Regular Research Paper
Probabilistic Analysis of Contracting Ebola Virus Using Contextual Intelligence

Arjun Gopalakrishnan and Krishna Kavi
Department of Computer Science, University of North Texas, USA

Abstract

The West African countries witnessed an "extraordinary" outbreak of the Ebola virus in August 2014. It was declared to be a Public Health Emergency of International Concern (PHEIC) by the World Health Organization (WHO). Due to the complex nature of the outbreak, Centers for Disease Control and Prevention (CDC) has created interim guidance for monitoring people potentially exposed to Ebola and for evaluating their intended travel and restricting the movements of carriers when needed. Tools to evaluate the risk of individuals and groups of individuals contracting the disease could mitigate the fear and anxiety. Our goal is to understand and analyze the nature of risk an individual would posses when he/she comes in contact with a carrier. This paper presents a tool that makes use of contextual data intelligence to predict the risk factor of individuals who come in contact with the carrier.

Keywords: Ebola, iDid app, Contextual intelligence, Susceptibility, Risk factor, Bayes theorem

1 Introduction

When the efforts to prevent a disease fails and an outbreak occurs, the resulting distribution of cases may take various forms that are called epidemic curves. These epidemic curves project the nature of a disease outbreak within a population that are potentially at risk of contracting the disease [12]. Although they indicate the nature of an outbreak, they do not provide sufficient data to understand the chances that a particular individual gets affected by a disease outbreak. To minimize the spread of an epidemic such as the Ebola virus, we need effective contact tracing. The problem is complex as the contact tracing must be done retroactively after a patient is diagnosed with the disease. Any lapse in the tracing could fail to track the citizens at risk. Ad hoc tracing, relying on the infected carrier's recollection of places visited and people met, may lead to inaccurate findings. We need a contextually intelligent application that can keep track of both the individuals' movement and the carrier's movements to identify if the individual is at low risk or at high risk of contracting the Ebola virus. This paper provides a means of tracing and analyzing the risk of contracting the Ebola virus using contextual intelligence and other contributing factors which are also discussed in detail in the following sections.

2 Related Work

There have been several attempts to create right mathematical models that can predict the nature of a disease spread and also monitor individuals health on a daily basis. Achrekar et.al [2] presented a method for gathering twitter data to curb large scale spread of epidemic diseases. This paper presented the Social Network Enabled Flu Trends(SNEFT) architecture as a continuous data collection engine which combines the detection and prediction capability on social networks to discover real world flu trends. Johnson et.al [9] provided a means to calculate a susceptibility ratio using mathematical models. The SIR Model (the number Susceptible, Infectious, or Recovered (immune)) is used in epidemiology to compute the degree of susceptibility for an infected/recovered group of people in a population. This model is an appropriate one to use under the following assumptions.

(1) The population is fixed.
(2) The only way a person can leave the susceptible group is to become infected.The only way a person can leave the infected group is to recover from the disease. Once a person has recovered, the person received immunity
(3) Age, sex, social status, and race do not affect the probability of being infected.
(4) Members of the population mix homogeneously (have the same interactions with one another to the same degree).

IBM has pressed Data Analytics, Mobility and Cloud Computing Technology into service to bring the spread of Ebola in Sierra Leone under control. IBM has deployed SoftLayer cloud technology to set up an Ebola Open Data Repository, to provide governments, aid agencies and researchers with free and open access to the data. [10].

HealthMap, a team of researchers, epidemiologists and software developers at Boston Children's Hospital founded in 2006, is an established global leader in utilizing online informal sources for disease outbreak monitoring and real-time surveillance of emerging public health threats. The freely available Web site 'healthmap.org' and mobile app 'Outbreaks Near Me' deliver real-time intelligence on a broad range of emerging infectious diseases for a diverse audience including libraries, local health departments, governments, and international travelers. It achieves a unified and comprehensive view of the current global state of infectious diseases and their effect on human health [8]. These works can help in preventing the virus from becoming an outbreak, and even provide means to avert the disease spread but at an individual level, it still remains a mystery as to how far they are exposed to the virus.

This paper focuses on monitoring each individual and their exposure to the disease to predict their chances of contracting the disease thereby allowing the user to understand their chances of contracting any virus that they maybe prone to such as the Ebola virus.

3 Factors of disease susceptibility

When studying disease outbreaks and their nature, any infectious disease involves four important factors that cause an individual to be susceptible to the disease. They are:

(1) Time of Exposure
(2) Proximity of Carrier
(3) Carrier Status
(4) Individual Medical History

3.1 Time of exposure

The time of exposure provides us information about the amount of time spent by an individual with the carrier. Certain communicable diseases have a higher rate of susceptibility even if the exposure is for a short duration. Based on research, we can say the exposure rate for contracting the Ebola virus (or any other infectious disease) is dependent on the duration of contacts made by the individual with a carrier.

3.2 Proximity of carrier

The proximity or the nature of contact with the infected individual also plays a vital role in the probability of contracting any disease. The distance of the carrier from an individual can indicate the risk of contracting the disease.

When an individual comes in contact with the carrier on just one occasion, their probability of contracting Ebola will be less when compared to the individual who comes in contact on multiple occasions. In cases like the Ebola virus, actual physical contact is needed to contract a disease while in other cases (such as flu) no physical contact is needed.

3.3 Carrier Status

The Ebola virus is a disease which has an incubation period of 21 days during which time the infected carrier can spread the disease. Thus an individual's probability of contracting from the carrier varies depending on the day within the infectious period of the carrier. The carrier status refers to the day on which an individual comes in contact with the carrier.

3.4 Individual Medical History

An individual's medical history is important to see how the immune levels of the person could either strengthen or weaken the prospect of contracting a disease. Medical records can be used to gather information to predict how prone an individual is to contracting the Ebola virus.

4 Other factors

Besides these primary factors, there are a few incidental factors that contribute to an individual's susceptibility.

4.1 Environment of Individual and Facilities

The environment or the locality play a vital role in analyzing the risk: if the individual is located in a place with very good health facilities and has generally good hygiene, then their risk of contracting or spreading the disease is less when compared to an individual living in an environment with few medical facilities and having generally poor hygiene.

4.2 Population Demography

The implications of demographic changes for the spread and control of infectious diseases are not fully understood. But an individual's susceptibility can be studied based on the population structure and the marked effect it can have on any disease transmission. A population with more carriers can indicate higher risk for any individual located in that area. Suppose an individual is located in Sierra Leone, then he/she has a higher risk of contracting Ebola than an individual living in New York.

Another factor relates to cultural habits including touching and treating infected persons. This can vary from one environment to another.

4.3 Heat Zones

The Global Surveillance Network developed by the CDC in 1995 was based on the concept of a data collection network for the surveillance of travel related morbidity. The goal was to direct clinics to be ideally situated so as to effectively detect geographic and tem-

poral trends in morbidity among travelers, immigrants and refugees [7]. Such a concept would be useful for tracking carriers of the Ebola virus who can be monitored and thus provide a heat zone to indicate how a particular area is geo-fenced based on the nature of the affected population. Using data collected by

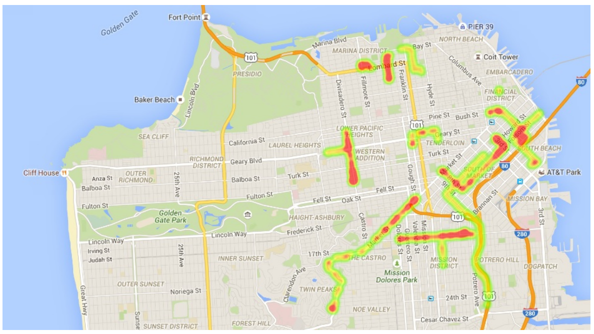


Figure 1: Heat zones

the CDC based on the survellance network established, heat zones can be constructed. Figure 1 shows a heat zoned area in California, where the traces of a carrier's movement is shown on the map. A zone's opacity indicates the area's degree of risk relative to other area. This could be useful for an individual travelling to an address that is close to the hot zones.

5 Implementation of our Application

5.1 Contextual data intelligence and iDid Inc app

As stated previously, to accurately track diseases, it is necessary to track the daily activities of both the carrier and the individual in question over the previous days and weeks. For example, since an Ebola infected carrier is contagious for 21 days, it would be necessary to discover all the locations the carrier visited and the people the carrier met. This information can then be used by individuals to correlate their own movements over the past 21 days to assess their risks. In many cases it is possible to predict past activities and behaviors of individuals based on current or future activities. However, the process of collecting data on an individual's activity and maintaining a repository is a tedious task: it must be collected in real-time to have an active monitoring scheme. It is also necessary to collect geographical information (GPS data) as well as the nature of the activity.

For example, GPS can be used to track the locations a person visits and perhaps provide contextual information to determine the duration of each visit.

Case 1: If a person visits a restaurant, it is reasonable to assume he/she will spend considerable time there to order food and consume it on the premises, which increases his/her chances of coming in contact with a carrier.

Case 2: If a person drives between locations, there is very little possibility of direct or very close contact with a carrier (unless the carrier is inside the vehicle).

Similar contextual information can be used to assess the risk of contracting or spreading infectious diseases. We use an application developed by iDid Inc. [1] to track activities carried out by individuals on a daily basis and generate reports of those activities at the end of the day.

5.2 Integration of iDid Inc and risk factor generation app

Data collected on individuals using the iDid app can be used to track their movements on a routine basis and log this information in a database. The database contains information such as duration spent at a place, time of visit, time of travel to a new location, number of contacts made with the individuals at the new location, etc. Data is stored in a backend repository and a users future schedule is predicted by analyzing his/her currently available data.

The data thus collected can be correlated with the data collected for the carrier (assuming such data exists) to generate discrete susceptibility ratio graphs (indicating probability of contracting the disease).

5.3 Creation of a web application

The purpose of the application is to provide an individual with an estimate of their risk or how susceptible they are to contracting the disease. The application collects information from the back-end which contains the individuals daily activities as described above. In addition, the individual's medical records can be used to improve the accuracy of the predictions.

A composite risk factor is generated based on the data collected and on the various factors described previously.

5.4 Data Privacy and Security

The information required to calculate the probability of contracting the disease is personal and often covered by HIPAA. Therefore, data privacy is considered to be of paramount importance. To protect a user's privacy in our system, the data for a given user will be maintained in their own smartphone or their personal storage spaces, often protected with passwords and data encryption. The iDid app uses calendar selected by the user when he/she installs the app and a private database is maintained to log the completed activities such as drives, flights, places visited, sleep, etc. Our application does not save user contextual information or medical history, but uses the data only to compute risk factors. The iDid app also uses the Google Maps

API(s) [11].

With the users permission, the risk probabilities are used to track disease spread in a population, without any information that can be used to identify the individual.

6 Bayesian Analysis of Data

We use Bayesian probabilities to assess the risk of contracting a disease utilizing contextual information and medical history [13].

6.1 Quick Bayesian Risk Calculator

$$P(B_1|A) = \frac{P(B_1 \cap A)}{P(A)} = \frac{P(B_1)P(A|B_1)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)}$$

Figure 2: Bayesian Analysis Formula

The Bayesian network for determining risk describes the probability of an event, based on the conditions that might relate to that event. If an individual travels to the same places as the carrier and engages in activities that require them to spend time in proximity to the carrier, then their risk factor depends on the total time spent wit the carrier doing the activities.

Instance: Suppose a carrier is sitting at Starbucks at 10 am and another patron enters the store to have a cup of coffee. One could assume the patron might spend 15-20 minutes in the store and so their risk factor is likely higher than it is for an individual who drives past the same Starbucks and does not come in close quarters with the carrier. This Bayesian network involves knowing the type of activity, place, time spent by an individual, by which we can calculate the risk factor instantly at that given time.

6.2 Bayesian Analysis of Medical record

An individual's medical history contributes to a fair share of the person's risk to a disease. The information they provide could lead to an accurate analysis of understanding the individual's probability of conracting the Ebola virus.

6.3 Contextual Intelligence Probability Calculator (CIPC)

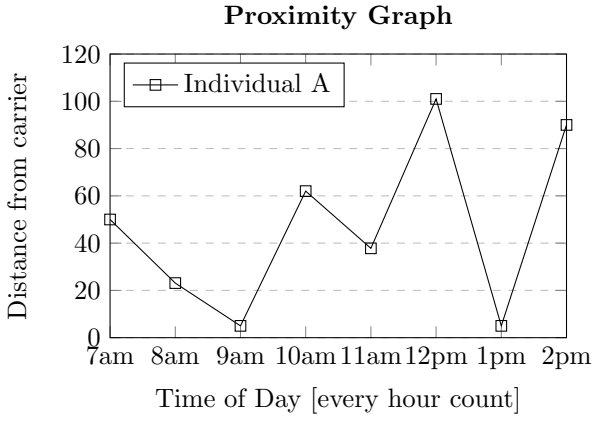
The goal is to provide the individual with objective data to assess their risk. The CIPC will yield a precise risk value which maybe high or low depending on the individual's susceptibility. It is calculated by the integration of probability values calculated from the Quick Bayesian calculator and the medical history Bayesian value, along with the probability values obtained from the factors that contribute to the risk, such as time of exposure, number of contacts made and the environment.

7 Graphical Representation of Individual Monitoring

7.1 Proximity Graph-

An individual's proximity is monitored to understand their closeness to the carrier. By monitoring the carrier and individual contextually, the data gets stored on an hourly basis.

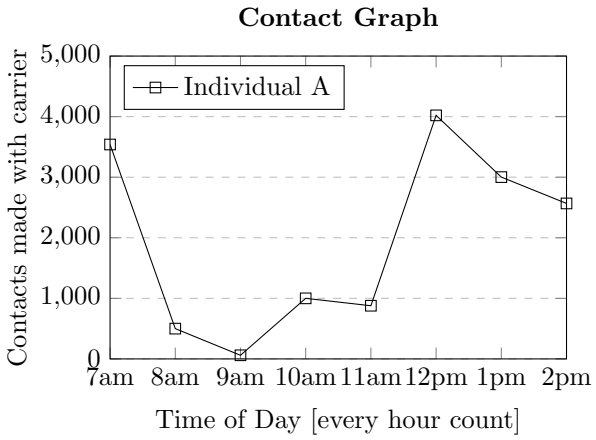
The proximity graph could indicate an individual A, whose distance from the carrier is being tracked from 7am to 12pm. The tracking yields a line graph that shows the individual's proximity to the carrier. Using the data collected, we can provide a probabilistic estimate of how prone the individual is to infection in terms of their distance from the carrier. This value is integrated along with the CIPC to calculate the overall risk factor.



7.2 Contact Graph-

The Ebola virus in particular is a communicable disease, as we could infer from the earlier discussion that the virus spreads swiftly through physical contacts, transfusion of blood, etc. Thus, examining the number of contacts an individual makes with the carrier would be highly beneficial to our objectives.

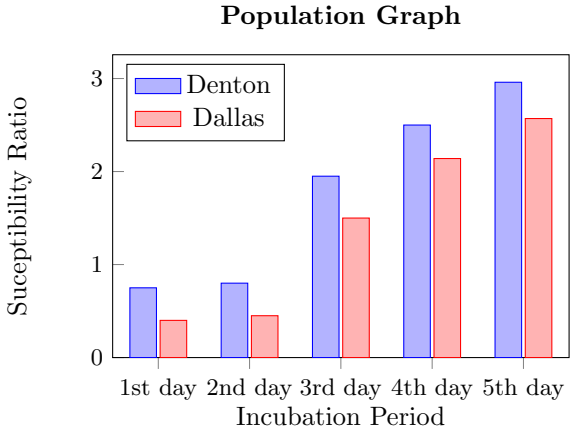
The Contact graph shows an individual A who has made contact with a carrier at various times during a day. The day of exposure, along with the proximity of the individual to the carrier, aids in recognizing the individual's risk factor. Realistically, the time spent with the carrier could be monitored and its significance can be judged on the basis of how close the individual was to the carrier during that time. We can then track the number of contacts and provide a likelihood value to help us derive an overall risk factor.



7.3 Population graph-

The demographic structure of a population is a key determinant of patterns of contact and hence of infectious disease spread, with implications for the design of effective control measures [6]. The population distribution in a given place can alter the risk for an individual. The increased risk maybe associated with a household, the impact of health facilities in the locality, the age of the population, etc.

The sample graph shows the population of Dallas and Denton, where the contrasting demography between the two cities indicates how an individual is at a higher risk if he/she comes in contact with a carrier in Denton as opposed to Dallas. This variation is an implication of how the faclities at Dallas are better than Denton. The awareness and treatment in Dallas is expeditious compared to Denton.



8 Working of the model

(1) **Individual/Carrier Movement:** The individual/carrier movement is tracked in real time by actively monitoring them using the iDid app. The data is stored and the tracking data on previous movements of individuals/carriers is used to calculate the risk factor of contracting the Ebola Virus.

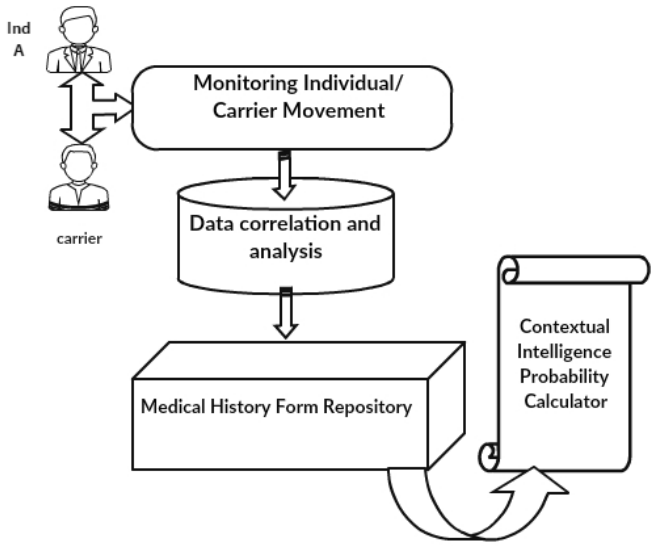


Figure 3: Framework of Application

(2) **Data Correlation:** The data is collected from the individuals and stored in the repository. They are correlated based on the factors that help to analyze the risk of contracting the Ebola virus.

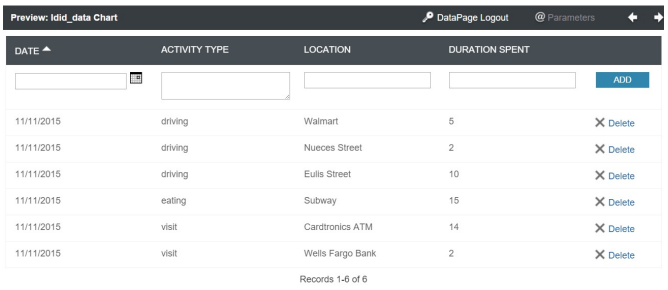
(3) **Medical History Form Repository:** An individual's medical history will support a more accurate analysis of their risk of contracting Ebola from a carrier. Using Bayesian Data analysis, we can predict a more accurate value of their risk.

(4) **CIPC:** The Contextual Intelligence Probability Calculator computes a composite risk by convolving the value generated from the medical history of the individual and the conditional probability calculated from the individual's movement with respect to the carrier at a given moment in time. The type of activity, place of visit and time spent at a place while the carrier is also in motion are correlated along with the main contributing factors such as day of exposure and number of contacts made. The conditional probability obtained is used to provide an accurate value indicating whether the individual is at high or low risk.

9 Experimental Results

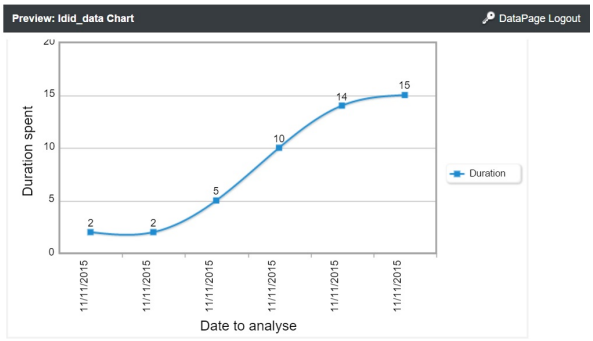
To evaluate the application, a series of tests were conducted and the results were analyzed. Due to the unavailability of real epidemic data, the tests were based on statistical and experimental data sets. The database containing the sample test data is created by tracking the movement of individuals and a mock-up carrier using the iDid application. The following is a series of data collected using the iDid app and stored in the repository.

The sample data shows six rows corresponding to



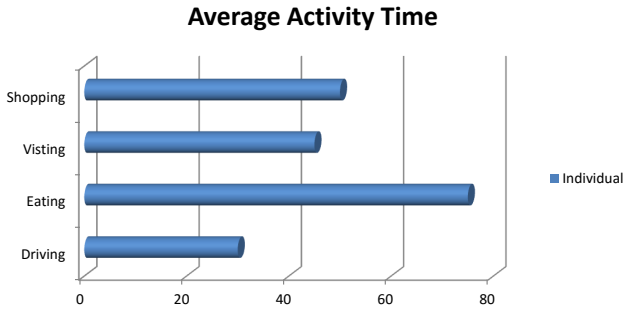
the movement of an individual with respect to the carrier. The user can add/delete rows depending on the correctness of the location. The location can be modified and the values will be updated accordingly.

The application will render a graph showing their proximity to the carrier and the risk probability value.



9.1 Average Activity Time

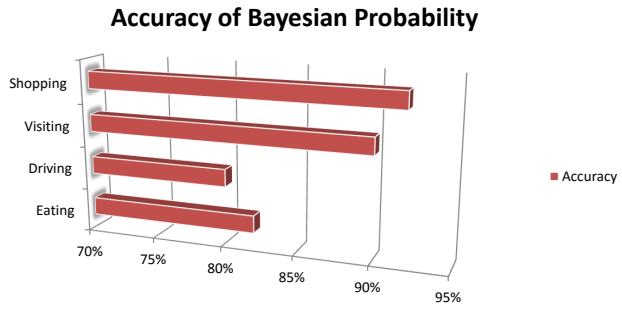
Based on analysis of time spent on various activities, we compute an average time for each activity.



This chart describes the average time an individual spends on a particluar activity. This may be helpful for estimating the duration of untracked activities for which empirical data is not otherwise available. This could reduce the number of false positives when predicting a risk value.

9.2 Accuracy of CIPC(Contextual Intelligence Probability Calculator)

The application monitored a series of individuals located at different proximities to the carrier, then probability values were assigned based on the contributing factors such as time of exposure, number of contacts made and medical history. The Bayesian Conditional probability was measured based on movements with respect to their activities.



This chart describes the accuracy of Contextual intelligence Bayesian analysis based on the different factors monitored for an individual with respect to the carrier's movement.

10 Future improvements

Although the thrust of this paper has been on epidemiological research relating to the factors that contribute to contracting the Ebola virus, there are several other communicable diseases that could affect individuals. One of the main reasons that the Ebola virus spread across Africa was due to the lack of awareness among the population about the severity of the epidemic. There are similar deadly diseases for which people fail to understand the risk factors and unwittingly become carriers. An epidemic of cholera infections was documented in Haiti for the first time in more than 100 years in October 2010. Cases have continued to occur, raising the question of whether the microorganism has established environmental reservoirs in Haiti [3].The patterns of cholera transmission and the seasonality of cholera in an environment is largely based on water contamination, poor sanitation facilities and inadequate hygiene.

Our application described in this paper can be extended to predict contaminated water locations and the probability of individual's susceptibility to Cholera based on location and climatic conditions. In May 2015, the Pan American Health Organization (PAHO) issued an alert regarding the first confirmed Zika virus infection in Brazil and on Feb 1, 2016, the World Health Organization (WHO) declared the Zika virus a public health emergency of international concern (PHEIC). Local transmission has been reported in many other

countries and territories [5]. The Zika virus will likely continue to spread to new areas. Our approach to Ebola risk analysis can be used to analyze risk posed by other diseases such as Zika, by changing the factors that play a role in the spread of the disease, probability curves and other external factors described in this paper.

11 Conclusion

In this paper we described a framework that can be used to assess the risk posed to individual by relating contextual information that tracks the activities of the individual and correlates this data with that of a carrier. We relied on iDid app and demonstrated how our system works. Since actual contextual information on any specific carriers is unavailable, we used made up data and provided users with privacy of data. We used Bayesian models to combine the risks emanating from several factors into a single risk value. We plan to extend our study to model other infectuous diseases

In [4], the report states that in 2014, a team of researchers from Virginia Tech Institute tried to create a model and characterize the nature of the disease outbreak in West Africa. But the research yielded results that did not prove to be accurate. Nevertheless, it can be a stepping stone to understand and analyze an individual's behavior and their movement around the carrier to statistically predict the nature of a disease outbreak. If each individual is able to understand their risk factor and susceptibility to the disease, it could mitigate the possibility of a disease outbreak.

This application reported the ways in which mobile and wireless technologies can be used to implement the vision of pervasive healthcare.

12 Acknowledgements

This research is supported in part by the NSF grant 1513369. The authors also acknowledge the help of Davis Struble for his editorial comments.

References

[1] *Idid Inc.* http://idid-inc.com/businesses/, 2013.

[2] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. *Predicting Flu Trends using Twitter Data.* 2011.

[3] Meer T Alam, Thomas A. Weppelmann, and Chad D. Weber. *Monitoring Water Sources for Environmental Reservoirs of Toxigenic Vibrio cholerae O1, Haiti.* Center for Disease Control and Prevention, 2014.

[4] David Brown. *How Computer Modelers Took On the Ebola Outbreak.* spectrum.ieee.org, 2015.

[5] CDC. *Zika virus and spread.* Center for Disease Control and Prevention, 2016.

[6] Nicolas Geard, Kathryn Glass, and James M McCaw. *Probabilistic graphic models applied to identification of diseases.* Epidemics Volume 13, December 2015, Pages 5664, 2015.

[7] David L Heymann and Guenael R Rodier. *Global Surveillance of Communicable Diseases.* 1998.

[8] Boston Children's Hospital. *Health Map Organisation.* http://www.healthmap.org/site/about, 2006.

[9] Teri Johnson. *Mathematical Modeling of Diseases: Susceptible-Infected-Recovered (SIR) Model.* University of Minnesota, Morris, 2009.

[10] IBM Africa Research Labs. *IBM applies data analytics, mobile technology and cloud computing to help fight the Ebola outbreak in West Africa.* 2014.

[11] Google Maps. *Google Timeline.* Google, 2014.

[12] AFMC Public Health Educators' Network. *Patterns of disease development in a population: the epidemic curve.* 2007.

[13] Renato Cesar Sato and Graziela Tiemy Kajita Sato. *Probabilistic graphic models applied to identification of diseases.* Einstein (So Paulo) vol.13 no.2, 2015.