Computational epidemiology: Bayesian disease surveillance

Kaja Abbas, Armin R. Mikler, Amir Ramezani, Sheena Menezes

University of North Texas, Denton TX 76203, USA

Email: kaja@cs.unt.edu , mikler@cs.unt.edu, ar0116@unt.edu, srm0034@unt.edu

Abstract

Disease monitoring plays a crucial role in the implementation of public health measures. The demographic profiles of the people and the disease prevalence in a geographic region are analyzed for inter-causal relationships. Bayesian analysis of the data identifies the pertinent characteristics of the disease under study. The vital components of control and prevention of the disease spread are identified by Bayesian learning for the efficient utilization of the limited public health resources. Bayesian computing, layered with epidemiological expertise, provides the public health personnel to utilize their available resources optimally to minimize the prevalence of the disease. Bayesian analysis is implemented using synthetic data for two different demographic and geographic scenarios for pneumonia and influenza, that exhibit similar symptoms. The analysis infers results on the effects of the demographic parameters, namely ethnicity, gender, age, and income levels, on the evidence of the prevalence of the diseases. Bayesian learning brings in the probabilistic reasoning capabilities to port the inferences derived from one region to another.

1. Introduction

Computational epidemiology forges the epidemiological and computational sciences in the study of diseases. Algorithms are used to develop computational tools to interface with large databases for the retrieval of pertinent information. Data mining techniques are applied on the data to reveal the correlation of the different parameters resulting in the varied epidemiological events. The proposed Bayesian disease surveillance model shall uncover the implicit features from the explicit characteristics and identify the points of control, prevention, and surveillance of diseases.

The demographics and the cases affected by a disease in a geographic region are recorded on to a database. Once the data set is available, Bayesian algorithms can be applied to learn the correlations between the different demographic parameters and their effects on the disease prevalence. The Bayesian network can be analyzed to identify the critical features of the demographic parameters that are affected by the disease. These features can be ordered by their levels of significance. In general, knowledge learned from a disease outbreak in a given area can not be readily ported on to a different geographic area due to variations in demographics.

The constrained resources of the public health departments have to be optimally allocated.[9] The most pertinent demographic parameters for control of the disease are learned from Bayesian analysis. This shall be an invaluable asset to the health department for efficient utilization of their resources.

The Bayesian network identifies the control features for a specific demographic and geographic scenario. The learned Bayesian network can be imported on to a different demographic and geographic scenario to predict the incidence of the disease. The Bayesian network shall adapt to the different scenario, and the corresponding control aspects of the diseases can be identified.

Section 2 gives an overview of past epidemics and traditional mathematical approaches to modeling of epidemics. Section 3 discusses the conceptual reasoning of Bayesian learning and illustrates the demographic analysis of diseases using Bayesian networks, along with the inferred results for two different geographic and demographic settings. Section 4 discusses related work in Bayesian learning techniques employed over the domain of public health. Section 5 concludes the paper along with directives for future work.

2. Epidemiology

2.1. History

The human race has exacted astronomical casualties to epidemic outbreaks. The 14th century Europe lost a quarter of its 100 million population to Black Death. The fall of the Aztecs empire in 1521 was due to smallpox that eradicated half of its 3.5 million population. The pandemic influenza of 1918 caused a fatality of 20 million in twelve months. More recently, the Severe Acute Respiratory Syndrome (SARS) outbreak of 2003 highlighted the rapid spread of an epidemic at the global level. The outbreak, emanating from a small Guangzhou province in China, spread around the world requiring a concerted response from public health administrations around the world and World Health Organization (WHO) to curtail the epidemic.[11]

2.2. Mathematical Epidemiology

The early 20th century laid the foundations of the mathematical theory of epidemics.[17] The initial work from 1900 to 1930 in epidemiology had a deterministic character. The probabilistic aspects of the different processes were not included. From 1930, binomial distributions were used to represent the incidence of specific diseases, thereby, stochastic modeling found its roots in the study of epidemics. The deterministic and the stochastic models are based on the principles of susceptibles, infectives and removals, namely, the SIR model. Susceptibles are those individuals in the population who can be infected by the disease under investigation. Infectives are those individuals that have become infected by the disease and are capable of transmission of the disease pathogen to the susceptibles. Removals encompass all individuals who are no more able to transmit the infection onto susceptibles, and are either in a state of recovery, fully recovered, or expired on contraction of the diseases. It is often assumed that removals acquire lifelong immunity. However, in more complex models, surviving removals on recovery will be included in the susceptibles. Epidemiological triangle model[14,20] for infectious diseases comprises of the host, environment, and agent. The inter-relationships between the above parameters are included in the evaluation and analysis of the diseases.

2.3. Deterministic and Stochastic Models

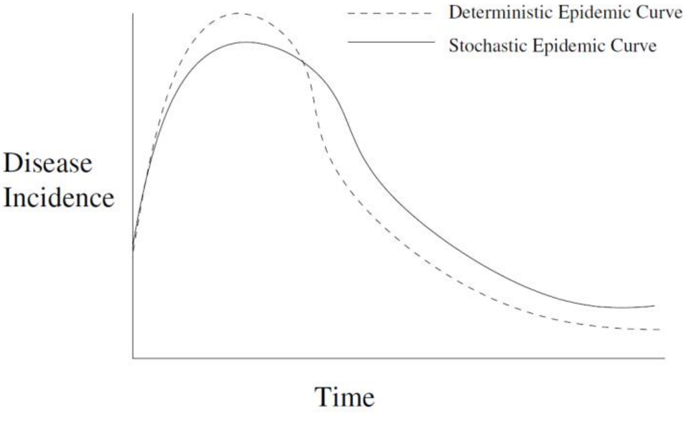

Figure 1. Epidemic curves for deterministic and stochastic models

The deterministic model used basic differential equations to model the epidemic progression. The behavior of stochastic models is similar to deterministic models for larger number of susceptibles and infectives. The stochastic models differ from the deterministic models, in that they provide a closer real life reasoning and modeling of the spread of infectious diseases, by introducing the probability metrics into the deterministic model. The epidemic curve illustrates the incidence of a disease, that is, the number of new infectives in a population of a given area over a period of time. Figure 1 shows the epidemic curves of both the models. From a computational perspective, stochastic models need high performance computing to realize their full potential. The stochastic models and the analysis of the epidemic curves have proved to be a primary mathematical framework for visualizing infectious diseases outbreaks, even in the 21st century.[2] The stress for alternative and multiple models for epidemiological thinking and reasoning has been highlighted.[18]

3. Bayesian learning

3.1. Bayesian principles

Bayesian philosophy incorporates the capabilities of probabilistic reasoning, and reasoning under uncertainty. Bayes' theorem updates the existing belief in a hypothesis, given the evidence, by use of probability distributions. In other words, the prior probabilities of a hypothesis transforms to posterior probabilities, taking into account the evidence. There are several good resources of study available for Bayesian artificial intelligence.[12,13,15]

A Bayesian network models the different parameters of the domain in the form of nodes, with directed links between them reflecting their dependencies. The nodes are associated with the corresponding probability distributions for their beliefs. The network can be built from a data set that contains a number of records with data for the different parameters.

The flow of information in the Bayesian network leads to varied inference methodologies. Diagnostic reasoning involves backward reasoning from the effects to the causes. Predictive reasoning uses the information of the causes leading up to the effects. Inter-causal reasoning is the role of mutual causes on a common effect. The above types of reasoning can be combined in any desired ways.

The SIR model is oriented towards the study of infectious diseases. Bayesian learning can be incorporated for the analysis of all diseases in varied demographic and geographic scenarios.

3.2. Applications

Bayesian learning has been successfully applied in the areas of medical diagnosis, weather forecasting, gaming, and fault diagnosis in various domains. Pathfinder[10] is a Bayesian expert system used for lymph-node pathology diagnosis. Hailfinder[1] is a weather forecasting system for severe summer hail applied in northeastern Colorado. The chip producer, Intel, uses Bayesian networks for fault diagnosis of semi-conductor chips.

3.3. Bayesian analysis

The Bayesian network (Figure 2) analyzes the effects of the demographic parameters on the incidence of symptoms and the related diseases in a geographic area. The demographic parameters are ethnicity, gender, age, and income; symptoms are cough and fever; and diseases are pneumonia and influenza. The Bayesian network illustrates the predictive reasoning of the demographics on the prevalence of diseases. Table 1 defines the list of symbols used for the parameters and their corresponding values. Severe disease outbreaks in two smaller geographic areas are considered. The artificial data has been synthetically generated and is not reflective of any real demographic and geographic settings.

3.3.1. Scenario I

Table 2 includes the beliefs for the demographic parameters in geographic area I. The probability distributions for the symptoms and diseases are given in Table 3. For example, P(F/G,A,I) refers to the conditional probability of fever, given the evidence of gender, age, and income. The Bayesian network is analyzed to derive useful inferences on the prevalence of pneumonia and influenza. The population affected by pneumonia and influenza are 11.21% and 8.84% respectively.

The values of each of the demographic parameters are ordered in their levels of significance on the outcome of the two diseases (Figure 3). The relative levels of significance within each demographic parameter are derived by setting each individual value to absolute unity, that is (P(value) = 1). All other parameters' beliefs are kept the same as before, and the proportional changes in the people affected by the diseases are inferred. Hence, by raising the whole community to be Hispanics, a decrease in pneumonia (-3.08%) and influenza (-3.89%) infections are observed.

Among the different ethnicities, the Black ethnic subgroup is observed to be at most risk for both pneumonia and influenza, followed by Asians, Whites and Hispanics. A similar analysis for the other demographic parameters gives a suit of significant results. In case of gender, females are at a relatively higher risk in comparison to males for both pneumonia and influenza. For age, children exhibited a higher risk in comparison to adults for both the diseases. The lower income people are inferred to be more likely infected by both diseases, compared to the higher income people.

The demographic parameters can be further combined and ordered in their levels of importance in the spread of diseases. Based on the artificial data for area I, lower income Black female children are at the higher end of the risk spectrum, while higher income Hispanic male adults are at the lower end of the spectrum for pneumonia. The spread of influenza also exhibited similar results in geographic area I.

3.3.2. Scenario II

Table 4 defines the demographic probability distributions of geographic area II. Similar to area I, area II is experiencing an epidemic of pneumonia and influenza. The prevalence of the two diseases, pneumonia and influenza, are currently unknown. The posterior probability distributions for the symptoms and the diseases of area I are ported into the Bayesian learning process for area II. The developed Bayesian network can then be analyzed for the role of demographics on the two diseases.

The infected population of pneumonia and influenza are 12.06% and 9.67% respectively. Figure 4 shows the relative levels of significance of the values of each demographic parameter. Considering ethnicity for both the diseases, Hispanics exhibit the highest risk, followed by Asians and Whites, while Blacks have the least risk. In case of gender, females show higher risk to both diseases in comparison to males. Children exhibited lower risk to pneumonia compared to adults, while adults have a lower risk to influenza in comparison to children. The lower income groups are observed to be more prone to both diseases in relation to the higher income groups.

On analysis of the synthetic data for area II, lower income Hispanic female adults have a higher risk of pneumonia infections, while higher income Black male children have a lower risk. For influenza, lower income Hispanic female children have a higher risk, while higher income Black male adults are at a lower risk. Although the data is hypothetical, the differences in the critical risk groups of the two areas indicates the significance of analyzing the demographic parameters. For instance, if the surveillance and preventive measures developed for area I had been applied to area II, the more critical groups of area II would have been less addressed.

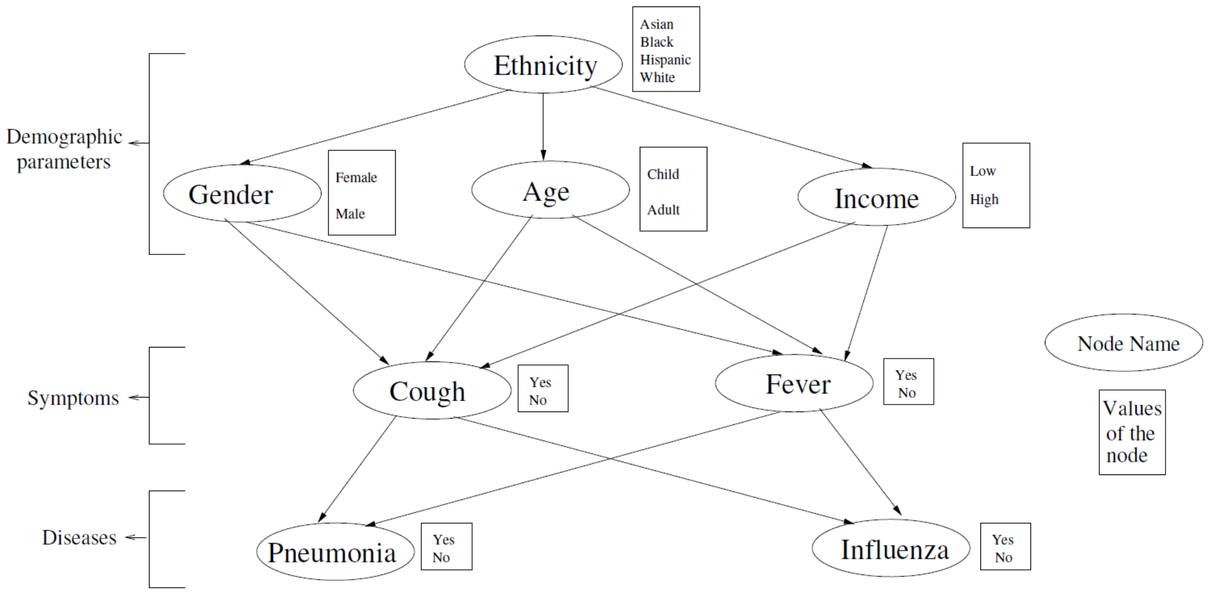

Figure 2. Bayesian network for demographic analysis of diseases

Table 1. Symbols for parameters and parameter values

Parameter	Symbol	Parameter Value	Symbol
Ethnicity	E	Asian	As
Gender	G	Black	Bl
Age	A	Hispanic	Hi
Income	I	White	Wh
Cough	C	Female	Fe
Fever	F	Male	Ma
Pneumonia	P	Child	Ch
Influenza	IN	Adult	Ad
		Low	Lo
		High	Hi
		Yes	Ye
		No	No

Table 2. Probability distributions of demographics for scenario I

E	P(E)	G	E	P(G/E)	A	E	P(A/E)	I	E	P(I/E)
As	0.15	Fe	As	0.55	Ch	As	0.25	Lo	As	0.40
Bl	0.20	Fe	Bl	0.60	Ch	Bl	0.15	Lo	Bl	0.55
Hi	0.30	Fe	Hi	0.46	Ch	Hi	0.30	Lo	Hi	0.30
Wh	0.35	Fe	Wh	0.56	Ch	Wh	0.23	Lo	Wh	0.35
		Ma	As	0.45	Ad	As	0.75	Hi	As	0.60
		Ma	Bl	0.40	Ad	Bl	0.85	Hi	Bl	0.45
		Ma	Hi	0.54	Ad	Hi	0.70	Hi	Hi	0.70
		Ma	Wh	0.44	Ad	Wh	0.77	Hi	Wh	0.65

Table 3. Probability distributions of symptoms and diseases

G	A	I	P(C/G,A,I)	G	A	I	P(F/G,A,I)
Fe	Ch	Lo	0.35	Fe	Ch	Lo	0.75
Fe	Ch	Hi	0.25	Fe	Ch	Hi	0.40
Fe	Ad	Lo	0.88	Fe	Ad	Lo	0.44
Fe	Ad	Hi	0.05	Fe	Ad	Hi	0.85
Ma	Ch	Lo	0.15	Ma	Ch	Lo	0.54
Ma	Ch	Hi	0.85	Ma	Ch	Hi	0.64
Ma	Ad	Lo	0.54	Ma	Ad	Lo	0.27
Ma	Ad	Hi	0.64	Ma	Ad	Hi	0.20
	C	F	P(P/C,F)		C	F	P(IN/C,F)
	Ye	Ye	0.30		Ye	Ye	0.25
	Ye	No	0.10		Ye	No	0.05
	No	Ye	0.10		No	Ye	0.10
	No	No	0.001		No	No	0.00

Table 4. Probability distributions of demographics for scenario II

E	P(E)	G	E	P(G/E)	A	E	P(A/E)	I	E	P(I/E)
As	0.20	Fe	As	0.60	Ch	As	0.15	Lo	As	0.55
Bl	0.15	Fe	Bl	0.55	Ch	Bl	0.25	Lo	Bl	0.40
Hi	0.35	Fe	Hi	0.75	Ch	Hi	0.60	Lo	Hi	0.75
Wh	0.30	Fe	Wh	0.40	Ch	Wh	0.46	Lo	Wh	0.64
		Ma	As	0.40	Ad	As	0.85	Hi	As	0.45
		Ma	Bl	0.45	Ad	Bl	0.75	Hi	Bl	0.60
		Ma	Hi	0.25	Ad	Hi	0.40	Hi	Hi	0.25
		Ma	Wh	0.60	Ad	Wh	0.54	Hi	Wh	0.36

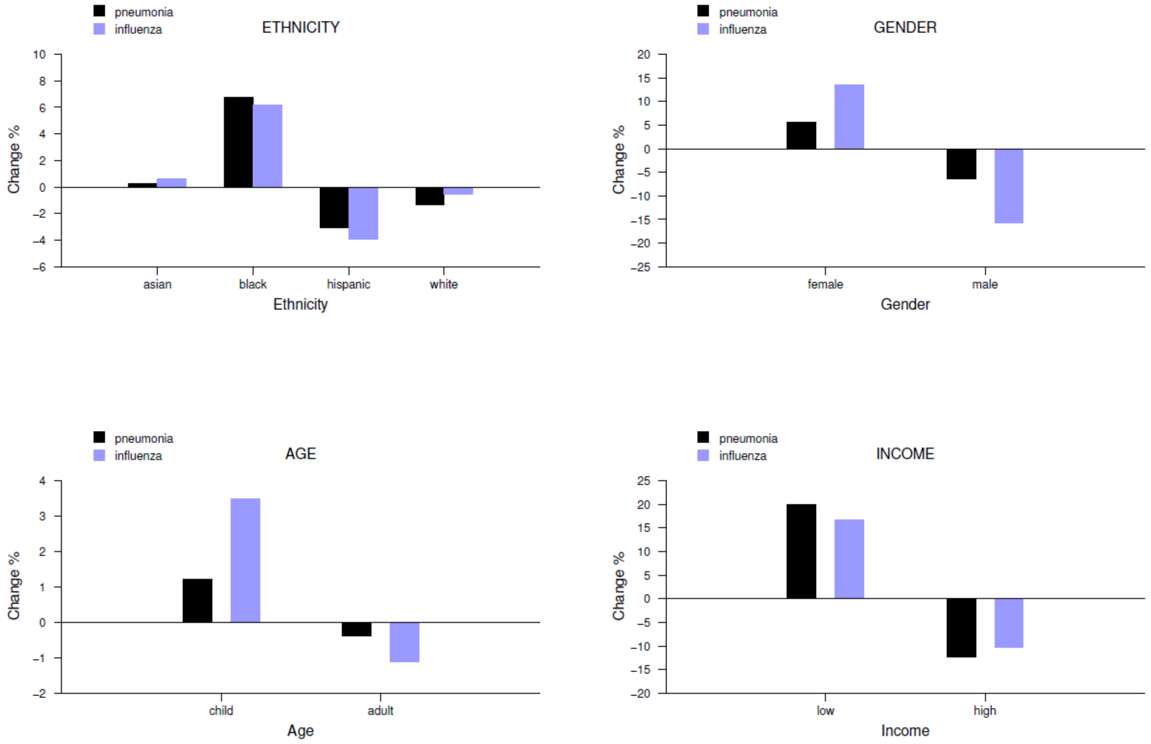

Figure 3. Bayesian analysis for geographic area I

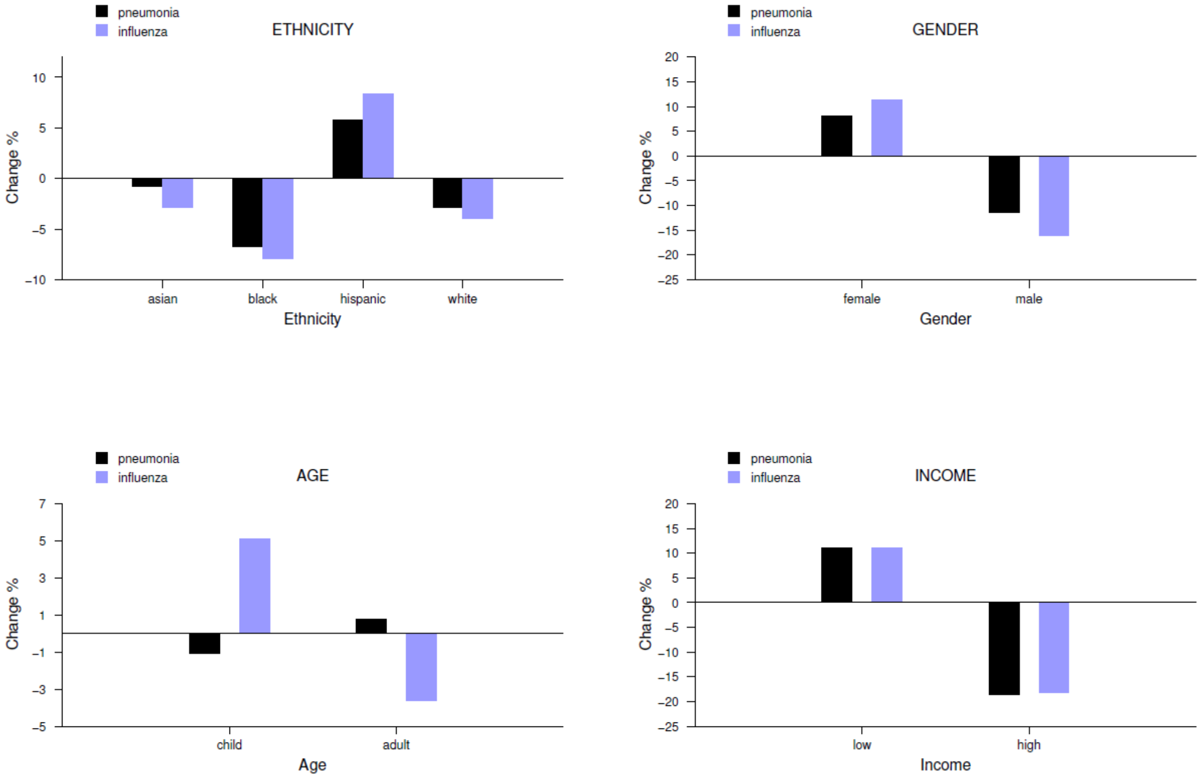

Figure 4. Bayesian analysis for geographic area II

4. Related work

Microfilariae had been studied in the Amazonian focus of onchocerciasis (river blindness) to identify the communities that need priority ivermectin treatment.[6] A Bayesian hierarchical model for human onchocerciasis was developed to investigate the role of individual and community characteristics in the infection. The model aids in research and control planning for the public health department as well as in its policy decision making. Bayesian analysis for health technology assessment has been investigated,[19] and highlights the practical advantages of the Bayesian approach in handling complex interrelated problems. Bayesian classifiers are used in the Real-time Outbreak and Disease Surveillance (RODS) system,[21] a computer based public health surveillance system that detects disease outbreaks. RODS had been used in the 2002 Winter Olympics. Pennsylvania and Utah currently use RODS for public health surveillance.

In social science hypothesis testing, the increase in independent variables for regression models leads to misleading errors, while Bayesian approximation reduces the uncertainty in error.[16] Bayesian concepts are used to calculate the risks of leukaemia following chemotherapy for Hodgkin's disease, based on case-control studies.[3] Bayesian monitoring of critical factors in cancer related clinical trials, such as toxicity and quality of life measures, led to higher accuracy.[8]

An epidemiological model using Bayesian analysis has been developed for the disease, plasmodium falciparum malaria, in Ndiop, Senegal.[5] The incidence of cancer in multiple cities had been collected from the survey data from the state of Sao Paulo, Brazil.[4] A correlation analysis using Bayesian methods between the multiple cancer sites estimated the cancer rate in a given area. The results had a better precision compared to the usual methods.

5. Conclusion

The probabilistic reasoning Bayesian methodology aids in the identification of the critical points for control, prevention, and surveillance of diseases. The limited resources of the public health department can be aptly used in the order of the identified high risk groups to derive the best gains. The Bayesian network learned from a specific demographic and geographic settings for a disease outbreak can be transferred to different demographic and geographic settings. The analysis of the adapted network helps in the identification of the control points for different demographic and geographic settings.

The critical groups identified at higher levels of risk are different for the two geographic regions. This underlines the significance of analyzing the demographics to discover the higher risk spectrum of the population. The high risk groups are to be accorded prime attention to curtail the epidemic. Consequently, it is imperative to develop more tools that allow epidemiologists to extrapolate the findings across multiple geographic regions, thereby allocating the public resources efficiently.

Future work includes the correlation between Bayesian networks learned from different demographic and geographic settings to predict the spatial flow of the disease. Also, disease characteristic features for the rate of spread of the disease shall be incorporated to determine the temporal flow of the disease.

Acknowledgements

We would like to thank Fabio G. Cozman for the use of his software package, JavaBayes,[7] to analyze the Bayesian networks.

References

1. B. Abramson, J. Brown, W. Edwards, A. Murphy, and R. Winkler, *International Journal of Forecasting*. **Hailfinder: A Bayesian system for forecasting severe weather**, 12(1): 57-72 (1996).
2. J. L. Aron, *Infectious Disease Epidemiology: Theory and Practice*. **Mathematical modeling: The dynamics of infection**, Aspen Publishers, MD, (2000).
3. D. Ashby, J. Hutton, and M. McGee, *The Statistician*. **Simple Bayesian analyses for case-control studies in cancer epidemiology**, 42:385-397, (1993).
4. R. Assuncao, and M. Castro, *International Journal of Epidemiology*. **Multiple cancer sites incidence rates estimation using a multivariate Bayesian model**, 33: 508-516, (2004).
5. N. Cancre, A. Tall, C. Rogier, J. Faye, O. Sarr, J. Trape, A. Spiegel, and F. Bois, *American Journal of Epidemiology*. **Bayesian analysis of an epidemiologic model of Plasmodium falciparum malaria infection in Ndiop, Senegal**, Volume 152, Issue 8 760-770, (2000).
6. H. Carabin, M. Escalona, C. Marshall, S. Vivas-Martinez, Carlos Botto, Lawrence Joseph, and Maria-Gloria Basanez, *Bulletin of the World Health Organization*. **Prediction of community prevalence of human onchocerciasis in the Amazonian onchocerciasis focus: Bayesian approach**, Volume 81, Number 7, 473-550, (2003).
7. F. Cozman. **JavaBayes**, http://www-2.cs.cmu.edu/javabayes/JavaBayes.
8. P. Fayers, D. Ashby, and M. Parmar, *Statistics in Medicine 16*. **Tutorial in biostatistics: Bayesian data monitoring in clinical trials**. 1413-1430, (1997).
9. L. Gordis, W.B. *Saunders Company*. **Epidemiology**, (2000).
10. D. Heckerman, *MIT Press*. **Probabilistic similarity networks**, Cambridge, MA, (1991).
11. D.L. Heymann and G. Rodier, *Emerging Infectious Diseases*. **Global surveillance, national surveillance, and SARS,** Volume 10, Number 2, February (2004).
12. K. Korb, and A. Nicholson, *CRC Press*. **Bayesian artificial intelligence**, FL, (2004).
13. R. Neapolitan, *Pearson Prentice Hall Series*. **Learning Bayesian networks**, NJ, (2004).
14. K. Nelson, C. Williams, and N. Graham, *Aspen Publishers*. **Infectious disease epidemiology**, 51-53, (2001).
15. J. Pearl, *Morgan Kaufmann*. **Probabilistic reasoning in intelligent systems**, San Mateo, CA, (1988).
16. A. Raftery, *Sociological Methodology*. **Bayesian model selection in social research**, 111-196, (1995).
17. R. Ross, *Proc. Roy. Soc*. **An application of the theory of probabilities to the study of a priori pathometry**, I, A, 92, 204-230, (1916).
18. K. Rothman, *Oxford University Press*. **Epidemiology An introduction**, 1-7, (2002).
19. D. Spiegelhalter, J. Myles, D. Jones, and K. Abrams, *BMJ*. **An introduction to Bayesian methods in health technology assessment**, Volume 319, 508-512, 21 August (1999).
20. T. Timmreck, *Jones and Bartlett Publishers*. **An introduction to epidemiology**, (2002).
21. F. Tsui, J. Espino, V. Dato, P. Gesteland, J. Hutman, and M. Wagner, *Journal of the American Medical Informatics Association*. **A real-time public health surveillance system**, 10(5): 399-408, September (2003).