Geovisualization for cluster detection of Hepatitis A & E outbreaks in Ahmedabad, Gujarat, India

Carl Hughes
Department of Geography, McGill University
+1 (514) 398 4111
carl.hughes@mail.mcgill.ca

Vinayak S. Naik
Indraprastha Institute of Information Technology (Delhi)
+91 (11) 2690 7455
naik@iiitd.ac.in

Raja Sengupta
Department of Geography, McGill University
+1 (514) 398 -5316
raja.sengupta@mcgill.ca

Deepak Saxena
Indian Institute of Public Health Gandhinagar
+91 (079) 4024 0444
ddeepak72@iiphg.org

ABSTRACT

In this paper, we describe a waterborne disease surveillance system for the city of Ahmedabad, Gujarat, India. The proposed system utilizes geocoded disease cases collected using android tablets in an open-source webGIS. Cluster and hot-spot analysis is automated in python and the results get pushed to a cloud-based database for subsequent web-based geovisualization. The end-user is able to interact with the geovisualization module to display individual and aggregated disease data along with related attributes and the locations of any hot-spots. This system meets the need for cost-effective, near-real time disease surveillance in developing countries.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences– *health*

K.4.1 [**Computing Milieux**]: Computers and Society– *public policy issues*

General Terms

Management, Design, Experimentation, Human Factors.

Keywords

Geoweb, disease surveillance, mHealth, India, hepatitis, cluster detection, webGIS.

1. INTRODUCTION

A large segment of India’s populace faces healthcare issues in terms of access, quality, and outcomes. This is reflected in the higher morbidity and infant mortality rates and lower life expectancy for underserved (urban poor, rural and periurban) populations [10]. Such a high disease burden can be attributed to the limited number of trained medical staff, poor access to healthcare facilities, and distance to and costs incurred for seeking medical treatment [1, 15, 16]. Waterborne diseases contribute significantly to this disease burden. World Health Organization (WHO) estimates that 4 000 children die each day around the world as a result of diseases caused by ingestion of contaminated water, mostly in Africa and Asia [16]. [2] also suggest that such serious public health implications are a result of poor urban water management and limited health monitoring.

Disease surveillance is used to detect emerging geographical clusters of disease caused by the random occurrence of risk factors [9]. Effective surveillance can be used to identify a previously unknown disease, risk factor or local existence of known risk factors. [3] argues that developing countries lack adequate surveillance systems mainly due to limited resources, and argues for a fundamental overhaul that capitalizes on real time data updates as well as becoming more cost-effective. The importance of this will inevitably increase the capacity for governments in developing countries, such as India, to better monitor emerging and seasonal outbreaks of communicable disease in a more resourceful and active manner than previous methods. [14] asserts that due to the massive burden of disease in India, it has been difficult to detect, diagnose and control outbreaks of disease until they are quite large. By introducing better automated surveillance mechanisms, interventions to target the risk factors can be implemented more rapidly and accurately with regards to location. In addition, health education programs can better target vulnerable populations according to their specific risk factors.

It is within this context that innovative approaches to healthcare must be investigated for increasing efficiency and efficacy for these underserved communities. With the expansion of mobile phone coverage and user adoption in developing countries, the opportunity for using mobile phones as a health tool (mHealth) has been made possible [11]. [6] identify four types of interventions for which mobile phones are currently being used in developing countries: prevention, disease surveillance, disease management, and patient compliance. Such interventions rely on the widespread availability of mobile phones, data storage, and transmission capabilities including location information and providing quick and cheap forms of direct communication [5]. mHealth initiatives currently being developed and tested make use of short message service (SMS), calling, and mobile phone based applications to record health information. Furthermore, such information when collected at a centralized location (e.g., stored in a database located on a cloud) can be used effectively for data analysis in an automated disease surveillance geovisualization. This system will carry out commonly used disease surveillance methods, such as detecting spatial and temporal clusters [7] and can be used by decision makers in designing specific interventions to control outbreaks.

HealthGIS'14, November 04-07 2014, Dallas/Fort Worth, TX, USA
Copyright 2014 ACM 978-1-4503-3136-4/14/11...$15.00
http://dx.doi.org/10.1145/2676629.2676631

2. STUDY SITE AND CONTEXT

The city of Ahmedabad, Gujarat, India, has a population of about 5.6 million. About 90% of its population is covered by municipal supply of drinking water and sanitation services. In spite of this high coverage, water- and vector-borne diseases continue to recur [12]. In particular, Gastroenteritis and Jaundice (Hepatitis A and E) are two diseases that have frequent and significant outbreaks related to the poor quality of drinking water supply (including municipal piped and ground water supply)[15]. Specifically, viral hepatitis is a major health concern in India [12]. For Ahmedabad, in 2008, 233 cases of hepatitis E infection were identified, with a case rate of 10.9/1000 population [4]. Environmental investigation confirmed sewage contamination of municipal water as the cause for that particular outbreak. Another source of contamination in Ahmedabad is through the use of illegal water pipes that are not properly monitored and maintained [12]. Geocoding the locations of cases is critical in addressing the issue as a cluster of cases can help identify the contamination source. As such, hepatitis A and E are inherently spatial in their distribution and should cluster about contamination sources. Therefore, recognition and early detection of the location of the clusters followed by specific control measures can help reduce outbreak severity and prevent deaths. Such action must also be rapid, accurate and cost-effective.

Currently, disease surveillance in Ahmedabad is carried out through the Integrated Disease Surveillance Programme (IDSP). Syndromic, presumptive and laboratory confirmed cases are reported to a centralized, city-wide database for 20 diseases including hepatitis A and E. Cases reports, including the patient's personal information and address, are supposed to be submitted at the end of each week by government appointed community health activists called link-workers. Quality assurance of the collated data poses a massive barrier for effective use of the system. Instances of multiple entries, missing data, including addresses or disease information, and recording incorrect addresses, often replacing the patient's home address with the hospital or lab address, render much of the data incorrect and unusable. There are also reports of link-workers not knowing how to submit reports and regularly missing the task all-together. Furthermore, the IDSP does not have a clear or standard procedure for analyzing the data to identify outbreaks in a timely manner [14]. In its current state, the IDSP lacks the proper protocols and structure to comprehensively monitor hepatitis A and E in Ahmedabad.

3. METHODOLOGY

We implemented a surveillance system which consists of tablet-based data collection, cloud-based database storage and an automated analysis system for detecting outbreaks of Jaundice Figure 1 presents the system overview. This tool utilizes completely free and mainly open-source software (ODK, QGIS, and pysal) as the core of the system. New methodologies for testing spatial-temporal clusters will continue to be developed and can be easily implemented within the existing system to increase the power of the tool.

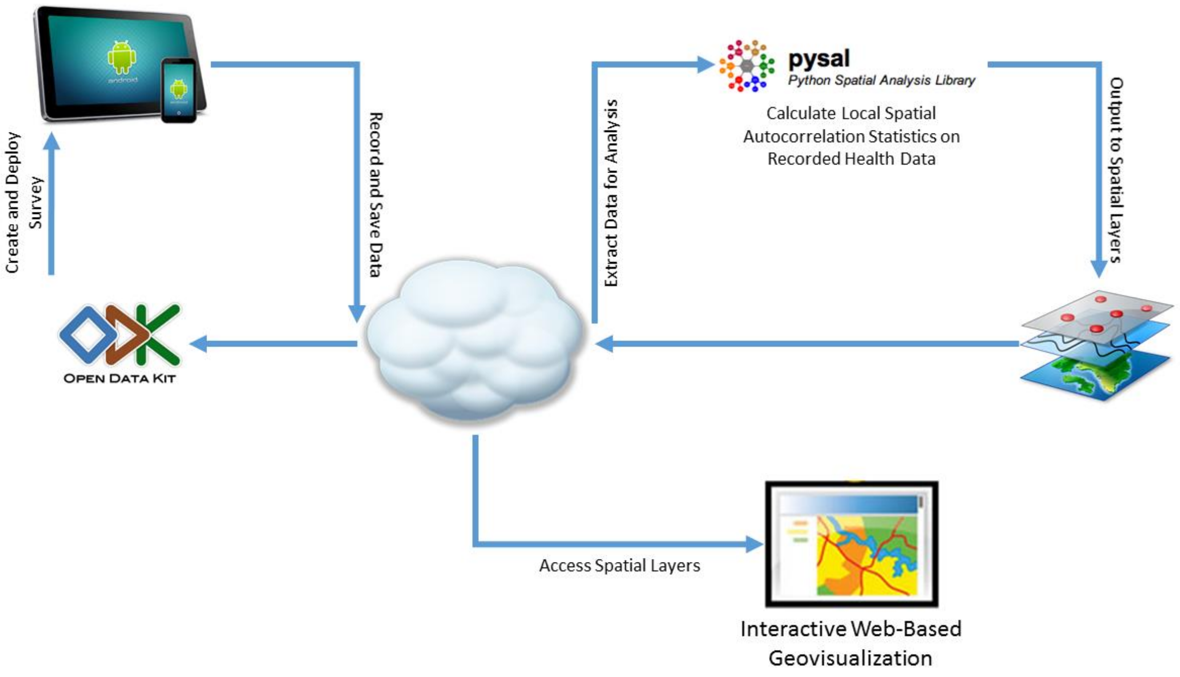

Create and Deploy Survey
Record and Save Data
Extract Data for Analysis
pysal
Python Spatial Analysis Library
Calculate Local Spatial Autocorrelation Statistics on Recorded Health Data
Output to Spatial Layers
OPEN DATA KIT
Access Spatial Layers
Interactive Web-Based Geovisualization

Figure 1. Disease surveillance system structure and framework

3.1.1 Data Collection

As a first step, the Open Data Kit (ODK) software toolbox is being used to build the tablet interface for data collection. Specifically, several data enabled tablets are tailored as a mobile platform to collect information about an affected individual, including the location of the residence, as well as communal water supply locations for those without piped water supply. Other information, such as date of onset and type of symptoms, date of diagnosis, type of doctor who diagnosed the patient, demographic details and socio-economic conditions of surrounding areas, are also captured in the ODK form. The form is first created as an XML file and is pushed into a central ODK aggregate account. Each of the tablets can connect to this account using the secure credentials to access any of the forms pushed into said account. A field worker then interviews a person with a confirmed case of Jaundice as reported to the existing monitoring system, the IDSP, as well as all households within a 10 dwelling radius and using the snowball sampling method. The fieldworker inputs the data using the touch-screen interface of the tablet, including, recording the location which is accessed through the internal GPS of the tablet. A separate form is collected for each individual indicating any of the following symptoms within the past month: jaundice, diarrhea, vomiting, abdominal pain, dark urine, high fever or loss of appetite. A confirmed case is only indicated if the individual has a diagnosis card from a doctor indicating the presence of Jaundice.

Once the data for an individual is collected, the form is submitted electronically using the mobile internet networks (2G/3G) by the field worker. If the tablet does not have internet access at the dwelling, the form can be saved locally and submitted once signal is regained. As soon as the form is submitted, the individual's data is immediately streamed to a centralized cloud-based database.

3.1.2 Spatial Layer Creation

The next step is to convert the patient data into informative and descriptive KML layers for use in the final geovisualization. This process is done using a python program that converts the database into a set of KML layers based on the Jaundice confirmation status. For the privacy and confidentiality of the data, this is done locally and the output remains on the local computer. Initially, the python program executes a custom made Java application that accesses and downloads the latest dataset on the local machine. Using the SimpleKML module, the data, including the home location of the patient is iteratively geocoded (by capturing the latitude and longitude values obtained from the tablet's GPS) and written to a KML file. Based on the rest of the ODK inputs (date of doctor diagnosis, symptoms etc.) a custom KML attribute is created for each patient describing their health profile. The date of diagnosis is also converted into a timestamp element in the KML. The status of Jaundice for each individual is also written as an extended data element within the KML to be accessed for attribute filtering within the geovisualization. The KML file is then saved locally.

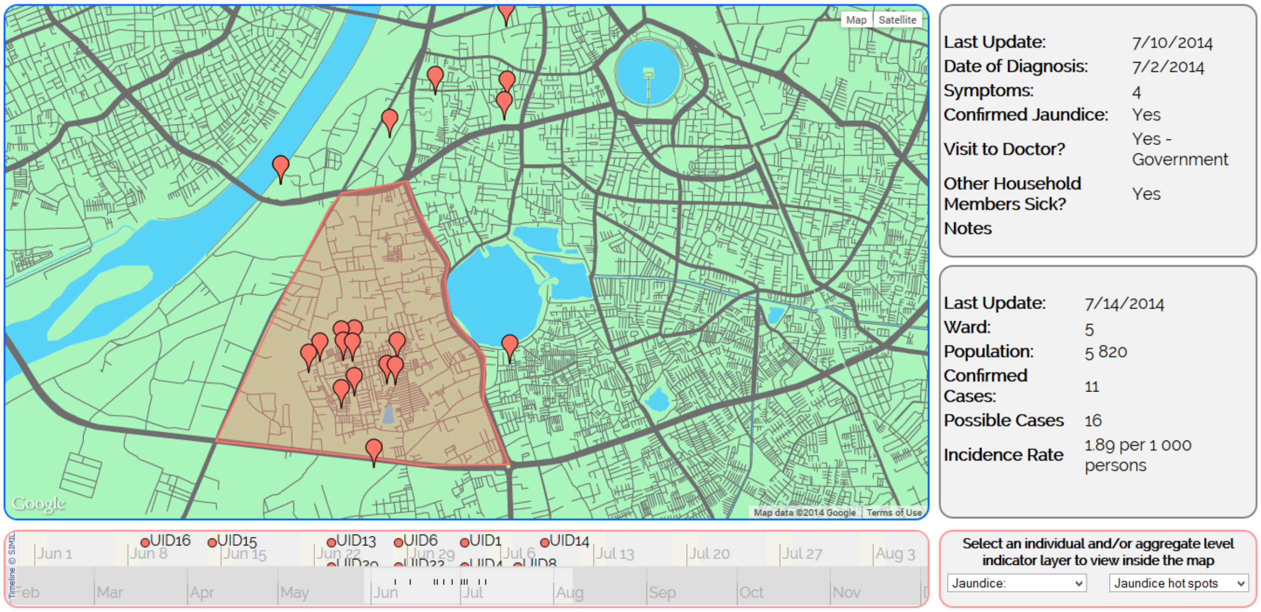
3.1.3 Spatial Analysis

Next, the data is analyzed for clustering and output as KML layers indicating, if any, the location of clusters. Currently, only the established spatial autocorrelation techniques, Global and Local Moran's I, are calculated using pysal. Pysal is the open-source python module of Geoda, a powerful geo-statistical software. New spatio-temporal cluster analysis techniques are currently being developed and will be implemented into the analysis as they are finalized, including a spatial scan statistic and Bayesian method. First, aggregation of cases by municipal ward is done automatically using PyQGIS. All cases diagnosed within a one week period are considered for each possible cluster analysis. The results of the analysis, statistically significant clusters (high-high or high-low) are output into a KML layer. The hot spot analysis will identify spatial units with a statistically significant high rate of incidence compared to what is expected. For comparison, all wards are included in the KML file with ward level data on the number of cases, as well as general demographic information written as text attributes for the KML features. The date of the spatial analysis is also used as a timespan element in the KML.

3.1.4 Geovisualization

Finally, the distribution of affected individuals and statistical hot-spots are presented to end-users in an interactive geovisualization (Figure 2). The geovisualization can help decision makers determine an appropriate course of action, including selection of sites for follow up testing of water quality and to formulate policy suggestions.

The geovisualization is an HTML based web-page that uses the Google Maps API with a customized map style for the base map. The interactive functionality for temporal and attribute filtering is enabled through the TimeMap JavaScript package. This package provides a customizable and interactive timeline that reads the timestamp and timespan elements of features within the KML. The temporal period, range and scale can be controlled by the user through basic mouse interaction. Furthermore, TimeMap enables an attribute filter based on the extended data tags for each KML feature. A simple drop down menu allows the user to turn on or off features according to the status of jaundice diagnosis (confirmed – public doctor, confirmed - private doctor or possible yet undiagnosed) or other attribute including presence of symptoms, age, gender or if other household members have hepatitis-like



Last Update: 7/10/2014
Date of Diagnosis: 7/2/2014
Symptoms: 4
Confirmed Jaundice: Yes
Visit to Doctor? Yes - Government
Other Household Members Sick? Yes
Notes

Last Update: 7/14/2014
Ward: 5
Population: 5 820
Confirmed Cases: 11
Possible Cases 16
Incidence Rate 1.89 per 1 000 persons

Select an individual and/or aggregate level indicator layer to view inside the map

Jaundice: ▾ Jaundice hot spots ▾

Figure 2. Sample screenshot of working geovisualization interface indicating a hot spot and individual cases of jaundice

symptoms. The same extended data tags are also used to apply appropriate symbology to the KML features. The user can turn on or off both the individual level and aggregated layers as is required. As new patient data is streamed to the database (i.e. a new bout of Jaundice) and the geovisualization is re-opened, the KML will subsequently be updated reflecting any changes in the data. Additionally, the new data point will be considered in the spatial analysis. Each new point on this dynamic online map will thus represent new disease cases. Figure 2 shows an example scene of the proposed geovisualization.

4. RESULTS AND DISCUSSION

By capitalizing on the plethora of open-source technology, webGIS tools and the widespread availability and cost of mobile phones, this system can provide a cheap, accurate, powerful and reliable disease surveillance tool.

4.1.1 Cost

In low-resource settings, such as India, the cost of disease surveillance often poses as an insurmountable barrier for effective public health monitoring. To overcome this challenge, we propose a system that can be implemented with minimal costs. The cost of mobile phones or tablets and their maintenance, as well as the wages of field workers are the only associated monetary costs. Even so, the price for mobile phones and tablets in India continue to decrease with fierce market competition and innovation. Furthermore ODK can run on any Android device running Android 2.2 or higher, meaning even the most basic, and thus, cheapest mobile technology are adequate for this purpose. In order to implement the system at a city scale, enough fieldworkers must be employed to reach all parts of the city, or at least the most problematic wards. For the example of Ahmedabad, labour costs can be considerably reduced by relying on the already existing municipal link-workers (i.e., community health activists hired by the municipal government to provide basic health education and advice to residents in mainly slum or informal settlement areas of the city). They are also responsible for submitting weekly reports to the IDSP on confirmed cases of certain diseases, including jaundice, as reported by public doctors within their community boundary. They are not trained medical professionals, but have received some basic training and are usually older women and longtime residents of the community they serve meaning they have substantial social capital. These link-workers could prove invaluable in mobile phone based disease surveillance due to their local knowledge and current activities placing them in the most affected areas in the first place. Thus, they could be provided with the ODK enabled devices for data collection. This system can also benefit from the existing social networks of the link-workers to monitor as large a population as possible and reduce any biases in the analysis from undetected cases.

4.1.2 Accuracy

The current disease reporting system for the city of Ahmedabad relies on weekly reports submitted by link-workers and public health providers on confirmed cases of jaundice and other diseases. The reporting is done periodically and often cases are missed or incorrectly reported. Subsequently, entries without the patients' home location, incomplete or incorrect home locations (often the hospital or doctors address is listed instead) are entered. This is often consolidated with multiple entries of the same person and possible missing entries, errors in disease diagnosis or other important information from data entry issues.

Geocoding private addresses using traditional GIS-based methods in India becomes almost impossible owing to the unstructured, frequently changing and relative (i.e. opposite hotel X) address structure. It is more complicated within slum-like areas where there is no hierarchical data structure and there is a lack of reference datasets or GIS infrastructure to validate the addresses. Even an attempt to geocode an address within Google Maps cannot be verified with any degree of accuracy in lieu of a field measurement. Often, when a complete address is provided there are multiple societies or apartment buildings with the same name, the street name can have several different spellings or, as is common, the point of reference in the address may be a colloquially termed feature. Obviously, not knowing the exact location of jaundice cases poses a huge challenge in effectively monitoring outbreaks.

Utilizing the accurate GPS capabilities of mobile technology, we were successful in overcoming this barrier in recording the exact coordinates of the individual's residence while in the field. This process is rapid (under 30 seconds), does not require any technical expertise as ODK is able to access the device's internal GPS data, and is extremely accurate. Even when collecting GPS coordinates in slum-like areas with narrow alleyways and a high density of informal residences, the accuracy of the GPS data was between two and ten metres. In collecting the data directly from the field, the steps of copying a paper based report to electronic form and geocoding the location information are not necessary and, as a result, the accuracy of the database vastly improved.

4.1.3 Utility and Reliability

The utility of this system comes from its ability to automate the main processes of disease surveillance in a near-real time manner. Through the ODK XML form structure, logic can be used for conditional questioning, as well as many types of data input including photos, videos and audio recordings. By storing the database in the cloud, the data is secure, can be accessed remotely and is not limited in its size. The system can accurately analyze the data spatially by automating the process within python code. Having the ability to easily implement alternative methods into the existing structure is one of the systems key features. The drawback of limiting the analysis to the local and global Moran's I method is that it cannot account for population differences and is at risk of reporting false positives. In the future, the spatial-scan statistic and Bayesian methods should be tested in the system as they may better account for the predicted pattern of hepatitis A and E cases and small population problem.

Converting the data and analysis results into the KML spatial format makes it relatively easy to perform additional advanced GIS analysis on the data, as well as the possibility of implementing a truly open-source system through the use of the OpenStreetMaps API for the geovisualization. Having a near-real time surveillance system will also prove extremely effective in providing quick public health responses. As soon as data is collected in the field and submitted, it is included in the final geovisualization. When intervening in outbreaks of hepatitis A and E, a delay of even one day can result in new cases unnecessarily. Instead, public health officials can correctly identify the location of outbreaks to intervene in an attempt to control the contamination at its source, while warning residents in the area of the danger.

5. INTENDED USE AND BENEFITS

This proposed disease surveillance system has the potential to provide an opportunity for public health agencies, government departments and NGOs to better monitor disease outbreaks and case distribution for more effective intervention. Decision makers will be able to easily interact with the geovisualization to load specific descriptive spatial layers, compare results temporally by defining a particular time period to overlay data and analyze aggregate level data such as incidence rate and progress that has been made in attempts at lowering the incidence rate. From here, decisions can be made regarding a specific intervention including the allocation of medical resources, increased health education programs or the repair of contaminated water pipes.

As the system utilizes completely free software, technical implementation will come at minimal cost. The spatial methods that are automated within the system have been selected as per conventional disease surveillance and epidemiological methods [2,8]. The novelty of automating this process is that it bypasses the need for the technical expertise and analysis from a GIS professional which can be a limiting factor in a resource-scarce setting [13]. The only requirements for implementation include a computer with default market specifications and an internet connection and one or more (preferably data-enabled) tablet devices. If the tablet does not receive a data connection, e.g., while collecting data as in urban settings with low signal strength or in rural area, the forms can still be collected and uploaded once it receives a signal or Wi-Fi connection, or directly copied via a USB cable to a computer.

In capturing the raw data electronically and directly uploading it to a cloud-based server, a comprehensive database can be maintained more effectively and efficiently than traditional paper-based reporting. This avoids the tedious task of data entry and validation from paper based surveys. The potential for human error is limited to the initial data input from the field. Although the collection interface is fairly simple (only one question per page), training must be provided to data collectors to ensure familiarity with the android device, ODK application and data inputs and submission. Fortunately, the existing widespread usage rates of smart-phone technology in Ahmedabad provides a positive starting point for this task. Once the data is submitted it can also be validated automatically or by person to ensure the inputs make sense and are in the appropriate format, value range or not conflicting with another value. ODK has enabled easy integration of data submissions into an existing database through direct streaming compatibility with common tools such as REDCap, and jSON publishers. New results are simply appended to the main database. Data security and confidentiality can also be ensured by restricting access to the main database and subsequent geovisualization through granting secure login credentials for authorized users. The tablet devices can also be password protected to prevent access to un-submitted forms as well as preventing the use of games, music and other non-work related applications by fieldworkers.

Finally, the use of this system is not limited to any particular disease. In theory, it could be used for any type of spatial monitoring including, for example, environmental groups locating endangered bird species nesting sites or by civic participation organizations, for example, reporting the location of derelict public infrastructure in need of repair. Furthermore, governments or community or activist groups could open up data submission for their initiatives to citizen participants as the system provides an efficient and direct method for collecting and utilizing VGI. The data contributor need only to have access to an android based device and connect it to the project's ODK account. Thus, in addition to a disease surveillance tool, this system could be used for a disease reporting tool from citizens to further increase public health monitoring.

6. ACKNOWLEDGMENTS

We would like to thank the city of Ahmedabad for allowing us to test our proposed system. We would also like to thank The National Geographic's Committee on Research and Exploration, MITACS and the Indian Ocean World Centre and the Social Sciences and Humanities Research Council (SSHRC) for their generous financial support.

7. REFERENCES

[1] Agarwal, S., and Lau, C. 2010. Remote health monitoring using mobile phones and web services. *Telemedicine and e-Health*. 16, 5 (Jun, 2010), 603-607. DOI=10.1089/tmj.2009.0165

[2] Alirol, E., Getaz, L., Stoll, B., Chappuis, F., and Loutan, L. 2011. Urbanisation and infectious diseases in a globalised world. *The Lancet infectious diseases*. 11, 2 (Feb, 2011), 131-141. DOI= http://dx.doi.org/10.1016/S1473-3099(10)70223-1

[3] Butler, D. 2006. Disease surveillance needs a revolution. *Nature.* 440, 7080, 6-7. DOI=10.1038/440006a

[4] Chauhan, N. T., Prajapati, P., Trivedi, A.V., and Bhagyalakshmi, A. 2010. Epidemic investigation of the jaundice outbreak in Girdharnagar, Ahmedabad, Gujarat, India. *Indian Journal of Community Medicine*. 35, 2 (Apr 2010), 294–297. DOI=10.4103/0970-0218.66864

[5] Chhabra, E. 2013. Ubiquitous across globe, cellphones have become tool for doing good. In *The New York Times. Retrieved* (Nov 8, 2013) from http://www.nytimes.com/2013/11/08/giving/ubiquitous-across-globe-cellphones-have-become-tool-for-doing-good.html

[6] Déglise, C., Suggs, L. S., and Odermatt, P. 2012. SMS for disease control in developing countries: A systematic review of mobile health applications. *Journal of Telemedicine and Telecare*. 18, 5 (Jul, 2012), 273-281. DOI= 10.1258/jtt.2012.110810

[7] Elliot, P., Wakefield, J. C., Best, N. G., and Briggs, D. J. 2000. *Spatial epidemiology: methods and applications*. Oxford University Press., Oxford, UK.

[8] Elliott, P., and Wartenberg, D. 2004. Spatial epidemiology: current approaches and future challenges. *Environmental health perspectives*. 112, 9 (Jun, 2004), 998-1006. DOI=10.1289/ehp.6735

[9] Kulldorff, M. 2001. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 164, 1, 61-72. DOI=10.1111/1467-985X.00186

[10] Patil, A. V., Somasundaram, K., and Goyal, R. 2002. Current health scenario in rural India. *Australian Journal of Rural Health*, 10, 2 (Jul, 2002), 129-135. DOI= 10.1046/j.1440-1584.2002.00458.x

[11] Patnaik, S., Brunskill, E., and Thies, W. 2009. Evaluating the accuracy of data collection on mobile phones: A study of forms, SMS, and voice. In *Proceedings of The Information and Communication Technologies and Development (ICTD), 2009 International Conference*. (Doha, April 17 - 19, 2009). IEEE, 74-84. DOI= 10.1109/ICTD.2009.5426700

[12] Saravanan, V. S., Mavalankar, D., Kulkarni, S., Nussbaum, S., and Weigelt, M. 2014, *Metabolized-water breeding diseases in urban India.* Working Paper. University of Bonn. ISSN=1864-6638

[13] Singh, A., Naik, V., Lal, S., Sengupta, R., Saxena, D., Singh, P., and Puri, A. 2011. Improving the efficiency of healthcare delivery system in underdeveloped rural areas. In *Communication Systems and Networks (COMSNETS), 2011 Third International Conference.* (Bangalore, January 4 - 11, 2011). IEEE, 1-6. DOI= 10.1109/COMSNETS.2011.5716519

[14] Suresh, K. 2008. Integrated Diseases Surveillance Project (IDSP) through a consultant's lens. *Indian journal of public health*. 52, 3, 136-143.

[15] Tandon BN, Gandhi BM, Joshi YK, Irshad M, and Gupta H. 1985. Hepatitis virus non-A, non-B: The cause of major public health problem. *India. Bulletin of the World Health Organization*. 63, 5, 931–934.

[16] World Health Organization. 2009. *Global health risks: mortality and burden of disease attributable to selected major risks*. World Health Organization., Geneva, CH.