of diseases and their outbreaks as well as a large number of socioeconomic attributes into consideration. In summary, our major contributions are as follows:

- We accumulated and cleaned a data set regarding disease outbreak reports by scrapping WHO: Disease Outbreak News [104], CDC [15, 105], from various informative websites such as Malacards, Mayoclinic [30, 68], and from various other sources.
- We considered 126 disease outbreak attributes derived from the characteristics of the 72 diseases and 192 socioeconomic factors. We performed large-scale point-biserial and spearman correlation analysis on the accumulated data to quantify the strength and nature of disease outbreak attributes and socioeconomic indicators relationships. We summarize the nature and prevalent trends between the relationship of the characteristics of different disease outbreaks with different socioeconomic factors in a systematic way and made them available which will benefit systems in predicting the effects of disease outbreaks over socioeconomic factors.
- We find the primary socioeconomic factors for a disease outbreak exhibiting a particular character. These make our study a novel contribution to the field of disease outbreak analysis and prediction.

2 RELATED WORK

Recently, the healthcare sector saw a rapid increase in the number of applications of data mining and machine learning techniques because of the availability of data. Several notable studies were done on disease outbreak prediction and analysis, and some of them addressed the notion of the relationship between socioeconomic factors and characteristics of different disease outbreaks.

The first plague, which was called the 'Black Death' killed up to 50 million people in Europe and the Mediterranean alone, making it the second-worst pandemic in human history in absolute terms, behind the 1917-1919 Spanish Influenza that killed between 50 and 100 million people. From the very beginning, relative to the rich, the poor were very vulnerable to plagues because of their relatively unhealthy living areas and very poor treatment during the epidemics [3]. The 1918 Spanish Influenza pandemic holds a particular place in medical history; it wiped out an estimated 1% of the global population and earned the dubious honor of being coined the "mother of all pandemics" [63]. A seminal study revealed more than 30-fold variation in 1918 pandemic excess mortality rates across a sample of 20 countries, with socioeconomic factors explaining a significant fraction of the observed variation. The role of socioeconomic disparities on influenza mortality has remained a subject of debate in the literature and is often confounded by the timing of arrival of the pandemic virus in a given locale, climatic conditions, or population density [17, 65]. A systematic review presents a meta-analysis on the association between socioeconomic status and disease outcomes in the last 5 influenza pandemics to document whether and to what extent there is an association between indicators of socioeconomic status (e.g. income, education) and pandemic outcomes (infection, hospitalizations, mortality) in the last five influenza pandemics (1889, 1918, 1957, 1968, 2009) [57]. Different socioeconomic status and morbidity of the disease indicators came into light through a rigorous data collection process named as pre-registered study protocol. The findings showed that lower socioeconomic status groups have the highest risks of the three considered pandemic outcomes (infection, hospitalizations, mortality) [59, 75, 96]. Another spatially refined study pieced together historical maps of pneumonia and influenza deaths reported during the lethal wave of the pandemic in Chicago in October–November 1918, together with archival census tract data, to analyze the relationship between pandemic mortality and sociodemographic variables (including illiteracy rate, homeownership, unemployment, population density, and age). Pneumonia and influenza mortality rates were found to increase on average by 32% for every 10% increase in illiteracy rates. The findings align with contemporary studies demonstrating how limited literacy and educational achievement hamper access to preventive services [35]. In a study, both GAM and MLR statistical techniques were employed to model the influences of meteorological and socioeconomic conditions on the interannual variability of cholera. It has been shown that increases in temperature, rainfall, poverty, and population density may increase both cholera cases and deaths, while improvement of drinking water and adult literacy might reduce the risk of contracting the disease [53]. Researchers [12] also reviewed and documented epidemiological and socioeconomic data on the outbreak of Cholera in Uganda. They found that access to safe water, sanitation, and hygiene are the main factors for the outbreaks, and people who lack these are at high risk of getting affected. They also found illiteracy and poverty as other reasons for the spread of the outbreak. A study [22] on the Cholera outbreak in Southern Ghana collected data on socioeconomic, household hygiene, food, and water exposures. They found that age below 18, education below tertiary, exclusive household toilet facility, cold/warm food, homemade food, and community pipe-borne water were associated with the outbreak. Researchers [8] worked on finding socioeconomic and environmental factors that affect the outbreaks of a mostly neglected yet epidemic-prone zoonotic disease Leptospirosis. They identified municipalities with lower socioeconomic status, i.e. lower household quality, low hygiene, higher extreme poverty, and illiteracy rates, as the main drivers for the outbreaks. Animal-borne infectious diseases have likely been precipitated by a complex interplay of changing ecological, epidemiological, and socioeconomic factors. One recent study develops an Environmental-Mechanistic model that captures elements of each of these factors, to predict the risk of Ebola virus disease (EVD) across time and space [78]. Another study also shows the significant level of severity of Ebola alongside the Zika virus among marginal people in low-income countries [48]. An analysis of the cases between 2007 and 2017 of Dengue, Zika, and chikungunya caused by arboviruses and transmitted by the mosquito Aedes aegypti, confirmed three distinct Colombian municipalities (Bello, Cúcuta, and Moniquirá) as three different ecosystems given their contrasted geographic, climatic, and socioeconomic profiles. Socioeconomic factors such as barriers to health and childhood services, inadequate sanitation, poor housing, and poor water supply in those areas were the fuels of disease transmission [64].

An early disease outbreak detection algorithm called WSARE which uses healthcare data as well as information regarding demographics, symptoms, etc. was discovered in a study [106]. Researchers [72] used GoogleTrends2 [14, 33] to develop *FluBreaks: an early warning system* for flu epidemics. Datasets from news and internet media regarding dengue outbreaks were collected and used for outbreak prediction [2]. A dengue outbreak prediction system in Malaysia by augmenting rainfall data with dengue case data is built which improved outbreak detection accuracy [110]. None of these studies work on finding relationships between socioeconomic factors and characteristics of disease outbreaks. Rather these studies could possibly be improved if socioeconomic data were augmented [110]. There are, however, some works that address the aforementioned relationships but are limited by their local nature, the number of socioeconomic factors they worked with, or both. Further, a dengue outbreak detection mechanism was built in a study [95]. In this work, the authors investigated the impact of different socioeconomic factors on dengue outbreaks which they found in their dataset. Researchers [7] conducted a study on Malaria, Diarrhoea, and Pneumonia and their outbreaks to understand the impact of different socioeconomic and environmental factors on the outbreaks of these diseases. A similar study took place to review how socioeconomic factors drive the outbreaks of dengue, chikungunya, yellow fever, and Zika Virus and found a large variability regarding the relationship between socioeconomic factors and the most common Aedes-borne diseases [103]. A study [76] to find out the effects of socioeconomic and environmental factors on the outbreak of Dengue fever in China. They used statistical analysis to assess and detect the relevant factors and analyzed the impact of those on the smallest administrative unit. Their analysis identified six related factors representing urbanization, poverty, accessibility, and vegetation. The findings of a review-based study suggest a gap in the literature which indicates the need for additional research or a better circulation of current findings regarding the relationship between socioeconomic factors and the distribution of Ae. aegypti [42].

Researchers investigated whether low socioeconomic status might play a role in increased risk for infectious diseases. Therefore, they analyzed the association between educational level and net household income, and serum IgG concentration (presence of antibody) against measles, mumps, rubella, varicella, etc. collected within a national cross-sectional survey (2006/2007) using linear regression analyses among non-vaccinated individuals. The result indicates that a higher educational level was associated with higher IgG concentrations against measles and rubella compared to a low education level. In contrast, higher education level was associated with lower IgG concentrations against pneumococcus, MenC, and CMV compared to low education level [41]. A similar type of study took place to determine the variations in taking vaccination of mumps, rubella, etc. A notable finding was over the 16 years studied, higher levels of socioeconomic deprivation (Income, Employment, Education Skills and Training, Crime, and Living Environment) were consistently and strongly associated with lower uptake of MMR1 (measles, mumps, and rubella dose 1) and MMR2 vaccine. However, poorer educational attainment, lower levels of employment, and lower household income were also significantly associated with lower uptake of both MMR1 and MMR2 [43].

One recent study has shown that the COVID-19 pandemic has affected areas of the United States that are already institutionally underprivileged. These areas shared content with negative expression, prayers, and discussion of the CARES Act economic relief package while all the privileged areas were concerned with stocks, social distancing, and national-level policies [89] Another study illustrated that two social distancing measures, which are: travel distance and stay-at-home dwell time, have a statistical relationship with the growth rate of COVID-19 confirmed cases across U.S. states. The statistical variation of the two social distancing measures can be easily explained by socioeconomic and geographic factors, including age groups, state policies, population density, race and ethnicity, and median household income [31]. Though Japan was less damaged than Europe by the COVID-19 pandemic, one multidimensional review showed that socioeconomic crises were created by the pandemic as the people in Japan are also suffering from social isolation and the socioeconomic impacts of the pandemic. As a decrease in income leads to an unstable lifestyle in general, it might directly increase illness and decrease well-being, without the mediating effect of the fear of COVID-19 [90]. While Bangladesh was already in a COVID-19 crisis from the beginning of the pandemic, one study warned that dengue and natural disasters could worsen the situation, identifying key indicators of risks exposures to COVID-19 including congested urban-focused unsustainable vulnerability, demographic and social vulnerability, economic and physical vulnerability, and recurrent disaster vulnerability, which listed the 20 most vulnerable districts out of total 64 [16]. Another study reveals the possible socioeconomic and environmental impacts of COVID-19 and finds medical waste to have a large impact on the environment in recent years [9]. It also concludes that densely populated countries are at higher risk of COVID-19.

The aforementioned studies deal with finding environmental and socioeconomic factors for different disease outbreaks. However, most of these studies focused on an individual disease or a distinct type of disease. None of the studies were done on a global scale.

3 RESEARCH METHODS

In general, we can split our data-driven study into the steps shown in Figure 1. We accumulate, process, and combine required data according to problem definition, and then, perform analysis. Finally we summarize and interpret the results.

- **Data Accumulation.** A good data mining research depends on credible data. For our study, we needed data on socioeconomic conditions, data on disease outbreaks, and data on disease characteristics. WHO keeps track of disease outbreaks in different parts of the world, on a yearly basis. When an outbreak occurs, it publishes reports on the diseases involved [104]. We accumulated our disease outbreak data from this source. World Bank keeps track of different socioeconomic factors of different countries on a yearly basis. We collect the socioeconomic data from this data source [107]. We gleaned data regarding attributes of the diseases in consideration from various sources on the web. For example, we studied various fact sheets provided by WHO and CDC [15, 105]. Besides websites such as Malacards, Mayoclinic, etc. also alleviated the process [30, 68].
- **Data Cleaning and Transformation** Unprocessed data contain various errors such as misspelling, repetition, malformed contents, etc. The process of removing these errors is known as data cleaning. Besides, we need to adapt the data according to a particular data analysis procedure before we can use it in the analysis tool. This is called data transformation. Now, we assembled data on different socioeconomic and environmental factors, disease outbreaks, and disease characteristics. Disease outbreak data and disease characteristics data were scrapped and manually collected and therefore, needed cleaning to make them usable for our computer program.
- **Statistical Analysis** After data is prepared, different statistical and machine learning techniques can be used to extract the pattern/relationships we are interested in. From the nature of our data, we have determined that we can use simpler techniques and therefore, decided to use Pearson and spearman correlation analysis [27, 51]. In correlation analysis, a value of +1 indicates a strong positive relationship, and a -1 indicates a strong negative relationship. A correlation of 0 indicates no correlation. Our disease attributes data is dichotomous in nature, whereas Pearson correlation is defined for two continuous variables. Therefore, we used a variant of Pearson correlation, point-biserial correlation [37] which is defined for a continuous and a dichotomous variable. Although spearman correlation is usually used for ordinal dependent variables, we find that it can be used for binary variables too [83, 88]. Therefore, we additionally performed Spearman correlation.
- **Summarization and Analysis of Results** Result analysis is the final task of any data mining task. After we find the correlations between different variables, we interpret those. Because of our problem formulation, which will be discussed later, we obtained massive-sized correlation tables that needed to be broken down for analysis. We provide a detailed analysis of our obtained correlation later.

3.1 Data Collection and Preparation

We have selected disease and attribute both at a country level. World Bank provides data on different socioeconomic attributes for almost all of the countries in the world each year [107]. We have extracted 761 such attributes from this database about the socioeconomic conditions of 264 places that include both individual countries and general regions or groups of countries. The 264 places that we get from the World Bank data are at national and regional levels. These countries often have a shared history or regime, and therefore, the world bank prefers to analyze their indicators not only by individual countries but the regions as well [107]. Thus, the notion of regions comes considering the unavoidable historical perspectives, which present few changes in country borders. It appeared that the world bank prefers to analyze its indicators not only by individual countries but the region as well [107]. The indicators are self-explanatory and continuous in nature. Some of them are – access to clean fuels and technologies for cooking (% of the population), access to electricity, rural (% of rural population), etc. A list of these attributes as well as corresponding interactive distributions can be found in World Bank [108]. Although, World Bank has data on the years 1960 - the current, Disease Outbreak News only contains reports from the year 1995, and therefore, we are considering the data for those years only.

We use web scrapping to acquire data from **Disease Outbreak News** maintained by WHO. This data has three attributes: date, disease, and country/region. We modified the date formats in case of mismatches, fixed misspelled disease names, fixed country names and made them consistent with the World Bank dataset, and replaced different region names with corresponding countries. Besides, there were instances where some diseases were addressed in different names. We read the related articles and made those consistent. Usually, the aforementioned reports are created a few days after an outbreak. As we are interested only in the year of the disease outbreak, we process these report dates and keep only the year. As a result, our disease outbreak dataset contains information regarding ongoing disease outbreaks by year and country. In other words, if a country had an ongoing outbreak in a certain year, we have a row corresponding to the country, year, and disease. Figure 2, 3 and 4 depict the distribution of disease outbreak data by country, year, and disease. Disease Outbreak News does not have any entry before 1996. However, some of the entries in 1996 refer to ongoing outbreaks which started in 1995. Therefore, we added a few entries for those.

Finally, we need data on the characteristics of the diseases appearing in the outbreak dataset. In total, we considered 72 diseases such as yellow fever, zika, malaria, various strains of influenza (h1n1,

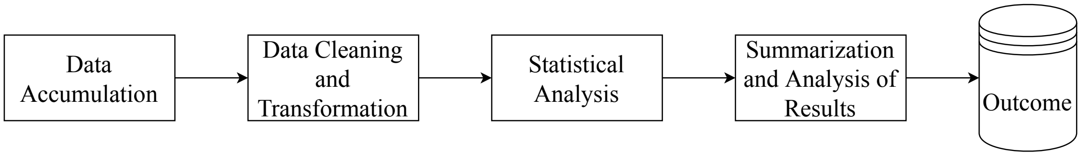

Data Accumulation
Data Cleaning and Transformation
Statistical Analysis
Summarization and Analysis of Results
Outcome

Figure 1: Brief overview of our study.

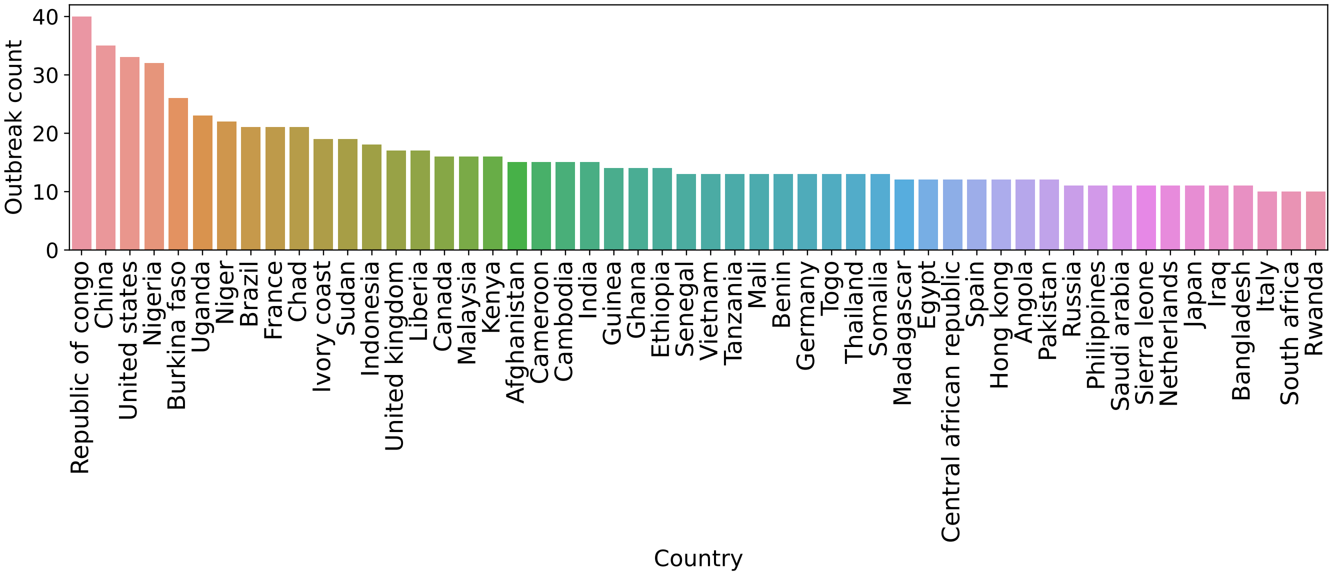


Figure 2: Disease outbreak data by country.

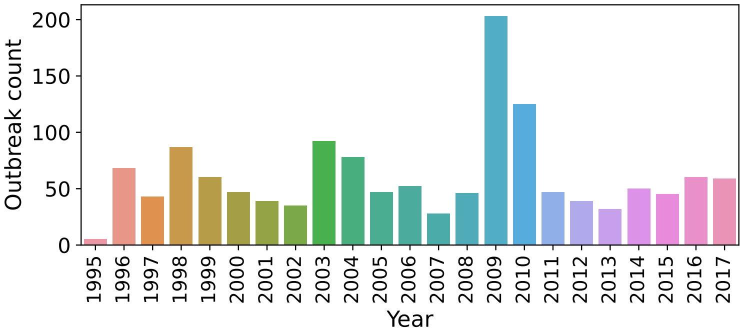


Figure 3: Disease outbreak data by year.

h5n1), etc. The 72 Diseases are the diseases that appeared in the WHO disease outbreak news [105]. There was no direct database for all the disease symptoms in one place. Therefore, we have manually collected those data. We garner data on 5 ordinal characteristics of a disease – affected organ systems, symptoms, transmission methods, carriers, and infectious agents. We acquired this information from fact sheets provided by WHO and CDC [15, 105], from various informative websites such as Malacards, Mayoclinic [30, 68], and from various other sources. Here we briefly describe these attributes and their considered values. Figure 5 shows the distribution of all the aforementioned disease attributes' values.

- Almost every disease that causes outbreaks affects one or more **organ systems** of the human body and therefore, this information can be considered as an attribute of the disease and the related outbreak. There are 11 major organ systems in the human body, which we considered in our study – cardiovascular, digestive, endocrine, urinary, integumentary, lymphatic, muscular, nervous, reproductive, respiratory, and

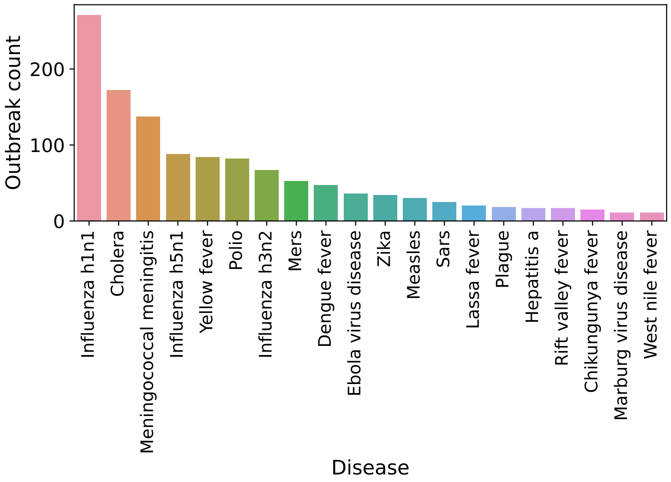


Figure 4: Disease outbreak data by disease.

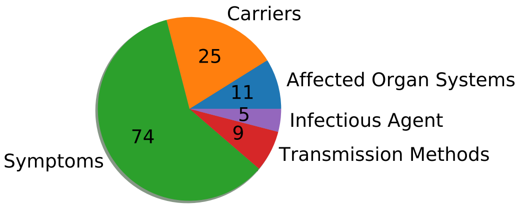


Figure 5: Considered number of types of each ordinal characteristics of a disease.

skeletal [20]. Besides, we added a N/A option that can be used in cases where information is not available, or if the disease does not affect any particular organ system.

- **Symptoms** are the indicators associated with a disease in general [66]. We considered a total of 74 symptoms and general problems that can occur during the diseases in consideration. Some of these symptoms are – abdominal pain, anorexia, arthralgia, bleeding, cellulitis, chest pain, etc.
- **Transmission methods** are the ways an outbreak usually spreads. We considered 9 types of transmission methods – air, animal to human, food contamination, indirect contact, mosquito to human, human to human, water contamination, pregnant women to baby, and genetics. We added N/A if we did not find any specific transmission method.
- **Carriers** are the animals that act as reservoirs for the disease of an outbreak and help spread it. We identified a total of 25 carriers that carry the 72 diseases we are dealing with. Some of these are – rodents, bats, camels, cattle, deer, dogs, mosquitoes, etc. We added N/A if we didn't find any specific carrier.
- **Infectious agents** are the substance that causes the disease. We found 5 such infectious agents – bacteria, virus, fungi, parasite, and prion. Most of the infectious diseases in consideration are either viral or bacterial. Only a few of the diseases are caused by fungi, parasites, or prions.

3.2 Statistical Analysis

After accumulating all three kinds of data from their corresponding sources, we need to combine them together for analysis. Figure 6 illustrates the steps for the whole process. First, we scrap WHO: Disease Outbreak News website for the disease outbreak data. This data contains information about diseases outside of our selected 72 as well as some outbreaks that are not infectious diseases at all (e.g. unexplained cluster of deaths, melamine poisoning, unidentified outbreak, etc.). Therefore, we remove these unwanted disease data. World Bank data is available on yearly basis only. We cannot use the day and month present in the 'Date' attribute in disease outbreak data and therefore, we discard those. As Disease Outbreak News usually produces multiple reports regarding ongoing outbreaks, we get a lot of duplicate rows containing the same year, country, and disease. Now, we only need to know each year, what outbreaks each country had. Therefore, we remove the duplicated rows to extract this information. This is our final disease outbreak dataset. It contains three attributes: Year, Country, and Disease (disease name). Next, we obtain data regarding 761 socioeconomic indicators (including Year, Country) from World Bank. After examining the values for these indicators we find that a good number of attributes have a lot of missing values and our program will not be able to work unless we impute them. There is no good threshold that we can use for data imputation. We find that the amount of data that can safely be imputed without introducing additional bias is still

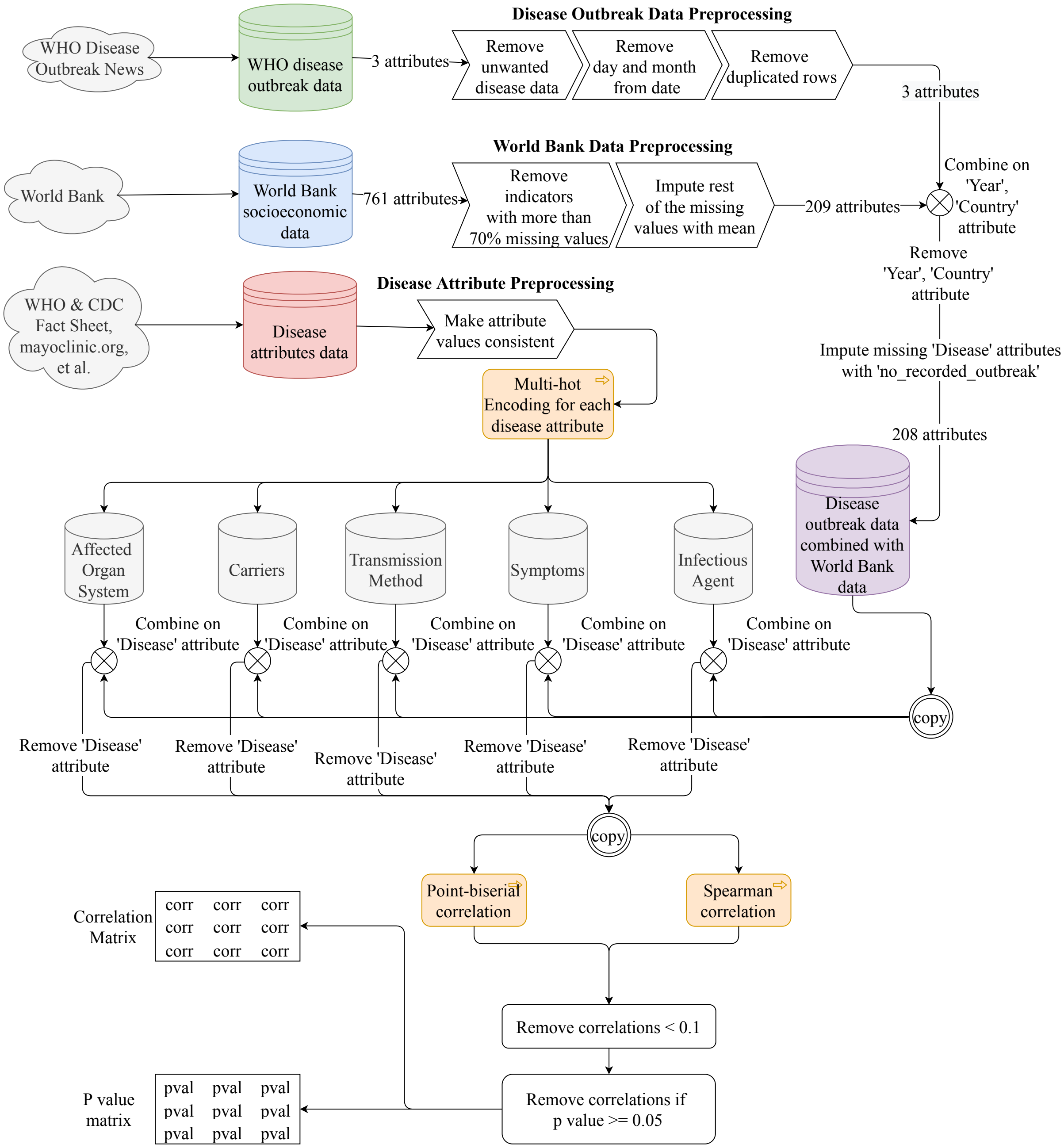
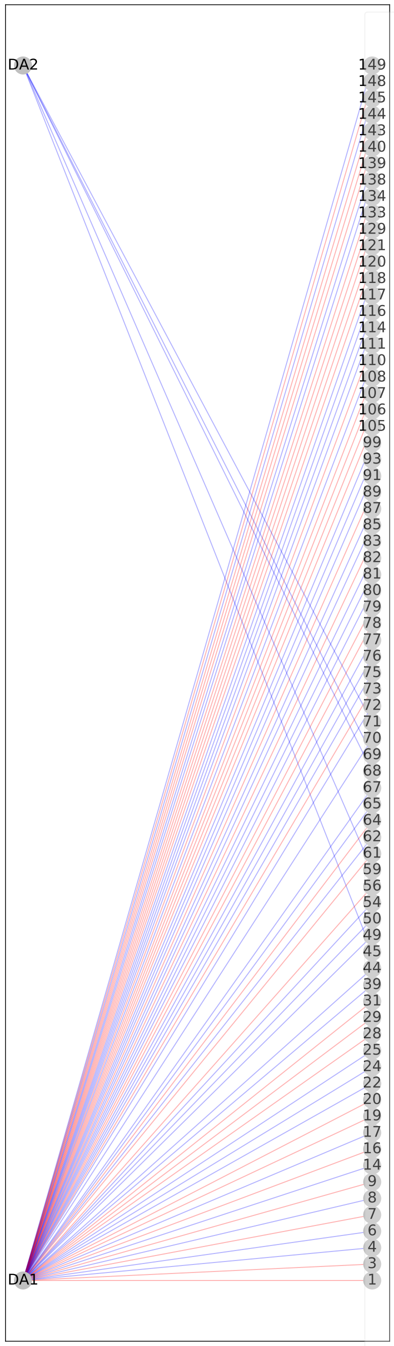

Disease Outbreak Data Preprocessing
WHO Disease Outbreak News
WHO disease outbreak data
3 attributes
Remove unwanted disease data
Remove day and month from date
Remove duplicated rows
3 attributes
World Bank Data Preprocessing
World Bank
World Bank socioeconomic data
761 attributes
Remove indicators with more than 70% missing values
Impute rest of the missing values with mean
209 attributes
Combine on 'Year', 'Country' attribute
Remove 'Year', 'Country' attribute
Impute missing 'Disease' attributes with 'no_recorded_outbreak'
208 attributes
Disease Attribute Preprocessing
WHO & CDC Fact Sheet, mayoclinic.org, et al.
Disease attributes data
Make attribute values consistent
Multi-hot Encoding for each disease attribute
Disease outbreak data combined with World Bank data
Affected Organ System
Carriers
Transmission Method
Symptoms
Infectious Agent
Combine on 'Disease' attribute
Combine on 'Disease' attribute
Combine on 'Disease' attribute
Combine on 'Disease' attribute
Combine on 'Disease' attribute
copy
Remove 'Disease' attribute
Remove 'Disease' attribute
Remove 'Disease' attribute
Remove 'Disease' attribute
Remove 'Disease' attribute
copy
Point-biserial correlation
Spearman correlation
Correlation Matrix
corr corr corr
corr corr corr
corr corr corr
Remove correlations < 0.1
Remove correlations if p value >= 0.05
P value matrix
pval pval pval
pval pval pval
pval pval pval

Figure 6: Schematic diagram showing all steps of our correlation analysis.

a topic for discussion and some suggest that up to 30% missing values can be imputed safely without biasing the data too much [79, 87, 100]. Therefore, in our study, we first remove all attributes that have more than 70% missing values. After that, we impute the rest of the indicators with their corresponding algebraic means [10]. After performing these steps, we are left with 209 attributes (including Year, and Country) of the original 761 attributes.

We merged the processed disease outbreak data with the 209 attributes' data using Year and Country attributes. As we were using the Year attribute for establishing an order in the dataset, and the Country dataset to help combine disease outbreak data and world bank data, after we got the merged dataset, we no longer needed them. Therefore, we removed Year and Country from the merged dataset. As a result, we are left with 208 attributes. This dataset contains a 'Disease' attribute, which contains a value if there's an outbreak in a particular country in a particular year, otherwise, it's missing. Therefore, we fill up these missing 'Disease' attribute values with 'no_recorded_outbreak'. Accordingly, after



DA1 - bacteria (+ve: 38, -ve: 32)

DA2 - virus (+ve: 5, -ve: 0)

1 - Life expectancy at birth total year (+ve: 0, -ve: 1)

3 - Employment in services male pct of male employment modeled ILO estimate (+ve: 0, -ve: 1)

4 - Vulnerable employment female pct of female employment modeled ILO estimate (+ve: 1, -ve: 0)

6 - Lifetime risk of maternal death pct (+ve: 1, -ve: 0)

7 - Mobile cellular subscriptions per 100 people (+ve: 0, -ve: 1)

8 - Forest rents pct of GDP (+ve: 1, -ve: 0)

9 - Employment in industry male pct of male employment modeled ILO estimate (+ve: 0, -ve: 1)

14 - Population ages 0-14 female pct of total (+ve: 1, -ve: 0)

16 - Access to electricity urban pct of urban population (+ve: 0, -ve: 1)

17 - Mortality rate adult male per 1000 male adult (+ve: 1, -ve: 0)

19 - School enrollment primary pct gro (+ve: 0, -ve: 1)

20 - Immunization measles pct of children ages 12-23 month (+ve: 0, -ve: 1)

22 - Prevalence of anemia among women of reproductive age pct of women ages 15-49 (+ve: 1, -ve: 0)

24 - Mortality rate under-5 per 1000 live birth (+ve: 1, -ve: 0)

25 - Employment in agriculture female pct of female employment modeled ILO estimate (+ve: 1, -ve: 0)

28 - Population ages 35-39 male pct of male population (+ve: 0, -ve: 1)

29 - Primary education pupils pct female (+ve: 0, -ve: 1)

31 - Population ages 40-44 male pct of male population (+ve: 0, -ve: 1)

39 - Renewable energy consumption pct of total final energy consumption (+ve: 1, -ve: 0)

44 - Adolescent fertility rate births per 1000 women ages 15-19 (+ve: 1, -ve: 0)

45 - Fixed telephone subscription (+ve: 1, -ve: 0)

49 - Prevalence of anemia among children pct of children under 5 (+ve: 1, -ve: 0)

50 - Death rate crude per 1000 people (+ve: 1, -ve: 0)

54 - Urban population growth annual pct (+ve: 1, -ve: 0)

56 - Access to electricity rural pct of rural population (+ve: 0, -ve: 1)

61 - Population ages 50-54 female pct of female population (+ve: 0, -ve: 1)

62 - Population ages 10-14 female pct of female population (+ve: 1, -ve: 0)

64 - Population ages 45-49 female pct of female population (+ve: 0, -ve: 1)

65 - Population ages 5-9 female pct of female population (+ve: 1, -ve: 0)

67 - Population ages 0-14 male pct of total (+ve: 1, -ve: 0)

69 - Population ages 65 and above female (+ve: 1, -ve: 0)

70 - Vulnerable employment male pct of male employment modeled ILO estimate (+ve: 1, -ve: 0)

71 - Urban population (+ve: 1, -ve: 0)

72 - Employment in agriculture pct of total employment modeled ILO estimate (+ve: 1, -ve: 0)

73 - Life expectancy at birth male year (+ve: 0, -ve: 1)

75 - Mortality rate neonatal per 1000 live birth (+ve: 1, -ve: 0)

76 - Population ages 0-14 pct of total (+ve: 1, -ve: 0)

77 - Mortality rate adult female per 1000 female adult (+ve: 1, -ve: 0)

78 - Wage and salaried workers female pct of female employment modeled ILO estimate (+ve: 0, -ve: 1)

79 - Employment in services pct of total employment modeled ILO estimate (+ve: 0, -ve: 1)

80 - Population ages 0-4 female pct of female population (+ve: 1, -ve: 0)

81 - Prevalence of anemia among pregnant women pct (+ve: 1, -ve: 0)

82 - Fertility rate total births per woman (+ve: 1, -ve: 0)

83 - Life expectancy at birth female year (+ve: 0, -ve: 1)

85 - Vulnerable employment total pct of total employment modeled ILO estimate (+ve: 1, -ve: 0)

87 - Self-employed female pct of female employment modeled ILO estimate (+ve: 1, -ve: 0)

89 - Population ages 35-39 female pct of female population (+ve: 0, -ve: 1)

91 - Prevalence of anemia among non-pregnant women pct of women ages 15-49 (+ve: 1, -ve: 0)

93 - Mortality rate infant per 1000 live birth (+ve: 1, -ve: 0)

99 - Self-employed male pct of male employment modeled ILO estimate (+ve: 1, -ve: 0)

105 - Employment in services female pct of female employment modeled ILO estimate (+ve: 0, -ve: 1)

106 - Wage and salaried workers total pct of total employment modeled ILO estimate (+ve: 0, -ve: 1)

107 - Population ages 15-64 male pct of total (+ve: 0, -ve: 1)

108 - Population ages 15-64 pct of total (+ve: 0, -ve: 1)

110 - Population ages 5-9 male pct of male population (+ve: 1, -ve: 0)

111 - Wage and salaried workers male pct of male employment modeled ILO estimate (+ve: 0, -ve: 1)

114 - Population ages 0-4 male pct of male population (+ve: 1, -ve: 0)

116 - Maternal mortality ratio modeled estimate per 100000 live birth (+ve: 1, -ve: 0)

117 - Self-employed total pct of total employment modeled ILO estimate (+ve: 1, -ve: 0)

118 - Birth rate crude per 1000 people (+ve: 1, -ve: 0)

120 - Population ages 15-64 female pct of total (+ve: 0, -ve: 1)

121 - Survival to age 65 female pct of cohort (+ve: 0, -ve: 1)

129 - Access to electricity pct of population (+ve: 0, -ve: 1)

133 - Population ages 40-44 female pct of female population (+ve: 0, -ve: 1)

134 - Population ages 50-54 male pct of male population (+ve: 0, -ve: 1)

138 - Age dependency ratio young pct of working-age population (+ve: 1, -ve: 0)

139 - Age dependency ratio pct of working-age population (+ve: 1, -ve: 0)

140 - Survival to age 65 male pct of cohort (+ve: 0, -ve: 1)

143 - Immunization DPT pct of children ages 12-23 month (+ve: 0, -ve: 1)

144 - Population ages 45-49 male pct of male population (+ve: 0, -ve: 1)

145 - Population ages 10-14 male pct of male population (+ve: 1, -ve: 0)

148 - Employment in industry pct of total employment modeled ILO estimate (+ve: 0, -ve: 1)

149 - Employment in agriculture male pct of male employment modeled ILO estimate (+ve: 1, -ve: 0)

59 - Population ages 65 and above male (+ve: 1, -ve: 0)

68 - Population ages 65 and above total (+ve: 1, -ve: 0)

Figure 7: Socioeconomic Indicators with their IDs.

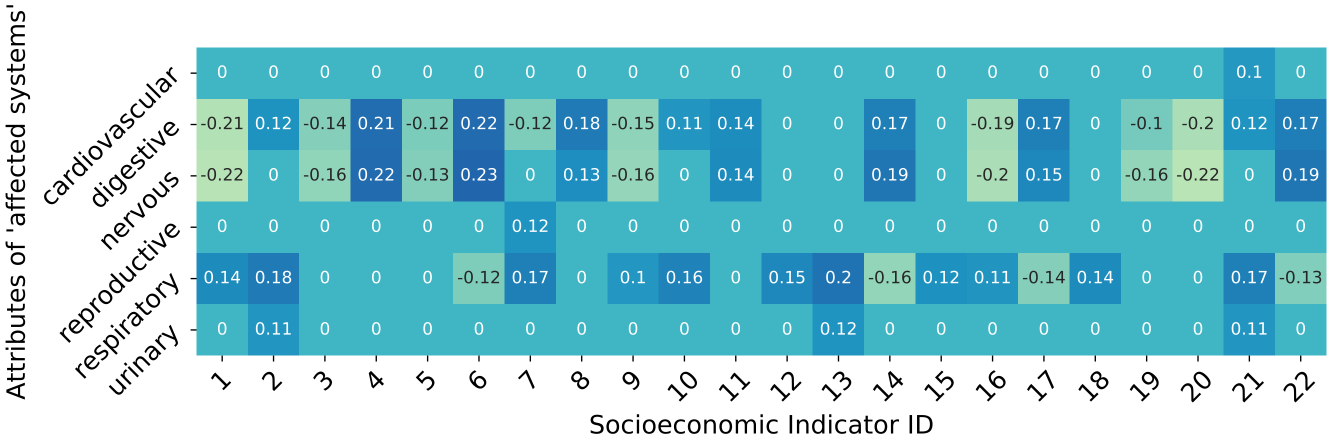


Figure 8: Point-biserial correlation between 'Affected Organ Systems' attribute's values with socioeconomic indicators.

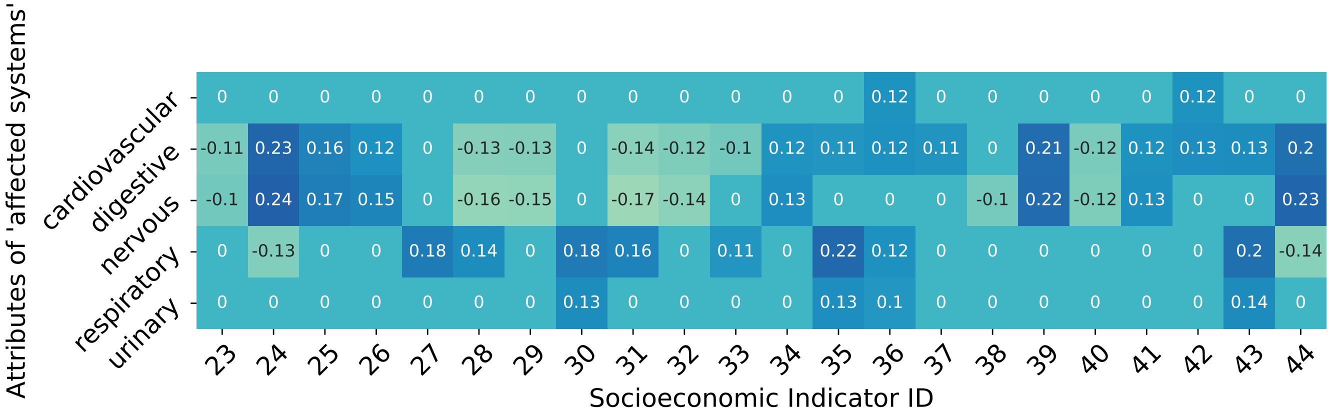


Figure 9: Point-biserial correlation between 'Affected Organ Systems' attribute's values with socioeconomic indicators.

gleaning the disease attributes data, we needed to make the attribute values consistent in the whole dataset. In our study, disease characteristics are categorical in nature and have more than two distinct values. Therefore, we cannot convert them to binary for the analysis. Instead, we use multi-hot encoding for them.

Figure 6 additionally shows the steps for our statistical analysis. We perform these analysis in *Python* using *NumPy, SciPy, Matplotlib*, and *Pandas* libraries [28, 38, 44, 60]. After multi-hot encoding of disease attributes, we remove the redundant 'Disease' column and perform point-biserial and spearman correlations. To be specific, we perform a correlation between each socioeconomic indicator and each of the values of a disease attribute. The output of this correlation is a matrix, where each row corresponds to a value of a certain outbreak attribute and each column corresponds to a socioeconomic indicator. Aside from the correlation matrix, we get a corresponding p-value matrix. The correlation matrix goes through two separate refinement steps. First, we replace all correlation values smaller than 0.1 with 0s as they are too weak [82]. Then, we remove all correlation values whose corresponding p-value is greater than or equal to **0.05** [29].

For any statistical analysis, the choice of p-value is of significant importance. It roughly indicates the probability of an uncorrelated system producing datasets that have a correlation at least as extreme as the one computed from the given datasets. The convention is to take a certain threshold for the p-value and if the computed p-value is less than the threshold, we reject the hypothesis that the dataset is uncorrelated (i.e. null hypothesis). Otherwise, we fail to reject the null hypothesis and as a result, our obtained correlation is said not to be statistically significant. The convention is to take one of 0.01, 0.05, 0.001, and 0.005 as the threshold. A p-value less than 0.05 is a standard level for deciding that the alternative hypothesis has evidence against the null hypothesis [29] and therefore, We use **0.05** as a threshold in our study.

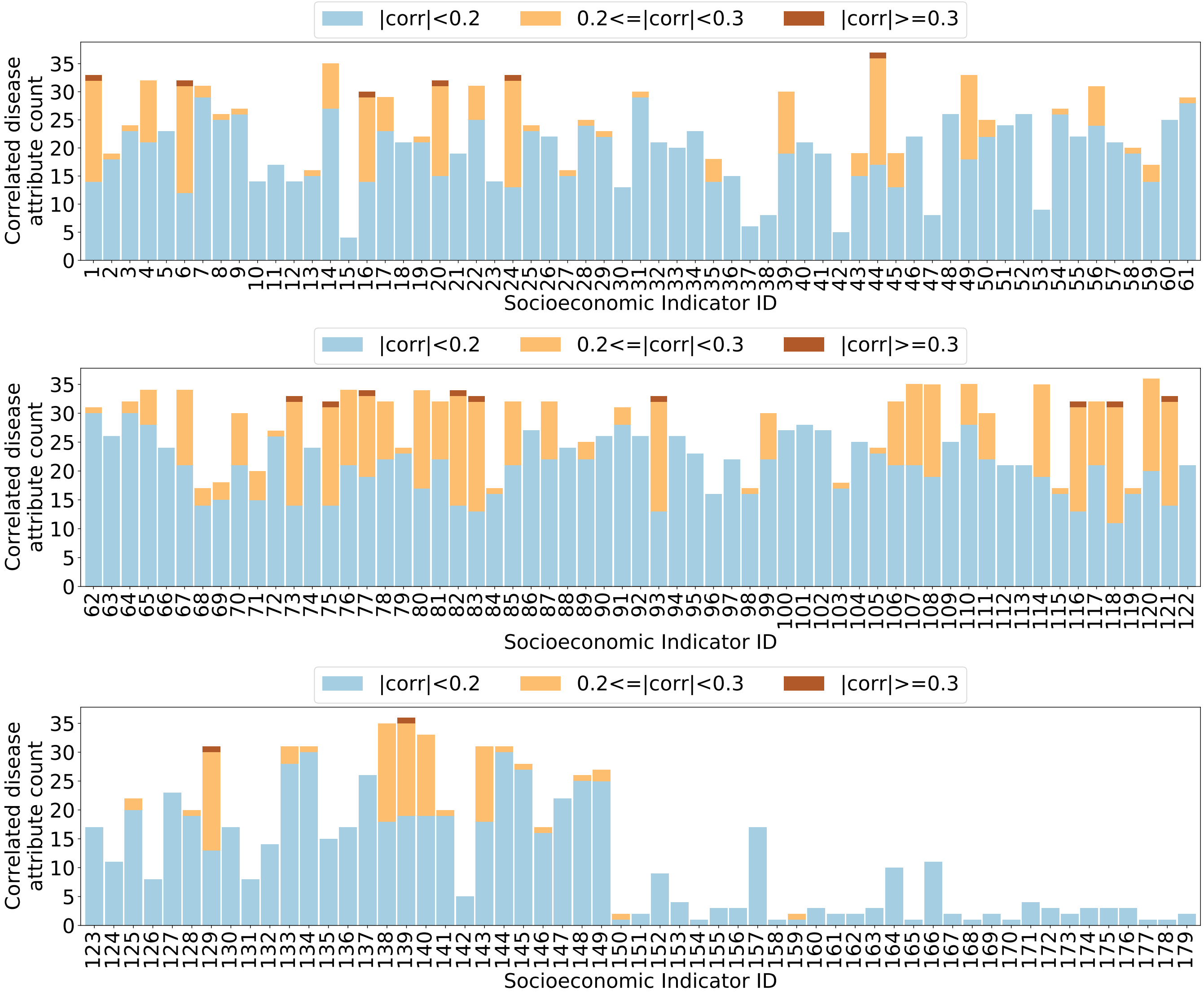


Figure 10: Bar plots showing distribution of point-biserial correlations between socioeconomic factors and outbreak attribute values.

Previously, our null hypothesis was – "there is no relationship between socioeconomic indicators and different attributes of disease outbreaks". However, as we multi-hot encode each of the characteristics of the diseases, the null hypothesis becomes the following – "there is no relationship between socioeconomic indicators and an attribute of disease outbreaks when the value of the attribute is *X*. Here, *X* is one of the possible values of the attribute in consideration. After completing the correlation analysis we get five separate correlation matrices for each type of correlation. These matrices are huge in size having up to 208 columns (maximum number of socioeconomic indicators) and 74 rows ("Symptom" disease characteristic can have 74 values).

4 RESEARCH FINDINGS

We obtained five different correlation matrices for each of the correlation types totaling ten matrices after our analysis. We also found that 192 out of 208 socioeconomic factors show a correlation in the case of point-biserial and spearman correlation. The other 16 socioeconomic attributes do not show any significant correlation. These 192 attributes are given in Figure 7. Here we analyze and discuss only the results of point-biserial correlation only.

We use heat maps (Figure 8 and 9) to illustrate the correlations. Figure 8 illustrates parts of one such matrix where we obtain a point-biserial correlation of the 'Affected Organ Systems' attribute's values with socioeconomic indicators. Apparently, the obtained matrices are massive in size even after our program removed redundant rows and columns (zero columns and rows). As socioeconomic indicators' names are long, we replace them with unique identifiers in our figures. In Figure 8, we can see that there are valid correlations between socioeconomic factors and outbreak attributes. Although the absolute values in these correlations are low, which may indicate a weak relationship, our obtained p-values affirm them to be highly significant.

To find the more significant socioeconomic indicators in case of disease outbreaks (e.g. more often correlated, shows higher correlation, etc.), we summarize the number of correlated disease attribute values for each of the socioeconomic indicators. We categorize the obtained correlations into three intervals: $abs(corr) < 0.2$, $0.2 \leq abs(corr) < 0.3$, and $abs(corr) \geq 0.3$ based on their absolute values and use bar plots to show the distributions. Figure 10 illustrates the bar plots containing the distribution of point-biserial correlations of socioeconomic indicators with outbreak attribute values. We can clearly see that, according to point-biserial correlation, 179 socioeconomic factors of the 208 show significant correlations and most of the correlations are less than 0.2 in absolute value. In fact, 87 of the 179 correlated socioeconomic factor does not show correlations greater than or equal to 0.2 in absolute value. Now, we illustrate the correlations of the other 92 socioeconomic factors and analyze them in detail as they show stronger relationships. Socioeconomic attributes obtained from the world bank have long names, and therefore, we encoded each with a unique identifier (Figure 7) and used those in our subsequent figures.

4.1 Affected Organ Systems

Figure 11 and 12 illustrate the point-biserial and Spearman correlation between different values of the 'Affected Organ Systems' disease attribute and various socioeconomic factors. Here, DAx indicates different values of the disease attribute in consideration. We additionally show some summary statistics regarding the number of positive and negative correlations shown by the value of the attribute. The blue lines correspond to positive correlations, while the red ones correspond to negative correlations. All of these correlations are greater than or equal to 0.2 in absolute value.

Accordingly, in Figure 11 we can see that, only 3 out of 11 values of the attribute 'Affected Organ Systems' show absolute correlations greater than or equal to 0.2. Below are the details.

- We notice that the disease outbreaks that cause problems for 'respiratory' organ systems (e.g. lungs) are positively correlated with 5 different socioeconomic factors: fixed telephone subscriptions (45), methane emissions (43), fisheries production (35), urban population (71), and the population of ages 65 and above (59, 68, 69).
- We also observe that disease outbreaks that cause problems for 'digestive' organ systems are associated positively with 13 socioeconomic attributes and negatively with 8. The positively correlated socioeconomic factors are vulnerable employment (4, 85), self-employment (87, 117), renewable energy consumption (39), adolescent and adult fertility (44, 82), birth rate (118), risk of maternal death and maternal mortality ratio (6, 116), and the mortality rate of children (24, 75, 93) while the negatively correlated socioeconomic factors are the percentage of wage and salaried workers (78, 106), electricity access (129), life expectancy (1, 73, 83), and survival to age sixty-five (121), and immunization (20).
- Lastly, we observe that disease outbreaks that cause problems for 'nervous' organ systems are positively associated with 21 socioeconomic factors, while negatively associated with 14. The positively associated factors represent vulnerable employment (4, 70, 85), self-employment (87, 99, 117), renewable energy consumption (39), adolescent and adult fertility (44, 82), birth rate(118), the prevalence of anemia among children and pregnant women (49, 81), risk of maternal death and maternal mortality ratio (6, 116), the mortality rate of children (24, 75, 93), the population of ages 0-4 (80, 114), and age dependency ratio (138, 139). The negatively correlated factors represent the percentage of wage and salaried workers (78, 106, 111), electricity access (56, 129), life expectancy (1, 73, 83), the population of ages 15-64 (107, 108, 120), survival to age sixty-five (121), and immunization (20, 143).

4.2 Transmission Methods

Figure 13 and 14 shows the correlations between values of 'Transmission methods' and various socioeconomic factors. Apparently, only 3 out of 9 values of 'Transmission methods' show a significant correlation with absolute value greater than or equal to 0.2 (Figure 13). Below are the details.

- We observe that the value 'animal to human', which indicates whether the disease is spread from animals to humans via contact, is positively correlated with 19 socioeconomic factors. These socioeconomic attributes are related to urban population (71), labor force (84), fixed telephone subscription (45), cereal production (2, 141), fisheries production (13, 35), CO_2 and methane emission (27, 43, 128), and population (58, 59, 68, 69, 98, 103, 115, 119, 146).
- Besides, the disease outbreaks that transmit via 'food contamination' and 'water contamination' are positively correlated with 29 and 28 socioeconomic factors in order, while both show a negative correlation with 15. The positively correlated factors for both of the values are almost the same representing self-employment (87, 99, 117), vulnerable employment (4, 70, 85), adolescent and adult fertility (44, 82), birth rate(118), the prevalence of anemia (22, 49, 81), renewable energy consumption (39), risk of maternal death and maternal mortality ratio (6, 116), the mortality rate (17, 24, 75, 77, 93), population (14, 65, 67, 76, 80, 110, 114), and age-dependency ratio (138, 139). 'Water contamination' does not show any correlation with an adult mortality rate (17). Both attributes show a negative correlation with the same factors which are related to the percentage of wage and salaried workers (78, 106, 111), electricity access (16, 129), survival to age sixty-five (121, 140), life expectancy (1, 73, 83), the population of ages 15-64 (107, 108, 120), and immunization (20, 143).

4.3 Infectious Agents

Figure 15 and 16 depicts the correlation of different values of the attribute 'Infectious agents' with the socioeconomic factors. In

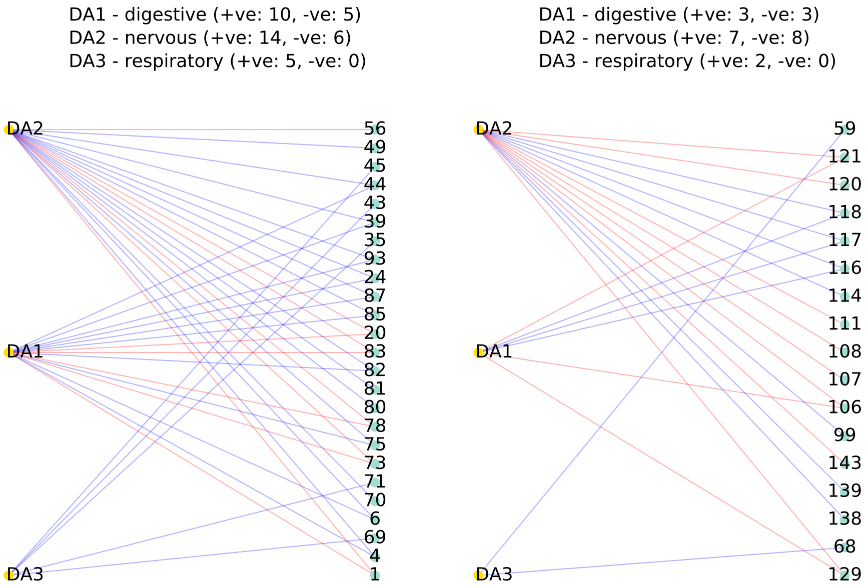


Figure 11: Point-biserial correlation between different values of 'Affected Organ Systems' atrribute and various socioeconomic factors.

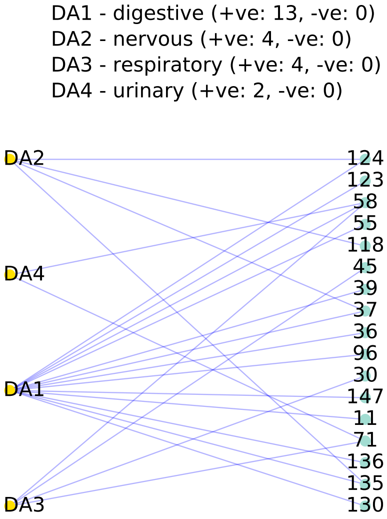


Figure 12: Spearman correlation between different values of 'Affected Organ Systems' atrribute and various socioeconomic factors.

Figure 15, the attribute has 5 distinct values, only 2, 'bacteria' and 'virus', show significant correlations with an absolute value of at least 0.2. Followings are the details.

- We observe that the disease outbreaks caused by viruses are positively correlated with 5 socioeconomic factors. They represent fixed telephone subscription (45), a population of ages 65 and above (59, 68, 69), and an urban population (71).
- The bacterial disease outbreaks are positively associated with 38 factors and negatively associated with 32 factors. The positively correlated factors are related to agricultural employment (25, 72, 149), vulnerable employment (4, 70, 85), self-employment (87, 99, 117), renewable energy consumption (39), forest rents (8), urban population growth (54), adolescent and adult fertility (44, 82), birth rate(118), prevalence of anemia (22, 49, 81, 91), risk of maternal death and

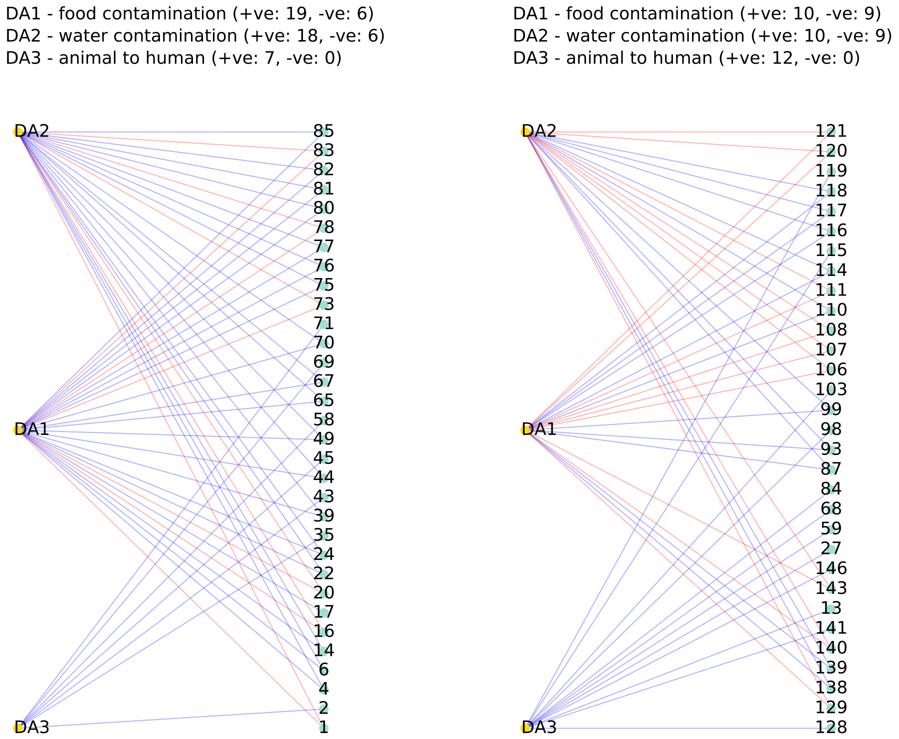

DA1 - food contamination (+ve: 19, -ve: 6)
DA2 - water contamination (+ve: 18, -ve: 6)
DA3 - animal to human (+ve: 7, -ve: 0)
DA2
DA1
DA3
85
83
82
81
80
78
77
76
75
73
71
70
69
67
65
58
49
45
44
43
39
35
24
22
20
17
16
14
6
4
2
1
DA1 - food contamination (+ve: 10, -ve: 9)
DA2 - water contamination (+ve: 10, -ve: 9)
DA3 - animal to human (+ve: 12, -ve: 0)
DA2
DA1
DA3
121
120
119
118
117
116
115
114
111
110
108
107
106
103
99
98
93
87
84
68
59
27
146
143
13
141
140
139
138
129
128

Figure 13: Pearson correlation between different values of 'Transmission methods' attribute and various socioeconomic factors.

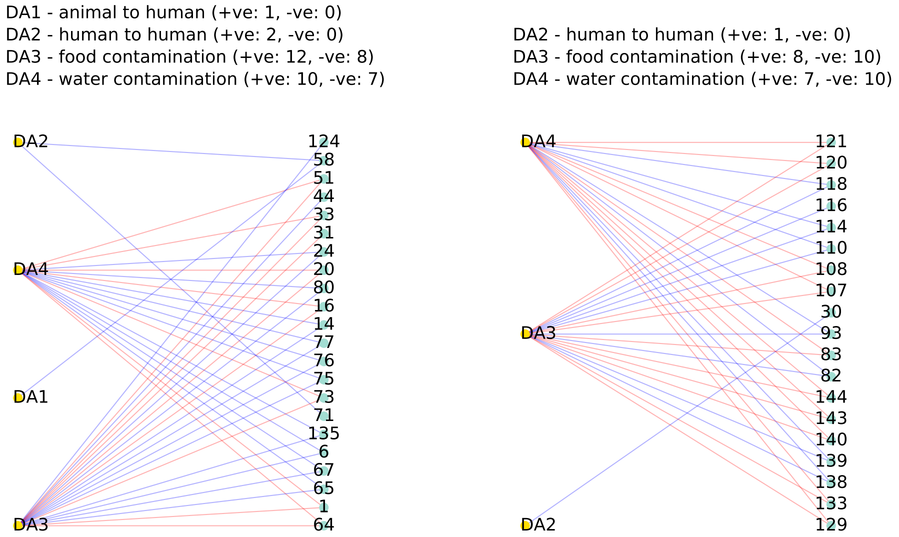

DA1 - animal to human (+ve: 1, -ve: 0)
DA2 - human to human (+ve: 2, -ve: 0)
DA3 - food contamination (+ve: 12, -ve: 8)
DA4 - water contamination (+ve: 10, -ve: 7)
DA2
DA4
DA1
DA3
124
58
51
44
33
31
24
20
80
16
14
77
76
75
73
71
135
6
67
65
1
64
DA2 - human to human (+ve: 1, -ve: 0)
DA3 - food contamination (+ve: 8, -ve: 10)
DA4 - water contamination (+ve: 7, -ve: 10)
DA4
DA3
DA2
121
120
118
116
114
110
108
107
30
93
83
82
144
143
140
139
138
133
129

Figure 14: Spearman correlation between different values of 'Transmission methods' attribute and various socioeconomic factors.

maternal mortality ratio (6, 116), mortality rate (17, 24, 75, 77, 93), death rate (50), population (14, 62, 65, 67, 76, 80, 110, 114, 145), and age-dependency ratio (138, 139). The negatively correlated factors are about employment in service

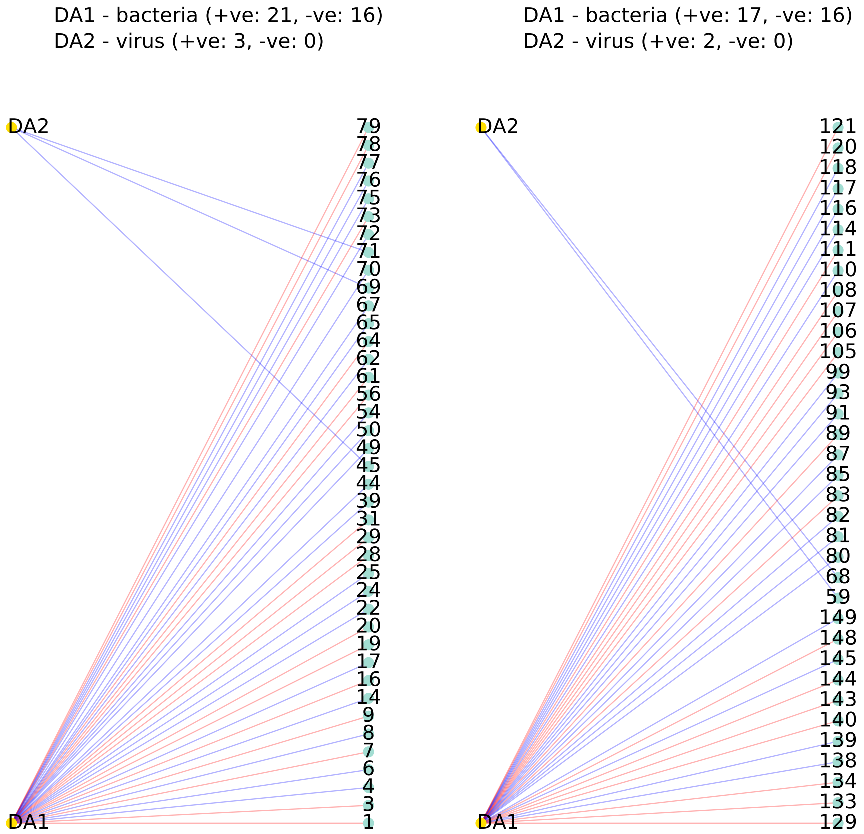


Figure 15: Point-biserial correlation between different values of 'Infectious Agents' attribute and various socioeconomic factors.

(3, 79, 105), employment in industry (9, 148), the percentage of wage and salaried workers (78, 106, 111), mobile cellular subscription (7), electricity access (16, 56, 129), primary education (19, 29), life expectancy (1, 73, 83), survival to age sixty-five (121, 140), population (28, 31, 61, 64, 89, 107, 108, 120, 133, 134, 144), and immunization (20, 143), .

4.4 Symptoms

Figure 17 and 18 depicts the correlation of different values of the attribute 'Symptoms' with the socioeconomic factors. We can see in Figure 17 that only 20 values out of 74 show a significant correlation with absolute values greater than or equal to 0.2.

Apparently, each of 57 correlated socioeconomic factors either shows positive correlations or negative correlations only and not both. Positively correlated socioeconomic attributes are about urban population (71), age-dependency ratio (138, 139), self-employment (87, 99, 117), agricultural employment (149), vulnerable employment (4, 70, 85), fisheries production (35), renewable energy consumption(39), fixed telephone subscriptions (45), mobile cellular subscription (7), NO2 and methane emissions (43, 125), adolescent and adult fertility (44, 82), birth rate(118), prevalence of anemia (22, 49, 81, 91), risk of maternal death and maternal mortality ratio (6, 116), mortality rate (17, 24, 75, 77, 93), death rate (50), and population (14, 65, 67, 76, 80, 110, 114). The negatively correlated attributes are related to the percentage of waged and salaried workers (78, 106, 111), electricity access (16, 56, 129), life expectancy (1, 73, 83), survival to age sixty-five (121, 140), population (64, 89, 107, 108, 120, 133), and immunization (20, 143).

4.5 Carriers

Figure 19 and 20 illustrates the correlation between different values of outbreak attribute 'Carriers' and various socioeconomic factors.

In Figure 19, we can see that only the value 'Camel' shows a significant correlation with an absolute value greater than or equal to 0.2 with two socioeconomic factors. The other 24 values show a correlation less than that or do not show any significant correlation at all. Here, the outbreaks, where 'Camel' is a carrier, are positively correlated with a male population (150) and are negatively correlated with the female population (159).

4.6 Population

We find that various socioeconomic factors related to the population show both positive and negative correlations. Hence, we investigate population-related socioeconomic factors' relationships

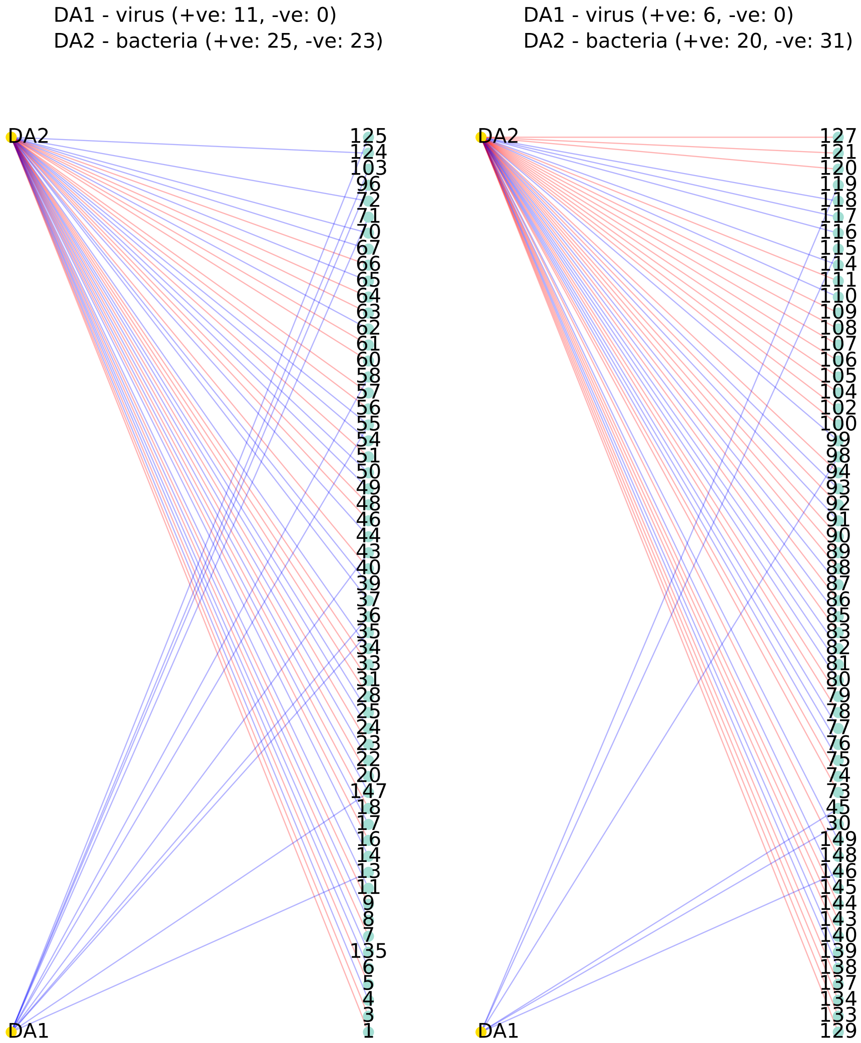


Figure 16: Spearman correlation between different values of ‘Infectious Agents’ attribute and various socioeconomic factors.

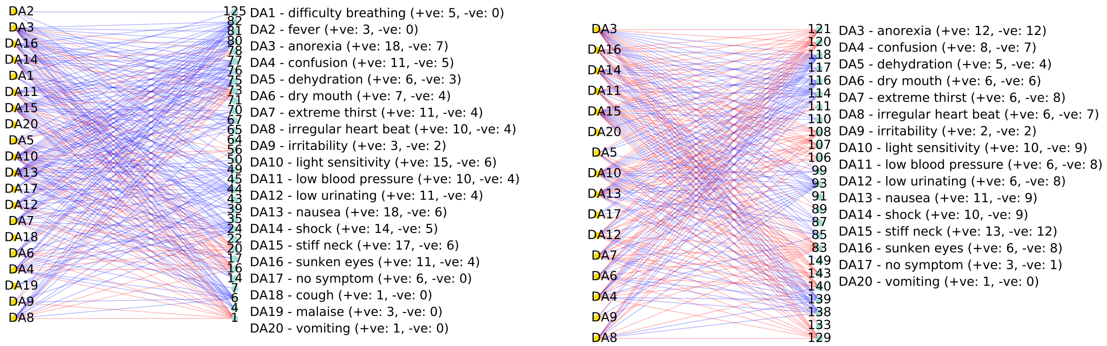

DA2
DA3
DA16
DA14
DA1
DA11
DA15
DA20
DA5
DA10
DA13
DA17
DA12
DA7
DA18
DA6
DA4
DA19
DA9
DA8
125
82
81
80
78
77
76
75
73
71
70
67
65
64
56
50
49
45
44
43
39
35
24
22
20
17
16
14
7
6
4
1
DA1 - difficulty breathing (+ve: 5, -ve: 0)
DA2 - fever (+ve: 3, -ve: 0)
DA3 - anorexia (+ve: 18, -ve: 7)
DA4 - confusion (+ve: 11, -ve: 5)
DA5 - dehydration (+ve: 6, -ve: 3)
DA6 - dry mouth (+ve: 7, -ve: 4)
DA7 - extreme thirst (+ve: 11, -ve: 4)
DA8 - irregular heart beat (+ve: 10, -ve: 4)
DA9 - irritability (+ve: 3, -ve: 2)
DA10 - light sensitivity (+ve: 15, -ve: 6)
DA11 - low blood pressure (+ve: 10, -ve: 4)
DA12 - low urinating (+ve: 11, -ve: 4)
DA13 - nausea (+ve: 18, -ve: 6)
DA14 - shock (+ve: 14, -ve: 5)
DA15 - stiff neck (+ve: 17, -ve: 6)
DA16 - sunken eyes (+ve: 11, -ve: 4)
DA17 - no symptom (+ve: 6, -ve: 0)
DA18 - cough (+ve: 1, -ve: 0)
DA19 - malaise (+ve: 3, -ve: 0)
DA20 - vomiting (+ve: 1, -ve: 0)
DA3
DA16
DA14
DA11
DA15
DA20
DA5
DA10
DA13
DA17
DA12
DA7
DA6
DA4
DA9
DA8
121
120
118
117
116
114
111
110
108
107
106
99
93
91
89
87
85
83
149
143
140
139
138
133
129
DA3 - anorexia (+ve: 12, -ve: 12)
DA4 - confusion (+ve: 8, -ve: 7)
DA5 - dehydration (+ve: 5, -ve: 4)
DA6 - dry mouth (+ve: 6, -ve: 6)
DA7 - extreme thirst (+ve: 6, -ve: 8)
DA8 - irregular heart beat (+ve: 6, -ve: 7)
DA9 - irritability (+ve: 2, -ve: 2)
DA10 - light sensitivity (+ve: 10, -ve: 9)
DA11 - low blood pressure (+ve: 6, -ve: 8)
DA12 - low urinating (+ve: 6, -ve: 8)
DA13 - nausea (+ve: 11, -ve: 9)
DA14 - shock (+ve: 10, -ve: 9)
DA15 - stiff neck (+ve: 13, -ve: 12)
DA16 - sunken eyes (+ve: 6, -ve: 8)
DA17 - no symptom (+ve: 3, -ve: 1)
DA20 - vomiting (+ve: 1, -ve: 0)

Figure 17: Point-biserial correlation between different values of ‘Symptoms’ attribute and various socioeconomic factors.

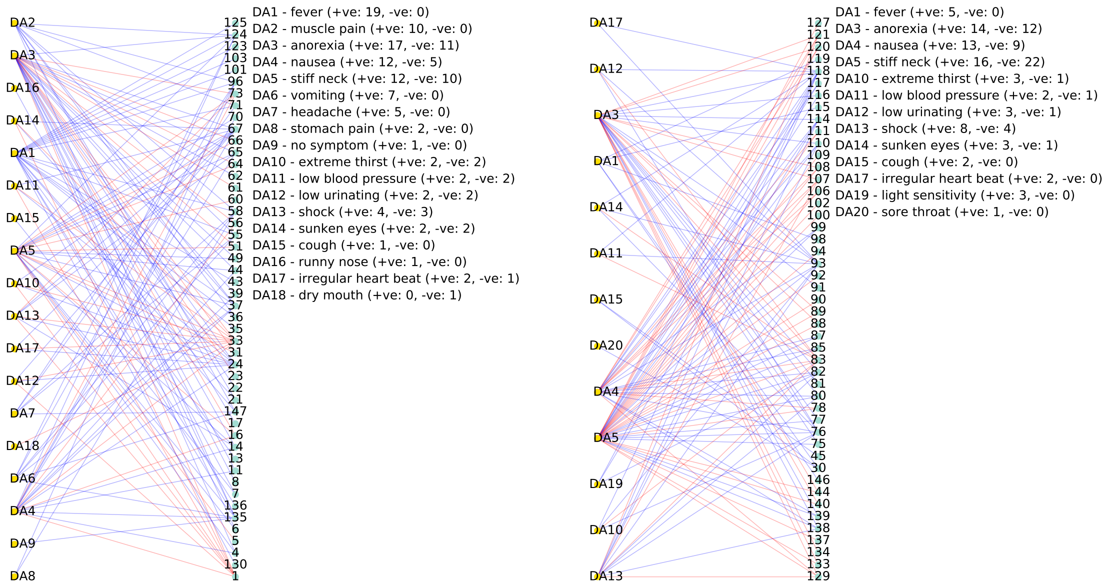


Figure 18: Spearman correlation between different values of 'Symptoms' attribute and various socioeconomic factors.

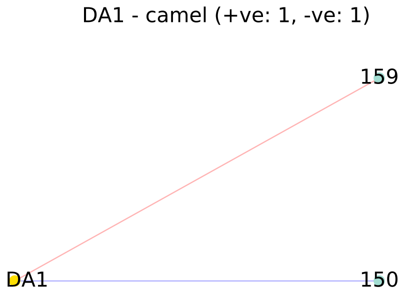


Figure 19: Point-biserial correlation between different values of 'Carriers' attribute and various socioeconomic factors.

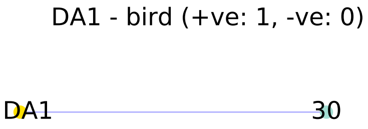


Figure 20: Spearman correlation between different values of 'Carriers' attribute and various socioeconomic factors.

with various outbreak attributes' values. Figure 21 depicts the correlations between different disease outbreak characteristics and socioeconomic factors related to population. Such socioeconomic indicators can be categorized into percentage and total population by age range and gender (123, 14, 136, 67, etc.), percentage and total population by gender (103, 159, 146, 150), total population (58, 34), rural population (96, 168, 122), urban population (71, 54, 32), and refugee population (153). The followings are the details.

- We observe that indicators related to the percentage-based population of the age ranges, 0-14, 0-4, 5-9, and 10-14 (14, 67, 76, 80, 114, 62, 145, 65, 110) show only significant positive correlations with different disease outbreak characteristics.
- Besides, indicators relating to the total population of the age ranges 15-64, 65 and above (119, 98, 115) also show significant positive correlations only, whereas percentage-based

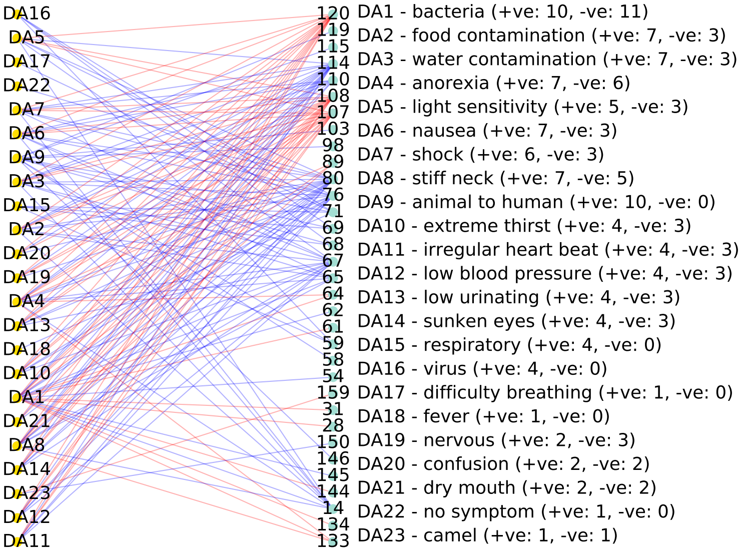

DA16
DA5
DA17
DA22
DA7
DA6
DA9
DA3
DA15
DA2
DA20
DA19
DA4
DA13
DA18
DA10
DA1
DA21
DA8
DA14
DA23
DA12
DA11
120
119
115
114
110
108
107
103
98
89
80
76
71
69
68
67
65
64
62
61
59
58
54
159
31
28
150
146
145
144
14
134
133
DA1 - bacteria (+ve: 10, -ve: 11)
DA2 - food contamination (+ve: 7, -ve: 3)
DA3 - water contamination (+ve: 7, -ve: 3)
DA4 - anorexia (+ve: 7, -ve: 6)
DA5 - light sensitivity (+ve: 5, -ve: 3)
DA6 - nausea (+ve: 7, -ve: 3)
DA7 - shock (+ve: 6, -ve: 3)
DA8 - stiff neck (+ve: 7, -ve: 5)
DA9 - animal to human (+ve: 10, -ve: 0)
DA10 - extreme thirst (+ve: 4, -ve: 3)
DA11 - irregular heart beat (+ve: 4, -ve: 3)
DA12 - low blood pressure (+ve: 4, -ve: 3)
DA13 - low urinating (+ve: 4, -ve: 3)
DA14 - sunken eyes (+ve: 4, -ve: 3)
DA15 - respiratory (+ve: 4, -ve: 0)
DA16 - virus (+ve: 4, -ve: 0)
DA17 - difficulty breathing (+ve: 1, -ve: 0)
DA18 - fever (+ve: 1, -ve: 0)
DA19 - nervous (+ve: 2, -ve: 3)
DA20 - confusion (+ve: 2, -ve: 2)
DA21 - dry mouth (+ve: 2, -ve: 2)
DA22 - no symptom (+ve: 1, -ve: 0)
DA23 - camel (+ve: 1, -ve: 1)

Figure 21: Correlation between different values of 'Population' attribute and various socioeconomic factors.

population by age ranges 15-64, 35-39, 40-44, 45-49, and 50-54 (120, 107, 108, 89, 28, 133, 31, 64, 144, 61, 134) show only negative correlations only.

- We also observe a significant positive correlation with factors relating to the total, male, and female population (103, 146, 58). However, the percentage-based female population (159) shows a negative correlation different from the positive correlation shown by the percentage-based male population (150). Interestingly, they are both related to disease outbreaks for which 'camel' is a carrier.
- Finally, we find factors related to the urban population (54, 71) are positively correlated with various attribute values of the disease outbreaks.

5 DISCUSSION

In this section, we discuss the trends we discover from our study and explain how we extend existing work done on the relationship between disease outbreaks and socioeconomic indicators. In parallel, we answer the research questions we set to explore earlier in this paper.

5.1 Effect of Socioeconomic Indicators on Disease Outbreaks

From our analysis, we see that various socioeconomic indicators are either showing positive correlations or negative correlations. In this regard, we observe the following trends after our analysis.

5.1.1 *Employment, Literacy, and Industrialization.*

- With the increase in employment related to agriculture, bacterial outbreaks increase in number. However, as employment in service and industry increases, bacterial disease outbreaks decrease in number. These support prior work [4], where researchers find that the risk for the bacterial disease named cholera in a district is negatively associated with high urbanization levels in the district. One prior study [18] also argues that contemporary processes of extended urbanization, which include suburbanization, post-suburbanization, and peri-urbanization, may result in increased vulnerability to infectious disease spread. Besides, another study identifies the most important factors such as the socioeconomic level, climate and environment, and urbanization level as the cause of the spread of a bacterial disease, meningococcal meningitis [98]. Thus, our findings contribute to the literature extending outcomes reported by these prior researches.
- We find that employment in agriculture and occurrences of outbreaks showing stiff neck as symptoms increase at the same time. This finding contributes to previous research [40, 73, 98] where researchers identify the most important factors such as the socioeconomic level and social behavior as the cause of the spread of meningitis, which shows stiff neck as a symptom.
- As cereal production and total labor force increase, disease outbreaks that spread from animals to humans (zoonotic diseases) also increase. This finding extends prior work [1, 26, 36] on the association of socioeconomic factors, for example, literacy, household income, social influence, knowledge gap, risk perceptions, etc., with zoonotic diseases such as rabies [26], swine flu [1], and bird flu [36].
- From prior work [6, 99], we see a distinct difference in exposure rates of hepatitis, a digestive system disease in populations belonging to high and lower-middle socioeconomic status. Our findings extend this as we find that, with an increase in the percentage of wage and salaried workers, disease outbreaks causing digestive and nervous system problems decrease.

- Outbreaks, which affect the organs of the digestive system (e.g., liver) or nervous system (e.g., spinal cord), increase with an increase of values of socioeconomic indicators such as percentage of vulnerable employment (male, female, and total) and percentage of self-employment (male, female, and total). We also find that food-borne, water-borne, and bacterial disease outbreaks fall in number with the increase in wage and salaried workers. Moreover, disease outbreaks showing symptoms such as anorexia, light sensitivity, nausea, shock, stiff neck, etc., fall in number, as the percentages of wage and salaried workers increase. These findings extend prior studies [42, 76] on finding out the effects of socioeconomic and environmental factors on the outbreak of Dengue fever (which shows as anorexia, light sensitivity, nausea, and shock as symptoms). They identified six related factors representing urbanization, poverty, accessibility, and vegetation associated with transmissibility.
- Health literacy is a must to respond correctly in the time of any health crisis, pandemic, epidemic, or in a short, disease outbreak. Individuals having proper health literacy have the ability to find, understand, and use information and services to inform health-related decisions and actions for themselves and others. Thus, literacy rate accelerates health literacy [39, 58]. We find that if the literacy rate increases, the attributes related to disease spread decrease and this phenomenon extends the studies [39, 58] related to pandemics and epidemics. In recent times, lessons learned from COVID-19 such as quick hospitalization of elderly patients, using masks, following lockdown and government guidelines, taking vaccination, etc. also provide the same insights [69, 109].
- As mobile cellular subscription per 100 people increases, bacterial disease outbreaks decrease. However, at the same time, outbreaks show "cough" as symptom increases. In this regard, from a previous study [91], we get the conclusion that the usage of cellular phones is accountable for the development of diseases such as brain tumors, male infertility, and hearing function. In another study [102], Internet use was found to have a direct positive relation to subjective health. In this era of the Internet, bulk use of electronic devices drove us to search the impacts of cellular subscriptions on disease outbreaks. Thus, our findings extend the aforementioned prior studies.
- From prior work [50, 52, 101], researchers find that due to the greenhouse effect and the subsequently increased global temperature, the prevalence of parasitic diseases such as dengue and yellow fever would exacerbate. They also suggest that global warming will cause changes in the epidemiology of infectious diseases and vector-borne diseases such as malaria, dengue, plague, and viruses will become more common. In our study, we also find that with CO_2 and methane emission, the number of fixed telephone subscriptions and fisheries production (captured and total) increase. Besides, in the same case, disease outbreaks that affect parts of the respiratory organ systems (e.g., lungs) or spread from animals to humans, or show symptoms such as malaise and difficulty breathing increase. Thus, our findings contribute to prior work.

5.1.2 *Life Expectancy and Mortality.*

- Life expectancy of population decreases as outbreaks cause digestive and nervous system problems to increase. Besides, food-borne, water-borne, and bacterial disease outbreaks also decrease life expectancy which resembles with previous studies where researchers found that water-borne [53] and bacterial diseases [19, 86, 93, 97] increase mortality rate. Accordingly, we also discover disease outbreaks show symptoms such as anorexia, confusion, dehydration, dry mouth, extreme thirst, irregular heartbeats, irritability, light sensitivity, low blood pressure, low urinating, nausea, shock, stiff neck, or sunken eyes, further decrease life expectancy. Researchers find the higher transmission of these symptoms showing diseases such as dengue, Zika, and chikungunya with poverty, population density and no access to improved water sources from previous work [48, 64]. Our findings extend these researches by discovering positive associations with life expectancy because of numerous symptoms of disease outbreaks.
- Several studies [32, 81, 85] show the most frequent causes of maternal mortality were preeclampsia, thromboembolism, sepsis, obstetric hemorrhage, and cardiovascular disorders. Our study finds that outbreaks showing symptoms such as dehydration, dry mouth, extreme thirst, irregular heartbeats, low blood pressure, low urinating, or sunken eyes increase the lifetime risk of maternal death, maternal mortality ratio, the prevalence of anemia (among children, pregnant women, and women of reproductive age), the mortality rate (under-5, infant, and neonatal). Thus, we extend prior work. Furthermore, we discover that as an adolescent and adult fertility rates and birth rates increase, outbreaks showing vomiting as a symptom increase, whereas outbreaks show irritability or malaise as symptoms increase with an increased lifetime risk of maternal death and maternal mortality ratio. These contribute to prior work [17, 53, 57, 59, 65, 75, 96] on relationship between mortality of population and disease outbreak attributes.
- As adolescent and adult fertility rate, birth rate, the lifetime risk of maternal death, maternal mortality ratio, renewable energy consumption, the mortality rate (under-5, infant, and neonatal), the prevalence of anemia (among children, pregnant women, and women of reproductive age), and the age-dependency ratio of the young and total percentage of working-age population increase, outbreaks, which affect the digestive (e.g. liver), or nervous (e.g. spinal cord) organ systems increase. These findings support prior work [5, 92] where researchers show that anemia is a common factor of hepatitis (a liver disease) and dementia (a nervous system disease). Besides, outbreaks that transmit via food or water contamination, or bacterial in nature, or show symptoms such as anorexia, confusion, light sensitivity, nausea, shock, or stiff neck also increase with the increase in renewable energy consumption, the mortality rate (under-5, infant, and neonatal), the prevalence of anemia (among children, pregnant women, and women of reproductive age), and the

age-dependency ratio of the young and total percentage of the working-age population.

- In prior studies [56, 61, 84], researchers find a substantially higher mortality rate with the increase in cholera, bacterial pneumonia, and bacterial meningitis. Moreover, there is an independent incremental association between delays in administrating antibiotics and mortality from adult acute bacterial meningitis [74]. Our findings contribute to these studies as we discover that the death rate per 1000 people increases, as outbreaks showing symptoms such as anorexia, or nausea increase, and bacterial outbreaks increase in number.

5.1.3 Population. We observed some interesting relationships between populations of different age groups, urban populations, gender, and various disease outbreak attributes. We present the relationships below.

- Geographical analysis and tracking of the spread of epidemics and other diseases present an important issue, which is of great concern to healthcare professionals all over the world. In this regard, we try to depict correlations between attributes of different disease outbreaks and population-related attributes such as urban population, rural population, refugee population, etc. Many prior studies also tried to explore such relationships [24, 49, 67]. We found a positive correlation between the factors related to the urban population and various disease outbreaks attributes. For example, viral outbreaks, outbreaks showing symptoms such as difficulty breathing, or fever increase with the total urban population. These findings contribute to the literature having existing studies [35, 57], where it was found that respiratory diseases with difficulty breathing such as influenza transmissibility and mortality rate increase with the increase in population density. In another study [25], researchers investigate the role of demographic patterns, urbanization, and comorbidities on the possible trajectories of COVID-19 in Ghana, Kenya, and Senegal. They found that compared with Europe, Africa's younger and rural population may limit the severity of the epidemic. A similar type of result came out for Bangladesh as well [77]. In our study, we also tried to find out the impact of rural and urban populations on different disease outbreak attributes.
- Besides, in our study, we observe different effects of age groups of population on disease outbreaks as follows.
  - As the percentage population over the age groups 0-14 increases, bacterial, food-borne, and water-borne disease outbreaks increase. Prior research studies also report findings similar to this. For example, existing studies [12, 53] find that, as population density and percentages of the population with absolute poverty increase, water-borne diseases such as cholera incidence and mortality rate increase. Besides, outbreaks showing the symptoms of anorexia, light sensitivity, nausea, shock, stiff neck, extreme thirst, irregular heartbeats, low blood pressure, low urinating, sunken eyes, confusion, or dry mouth also increase similarly. At the same time, outbreaks affecting nervous systems rise in number. Besides, another study shows that infectious disease mortality is relatively high in age group 5–9, reaches a minimum in adolescence (age group 10–19), then rises with age, with the growth rate gradually slowing down from approximately age 75 [54].
  - For the population age group of 15-64, we observe a mixed correlation. We see that, as the male, female, and total population of the age group 15-64 increase in absolute number, disease outbreaks that spread from animals to humans increase. On the other hand, as the population percentage of the 15-64 age group increases, the occurrences of outbreaks that have an effect on the nervous system, are bacterial in nature or spread via food or water, or show symptoms such as anorexia, extreme thirst, irregular heartbeats, light sensitivity, low blood pressure, low urinating, nausea, shock, stiff neck, sunken eyes, confusion, dry mouth, etc., decrease.
  - Bacterial disease outbreaks decline, as the population of the age group 35-54 increases. Besides, a reduction in disease outbreaks showing anorexia, or stiff neck as symptoms is noticed, when the female population of the age group 35-49 increases. Viral disease outbreaks increase, as the population of 65 and above increases. The same goes for disease outbreaks spreading from animals to humans, or affecting respiratory organ systems. As the total population increase, outbreaks that transmit from animals to humans also increase. Besides, outbreaks, where camels play the role of a carrier, increase as the male population increases, and decrease as the female population increases.

Characterizing disease outbreaks according to the different age groups is a common trait in the domain of medical research. Our determination of the correlation between different age groups with disease outbreaks supports some prior studies related to the recent pandemic COVID-19. Here, one research study tries to characterize symptom patterns amongst young children [94]. Another study supports our work by classifying susceptibility rates according to age disparity and different age groups [21]. Besides, both age and gender present variables that influence the clinical outcomes of COVID-19. Individuals in the 0–40 age range and females under 60 are significantly less likely to develop a severe condition and die, whereas males equal to or above 60 are more likely at risk of severe disease and death. This is reflected in ICU admissions as already reported in the literature [13]. Another prior study reveals that for almost all infections over school-age children have the least severe disease, and severity starts to rise long before old age [34]. Thus, our findings contribute to these previous researches.

5.1.4 Immunization.

- In a prior study [70] where researchers used regression models to evaluate the achievements of China's immunization program between 1950 and 2018 show that most of the 11 vaccine-preventable diseases exhibited dramatic declines in morbidity after their integration into the Expanded Program on Immunization (EPI), while varicella and paratyphoid fever, which were not integrated into the EPI, showed increased morbidity. Researchers [23, 46, 47] further find, that older adults and children receiving the influenza vaccine may have a lower risk of influenza. Another previous research [71] assessed receipt of flu immunization (2014–2019) by sickle cell disease (SCD) status among all Michigan children <18 years of age using the statewide immunization registry, logistic regression model. The researchers estimated that children with SCD had higher annual flu immunization rates than those without SCD, but >50% remain unimmunized. Vaccination reduced the overall attack rate to 4.6% from 9.0% without vaccination, over 300 days in United States [62]. In general, the immunization process stimulates the prevention of cancer, reduces the secondary infection of any disease, prevents antibiotic resistance, generates herd immunity, and provides cost-effective preparedness for outbreaks [80]. Our findings extend these prior researches as we discover that as more and more people get immunized (DPT and measles), and receive access to electricity, disease outbreaks that cause digestive and nervous system problems decrease.

- We find that food-borne, water-borne, and bacterial disease outbreaks also fall in number as more and more people get immunized. Accordingly, the number of disease outbreaks showing symptoms such as anorexia, confusion, dehydration, dry mouth, extreme thirst, irregular heartbeats, light sensitivity, low blood pressure, low urinating, nausea, shock, stiff neck, sunken eyes decreases, with the increase in immunization and electricity access. Moreover, lower uptake of vaccines was significantly associated with poorer educational attainment, lower levels of employment, and lower household income in prior studies [41, 43]. The association between socioeconomic determinants and vaccine hesitancy/refusal was also investigated in a prior work [11] where researchers find rising levels of perceived economic hardship were associated with vaccine hesitancy and lower parental education was significantly associated with vaccine refusal. Our findings contribute to these prior works on immunization and socioeconomic indicators. These findings can serve as warnings, and further explanations of socioeconomic inequities are needed in universal healthcare systems.

5.2 Challenges

We faced significant challenges during the data accumulation, analysis, and result summarization phase. First, as our study is heavily data-dependent, we needed to find credible sources of three types of data – regarding the occurrences of disease outbreaks, attributes of the infectious diseases in consideration, and various socioeconomic factors. After we chose WHO: Disease Outbreak News [104] as the source for our disease outbreak data, we scraped the website and created a raw dataset. Our first challenge was to clean this raw dataset and make it consistent so that automated tools can use it. We found that sometimes Disease Outbreak News labels the same disease by a different name. For example, *meningococcal meningitis* and *meningococcal disease* are referring to the same disease caused by *Neisseria meningitidis* bacteria, but sometimes it's referred to as meningitis, and other times meningococcal meningitis. Besides, sometimes outbreaks were marked with a generic names at the beginning of the outbreak, and later more reports on the same outbreak would indicate the name of the disease. Moreover, Disease Outbreak News often publishes reports that indicate multi-country outbreaks, and we could not scrap efficiently with any automated tools. Hence, We went through all of the articles published by Disease Outbreak News, fixed the aforementioned issues, and made the dataset consistent.

After accumulating and fixing the disease outbreak data, we needed to glean the attributes of the diseases appearing in the disease outbreak dataset. We found no centralized source of information for the attributes we considered. Therefore, we accumulated them from fact sheets provided by WHO and CDC, different articles published on the disease, and various informational websites such as Malacards, Mayoclinic, etc. [30, 68] As this information was manually collected, we needed to go through each of them and make various attributes consistent throughout the dataset. We obtained the data on different socioeconomic factors from the World Bank database [107]. After exploring the data, we found that country names are often inconsistent between this dataset and the outbreak dataset, and therefore, we made them consistent. We also notice that a significant number of attributes in the obtained dataset have a lot of missing values. Hence, we researched how to deal with them and finally removed the attributes which have more than 70% missing values [79, 87, 100], and imputed the rest.

After data accumulation and cleaning were over, we faced more challenges regarding formulating our problem which would allow us to use simple correlation analyses. We found that to use such techniques, we have to relate each of the values of various disease attributes with each of the socioeconomic factors. Therefore, we modified our null hypothesis accordingly. After calculating the correlation in an aforementioned way, we ended up with huge correlation matrices, and we faced our final challenge during the summarization and presentation of these matrices. We divided the obtained correlations into three separate groups and created bar plots to visualize the distribution of each group as shown in Figure 10. We find that a significant amount of the correlations are less than 0.2, and hence, we decided to do an exhaustive analysis of the correlations equal to and above 0.2. We chose to illustrate the aforementioned correlations in the graphical manner as discussed in the relevant section.

Moreover, please note that, due to data scarcity, our attributes are aggregated monthly, quarterly, or annually and we have taken them on a country basis. This coarse granularity of data aggregation might have introduced losing a lot of valuable in-depth insights if they existed. Therefore, it also resulted in revealing weak correlations. If we could accumulate very fine-grained data (at city-level as well as daily or weekly data), we could expect to have a stronger correlation within a window of time, if not all the time. This is why our study is important in the sense that even the coarse-grained data that we can get still point to meaningful correlation at a certain level. Also, we acknowledge that we could have done multivariate linear regression. However, we faced challenges in choosing a combination of attributes to do multivariate analysis. We did not have a foundation for such a categorical data analysis at large, which needs another research effort in the future.

6 CONCLUSION

To make systems that can anticipate disease outbreaks more precisely, we need to consider all the factors contributing to them. In this study, we quantified the relationships between different diseases' characteristics and the socioeconomic factors that may play a role in creating their outbreaks. We accumulated the relevant data from various online sources such as WHO, CDC, World Bank, etc. After that, we combined them according to our problem definition and performed a correlation analysis to quantify the strength and nature of their relationships. Our analysis shows that, like socioeconomic factors regarding CO_2 and methane emission, fisheries production, urban population, cereal production, telephone and cellular subscription, etc. increase in value, outbreaks showing various disease attributes we considered, increase in number. We also find that increase in values of socioeconomic factors related to immunization, electricity access, education, employment in service and industry, etc. decreases the number of such disease outbreaks.

These correlations can help build better disease outbreak surveillance and forecasting systems by providing suitable socioeconomic indicators to augment disease outbreak data. Besides, our research can help guide more incisive studies on the apparent correlations we find between different disease outbreak characteristics and socioeconomic factors. Moreover, systems for predicting the effect of disease outbreaks on socioeconomic factors of a community can also be benefited from the information we derived from our obtained correlations.

ACKNOWLEDGMENTS

The work has been conducted at and supported by Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh.

REFERENCES

[1] Anju Damu Ade, Bhavani Yamasani, and Nagaraj Kondagunta. 2018. Awareness of H1N1 influenza (swine flu) among rural population of Chittoor district, Andhra Pradesh. *International Journal of Community Medicine and Public Health* 5, 12 (2018), 1.

[2] Talal Ahmad, Nabeel Abdur Rehman, Fahad Pervaiz, Shankar Kalyanaraman, Maaz Bin Safeer, Sunandan Chakraborty, Umar Saif, and Lakshminarayanan Subramanian. 2013. Characterizing dengue spread and severity using internet media sources. In *Proceedings of the 3rd ACM Symposium on Computing for Development*. ACM, 18.

[3] Guido Alfani and Tommy E. Murphy. 2017. Plague and Lethal Epidemics in the Pre-Industrial World. *The Journal of Economic History* 77, 1 (2017), 314–343. https://doi.org/10.1017/S0022050717000092

[4] Mohammad Ali, Sanjukta Sen Gupta, Nisha Arora, Pradeep Khasnobis, Srinivas Venkatesh, Dipika Sur, Gopinath B Nair, David A Sack, and Nirmal K Ganguly. 2017. Identification of burden hotspots and risk factors for cholera in India: An observational study. *PloS one* 12, 8 (2017), e0183100.

[5] Alexander Andreev, Burak Erdinc, Kiran Shivaraj, Julia Schmutz, Olga Levochkina, Dhrity Bhowmik, Fady Farag, Kelli M Money, Louis H Primavera, Vladimir Gotlieb, et al. 2020. The association between anemia of chronic inflammation and Alzheimer's disease and related dementias. *Journal of Alzheimer's Disease Reports* 4, 1 (2020), 379–391.

[6] VA Arankalle, MS Chadha, SD Chitambar, AM Walimbe, LP Chobe, and SS Gandhe. 2001. Changing epidemiology of hepatitis A and hepatitis E in urban and rural India (1982-98). *Journal of viral hepatitis* 8, 4 (2001), 293–303.

[7] Zaheer Babar, Abdul Mannan, Faisal Kamiran, and Asim Karim. 2015. Understanding the Impact of Socio-Economic and Environmental Factors for Disease Outbreak in Developing Countries. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 124–131.

[8] Jorge Bacallao, Maria Schneider, Patricia Najera, Sylvain Aldighieri, Aida Soto, Wilmer Marquiño, Carlos Sáenz, Eduardo Jiménez, Gilberto Moreno, Octavio Chávez, et al. 2014. Socioeconomic factors and vulnerability to outbreaks of leptospirosis in Nicaragua. *International journal of environmental research and public health* 11, 8 (2014), 8301–8318.

[9] M. Bashir, B. MA, and L. Shahzad. 2020. A brief review of socio-economic and environmental impact of Covid-19. *Air Quality, Atmosphere and Health volume* 13 (08 2020), 1403–1409. https://doi.org/10.1007/s11869-020-00894-8

[10] Dan Becker. 2021. Handling Missing Values. https://www.kaggle.com/dansbecker/handling-missing-values. Accessed on October 15, 2021.

[11] Chiara Bertoncello, Antonio Ferro, Marco Fonzo, Sofia Zanovello, Giuseppina Napoletano, Francesca Russo, Vincenzo Baldo, and Silvia Cocchio. 2020. Socioeconomic determinants in vaccine hesitancy and vaccine refusal in Italy. *Vaccines* 8, 2 (2020), 276.

[12] Godfrey Bwire, Aline Munier, Issaka Ouedraogo, Leonard Heyerdahl, Henry Komakech, Atek Kagirita, Richard Wood, Raymond Mhlanga, Berthe Njanpop-Lafourcade, Mugagga Malimbo, et al. 2017. Epidemiology of cholera outbreaks and socio-economic characteristics of the communities in the fishing villages of Uganda: 2011-2015. *PLoS neglected tropical diseases* 11, 3 (2017), e0005407.

[13] Carlo Vittorio Cannistraci, Maria Grazia Valsecchi, and Ilaria Capua. 2021. Age-sex population adjusted analysis of disease severity in epidemics as a tool to devise public health policies for COVID-19. *Scientific reports* 11, 1 (2021), 1–8.

[14] Herman Anthony Carneiro and Eleftherios Mylonakis. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases* 49, 10 (2009), 1557–1564.

[15] CDC. 2021. CDC Fact Sheets. https://www.cdc.gov/chronicdisease/resources/publications/aag.htm. Accessed on October 15, 2021.

[16] Promit Barua Chowdhury, Sorif Hossain, and Raaj Kishore Biswas. 2020. A combination of COVID-19 and dengue fever in Bangladesh: Preparedness of Bangladesh. *Journal of global health* 10(2) (2020). https://doi.org/10.7189/jogh.10.020314

[17] Gerardo Chowell and Cécile Viboud. 2016. Pandemic influenza and socioeconomic disparities: Lessons from 1918 Chicago. *Proceedings of the National Academy of Sciences* 113, 48 (2016), 13557–13559.

[18] Creighton Connolly, Roger Keil, and S. Harris Ali. 2021. Extended urbanisation and the spatialities of infectious disease: Demographic change, infrastructure and governance. *Urban Studies* 58, 2 (2021), 245–263. https://doi.org/10.1177/0042098020910873 arXiv:https://doi.org/10.1177/0042098020910873

[19] Robert J Corner, Ashraf M Dewan, and Masahiro Hashizume. 2013. Modelling typhoid risk in Dhaka Metropolitan Area of Bangladesh: the role of socio-economic and environmental factors. *International journal of health geographics* 12, 1 (2013), 1–15.

[20] Cassandra Crosby. 2021. Human Body Organ Systems. https://www.hillandponton.com/human-body-organ-systems/. Accessed on October 15, 2021.

[21] Nicholas G Davies, Petra Klepac, Yang Liu, Kiesha Prem, Mark Jit, and Rosalind M Eggo. 2020. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nature medicine* 26, 8 (2020), 1205–1211.

[22] BBK Davies-Teye, L Vanotoo, JB Yabani, and C Kwakye-Maclean. 2014. Socio-Economic Factors Associated With Cholera Outbreak In Southern Ghana, 2012: A Case-Control Study. *Value in Health* 17, 3 (2014), A128.

[23] Vittorio Demicheli, Tom Jefferson, Carlo Di Pietrantonj, Eliana Ferroni, Sarah Thorning, Roger E Thomas, and Alessandro Rivetti. 2018. Vaccines for preventing influenza in the elderly. *Cochrane Database of Systematic Reviews* 2 (2018).

[24] A.N. Desai, J.W. Ramatowski, and N. Marano. 2020. Infectious disease outbreaks among forcibly displaced persons: an analysis of ProMED reports 1996–2016, Attitudes, and Practices. *Conflict and Health* 14, 49 (2020). https://doi.org/10.1186/s13031-020-00295-9

[25] Binta Zahra Diop, Marieme Ngom, Clémence Pougué Biyong, and John N Pougué Biyong. 2020. The relatively young and rural population may limit the spread and severity of COVID-19 in Africa: a modelling study. *BMJ Global Health* 5, 5 (2020). https://doi.org/10.1136/bmjgh-2020-002699 arXiv:https://gh.bmj.com/content/5/5/e002699.full.pdf

[26] Betty Dodet, Amlan Goswami, Amila Gunasekera, Ferdinand de Guzman, Seemin Jamali, Cecilia Montalban, Wilfried Purba, Beatriz Quiambao, Naseem Salahuddin, Gadey Sampath, et al. 2008. Rabies awareness in eight Asian countries. *Vaccine* 26, 50 (2008), 6344–6348.

[27] Yadolah Dodge. 2008. *Spearman Rank Correlation Coefficient*. Springer New York, New York, NY, 502–505. https://doi.org/10.1007/978-0-387-32833-1_379

[28] Pearu Peterson Eric Jones, Travis Oliphant et al. 2001. SciPy: Open Source Scientific Tools for Python. (2001).

[29] Ronald Aylmer Fisher. 1992. Statistical methods for research workers. In *Breakthroughs in statistics*. Springer, 66–70.

[30] Mayo Foundation for Medical Education and Research (MFMER). 2021. Mayo Clinic. https://www.mayoclinic.org/. Accessed on October 15, 2021.

[31] Song Gao, Jinmeng Rao, Yuhao Kang, Yunlei Liang, Jake Kruse, Doerte Doepfer, Ajay K. Sethi, Juan Francisco Mandujano Reyes, Jonathan Patz, and Brian S. Yandell. 2020. Mobile phone location data reveal the effect and geographic variation of social distancing on the spread of the COVID-19 epidemic. *CoRR* abs/2004.11430 (2020). arXiv:2004.11430 https://arxiv.org/abs/2004.11430

[32] Labib Ghulmiyyah and Baha Sibai. 2012. Maternal mortality from preeclampsia/eclampsia. In *Seminars in perinatology*, Vol. 36. Elsevier, 56–59.

[33] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.

[34] J.R. Glynn and P.A.H Moss. 2020. Systematic analysis of infectious disease outcomes by age shows lowest severity in school-age children. *Scientific Data*

07 (2020), 329. https://doi.org/10.1038/s41597-020-00668-y

[35] Kyra H Grantz, Madhura S Rane, Henrik Salje, Gregory E Glass, Stephen E Schachterle, and Derek AT Cummings. 2016. Disparities in influenza mortality and transmission related to sociodemographic factors within Chicago in the pandemic of 1918. *Proceedings of the National Academy of Sciences* 113, 48 (2016), 13839–13844.

[36] Annick Guénel and Sylvia Klingberg. 2016. Press coverage of bird flu epidemic in Vietnam. In *Liberalizing, feminizing and popularizing health communications in Asia*. Routledge, 91–106.

[37] S Das Gupta. 1960. Point biserial correlation coefficient and its generalization. *Psychometrika* 25, 4 (1960), 393–408.

[38] Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. 2020. Array programming with NumPy. *Nature* 585, 7825 (2020), 357–362.

[39] K. Ho and G. D. Smith. 2020. A discursive paper on the importance of health literacy among foreign domestic workers during outbreaks of communicable diseases. *Journal of clinical nursing* 29, 23-24 (2020), 4827–4833. https://doi.org/10.1111/jocn.15495

[40] Abraham Hodgson, Thomas Smith, Sebastien Gagneux, Martin Adjuik, Gerd Pluschke, Nathan Kumasenu Mensah, Fred Binka, and Blaise Genton. 2001. Risk factors for meningococcal meningitis in northern Ghana. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 95, 5 (2001), 477–480.

[41] J. Hoes, A. Boef, M. J. Knol, H. E. de Melker, L. Mollema, F. van der Klis, N. Y. Rots, and D. van Baarle. 2018. Socioeconomic Status Is Associated With Antibody Levels Against Vaccine Preventable Diseases in the Netherlands. *Frontiers in public health* 6 (07 2018), 209. https://doi.org/10.3389/fpubh.2018.00209

[42] Whitney M. Holeva-Eklund, Timothy K. Behrens, and Crystal M. Hepp. 2021. Systematic review: the impact of socioeconomic factors on Aedes aegypti mosquito distribution in the mainland United States. *Reviews on Environmental Health* 36, 1 (2021), 63–75. https://doi.org/doi:10.1515/reveh-2020-0028

[43] D. HUNGERFORD, P. MACPHERSON, S. FARMER, S. GHEBREHEWET, D. SEDDON, R. VIVANCOS, and A. KEENAN. 2016. Effect of socioeconomic deprivation on uptake of measles, mumps and rubella vaccination in Liverpool, UK over 16 years: a longitudinal ecological study. *Epidemiology and Infection* 144, 6 (2016), 1201–1211. https://doi.org/10.1017/S0950268815002599

[44] John D Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in science & engineering* 9, 3 (2007), 90–95.

[45] Nor Azura Husin, Naomie Salim, et al. 2008. Modeling of dengue outbreak prediction in Malaysia: a comparison of neural network and nonlinear regression model. In *2008 International Symposium on Information Technology*, Vol. 3. IEEE, 1–4.

[46] Michael L Jackson, Jessie R Chung, Lisa A Jackson, C Hallie Phillips, Joyce Benoit, Arnold S Monto, Emily T Martin, Edward A Belongia, Huong Q McLean, Manjusha Gaglani, et al. 2017. Influenza vaccine effectiveness in the United States during the 2015–2016 season. *New England Journal of Medicine* 377, 6 (2017), 534–543.

[47] Tom Jefferson, Alessandro Rivetti, Anthony Harnden, Carlo Di Pietrantonj, and Vittorio Demicheli. 2008. Vaccines for preventing influenza in healthy children. *Cochrane Database of Systematic Reviews* 2 (2008).

[48] L. Kapiriri and A. Ross. 2020. he Politics of Disease Epidemics: a Comparative Analysis of the SARS, Zika, and Ebola Outbreaks. *Global Social Welfare* 7 (09 2020), 33–45. https://doi.org/doi.org/10.1007/s40609-018-0123-y

[49] Md Nuruzzaman Khan, M. Mofizul Islam, and Md Mashiur Rahman. 2020. Risks of COVID19 outbreaks in Rohingya refugee camps in Bangladesh. *Public Health in Practice* 1 (2020), 100018. https://doi.org/10.1016/j.puhip.2020.100018

[50] Atul A Khasnis and Mary D Nettleman. 2005. Global warming and infectious disease. *Archives of medical research* 36, 6 (2005), 689–696.

[51] Wilhelm Kirch (Ed.). 2008. *Pearson's Correlation Coefficient*. Springer Netherlands, Dordrecht, 1090–1091. https://doi.org/10.1007/978-1-4020-5614-7_2569

[52] Ichiro Kurane. 2010. The effect of global warming on infectious diseases. *Osong public health and research perspectives* 1, 1 (2010), 4–9.

[53] Gregor C Leckebusch and Auwal F Abdussalam. 2015. Climate and socioeconomic influences on interannual variability of cholera in Nigeria. *Health & place* 34 (2015), 107–117.

[54] Zhi Li, Peigang Wang, Ge Gao, Chunling Xu, and Xinguang Chen. 2016. Age-period-cohort analysis of infectious disease mortality in urban-rural China, 1990–2010. *International Journal for Equity in Health* 15, 1 (2016), 1–9.

[55] R. J. Littman. 2009. The plague of Athens: epidemiology and paleopathology. *Mt Sinai J Med* 76, 5 (Oct 2009), 456–467.

[56] Francisco J Luquero, Marc Rondy, Jacques Boncy, André Munger, Helmi Mekaoui, Ellen Rymshaw, Anne-Laure Page, Brahima Toure, Marie Amelie Degail, Sarala Nicolas, et al. 2016. Mortality rates during cholera epidemic, Haiti, 2010–2011. *Emerging infectious diseases* 22, 3 (2016), 410.

[57] Svenn-Erik Mamelund, Clare Shelley-Egan, and Ole Rogeberg. 2021. The association between socioeconomic status and pandemic influenza: Systematic review and meta-analysis. *PLOS ONE* 16, 9 (09 2021), 1–31. https://doi.org/10.1371/journal.pone.0244346

[58] U. Matterne, N. Egger, J. Tempes, C. Tischer, J. Lander, M. L. Dierks, E. M. Bitzer, and C. Apfelbacher. 2021. Health literacy in the general population in the context of epidemic or pandemic coronavirus outbreak situations: Rapid scoping review. *Patient education and counseling* 104, 2 (2021), 223–234. https://doi.org/10.1016/j.pec.2020.10.012

[59] José María Mayoral, Jordi Alonso, Olatz Garín, Zaida Herrador, Jenaro Astray, and Maretva et al. Baricot. 2013. Social factors related to the clinical severity of influenza cases in Spain during the A (H1N1) 2009 virus pandemic. *BMC public health* 13, 118 (02 2013). https://doi.org/10.1186/1471-2458-13-118

[60] Wes McKinney et al. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, Vol. 445. Austin, TX, 51–56.

[61] Naoya Miyashita and Yasuhiro Yamauchi. 2018. Bacterial pneumonia in elderly Japanese populations. *Japanese clinical medicine* 9 (2018), 1179670717751433.

[62] Seyed M Moghadas, Thomas N Vilches, Kevin Zhang, Chad R Wells, Affan Shoukat, Burton H Singer, Lauren Ancel Meyers, Kathleen M Neuzil, Joanne M Langley, Meagan C Fitzpatrick, and Alison P Galvani. 2021. The Impact of Vaccination on Coronavirus Disease 2019 (COVID-19) Outbreaks in the United States. *Clinical Infectious Diseases* 73, 12 (01 2021), 2257–2264. https://doi.org/10.1093/cid/ciab079 arXiv:https://academic.oup.com/cid/article-pdf/73/12/2257/41793489/ciab079.pdf

[63] David M Morens and Anthony S Fauci. 2007. The 1918 influenza pandemic: insights for the 21st century. *The Journal of infectious diseases* 195, 7 (2007), 1018–1028.

[64] Jasmine Morgan, Clare Strode, and J Enrique Salcedo-Sora. 2021. Climatic and socio-economic factors supporting the co-circulation of dengue, Zika and chikungunya in three different ecosystems in Colombia. *PLoS Neglected Tropical Diseases* 15, 3 (2021), e0009259.

[65] Christopher JL Murray, Alan D Lopez, Brian Chin, Dennis Feehan, and Kenneth H Hill. 2006. Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918–20 pandemic: a quantitative analysis. *The Lancet* 368, 9554 (2006), 2211–2218.

[66] Nature. 2021. Signs and symptoms. https://www.nature.com/subjects/signs-and-symptoms. Accessed on October 15, 2021.

[67] CJ Neiderud. 2015. How urbanization affects the epidemiology of emerging infectious diseases. *Infection Ecology and Epidemiology* 05, 27060 (2015). https://doi.org/iee.v5.27060

[68] Weizmann Institute of Science. 2021. Malacards. https://www.malacards.org/. Accessed on October 15, 2021.

[69] Leena Paakkari and Orkan Okan. 2020. COVID-19: health literacy is an underestimated problem. *The Lancet* 05, 05 (2020), 249–250. https://doi.org/10.1016/S2468-2667(20)30086-4

[70] Jinhua Pan, Yesheng Wang, Lingsheng Cao, Ying Wang, Qi Zhao, Shenglan Tang, Wenfeng Gong, Lei Guo, Zhixi Liu, Zexuan Wen, et al. 2021. Impact of immunization programs on 11 childhood vaccine-preventable diseases in China: 1950–2018. *The Innovation* 2, 2 (2021), 100113.

[71] Hannah K Peng, Kevin J Dombkowski, Gary L Freed, Susan E Creary, Dominic Smith, and Sarah L Reeves. 2021. Influenza immunization coverage of children with sickle cell disease. *Vaccine* 39, 39 (2021), 5538–5540.

[72] Fahad Pervaiz, Mansoor Pervaiz, Nabeel Abdur Rehman, and Umar Saif. 2012. FluBreaks: early epidemic detection from Google flu trends. *Journal of medical Internet research* 14, 5 (2012), e125.

[73] Line Pickering, Poul Jennum, Rikke Ibsen, and Jakob Kjellberg. 2018. Long-term health and socioeconomic consequences of childhood and adolescent onset of meningococcal meningitis. *European journal of pediatrics* 177, 9 (2018), 1309–1315.

[74] N Proulx, D Frechette, B Toye, J Chan, and S Kravcik. 2005. Delays in the administration of antibiotics are associated with mortality from adult acute bacterial meningitis. *Qjm* 98, 4 (2005), 291–298.

[75] J. Pujol, P. Godoy, N. Soldevila, J. Castilla, F. González-Candelas, J. M. Mayoral, J. Astray, S. Garcia, V. Martin, S. Tamames, M. Delgado, and A. Domínguez. 2015. Social class based on occupation is associated with hospitalization for A(H1N1)pdm09 infection. Comparison between hospitalized and ambulatory cases. *Epidemiology and infection* 144, 4 (08 2015), 732–740. https://doi.org/10.1017/S0950268815001892

[76] Xiaopeng Qi, Yong Wang, Yue Li, Yujie Meng, Qianqian Chen, Jiaqi Ma, and George F Gao. 2015. The effects of socioeconomic and environmental factors on the incidence of dengue fever in the Pearl River Delta, China, 2013. *PLoS neglected tropical diseases* 9, 10 (2015), e0004159.

[77] M.S. Rahman, A. Karamehic-Muratovic, M. Amrin, A.H. Chowdhury, M.S. Mondol, and U. Haque. 2021. COVID-19 Epidemic in Bangladesh among Rural and Urban Residents: An Online Cross-Sectional Survey of Knowledge, Attitudes, and Practices. *Epidemiologia* 2 (2021), 1–13. https://doi.org/10.3390/epidemiologia2010001

[78] David W. Redding, Peter M. Atkinson, Andrew A. Cunningham, Gianni Lo Iacono, Lina M. Moses, James L. N. Wood, and Kate E. Jones. 2019. Impacts of environmental and socio-economic factors on emergence and epidemic potential of Ebola in Africa. *Nature Communications* 10, 4531 (07 2019), 1776–1784. https://doi.org/10.1038/s41467-019-12499-6

[79] ResearchGate. 2021. What proportion of missing data is too big for multiple imputation in longitudinal data? https://www.researchgate.net/post/What_proportion_of_missing_data_is_too_big_for_multiple_imputation_in_longitudinal_data. Accessed on October 15, 2021.

[80] Charlene M. C. Rodrigues and Stanley A. Plotkin. 2020. Impact of Vaccines; Health, Economic and Social Perspectives. *Frontiers in Microbiology* 11 (2020). https://doi.org/10.3389/fmicb.2020.01526

[81] Khama O Rogo, John Oucho, and Philip Mwalali. 2006. Maternal mortality. *Disease and Mortality in Sub-Saharan Africa. 2nd edition* (2006).

[82] RsearchGate. 2021. What is the minimum value of correlation coefficient to prove the existence of the accepted relationship between scores of two of more tests? https://www.researchgate.net/post/What_is_the_minimum_value_of_correlation_coefficient_to_prove_the_existence_of_the_accepted_relationship_between_scores_of_two_of_more_tests. Accessed on October 15, 2021.

[83] RsearchGate. 2021. Which correlation coefficient is better to use: Spearman or Pearson? https://www.researchgate.net/post/Which_correlation_coefficient_is_better_to_use_Spearman_or_Pearson2. Accessed on October 15, 2021.

[84] Anne Schuchat, Katherine Robinson, Jay D Wenger, Lee H Harrison, Monica Farley, Arthur L Reingold, Lewis Lefkowitz, and Bradley A Perkins. 1997. Bacterial meningitis in the United States in 1995. *New England journal of medicine* 337, 14 (1997), 970–976.

[85] JM Schutte, EAP Steegers, NWE Schuitemaker, JG Santema, Karin de Boer, M Pel, G Vermeulen, Willy Visser, Jos van Roosmalen, and Netherlands Maternal Mortality Committee. 2010. Rise in maternal mortality in the Netherlands. *BJOG: An International Journal of Obstetrics & Gynaecology* 117, 4 (2010), 399–406.

[86] Balakrishnan Senthilkumar, Duraisamy Senbagam, and Moses Rajasekarapandian. 2014. An epidemiological surveillance of asymptomatic typhoid carriers associated in respect to socioeconomic status in India. *Journal of Public Health* 22, 3 (2014), 297–301.

[87] StatsStackExchange. 2021. How much missing data is too much? https://stats.stackexchange.com/questions/149140/how-much-missing-data-is-too-much-multiple-imputation-mice-r. Accessed on October 15, 2021.

[88] StatsStackExchange. 2021. Spearman's rho to correlate discrete with binary variables. https://stats.stackexchange.com/questions/82230/spearmans-rho-to-correlate-discrete-with-binary-variables. Accessed on October 15, 2021.

[89] Yihua Su, Aarthi Venkat, Yadush Yadav, Lisa B. Puglisi, and Samah J. Fodeh. 2021. Twitter-based analysis reveals differential COVID-19 concerns across areas with socioeconomic disparities. *Computers in Biology and Medicine* 132 (2021), 104336. https://doi.org/10.1016/j.compbiomed.2021.104336

[90] D. Sugawara, A. Masuyama, and T. Kubo. 2020. Socioeconomic Impacts of the COVID-19 Lockdown on the Mental Health and Life Satisfaction of the Japanese Population. (2020). https://doi.org/10.31234/osf.io/sndpm

[91] AK Suhag, RS Larik, GZ Mangi, M Khan, SK Abbasi, and H Madiha. 2016. Impact of excessive mobile phone usage on human. *J Comput Sci Syst Biol* 9, 6 (2016), 173–17710.

[92] Mark S Sulkowski. 2003. Anemia in the treatment of hepatitis C virus infection. *Clinical infectious diseases* 37, Supplement_4 (2003), S315–S322.

[93] Dipika Sur, Mohammad Ali, Lorenz Von Seidlein, Byomkesh Manna, Jacqueline L Deen, Camilo J Acosta, John D Clemens, and Sujit K Bhattacharya. 2007. Comparisons of predictors for typhoid and paratyphoid fever in Kolkata, India. *BMC public health* 7, 1 (2007), 1–10.

[94] Olivia V Swann, Karl A Holden, Lance Turtle, Louisa Pollock, Cameron J Fairfield, Thomas M Drake, Sohan Seth, Conor Egan, Hayley E Hardwick, Sophie Halpin, et al. 2020. Clinical characteristics of children and young people admitted to hospital with covid-19 in United Kingdom: prospective multicentre observational cohort study. *bmj* 370 (2020).

[95] Noor Diana Ahmad Tarmizi, Farha Jamaluddin, A Abu Bakar, Zulaiha Ali Othman, Suhaila Zainudin, and Abdul Razak Hamdan. 2013. Malaysia Dengue Outbreak Detection Using Data Mining Models. *Journal of Next Generation Information Technology (JNIT)* 4, 6 (2013), 96–107.

[96] D. L. Thompson, J. Jungk, E. Hancock, C. Smelser, M. Landen, M. Nichols, D. Selvage, J. Baumbach, and M. Sewell. 2011. Risk factors for 2009 pandemic influenza A (H1N1)-related hospitalization and death among racial/ethnic groups in New Mexico. *American journal of public health* 101, 9 (07 2011), 1776–1784. https://doi.org/10.2105/AJPH.2011.300223

[97] Lícia KAM Thörn, Ruth Minamisava, Simonne S Nouer, Luiza H Ribeiro, and Ana Lucia Andrade. 2011. Pneumonia and poverty: a prospective population-based study among children in Brazil. *BMC infectious diseases* 11, 1 (2011), 1–10.

[98] Emmanuel Tanko Umaru[1], Ahmed Nazri Muhamad Ludin[1] Mohammed Rafee, Majid[1] Soheil Sabri[1] Chingle Moses P, Wallace Enegbuma[1] Abdrazack Nelson Tajudeen A[1], and Malaysia[1] Bahru. 2013. Risk factors responsible for the spread of meningococcal meningitis: a review. (2013).

[99] Sunil R Vaidya, Bipin N Tilekar, Atul M Walimbe, and Vidya A Arankalle. 2003. Increased risk of hepatitis E in sewage workers from India. *Journal of occupational and environmental medicine* 45, 11 (2003), 1167–1170.

[100] Analytics Vidhya. 2021. What should be the allowed percentage of Missing Values? https://discuss.analyticsvidhya.com/t/what-should-be-the-allowed-percentage-of-missing-values/2456. Accessed on October 15, 2021.

[101] Yunling Wang. 2017. Introduction to parasitic disease. In *Radiology of Parasitic Diseases*. Springer, 3–3.

[102] Silje C Wangberg, Hege K Andreassen, Hans-Ulrich Prokosch, Silvina Maria Vagos Santana, Tove Sørensen, and Catharine E Chronaki. 2008. Relations between Internet use, socio-economic status (SES), social support and subjective health. *Health promotion international* 23, 1 (2008), 70–77.

[103] Ari Whiteman, Jose R. Loaiza, Donald A. Yee, Karen C. Poh, Alexandria S. Watkins, Keira J. Lucas, Tyler J. Rapp, Lillie Kline, Ayman Ahmed, Shi Chen, Eric Delmelle, and Judith Uche Oguzie. 2020. Do socioeconomic factors drive Aedes mosquito vectors and their arboviral diseases? A systematic review of dengue, chikungunya, yellow fever, and Zika Virus. *One Health* 11 (2020), 100188. https://doi.org/10.1016/j.onehlt.2020.100188

[104] WHO. 2021. Disease Outbreak News. https://www.who.int/csr/don/en/. Accessed on October 15, 2021.

[105] WHO. 2021. WHO Fact Sheets. https://www.who.int/news-room/fact-sheets. Accessed on October 15, 2021.

[106] Weng-Keen Wong, Andrew Moore, Gregory Cooper, and Michael Wagner. 2003. WSARE: what's strange about recent events? *Journal of Urban Health* 80, 1 (2003), i66–i75.

[107] WorldBank. 2021. World Bank Database. https://data.worldbank.org/. Accessed on October 15, 2021.

[108] WorldBankIndicators. 2021. Indicators. https://data.worldbank.org/indicator?tab=all. Accessed on October 15, 2021.

[109] Fedayi Yağar. 2021. Fear of COVID-19 and Its Association With Health Literacy in Elderly Patients. *Journal of Patient Experience* 8 (2021), 23743735211056506. https://doi.org/10.1177/23743735211056506 PMID: 34926799.

[110] Yuhanis Yusof and Zuriani Mustaffa. 2011. Dengue outbreak prediction: A least squares support vector machines approach. *International Journal of Computer Theory and Engineering* 3, 4 (2011), 489.