Developing an Early-Warning System for Cholera in African Countries using Socio-Economic Indicators and Statistical Models

Howard Lei*
California State University, East Bay, Hayward, California, USA

Farnaz Ganjeizadeh
California State University, East Bay, Hayward, California, USA

Shervin Baharmand
California State University, East Bay, Hayward, California, USA

Andishe Almasi
California State University, East Bay, Hayward, California, USA

The ability to produce an early warning system for disease outbreaks is crucial to knowing its medication demand, with implications on supply chain management for meeting the production needs of the medication. One way to produce medication demand forecasts for a disease using a low-cost and minimally-invasive framework is to leverage historical data for the disease, along with indicators correlated with the rise and fall of the disease, in a statistical framework. In this work, we apply various statistical classifiers used in machine learning, along with 30 socio-economic indicators, to predict Cholera infections across 10 African countries up to two years into the future. Each classifier represents a numerical method for making predictions of outcomes given a set of input factors. The classifiers include Logistic Regression, Perceptron, Support Vector Machine, Decision Tree, and Linear Regression. The classifiers are able to identify the majority of the countries and years where there is an increased risk of Cholera infection, while reducing the false identification of countries and year where there was no increased risk of infection. The latter allows for cost-savings for reducing medication production when there is little demand. The classifiers also allow us to better understand the importance of each indicator in the predictions. Results show that only a couple of indicators are needed to make accurate predictions. The Improved water source indicator is included amongst the important indicators, which supports the fact that the bacterium for Cholera is transmitted mainly through water.

* Corresponding Author. E-mail address: howard.lei@csueastbay.edu

I. INTRODUCTION

Epidemic outbreaks cause a global threat to humanity, and it is crucial to be prepared to face such outbreaks. An epidemic is the occurrence in a community or region of an illness, a specific health-related behavior, or other unexpected health-related events. The number of infections indicating the occurrence of an epidemic varies according to the agent, size, characteristics of the exposed population, and time and place of occurrence of the disease. Nadu (2012) stated that the representation of epidemics in a particular community is influenced by the distribution and characteristics of the disease, the social pattern and cultural behavior of people, the environmental indicators, and people's prior exposure to the disease. Watson, Gayer, and Connolly (2007) stated that communicable diseases are common in displaced populations with poor access to basic needs such as safe water and sanitation, adequate shelter, and primary health-care services. Assuring access to safe water and primary health care services are crucial, as are surveillance and early warning to detect epidemic-prone diseases known to occur in the disaster-affected area.

Apart from the health risks, a disease outbreak also has significant implications to businesses and economies. The 2016 Global Risks Report by the World Economic Forum states that disease outbreaks produce a response called "aversion behavior", which caused Singapore's economy to arrive at a standstill as a result of SARS in 2003, and Ebola in 2014 (World Economic Forum, 2016). Even individuals who have not been exposed to the disease will take actions such as staying home from work to avoid contracting the disease, resulting in the reduction of labor. Individuals may also alter behaviors such as taking fewer trips to the supermarket, which affects prices of goods. In the Ebola crises in Guinea, Liberia, and Sierra Leone, loss of life likely resulted in the business closures and tourism decreases. In these countries, the actual economic growth was significantly less compared to the expected growth. Furthermore, due to globalization, economic impacts in one region will increasingly affect other regions.

While it is virtually impossible to drastically alter the environmental and social economic conditions in a short amount of time to prevent an epidemic from occurring, it is possible to predict the amount of medication needed to combat the epidemic once it occurs. The forecasting of the amount of medication needed requires foreknowledge of when the outbreak would occur, and the number of individuals affected by the outbreak. Knowing the quantity of the medication that needs to be produced in response to an epidemic affects the demand scheduling, resource planning, and decision support in the public health administration for medication production.

The forecasting of epidemic medication demands and expenditures can be done using statistical models, as was done in (Tizzoni et al., 2012) and (D'sa, Nakagawa, Hill, and Tan, 1994). The forecasting of the amount of medication needed can be based on important environmental, social, and historical indicators related to the disease. A wide range of statistical models can be used, and the choice of an appropriate model depends on the type of data encountered. Data on infectious disease outbreaks and the contributing indicators can be collected from national and international health-care data repositories, such as the World Health Organization (WHO) website (World Health Organization, 2015) as well as the World Bank website (World Bank, 2015).

This work focuses on Cholera outbreaks in African countries, which have been drastically affected by the outbreaks. Cholera is an infectious disease caused by intestinal infection of the bacterium Vibrio Cholerae (Sack, Sack, Nair, and Siddique, 2004). Without adequate treatment, severe Cholera kills about half of the people affected.

Cholera is mainly associated with the contamination of water. Because it is difficult to completely remove the bacterium responsible for Cholera (the bacterium is part of the ecosystem surrounding the surface water of the planet), Cholera itself is difficult to completely remove. Hence, one solution for controlling Cholera is the effective management and control of the disease with medication, resources and facilities (Sack, Sack, Nair, and Siddique, 2004). Hence, this work attempts to predict the number of cases of Cholera in a given year, for purposes of forecasting the medication demands of Cholera to control the disease.

In our prior work, we attempted to predict the number of Cholera infections across 12 African countries over a period of 13 years, using the Linear Regression model. We examined a set of 16 total socio-economic indicators, and determined that the best results were obtained using 6 of the 16 indicators. However, the results for this prior work were such that the prediction attained by the regression model did not strongly correlation with the actual number of Cholera infections, and Cholera predictions were not made for future years. Nevertheless, among the 6 indicators, one of them relates to the percentage of rural population with access to improved water sources. Since Cholera is mainly associated with the contamination of water, it is expected that an indicator pertaining to improved water sources would be among the final set of indicators. In this work, we expand upon our prior work by using a total of 30 socio-economic indicators, and make predictions up to two years into the future across a period of over 20 years. We also examined five statistical classifiers – Logistic Regression, Perceptron, Support Vector Machine, Decision Tree, along with Linear Regression - which allows for making better predictions, and for gaining a better understanding regarding the importance of each indicator in making the predictions. Each classifier represents a numerical method for making predictions of outcomes given a set of input factors.

This paper is organized as follows: Section 2 describes prior and related work. Section 3 discusses the data collection process, and Section 4 describes the methodology. Section 5 describes the experiments and results, and Section 6 provides an analysis and discussion of the results. Section 7 contains a summary of the work, and describes future work.

II. PRIOR AND RELATED WORK

There has been prior work related to the forecasting of health risks associated with Cholera, as well as supply chain management associated with medications and pharmaceuticals. Fleming, Van der Merwe, and McFerren (2007) developed systems based on expert knowledge and historical data to identify favorable conditions for Cholera outbreaks. Constantin de Magny et al. (2008) used satellite and rainfall data for Cholera prediction in Kolkata, India, and Matlab, Bangladesh. It was discovered that there is a statistically significant relationship between Cholera infections, and chlorophyll-A concentration and rainfall anomalies, and that ocean and climate patterns are useful for predicting Cholera infections. Pascual, Chaves, Cash, Rodó, and Yunus (2008) used a time-series model incorporating the non-linear dynamics of the Cholera disease and levels of immunity in the population to develop an early-warning Cholera forecasting system in Bangladesh, while Jutla et al. in 2013 (Jutla et al., 2013) sought to understand the causes of Cholera outbreaks in Haiti after the 2010 earthquake.

Khoury and Loannidis (2014) discussed the use of big data systems to build predictive models to combat public health risks. Nasr-Azadani et al. (2015) used statistical multiple regressive models to attempt to estimate river discharge in the Bengal Delta region, as it was discovered that river discharges lead to cross-contamination as well as possible intrusion of bacteria, leading to Cholera. Pezeshki, Tafazzoli-Shadpour, Nejadgholi, Mansourian, and Rahbar (2016) used Artificial Neural Networks (ANNs) with identified risk factors to predict Cholera infection in villages. They were able to achieve 80% accuracy in forecasting Cholera infections in villages.

The following describes work involving the supply chain management of medications and pharmaceuticals. Chen, Mockus, Orcun, and Reklaitis (2012) developed simulations to help reduce clinical trial costs by improving the supply chain management for clinical trials. Reducing clinical trial costs is a major step towards reducing development costs for new drugs. Rossetti, Handfield, and Dooley (2011) examines factors affecting the purchasing and distribution of biopharmaceutical medications. Mustaffa and Potter (2009) investigates healthcare inventory management in Malaysia, with focus on medication distribution to clinics. The work found that urgent orders and stock availability are the main issues that must be handled. Shah (2004) discusses the need for balancing future capacity with anticipated demands for pharmaceutical products. This work also identifies the issue of lack of responsiveness in pharmaceutical product production, with an overall supply chain cycle of up to 300 days. Our work addresses these issues as we attempt to develop early-warning systems for predicting Cholera infections up to two years (730 days) in advance. The results of the different classifiers for our early warning system allows for the choice of either preferring to meet all future anticipated medication demands, or preferring to preserve future medication production capacity while meeting most demands.

III. DATA COLLECTION

The data collected for this work is from the World Bank website (World Bank, 2015), and the World Health Organization's Global Health Repository website (World Health Organization, 2015). Data for a total of 47 African countries and 42 socio-economic indicators across a period of 50 years were initially collected. Because of the sparseness of this data, with many missing socio-economic indicator values, we decided to filter the data down to 10 African countries and 30 socio-economic indicators across a period 24 years (1990-2013). In the filtered dataset, almost all socio-economic indicators values exist for all countries and years. African countries were selected because they are prone to Cholera outbreaks. According to the data, several of the African countries have experienced outbreaks of Cholera in the past 25 years. The countries used for Cholera prediction in this work are the following: Benin, Burundi, Cameroon, Democratic Republic of Congo, Ghana, Malawi, Mozambique, Nigeria, Tanzania, and Togo. Table 1 shows the socio-economic indicators that are used, where each indicator pertains to an individual country.

It is evident that the total population of a country would have a high impact on the number of reported cases of Cholera for that country, since countries with higher populations likely have more people who could be affected. Hence, we are interested in predicting the ratio of the number of reported cases of Cholera infections to the total population for each country. We will refer to this ratio as the "Cholera-to-Total-Population Ratio," with the abbreviation CTP-Ratio. The distribution of the CTP-Ratio across all years and countries are shown in Fig. 1.

According to Fig. 1, the CTP-Ratios range from 0 to 277. About 85% of the values are below 50, 81% of the values are below 40, 75% below 30, 67% below 20, and 52% below 10. Note that the total number of reported cases of Cholera for a given year and country range from 0 to 60,000, with 80% of the number of cases below 10,000, 70% below 5,000, and 60% below 3,000.

TABLE 1. SOCIO-ECONOMIC INDICATORS USED.

Socio-economic indicators used:
Population growth (annual %)
Surface area (sq. km)
Gross National Income (GNI), Atlas method
GNI per capita, Atlas method
Life expectancy at birth
Fertility rate (births per woman)
Adolescent fertility rate (births per 1,000 women ages 15-19)
Mortality rate (under age 5)
Immunization (measles) (% of children ages 12-23 months)
Prevalence of HIV, total (% of population ages 15-49)
Forest area (sq. km)
Improved water source (% of population with access)
Improved sanitation facilities (% of population with access)
CO2 emissions (metric tons per capita)
GDP at market prices (current US$)
GDP growth (annual %)
Inflation, GDP deflator (annual %)
Agriculture, value added (% of GDP)
Industry, value added (% of GDP)
Services, etc., value added (% of GDP)
Exports of goods and services (% of GDP)
Imports of goods and services (% of GDP)
Mobile cellular subscriptions (per 100 people)
Merchandise trade (% of GDP)
Net barter terms of trade index
External debt stocks
Total debt service (% of GNI)
Net migration
Foreign direct investment
Net ODA received per capita

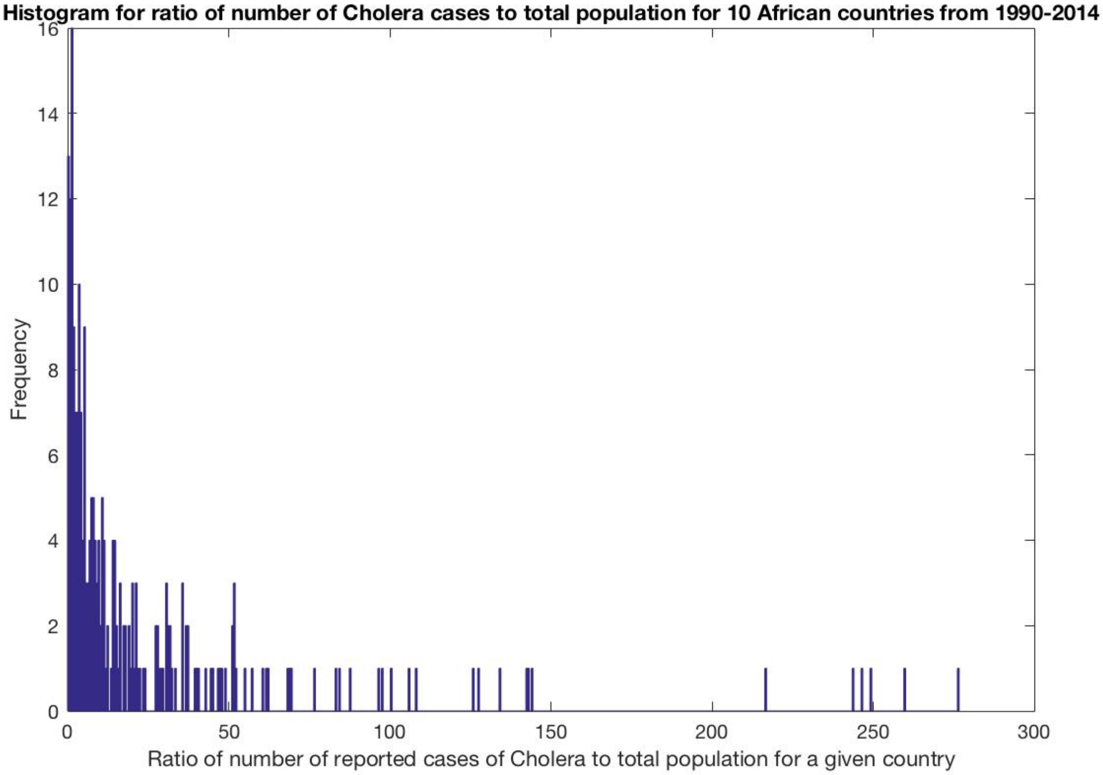

FIGURE 1. HISTOGRAM OF THE CTP-RATIO FOR THE 10 AFRICAN COUNTRIES IN THE DATASET.

IV. RESEARCH METHODOLOGY

Given the difficult nature of predicting the continuous-ranged CTP-Ratio values themselves, as they vary unpredictability from year to year, we decided to simplify the problem by dividing the CTP-Ratios into two classes. Based on Fig. 1, we decided to set the decision threshold at the small gap in the histogram around the CTP-ratio of 25. Many data points are clustered below the threshold of 25, which indicates low/noise-level Cholera infections for those countries and years. Data is more widely scattered above the threshold of 25, which indicates that some significant degree of infection has occurred for those countries and years. Hence, CTP-ratios below 25 (60% of data) are in the Low Cholera risk class, and ratios equal or above 25 (40% of data) are in the High Cholera risk class.

The freely-available Scikit-learn package (scikit-learn, 2017) is used to train five statistical classifiers - Logistic Regression, Perceptron, Support Vector Machine (SVM), Decision Tree, and Linear Regression classifiers – for predicting whether the CTP-Ratio for a given country and year falls into the Low or the High classes based on the socio-economic indicators. For the regression-based classifiers, the Low risk class labels are assigned a value of 0 and the High risk class labels are assigned a value of 1, and regression results greater or equal to 0.5 are categorized into the High class. The two-class classifiers serve to warn countries of whether or not a Cholera epidemic risk may arise in the future, and a more careful examination of the socio-economic indicators and the use of complementary information can help countries determine the specific amounts of medication needed. Hence, our classifiers serve as early-warning systems to alert countries when there is a risk of high cholera infections.

In our two-class classification framework, we focused on reducing two types of errors – the first is the false alarm error where a country and year with a CTP-Ratio in the Low class is classified as being in the High class; the second is the miss error where a country and year with a CTP-Ratio in the High class is classified as being in the Low class. We believe that reducing the second type of error is more important than reducing the first type of error, because it is more important to error on the side of having too much medication as opposed to not enough, even at the cost of revenue loss from producing more than is needed. Hence, in training the statistical models, we gave greater importance to the countries and years with High CTP-Ratios in the training data via replication of the High class training data. For most classifiers, this is mathematically equivalent to increasing the classifier weights associated with the High class data.

V. EXPERIMENTS AND RESULTS

A variety of experiments are performed using different classifiers, different combinations of socio-economic indicators, different weights given to the High CTP-Ratio training data, and for same-year, one-year ahead, and two-year ahead predictions. The same-year predictions give a basic understanding of how the indicators interact with the CTP-Ratios for a given country. The one-year ahead predictions use socio-economic indicators in a given year to predict the CTP-Ratio 1 year later for a given country, and the two-year ahead predictions use socio-economic indicators in a given year to predict the CTP-Ratio two years later for a given country.

5.1. Same-year predictions

The first set of experiments involves using socio-economic indicators for a given year to predict the CTP-Ratios for the same year. We began by using all 30 socio-economic indicators in the predictions using the Logistic Regression, Perceptron, Support Vector Machine (SVM), Decision Tree, and Linear Regression classifiers. For the Logistic Regression, Perceptron, SVM, and Linear Regression classifiers, the default cut-off points are used. The years 1990-2003 are used for training the classifiers, with a total of 135 data points (each data point being the set of socio-economic indicators for a particular country and year); the years 2004-2013 are used for evaluating the classifiers, with a total of 97 data points. The training data contains 91 data points in the Low CTP-Ratio class, and 44 in the High CTP-Ratio class. The evaluation data contains 75 data points in the Low CTP-Ratio class and 22 in the High CTP-Ratio class.

All socio-economic indicators are scaled to have standard normal distributions across the entire training data, and the parameters used for scaling the training data are used to scale the evaluation data. Because the use of all 30 socio-economic indicators for classification may result in over-fitting of the classifiers (since there are only 135 pieces of training data), we first decided to narrow down the list of indicators based on each indicator's Pearson's Correlation magnitude with respect to the same year's CTP ratio. We found 12 indicators with $p \leq 0.05$, and decided to select 10 of the 12 with the highest correlation magnitudes. Reducing the number of indicators also allows us to gain greater understanding of the importance of each individual indicator, and their effectiveness in the classification framework. We note that in this work, the use of alternative techniques for dimensionality reduction of the indicators (i.e. Principal Component Analysis) is avoided, because we are mainly interested in how the indicators themselves interact with the classifiers. The 10 indicators, along with the magnitudes of the correlation values, are shown in Table 2.

We then examined the importance of each of these 10 indicators used in the Logistic Regression and Linear Regression classifiers, and ran additional experiments using only the few indicators with highest importance. These two regression-based classifiers allow us to easily obtain the importance of the indicators by examining the classifier weights associated with each of the indicators. We note that for both the Logistic Regression and Linear Regression classifiers, the two indicators – Improved water source and CO2 emissions – had significantly higher importance weights than the other indicators across almost all experiments. Hence, these two indicators are also used standalone for all five classifiers. Table 3 shows the results on evaluation data for the same-year predictions in terms of the false alarm and miss errors, and their averages. The errors range from 0 (no errors) to 1 (100% error). The weights given to the training data belonging to the High CTP-Ratio class also varied from experiment to experiment, and Table 3 shows the results with training weights giving the lowest average false alarm and miss errors.

TABLE 2. SOCIO-ECONOMIC INDICATORS WITH THE 10 HIGHEST PEARSON'S CORRELATION MAGNITUDES WITH THE SAME YEAR'S CTP-RATIOS ACROSS ALL TRAINING DATA, ALONG WITH THE CORRELATION MAGNITUDE VALUES.

Socio-economic indicator:	**Pearson's correlation magnitude with CTP-Ratio**
Net ODA received per capita	0.33
Adolescent fertility rate (births per 1,000 women ages 15-19)	0.32
Prevalence of HIV, total (% of population ages 15-49)	0.30
Life expectancy at birth	0.27
CO2 emissions (metric tons per capita)	0.26
Improved water source (% of population with access)	0.26
Mortality rate (under age 5)	0.25
Industry, value added (% of GDP)	0.19
Exports of goods and services (% of GDP)	0.18
Services, etc., value added (% of GDP)	0.18

TABLE 3. SAME YEAR CTP-RATIO PREDICTION RESULTS USING FIVE DIFFERENT CLASSIFIERS, AND DIFFERENT COMBINATIONS OF SOCIO-ECONOMIC INDICATORS.

Classifier	Socio-economic indicators	Hit Rate	False Alarm Rate	Miss Rate	Avg False Alarm and Miss
Logistic Regression	All 30	0.09	0.04	0.91	0.48
Logistic Regression	10 w/ highest Corr.	0.64	0.22	0.36	0.29
Logistic Regression	Improved water source CO2 emissions	0.64	0.16	0.36	**0.26**
Linear Regression	All 30	0.09	0.04	0.91	0.48
Linear Regression	10 w/ highest Corr.	0.55	0.22	0.45	0.34
Linear Regression	Improved water source CO2 emissions	0.64	0.16	0.36	**0.26**
Perceptron	All 30	0.09	0.03	0.91	0.47
Perceptron	10 w/ highest Corr.	0.41	0.05	0.59	0.32
Perceptron	Improved water source CO2 emissions	0.55	0.11	0.45	**0.28**
SVM	All 30	0.27	0.05	0.73	0.39
SVM	10 w/ highest Pearson's corr.	0.64	0.22	0.36	0.29
SVM	Improved water source CO2 emissions	0.55	0.07	0.45	**0.26**
Decision Tree	All 30	0.00	0.00	1.00	0.50
Decision Tree	10 w/ highest Pearson's corr.	0.45	0.28	0.55	0.42
Decision Tree	Improved water source CO2 emissions	0.36	0.05	0.64	**0.35**

Examining the error rates in Table 3, we see that across all five classifiers, using just two socio-economic indicators (Improved water source and CO2 emissions) produced better average false alarm and miss results compared to using all 30 indicators. While this is understandable given that the use of all 30 indicators may have resulted in over-fitting of the classifiers, the use of the two indicators also outperformed the use of 10 indicators (a sufficiently low number to avoid over-fitting) with the highest Pearson's correlation with CTP-Ratio across the training data. For the Logistic Regression classifier, using the two indicators led to a 10.3% improvement over using the 10 indicators (0.26 vs. 0.29), and for the Perceptron classifier, using the two indicators led to a 12.5% improvement over using the 10 indicators. For the Linear Regression classifier, using the two indicators resulted in a 23.5% improvement over using the 10 indicators. For the SVM classifier, using the two indicators resulted in a 10.3% improvement over using the 10 indicators, while for the Decision Tree classifier, using the two indicators lead to a 16.7% improvement over using the 10 indicators.

The observed improvements when using just the two indicators suggest that there is discriminative value for the Improved water source and CO2 emissions indicators. Since the bacterium that causes Cholera is transmitted through water, it is not surprising that the Improve water source indicator holds discriminative power in our classification framework. The effect of CO2 emissions on Cholera infections has not been widely understood, and warrants future study. These results are merely for the same-year predictions, however, and in the next set of experiments, we will affirm their predictive power of the CPT-Ratios in future years.

5.2. One-year ahead predictions

We next performed experiments using the socio-economic indicators to make predictions of CTP-Ratios one year into the future using the same five classifiers. For the one-year ahead predictions, socio-economic indicators from years 1990-2003 are used to train the classifiers for predicting CTP-Ratios from 1991-2004. In this framework, the classifiers are trained such that the indicators from 1990 are used to predict the ratio in 1991; the indicators from 1991 are used to predict the ratio in 1992, and so on. The classifiers are then evaluated using the indicators from years 2004-2012 to predict the CTP-Ratios from years 2005-2013. Once again, we narrowed down the set of indicators by first selecting the 10 indicators with the highest Pearson's correlation magnitudes with the CTP-Ratios (all with $p \leq 0.05$) in the training data. However, the correlations are now obtained between the indicators for a given year with the CTP-Ratio in the following year. Table 4 shows the 10 indicators and the corresponding Pearson's correlation magnitudes.

Table 4 shows that 9 out of 10 indicators with highest Pearson's correlation magnitudes for the same-year prediction framework also have the highest Pearson's correlation magnitudes for the one-year ahead prediction framework. The only indicator that no longer appears is the Exports of goods and services (% of GDP) indicator, which is replaced by the Net barter terms of trade index indicator. The Logistic Regression and Linear Regression classifiers are once again used to determine the importance of each of the 10 indicators. It was determined that for one-year ahead prediction, a set of five socio-economic indicators had significantly higher importance weights compared to the other indicators for both classifiers – Adolescent fertility rate, Mortality rate, Improved water source, CO2 emissions, and Net barter terms of trade index. Experiments were performed using these five indicators, as well as with the two indicators – Improved water source and CO2 emissions – with lowest errors for the same-year prediction experiments. Table 5 shows the results for experiments using the two, five, and 10 indicators (with no classifier over-fitting).

Examining the error rates in Table 5, there are some combinations of classifiers and indicators that result in low false alarm errors and high miss errors, and other combinations resulting in low miss errors and high false alarm errors for the one-year ahead prediction. Depending on which error one wishes to minimize, one can select the classifier and indicator combination to reduce that type of error. Overall, the prediction results suggest that for the Logistic Regression, SVM, and Linear Regression classifiers, the use of only the two indicators – Improved water source and CO2 emissions – still produces the lowest average false alarm and miss errors overall.

For Logistic regression, these two indicators outperformed the use of the five indicators with the highest importance weights (0.28 vs. 0.32 for average false alarm and miss errors), and was comparable to the use of the 10 indicators with highest Pearson's correlation magnitude (0.28 vs. 0.27). For SVM, these two indicators outperformed the use of both the five indicators (0.24 vs. 0.33), as well as the 10 indicators (0.24 vs. 0.34). For Linear Regression, the results also show that using the two indicators outperformed the use of the five indicators (0.28 vs. 0.34) and the 10 indicators (0.28 vs. 0.32). When used with the Perceptron and Decision Tree classifiers, however, these two indicators did not achieve significantly better results compared to the use of the five and the 10 indicators. However, the average false alarm and miss errors for both the Perceptron and Decision Tree classifiers are higher than the error obtained when using the two indicators for the Logistic Regression, SVM, and Linear Regression classifiers. This shows the importance of the regression-based classifiers (where the SVM classifier is based on the Support Vector Regression technique) along with the two indicators in making the one-year ahead predictions.

TABLE 4. SOCIO-ECONOMIC INDICATORS WITH THE 10 HIGHEST PEARSON'S CORRELATION MAGNITUDES WITH THE FOLLOWING YEAR'S CTP-RATIOS ACROSS ALL TRAINING DATA, ALONG WITH THE CORRELATION MAGNITUDE VALUES.

Socio-economic indicator:	Pearson's correlation magnitude with CTP-Ratio
Adolescent fertility rate (births per 1,000 women ages 15-19)	0.32
Improved water source (% of population with access)	0.28
Net ODA received per capita	0.28
Life expectancy at birth	0.27
Net barter terms of trade index	0.27
Prevalence of HIV, total (% of population ages 15-49)	0.27
CO2 emissions (metric tons per capita)	0.26
Mortality rate (under age 5)	0.26
Services, etc., value added (% of GDP)	0.20
Industry, value added (% of GDP)	0.18

TABLE 5. ONE-YEAR AHEAD CTP-RATIO PREDICTION RESULTS USING FIVE DIFFERENT CLASSIFIERS, AND DIFFERENT COMBINATIONS OF SOCIO-ECONOMIC INDICATORS.

Classifier	Socio-economic indicators	Hit Rate	False Alarm Rate	Miss Rate	Avg False Alarm and Miss
Logistic Regression	10 w/ highest Corr.	0.89	0.43	0.11	0.27
Logistic Regression	5 with highest weights	0.79	0.42	0.21	0.32
Logistic Regression	Improved water source CO2 emissions	0.63	0.18	0.37	**0.28**
Linear Regression	10 w/ highest Corr.	0.74	0.37	0.26	0.32
Linear Regression	5 with highest weights	0.63	0.31	0.37	0.34
Linear Regression	Improved water source CO2 emissions	0.63	0.18	0.37	**0.28**
Perceptron	10 w/ highest Corr.	0.89	0.58	0.11	**0.35**
Perceptron	5 with highest weights	0.89	0.60	0.11	0.36
Perceptron	Improved water source CO2 emissions	0.68	0.52	0.32	0.42
SVM	10 w/ highest Corr.	0.68	0.36	0.32	0.34
SVM	5 with highest weights	0.63	0.30	0.37	0.33
SVM	Improved water source CO2 emissions	0.58	0.06	0.42	**0.24**
Decision Tree	10 w/ highest Corr.	0.74	0.54	0.26	0.40
Decision Tree	5 with highest weights	0.74	0.54	0.26	0.40
Decision Tree	Improved water source CO2 emissions	0.68	0.45	0.32	**0.39**

5.3. Two-year ahead predictions

We lastly performed experiments for making two-year ahead predictions, which is more appropriate for developing an early warning system that can make predictions multiple years in advance of an outbreak. Socio-economic indicators from years 1990-2003 are used to train the classifiers for predicting CTP-Ratios from 1992-2005, and evaluation is performed using the indicators from years 2004-2011 to predict the CTP-Ratios from years 2006-2013. Once again, the 10 indicators with the highest Pearson's correlations with respect to the CTP-Ratios (all with $p \leq 0.05$) are obtained, but now, the correlations are obtained using the CTP-Ratio values two years later. Table 6 shows the 10 indicators and the corresponding Pearson's correlation magnitudes.

Table 6 shows that 9 out of 10 indicators with highest Pearson's correlation magnitudes for the two-year ahead prediction framework also have the highest Pearson's correlation magnitudes for the one-year ahead prediction framework. The only indicator that no longer appears is the Services, etc., value added (% of GDP) indicator, which is replaced by the GDP growth (annual %) indicator. A comparison of Table 6 with Table 2 shows that 8 out of 10 indicators with highest Pearson's correlation magnitudes for the two-year ahead prediction framework also have the highest correlation magnitudes for the same-year prediction framework. The Logistic Regression and Linear Regression classifiers are once again used to determine the importance of each of the 10 indicators. For two-year ahead prediction, the following five socio-economic indicators had significantly higher importance weights compared to the other indicators for both classifiers – Adolescent fertility rate, Mortality rate, CO2 emissions, Net barter terms of trade index, and Net ODA received per capita. Experiments were performed using these five indicators, as well as with the two indicators – Improved water source and CO2 emissions – that were effective for the same-year and one-year ahead experiments. Table 7 shows the results for experiments using the two, five, and 10 indicators (with no classifier over-fitting).

**TABLE 6. SOCIO-ECONOMIC INDICATORS
WITH THE 10 HIGHEST PEARSON'S CORRELATION MAGNITUDES
WITH THE CTP-RATIOS TWO YEARS LATER ACROSS ALL TRAINING DATA,
ALONG WITH THE CORRELATION MAGNITUDE VALUES.**

Socio-economic indicator:	Pearson's correlation magnitude with CTP-Ratio
Life expectancy at birth	0.31
Adolescent fertility rate (births per 1,000 women ages 15-19)	0.30
Mortality rate (under age 5)	0.30
Improved water source (% of population with access)	0.28
Net barter terms of trade index	0.28
CO2 emissions (metric tons per capita)	0.26
Prevalence of HIV, total (% of population ages 15-49)	0.26
Net ODA received per capita	0.23
GDP growth (annual %)	0.21
Industry, value added (% of GDP)	0.20

TABLE 7. TWO-YEAR AHEAD CTP-RATIO PREDICTION RESULTS USING FIVE DIFFERENT CLASSIFIERS, AND DIFFERENT COMBINATIONS OF SOCIO-ECONOMIC INDICATORS.

Classifier	Socio-economic indicators	Hit Rate	False Alarm	Miss	Avg False Alarm and Miss
Logistic Regression	10 w/ highest Corr.	0.89	0.40	0.11	0.26
Logistic Regression	5 with highest weights	0.95	0.45	0.05	**0.25**
Logistic Regression	Improved water source CO2 emissions	0.58	0.08	0.42	**0.25**
Linear Regression	10 w/ highest Corr.	1.00	0.53	0.00	0.27
Linear Regression	5 with highest weights	0.89	0.40	0.11	0.26
Linear Regression	Improved water source CO2 emissions	0.58	0.08	0.42	**0.25**
Perceptron	10 w/ highest Corr.	0.79	0.40	0.21	0.31
Perceptron	5 with highest weights	1.00	0.50	0.00	**0.25**
Perceptron	Improved water source CO2 emissions	0.63	0.20	0.37	0.29
SVM	10 w/ highest Corr.	0.79	0.35	0.21	0.28
SVM	5 with highest weights	0.89	0.38	0.11	**0.24**
SVM	Improved water source CO2 emissions	0.58	0.08	0.42	0.25
Decision Tree	10 w/ highest Corr.	0.11	0.02	0.89	0.46
Decision Tree	5 with highest weights	0.74	0.18	0.26	0.22
Decision Tree	Improved water source CO2 emissions	0.68	0.08	0.32	**0.20**

Examining the error rates in Table 7, it is noteworthy that the two-year ahead predictions have overall lower average false alarm and miss errors compared to results for the same-year and one-year ahead predictions. However, this can be attributed to the fact that the evaluation data used for two-year ahead predictions differs from that used for the other predictions. Fewer years of CTP-Ratio evaluation data (years 2004-2011) are available for the two-year ahead predictions, compared to the one-year ahead and same-year predictions (years 2004-2012 and 2004-2013 respectively). Results show that for three of the five classifiers (Logistic Regression, Linear Regression, and Decision Tree), the two indicators – Improved water source and CO2 emissions – still produce the lowest average false alarm and miss errors. This is in spite of the fact that the Improved water source indicator is not even amongst the list of five indicators with the highest importance weights. The Decision Tree classifier performed particularly well for the two-year ahead predictions, with an average error of 0.20 using only the two indicators. For the other classifiers, the lowest average errors are 0.25. For the Perceptron and SVM classifiers, the set of five indicators with highest importance weights give the lowest average error.

VI. ANALYSIS AND DISCUSSION

The set of experiments for the same year, one-year ahead, and two-year ahead predictions using the five classifiers and different socio-economic indicator combinations illustrate several key points. First, it is possible to produce an early-warning system a couple years in advance of a Cholera epidemic using simple and freely available statistical analysis techniques. The two-year ahead prediction result using the Decision Tree classifier with the Improved water source and CO2 emissions indicators had a false alarm error of 0.08 and a miss error of 0.32. This means that the classifier is able to correctly predict around two-thirds of the Cholera outbreaks two years into the future. Furthermore, when it predicts that there is an outbreak, it is correct 92% of the time, which has cost-reducing implications for medication production.

Second, this work sheds light on the relative importance of each of the 30 socio-economic indicators. It is noteworthy that Improved water source and CO2 emissions have some inherent discriminative power even as they don’t even have highest Pearson’s correlations with the CTP-Ratio among the 10 indicators as shown in Tables 2, 4, and 6 for the same-year, one-year ahead, and two-year ahead prediction frameworks. Hence, these two indicators should serve as the foundation to developing an actual early-warning system. Given the plethora of socio-economic indicators found in the World Bank and World Health Organization websites, we have shed light on which indicators have inherent discriminative power for the number of Cholera infections.

Third, this work shows that for the two-year ahead prediction results, which are more relevant to that of an early warning system compared to the same-year and one-year ahead prediction results, the choice of classifier used does not lead to significantly different average false alarm and miss errors. The average errors range from 0.20 to 0.25 for all five classifiers. Nevertheless, the different classifiers and indicator combinations produce different individual false alarm and miss errors. This means that it is possible to choose the classifier and indicators to optimize for the type of error that’s desired. For example, the SVM classifier using the five indicators has a false alarm error that’s 0.30 higher, but a miss error that’s 0.21 lower, compared to the Decision Tree classifier using the two indicators. Hence, if one wishes to meet all anticipated future medication demands for every potential Cholera outbreak two years into the future, one would prefer the SVM classifier to the Decision Tree classifier, since the SVM classifier’s optimal result has a lower miss error. On the other hand, if one prefers to save costs on medication production and preserve future production capacity while still being prepared for most of the outbreaks two years into the future, then the Decision Tree classifier is preferable to the SVM classifier.

Finally, we note that for the same-year prediction results (Table 3), the optimal results for the Logistic Regression, Linear Regression, Perceptron, and SVM classifier all produced average false alarm and miss errors from 0.26 through 0.28, and these four are preferable to the Decision Tree classifier, with 0.35 average error. For the one-year ahead prediction results (Table 5), the Logistic Regression, Linear Regression, and SVM classifiers produced average errors from 0.24 through 0.28, which are preferable to the Perceptron and Decision Tree classifiers (average errors of 0.35 and 0.39, respectively). The optimal results for same-year prediction generally have lower false alarm and higher miss errors, while the results for the one-year ahead prediction have varying degrees of false alarm and miss errors.

VII. CONCLUSION

We attempted to construct simple statistical early warning systems for Cholera infections, for African countries with histories of Cholera outbreaks. The early warning systems are akin to medication demand forecasting systems. We investigated the use of five classifiers - Logistic Regression, Perceptron, Support Vector Machine, Decision Tree, and Linear Regression - to predict whether a country is at risk for a Cholera infection up to two years into the future for 10 African countries. A set of 30 socio-economic indicators for each of the 10 countries are investigated for use with the five classifiers, and the data was obtained from the World Bank website and the World Health Organization's Global Health Repository website. The results from the classifiers suggest that for the two-year ahead predictions, the choice of classifier does not lead to significantly different average error rates, but does lead to significantly different false alarm and miss error rates. Hence, certain classifiers are preferred over others based on whether one wishes to always meet all anticipated demand, or whether one wishes to save costs on medication production and preserve future production capacity while meeting most of the anticipated demand. The same applies for one-year ahead prediction, while for the same-year prediction, the optimal classifier results generally have lower false alarm and higher miss errors. The Decision Tree classifier used with only two socio-economic indicators – Improved water source and CO2 emissions – had the lowest average false alarm and miss error for two-year ahead prediction.

This work sheds light on the inherent discriminative power of the individual socio-economic indicators. It is noteworthy that use of only the two indicators lead to the best overall results, and confirms the notion that improved water quality is important in reducing the spread of the bacterium responsible for Cholera. Future work involves better understanding the effect(s) of CO2 emissions on Cholera outbreaks, and exploring additional countries – both African countries and those in other continents – to be used for training and evaluation. As additional data becomes available online, data from additional years can be incorporated. Having additional years of data will also facilitate the development of systems for making predictions more than two years into the future.

Lastly, this work is important from both humanitarian as well as a business and economic perspectives. Loss of life caused by disease outbreaks, while tragic in itself, also introduces aversion behavior for individuals which affect labor supply and price of goods. In severe cases, such as in the Ebola crises in 2014, there were business closures and decreases in tourism, resulting in significantly less actual economic growth compared to expected growth. While the focus of our work is on Cholera, the statistical classifiers and techniques can be used to build simple early-warning systems for a number of diseases.

REFERENCES

Chen, Y., Mockus, L., Orcun, S., and Reklaitis, G., "Simulation-optimization approach to clinical trial supply chain management with demand scenario forecast", *Computers & Chemical*

Engineering (published by Elsevier), 40, 2012, 82-96.

Constantin de Magny, G., Murtugudde, R, Sapiano, M.R.P., Nizam, A., Brown, C.W., Busalacchi, A.J., Yunus, M., Balakrish Nair, G., Gil, A.I., Lanata, C.F., Calkins, J., Manna, B., Rajendran, K., Bhattacharya, M.K., Huq, A., Bradley Sack, R., and Colwell, R., "Environmental signatures associated with cholera epidemics", *Proceedings of the National Academy of Sciences of the United States of America*, 105(46), 2008, 17676-17681.

D'sa, M.M., Nakagawa, R.S., Hill, D.S. and Tan, J.K., "Exponential smoothing method for forecasting drug expenditures", *American Journal of Health-System Pharmacy*, 51(20), 1994, 2581-1.

Fleming, G., Van der Merwe, M., McFerren, G., "Fuzzy expert systems and GIS for cholera health risk prediction in southern Africa", *Environmental Modelling & Software*, 22(4), 2007, 442-448.

Jutla, A., Whitcombe, E., Hasan, N., Haley, B., Akanda, A., Huq, A., Alam, M., Bradley Sack, R., and Colwell, R, "Environmental factors influencing epidemic cholera", *The American Journal of Tropical Medicine and Hygiene*, 89(3), 2013, 597-607.

Lei, H., Almasi, A., Jayachandran, P.K., Nordmeyer, D., Ganjeizadeh, F., and Hattar, H., "Predicting Medication Demands for Cholera in African Countries", 2015, Unpublished.

Khoury, M.J. and Loannidis, J.P.A, "Big data meets public health", *Science*, 346(6213), 2014, 1054-1055.

Mustaffa, N. and Potter, A., "Healthcare supply chain management in Malaysia: a case study", *Supply Chain Management: An International Journal*, 14(3), 2009, 234-243.

Nadu, S., "What are the features of epidemic diseases?", http://www.preservearticles.com/201101072772/features-of-epidemic-diseases.html (accessed December 21, 2015).

Nasr-Azadani, F., Unnikrishnan, A., Akanda, A., Islam, S., Alam, M., Huq, A., Jutla, A., and Colwell, R., "Downscaling river discharge to assess the effects of climate change on cholera outbreaks in the Bengal Delta", *Inter-Research Climate Research*, 64(3), 2015, 257-274.

Pascual, M., Chaves, L.F., Cash, B., Rodó, X., and Yunus, M.D., "Predicting endemic cholera: the role of climate variability and disease dynamics", *Inter-Research Climate Research*, 36(2), 2008, 131-140.

Pezeshki, Z., Tafazzoli-Shadpour, M., Nejadgholi, I., Mansourian, A., and Rahbar, M., "Model of cholera forecasting using artificial neural network in Chabahar City, Iran", *International Journal of Enteric Pathogens*, 4(1), 2016.

Rossetti, C., Handfield, R., and Dooley, K., "Forces, trends, and decisions in pharmaceutical supply chain management", *International Journal of Physical Distribution & Logistics Management*, 41(6), 2011, 601-622.

Sack, D.A., Sack, R.B., Nair, G.B. and Siddique, A.K., "Cholera", *Lancet* 363(9404), 2004, 223-33.

scikit-learn, "scikit-learn: Machine Learning in Python", http://scikit-learn.org/stable/ (accessed March 22, 2017).

Shah, N., "Pharmaceutical supply chains: key issues and strategies for optimization", *Computers & Chemical Engineering (published by Elsevier)*, 28(6-7), 2004, 929-941.

World Economic Forum, "The Global Risks Report 2016, 11th Edition", *World Economic Forum*, 2016.

Tizzoni, M., Bajardi, P., Poletto, C., Ramasco, J.J., Balcan, D., Goncalves, B., Perra, N., Colizza, V. and Vespignani, A., "Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm", *BMC medicine*, 10, 2012, 165.

Watson, J.T., Gayer, M. and Connolly, M.A., "Epidemics after natural disasters", *Emerging Infectious Diseases*, 13(1), 2007, 1-5.

World Health Organization, "Global Health Observatory," *World Health Organization*, http://www.who.int/gho/database/en/ (accessed December 20, 2015).

World Bank, "World Bank Open Data," *World Bank*, http://data.worldbank.org/indicator/all (accessed December 20, 2015).