Acute Diarrheal Syndromic Surveillance

Effects of Weather and Holidays

H.J. Kam[1]; S. Choi[2]; J.P. Cho[2]; Y.G. Min[2]; R.W. Park[1]

[1]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea; [2]Department of Emergency, School of Medicine, Ajou University School of Medicine, Suwon, South Korea

Keywords

Surveillance, diarrhea, forecasting, environment, emergency service hospital

Summary

Objective: In an effort to identify and characterize the environmental factors that affect the number of patients with acute diarrheal (AD) syndrome, we developed and tested two regional surveillance models including holiday and weather information in addition to visitor records, at emergency medical facilities in the Seoul metropolitan area of Korea.

Methods: With 1,328,686 emergency department visitor records from the National Emergency Department Information system (NEDIS) and the holiday and weather information, two seasonal ARIMA models were constructed: (1) The simple model (only with total patient number), (2) the environmental factor-added model. The stationary R-squared was utilized as an in-sample model goodness-of-fit statistic for the constructed models, and the cumulative mean of the Mean Absolute Percentage Error (MAPE) was used to measure post-sample forecast accuracy over the next 1 month.

Results: The $(1,0,1)(0,1,1)_7$ ARIMA model resulted in an adequate model fit for the daily number of AD patient visits over 12 months for both cases. Among various features, the total number of patient visits was selected as a commonly influential independent variable. Additionally, for the environmental factor-added model, holidays and daily precipitation were selected as features that statistically significantly affected model fitting. Stationary R-squared values were changed in a range of 0.651-0.828 (simple), and 0.805-0.844 (environmental factor-added) with $p<0.05$. In terms of prediction, the MAPE values changed within 0.090-0.120 and 0.089-0.114, respectively.

Conclusion: The environmental factor-added model yielded better MAPE values. Holiday and weather information appear to be crucial for the construction of an accurate syndromic surveillance model for AD, in addition to the visitor and assessment records.

Correspondence to:

Rae Woong Park, M.D., Ph.D.
Department of Biomedical Informatics,
Ajou University School of Medicine,
Wonchon-dong, Yeongtong-gu, Suwon, Gyeonggi-do,
442-749, Korea
Tel.: +82-31-219-5342
Fax: +82-31-219-4472
E-mail: veritas@ajou.ac.kr

Appl Clin Inf 2010; 1: 79–95
doi: 10.4338/ACI-2009-12-RA-0024
received: December 12, 2009
accepted: April 6, 2010
published: April 14, 2010

Citation: Kam HJ, Choi S, Cho JP, Min YG, Park RW. Acute diarrheal syndromic surveillance – effects of weather and holidays.
Appl Clin Inf 2010; 1: 79–95
http://dx.doi.org/10.4338/ACI-2009-12-RA-0024

1. Background

1.1 Syndromic Surveillance

Owing to global heightened concerns about possible bioterrorist attacks and emerging infectious diseases, syndromic surveillance has become a growing field. Syndromic surveillance protocols are currently being developed to serve several functions: early outbreak detection, monitoring of the size, spread, and tempo of outbreaks, monitoring disease trends, and providing the public with reassurance [1-2]. The principal objective of syndromic surveillance is to identify specific illness syndromes early-before confirmed diagnoses are made and reported to the public health authorities and then to mount a rapid response, thereby minimizing morbidity and mortality in such an event [2-5].

1.2 Current Surveillance Systems

A variety of sizes and forms of surveillance systems have been developed worldwide [6-7]. In this section, we compare the existing systems and several conditions to be satisfied with prolonged, simultaneous, and stable syndromic surveillance, and identify and explain the limitations of these current systems.

1.2.1 Automatic and Near-real Time Surveillance

In general, syndromic surveillance systems fall into three categories, according to the surveillance period and data collection technique employed [5]. First, a short-time, drop-in surveillance is a labor-intensive monitoring method, which is used for specific events. The Lightweight Epidemiology Advanced Detection and Emergency Response System (LEADERS) [8] is one example of such a drop-in surveillance system. Second, passive surveillance can be conducted via manual gathering or transmitted fax records, such as the Syndromic Surveillance Tally Sheet [9], the NHS Direct Service-based surveillance system used in England and Wales [10], and the system utilized in the Southeastern Virginia region [11]. The latter is an automated and (nearly) real-time surveillance system which operates via system networks. For continuous, stable, and in-time surveillance, automated surveillance structures are optimal, and can provide standardized real-time data and reduce labor costs. Recently, many systems have adopted an automatic or semi-automatic daily data transmission protocol scheme [12-15].

1.2.2 Public Use and Nation-wide or Scalable Regional Area

Some currently-operating surveillance systems are based on military hospitals, such as the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE) [16], or LEADERS [8]. There are also some systems based on a single hospital, minor hospital union [17-18], a small region [19], or a single metropolitan city such as New York [15, 20-21], Paris [22], Minneapolis [14], Washington [23], and Boston [24-25]. A truly nation-wide surveillance system requires a great deal of public and national data that can be scaled for the nation-wide surveillance of various syndromes; however, thus far the Taiwanese system [12] is really the only system that can be reasonably said to accomplish this.

1.2.3. Consideration of Living Environments for Proper Data Analysis

There have been several types of data source are available for surveillance such as Over-the-counter (OTC) pharmaceutical sales, ambulatory visits, and emergency department (ED) visits. In regions like the USA, a great many surveillance techniques are predicated on OTC pharmaceutical sales [26-28]. However, in regions including Taiwan and Korea, patients tend to visit ED rather than buying OTC drugs on weekends and holidays, owing to the lower cost of ED visits as compared to the enormous costs in countries such as the USA. Therefore, in such regions, a greater proportion of the entire population could be covered by an ED patient-based surveillance regime, as compared to an OTC drug data-based scheme.

Moreover, the characteristics of the data can vary considerably, according to the medical circumstances of each country and the data acquisition techniques employed. It has been previously reported that patient visitation patterns look different on weekdays and weekends than on regular

weekdays [14], and surges in visitation rates have also been noted on specific holidays [12]. In accordance with those observations, one should consider weekend and holiday information specifically, rather than simply the weekly number of patient visits. Additionally, a seasonal variance of patterns in syndromes has been reported. Several studies have been shown the importance of the seasonality or the climatic factors in the rotavirus epidemiology in the tropics [29] as well as in region such as Venezulela [30] and Saudi Arabia [31]. In Korea, a statistically significant difference in acute diarrheal syndrome according to changes in temperature [32] has also been identified. In a similar fashion, considering that many environmental factors can influence ED visits, a variety of factors should be reflected in syndromic surveillance; very few models have thus far been suggested that include those factors [33].

1.2.4 Standardized Code Structure

In many studies, the chief complaints or diagnosis in free text were used for syndromic surveillance [11, 13, 15, 34-35]. However, free textual data requires human intervention, which, in most cases, presents an obstacle to the provision of information [11]. It also prevents data integration among heterogeneous terminologies of different institutions. Therefore, a standardized code system is required, for a wide range of syndromic surveillance. For this reason, recent surveillance systems tend to use ICD or UMLS codes as a standardized code system [12, 14, 18, 36].

1.2.5 Low Construction and Maintenance Costs

For long-term, continuous, and automatic surveillance, participation and budgeting from many institutions, states, or federal governments are necessary. In particular, syndrome monitoring can prove extremely difficult because of high initial investment costs or un-automated methodologies, followed by high demands for manpower and maintenance costs. Here are some cases in which the cost of system construction and management has been mentioned, although such cases are rare. In the NHS Direct Surveillance system [10], as mentioned previously, the direct annual cost is reported to be approximately $280,000 (USD), including personnel expenses for participating researchers. In reference to the public health practices of the New York region [15], annual maintenance costs of $130,000 (USD) were reported. It is possible to reduce the construction and maintenance costs for a syndromic surveillance system when we utilize pre-existing information networks among hospitals or centralized data integration systems. The application of an automated surveillance model can reduce the time required for system construction, and can cut down on the costs associated with manual monitoring. Meanwhile, such a model would prove economic and efficient because of the automatic monitoring that takes place every 24 hours. Automated surveillance models or algorithms will be addressed in section 1.2.6.

1.2.6 Automatic Outbreak Detection Algorithms

Many statistically-based detection algorithms have been proposed for use in automatic surveillance systems [6]. Some of these were applied forms of basic regression or moving average (MA), some were widely used in quality management areas such as CUSUM, EWMA, and some others-such as ARIMA, SARIMA, Spatial-Scan, or Wavelet analysis-were utilized or integrated into a detection algorithm. In reality, however, so few cases of syndromic outbreaks have occurred that there are no available reference standards against which to compare the surveillance data [7]. Comparative studies among a variety of algorithms would be required for the development of an effective surveillance system; it is difficult to compare and evaluate various algorithms under identical conditions. Some comparative studies of surveillance algorithms have previously been conducted [18, 24, 37-38], and no algorithm has been developed that yields absolutely superior outcomes for every syndrome; thus, surveillance algorithms can be said to be currently in a “Warring States Period”.

1.3 Syndromic Surveillance in Korea

The use of syndromic surveillance in Korea initially began as an attempt to prepare for possible bioterror attacks during international events such as the World Cup football games in 2002 and the Busan Asian Games of 2002 [5]. The Korea Center for Disease Control and Prevention (KCDC) has operated sustainable syndromic surveillance systems, which have, since 2002, run by 125 sentinel EDs throughout Korea [39]. However, an important limitation of the current syndromic surveil-

lance system is that the process of gathering the data for syndromic surveillance is largely conducted manually. Therefore, an extra burden on medical teams, such as emergency physicians or nurses, in addition to patient care, always exists; the syndromic surveillance data must be collected and reported, and proper education must be implemented. As this system can induce a depreciation of reliability on data with fluctuating accuracy, and because active data collection protocols can bias physician's decisions, there is a clear need for the development of an automated and timely data collection system.

2. Objective

The principal objective of this study was to identify the environmental factors that affect the number of patients with acute diarrheal (AD) syndrome, and to construct a new prediction model with greater predictive power that includes the environmental factors discovered herein. The principal objective of this study was to develop an automatic syndromic surveillance system that satisfies the pre-acquisitions, such as automatic and near-real time, scalable, considering living environment, standardized, and inexpensive surveillance, as referenced in section 1.2. In order to confirm whether the included environmental factors actually improved the predictability with regard to the number of patients with AD, we developed and evaluated two regional surveillance models for the Seoul metropolitan area in Korea. We constructed AD syndrome models including holidays and weather information in addition to visiting and assessment records based on the National Emergency Department Information system (NEDIS).

3. Methods

The study flow is depicted in ▶Figure 1. More detailed explanations for each step and process will be addressed from sections 3.1 to 3.4.

3.1 Data Acquisition from NEDIS

NEDIS is a pre-constructed operating nationwide data collection system that contains every patient's ED visit information; the system operates nationwide, in approximately 170 emergency medical facilities in Korea [40]. The data of NEDIS is composed of 23 items, including patient information and hospital visits, initial assessments-including patients' chief complaints and vital signs-and medical consultations, including ED discharge diagnoses based on ICD-10 codes. These data are transmitted automatically every week to the National Emergency Medical Center (NEMC). A variety of data has been gathered daily from approximately 170 emergency medical facilities throughout the nation (Here, the participating emergency medical facilities include representative Emergency Medical Centers (EMCs) such as Regional EMCs, Local EMCs, and Specialty EMCs).

In this study, we developed two regional surveillance models based on the data of emergency medical facilities and the visiting patients (per day) for the Seoul metropolitan area in Korea. Seoul has a flat topography, which is not blocked out or segmented by mountains or ocean: The climate of the Seoul area is generally fairly constant among its sub-regions. Seoul is Korea's capital and largest metropolitan city, with an area of 605.33 km^2. Because approximately 20% of the entire population of the Korean peninsula lives in this single city, the impact from any variety of biological terrors will be severe here: intensive monitoring will be required for syndromic surveillance of the Seoul area. In order to monitor the patterns of AD in the Seoul metropolitan area, data on patient visitations between May 1, 2007 and April 30, 2009 was extracted from NEDIS. The number of participating emergency medical facilities in Seoul increased gradually from 24 to 31 during that period. We constructed a data mart including patient/institution, visiting records, early assessment, ED treatment information, and final diagnosis results in the NEDIS system after the de-identification of the patients.

Among the data obtained from NEDIS, we utilized coded ED discharge diagnosis to select out AD patients among the population of visiting patients: ED discharge diagnosis codes were based on

ICD-10. The reason that the ED discharge diagnoses on ICD-10 code are used-as opposed to the chief complaint, which is generally used in syndromic surveillance models-is that there is no nationally standardized code for chief complaints generally used among Korean institutions. The diagnostic codes for AD were classified by ED doctors, based on the information provided by the Centers for Disease Control and Prevention of America and ESSENCE (based on ICD-9), and the diagnostic codes were mapped on the basis of ED-discharge diagnosis information (by ICD-10), as shown in ▶Table I. The diagnostic codes, daily data regarding the number of institutions, and AD patients in the Seoul metropolitan area were processed on the basis of ED patient visitation information from NEDIS.

3.2 Additional Variables

Besides the daily number of institutions, total patient visits and ED patient visits, other variables expected to contribute to patients' visiting patterns were also prepared and processed.

3.2.1 Day of the Week and Holidays

As introduced in section 1.2.3, there is a trend toward increasing patient visitation on weekends and holidays. With the 7-day periodic seasonal effect, information on national holidays-including the biggest national holidays such as New Year's Day (on both solar and lunar calendars), the Korean Thanksgiving (August 15th of the lunar calendar) period, Memorial Day and Independence Day were collected to be reflected in the surveillance models as events (discontinuous incidences).

3.2.2 Weather Information

Actually, many kinds of syndromes have been identified reflective of a seasonal relationship with environmental effects such as temperature, humidity, or precipitation [41-48]. Thus, some of those factors were additionally considered in this study. Environmental factors could vary substantially; daily information about several weather components – such as average temperature, highest/lowest temperature, temperature difference, relative humidity, precipitation, average wind speed, and sunshine duration – were obtained from the Korean Meteorological Administration (KMA) [49], and applied for further analyses and model construction.

3.3 Modeling

As previously mentioned, we focused on what needed to be considered for an effective surveillance model, via in-depth analysis using a single algorithm. A seasonal ARIMA (SARIMA) time-series forecasting model was selected as the target algorithm for this study, considering the 7-day periodicity, significant autocorrelation from the patient sequence graphs, and other environmental factors to be reflected. ARIMA was considered a suitable representative of a susceptible model, because interventions or events can be readily included therein. Seasonal autoregressive integrated moving average (SARIMA) models extend basic ARIMA models and permit the incorporation of a repetitive pattern, such as the observed weekly pattern in the number of daily ED patients [33]. The structure of the SARIMA model is represented by the following notation: $(p,d,q) \times (P,D,Q)_s$, in which P, D, and Q, respectively, provide additional information regarding the seasonal autoregressive, differencing, and moving average components of seasonal level for the model [50]. More detailed information concerning model construction is also referenced [51].

The proposed surveillance models were constructed using SPSS software package Time Series analysis (Release 15.0.0, SPSS Inc., Illinois, USA). Firstly, two prediction models were constructed from May 1, 2007 to April 30, 2008 (1 year): (1) A simple model with total patient number, (2) An environmental factor-added model containing additional holiday and weather information, as well as the number of total patients. All of the variables applied for the construction of the environmental factor-added model were selected in accordance with section 3.2. Among the inserted variables, valid variables for the two ARIMA models were selected according to the values of parameter estimates with $p<0.05$ using the SPSS ARIMA module. The ARIMA model is utilized for short-term, rather than long-term prediction. Therefore, we gradually increased the data for model training with 1-month intervals (starting at May 1, 2008, which became the 13th month from the initial data) to the month just prior to the target prediction period, and constructed new best-fitting pre-

diction models for each of the following months, as shown in ▶Figure 2: For example, the ARIMA model based on the data for 12 months was utilized for the prediction of the 13th month (May, 2008) and the 13-month model (from May, 2007 to May, 2008) was used for the 14th month (June, 2008). The applied independent variables for each model were exactly the same; the formula of each of the ARIMA models can be changed according to the data updates, because the parameters and coefficients of each variable can be influenced by variations in the data.

3.4 Evaluation

3.4.1 Model Fitting Statistics (Stationary R-squared)

The stationary R-squared is the in-sample (the target period: May 1, 2007~April 30, 2008 for the 1st model training, and increases by 1 month for the following models) model goodness-of-fit statistics, with higher values representing better fit.

3.4.2 Mean Absolute Percentage Error (MAPE)

Those models were validated via comparisons of the MAPE values of the predicted numbers of patients and actual visiting patients with the syndrome. The MAPE is a scale-independent statistic that expresses the prediction error as a percentage. For a series of predicted values ($\breve{y}_1, \breve{y}_2, ..., \breve{y}_n$) and the corresponding series of observed values ($y_1, y_2, \ldots, y_n$) [33].

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|(y_t - \breve{y}_t)/y_t\right|.$$

A MAPE value of 0% indicated a perfect fit of the model to the test (prediction) dataset, which is indicative of a perfect prediction. Additionally, 'very precise', 'relatively precise' or 'relatively reasonable' predictions are those with MAPE values of up to 10%, up to 20%, and up to 50%, respectively.

4. Results

4.1 Data Characteristics

The characteristics of the data during the initial testing (2007.05~2008.04) and prediction period (2008.05~2009.04) in the Seoul metropolitan region were shown in ▶Table 2. Additionally, in ▶Figure 3, the sequence graph of actual visiting AD patients was also demonstrated. As anticipated, the daily pattern of visiting patients was characterized in terms of weekly seasonality. The number of visiting patients had a tendency to increase with passing time, and thus we applied the logarithm transformation: this is because the number of participating NEDIS facilities gradually increased, and the changes (variance) in patient numbers did not occur in a linear relationship with this tendency, as mentioned in section 3.1. Additionally, first-order seasonal differencing was also applied to remove the 7-day periodicity: a pattern graph that moved around the 0 point, as shown in ▶Figure 3(B). The ratio of AD patients to total ED patients evidenced a tendency to increase on weekends, particularly on Sundays, as is shown in ▶Figure 4, with box plots for each of the days of the week.

4.2 In-Sample Model Goodness of Fit

After reflecting the periodicity, the $(1,0,1)(0,1,1)_7$ ARIMA model evidenced adequate model fit (refer to the values of stationary R-squared below) for the daily number of AD patient visits over 12 months for both of the model cases. Among a variety of imported features, the total number of patient visits was selected as a commonly influential independent variable. Additionally, for the environmental factor-added model, holidays (events) and daily precipitation were selected as the features that statistically significantly affect model fit. The in-model goodness-of-fit values (station-

ary R-squared) for the models are shown in ▶Figure 5. The stationary R-squared values changed between 0.651 and 0.828 (for the simple model) and between 0.805 and 0.844 (for the environmental factor-added model) with $p<0.05$.

4.3 Post-sample Forecast Accuracy

The result and estimation of predictions of the numbers of visiting patients from May, 2008 to April, 2009 are shown in ▶Figure 6 with MAPE values: the MAPE values between the predicted and actual numbers of patients were expressed as cumulative means by monthly undated ARIMA models. The environmental factor-added model predicted the number of visiting patients more accurately than the other model. In ▶Figure 6, the difference between the two models was depicted via a linear graph: The most profound difference was noted in August, 2008 and the environmental factor-added model evidenced lower MAPE values – in other words, better predictability. The predictions of these models were on the borderline between 'very accurate' and 'relatively accurate'.

5. Discussion

5.1 Constructed Models and Their Performances

The values of in-model goodness-of-fit showed scant difference between the simple model and the weather-supplemented model. However, as additional factors-such as day of the week and weather information-were added, performance in terms of prediction period improved: the weather-supplement model evidenced the best predictive power for the number of AD patients. Accordingly, a model that included environmental factors such as precipitation or temperature differences would be expected to be better than the simple model.

5.2. Influence of Living Environments

The HPMG network (14) evidenced delayed results in an increased caseload on Monday, which was caused by the same factor. In our models, as shown in ▶Figure 3, the total number of visiting patients and those with AD in the ED fluctuated widely according to the days of the week, and evidenced sharp increases on weekends and special holidays: this appears to be the result of the closure of private clinics on weekends and holidays, particularly in Korea where medical treatment at an ED is relatively inexpensive. A national medical insurance is obligatory for every Korean. The cost for a medical treatment is the cheapest among the OECD member nations: for example, it costs around $50~100 (USD) for GI diseases such as diarrhea. Hence, the application of this periodicity into the surveillance model would be expected to generate better results.

We also demonstrated that the addition of certain environmental factors could improve the performance of a surveillance model with precipitation. It has been recently determined that there exist seasonal differences in the patterns of syndromes [29-32, 41-42, 52] and that precipitation significantly influences the occurrence of acute diarrheal disease [44-48]. Additionally, the number of total ED patients can be correlated with precipitation (Pearson's correlation coefficient: -0.105, $p<0.05$, data not shown herein). A $(1,0,0)(0,1,1)_7$ ARIMA model (with precipitation variable) was constructed for the prediction of the total numbers of ED patients based on 12 months of patient visitation data (May, 2007-April, 2008), with a stationary R-squared value of 0.718 and a MAPE value of 5.165. ▶Figure 4 shows that a rapid increase in precipitation occurred in June, 2008; additionally, we monitored subsequent increases in MAPE and in the difference between the two models in the following month (August, 2008). These phenomena can be explained by the effects of precipitation on the number of total ED patients and those with acute diarrhea: The changes in precipitation are reflected indirectly in the simple model; the changes were also directly reflected in the environmental factor-added model that included that information as a variable.

Considering those results, it can be surmised that the environment/weather affects infectious disease transmission, and also that environment/weather factors such as precipitation influence human behavior (including that of AD patients). Therefore, a good syndromic surveillance model

should attempt to take into account a variety of environmental factors, in addition to weather information.

5.3 The Efficiency and Effectiveness of an Automatic Surveillance System

An automatic system can provide confidence by minimizing the human errors that can occur in the reporting steps, and can also reduce the costs associated with manual monitoring. NEDIS can provide real-time data for an automated surveillance system with high report rates and accountability, continuous monitoring of patient fluctuations, and a scalable nation-wide surveillance system via multi-institutional information gathering.

5.4 Limitations

In this study, the ED-discharge diagnosis with code ICD-10 was used to monitor a distinct syndrome. However, for more accurate and timely surveillance, the chief complaints of UMLS codes can be employed for more accurate and more rapid syndrome prediction [53], as previously mentioned. Because the principal focus of this study was to confirm some of the effects of medical and living environments, we characterized some of the implications of these environmental effects. We are currently planning to generate more mature surveillance models that involve continuous data collection and monitoring. Additionally, toward a more accurate surveillance, a more predictable algorithm should be provided, which takes into consideration appropriate climatic and environmental variables. For that purpose, more research and investigation will be required for the development of more appropriate time-series algorithms and the effects of environmental factors.

The ability to predict the number of patients with a specific syndrome can be linked to the detection of abnormal patterns for a surge in the number of incidents. We constructed ARIMA prediction models to conduct syndromic surveillance for acute diarrhea: thus, our study could not be extended to the stage of detection or evaluation of alerts, owing to a lack of cases of acute diarrhea due to bio-terror (thus far). Although comparisons with other algorithms, such as CUSUM or EWMA which were mentioned in section 1.2.6, should be utilized for evaluation as a surveillance mode, this should be a matter for further studies, because it is beyond the scope of this research, which aims to evaluate the effects of weather and holidays. Additionally, surveillance algorithms such as CUSUM and EWMA are strategies for assessing unusual behavior of the time series, not modeling techniques: there is a limitation to be evaluated by the same criteria.

6. Conclusions

Toward the development of an effective automatic syndromic surveillance system, we developed and evaluated two regional surveillance models for AD with additional weather and holiday information in a major Korean metropolitan city prior to the construction of a nationwide surveillance system. According to those seasonal ARIMA models, the environmental factor-added model most closely predicted the number of visiting patients, as compared to the other two models. The results of this study showed that holiday and weather information could improve the performance of syndromic surveillance models for AD, in addition to the visitation and assessment records. Weather information can provide more accurate predictions for syndromic surveillance, and must be included for the construction of an accurate AD syndromic surveillance model. Also, in addition to its original objective, NEDIS can provide real-time data for an automated nationwide surveillance system via multi-institutional information gathering.

Acknowledgements

This research was supported, in part, by the strategic research project of the Korea Center for Disease Control and Prevention (KCDC) for a study of the strategy for introducing an automated syndromic surveillance system. Additionally, the authors thank the National Emergency Medical Center for cooperation on NEDIS data.

Clinical Relevance

Besides the original objective of NEDIS, the application of an automated nationwide surveillance system can be employed for syndromic surveillance with real-time data for thorough multi-institutional information gathering. Holiday and weather information might improve the performance of syndromic surveillance models for AD, in addition to visitation and assessment records.

Conflict of Interest

All of the authors of this work hereby declare that there were no conflicts of interest or any other relationships that could have inappropriately influenced this study.

Human Research / IRB section

This study employed the summarized number of patients of total visiting patients and those with acute diarrheal syndrome for each of the participating institutions, and utilized no identifications or patients' personal information. Therefore, this study is not applicable to a range of IRB reviews. Additionally, we obtained a privacy and confidentiality review process from NEDIS for data acquisition.

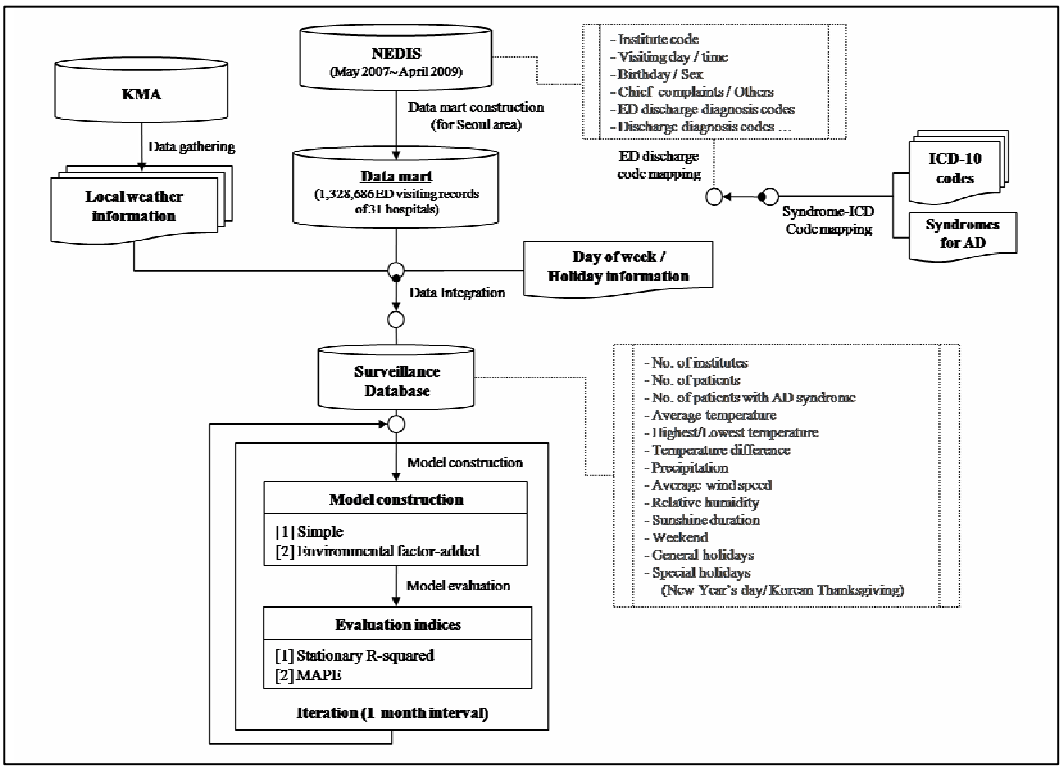

Fig. 1 Overview of model construction and evaluation process. NEDIS = National Emergency Department Information System; ED = Emergency Department; KMA = Korea Meteorological Administration; MAPE = Mean Absolute Percentage Error; AD = Acute Diarrhea.

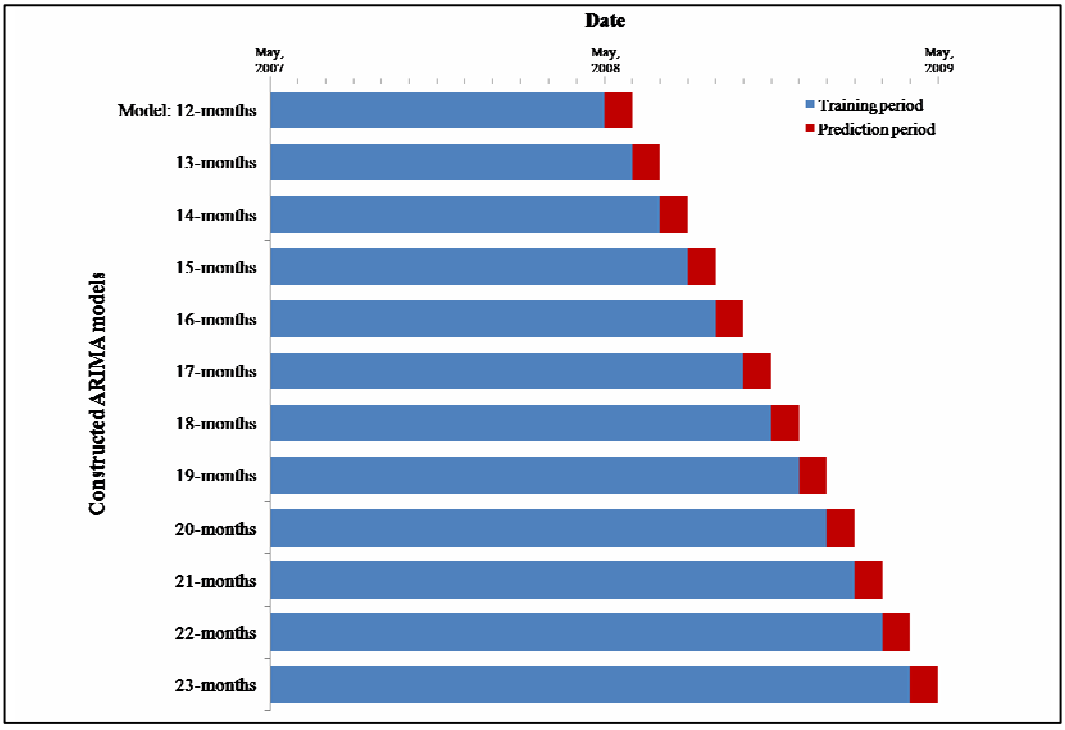

Fig. 2 A schematized method for the gradual period extensions. We gradually increased the data for model training with 1-month intervals (starting at May 1, 2008, which became the 13th month from the initial data) to the month just prior to the target prediction period, and constructed new best-fitting prediction models for each of the following months.

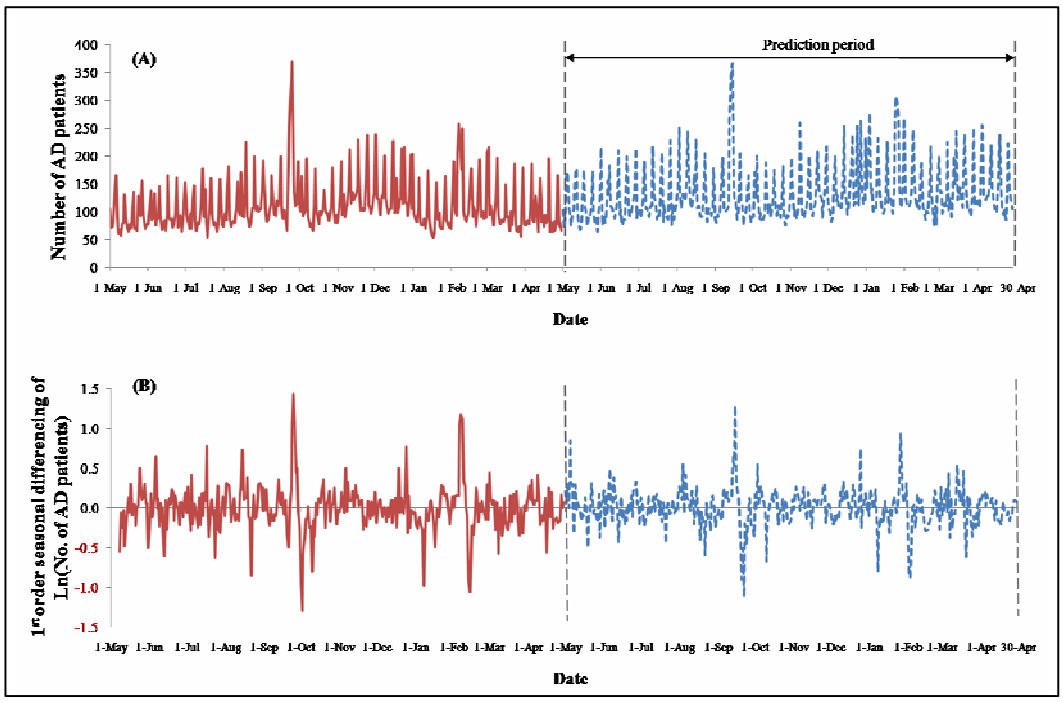

Fig. 3 Sequence graph of actual visiting AD patient (A), and after applying the first-order seasonal differencing (B). The number of visiting patients had a tendency to increase with passing time, and thus we applied the logarithm transformation. Additionally, first-order seasonal differencing was also applied to remove the 7-day periodicity: a pattern graph that moved around the 0 point, as shown in graph (B). AD = Acute Diarrhea.

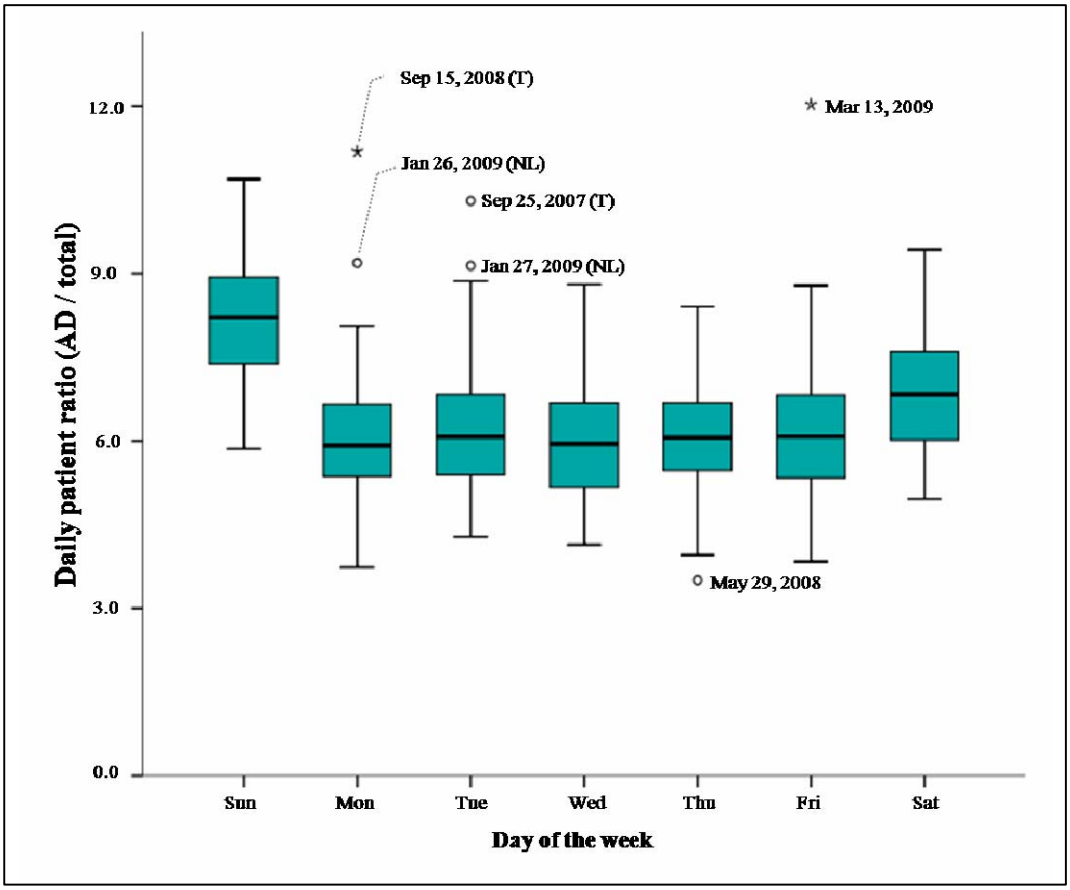

Fig. 4 Ratio of AD patients by day of the week. The ratio of AD patients to total ED patients evidenced a tendency to increase on weekends, particularly on Sundays. AD = Acute Diarrhea; T = Korean Thanksgiving period (lunar); NL = New Year's Day (lunar).

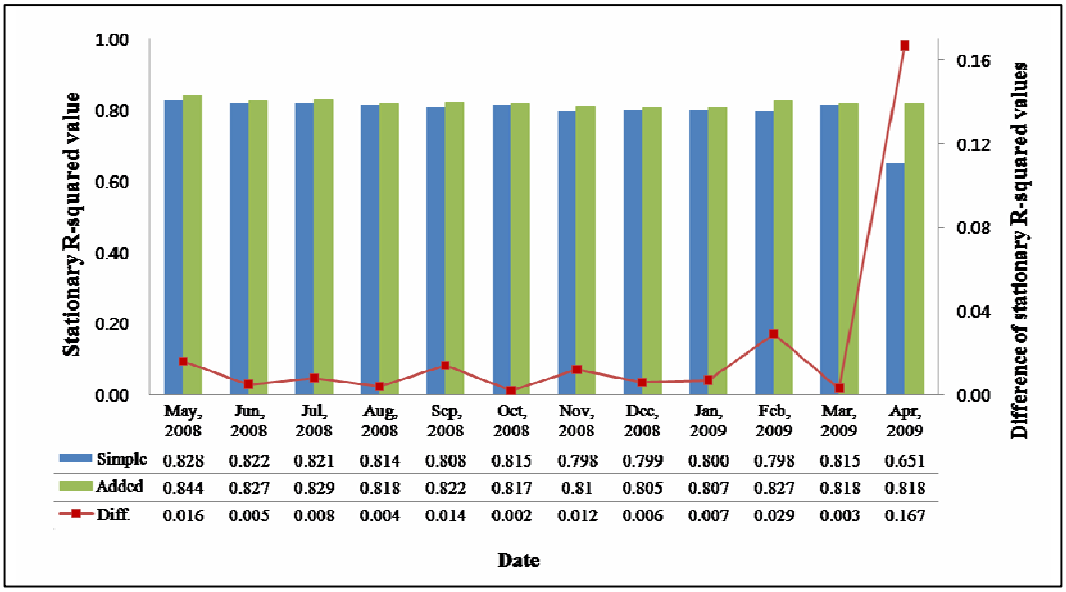

Fig. 5 In-model goodness-of-fit values (stationary R-squared) for the constructed models. The stationary R-squared values changed between 0.651 and 0.828 (for the simple model) and between 0.805 and 0.844 (for the environmental factor-added model) with $p<0.05$.

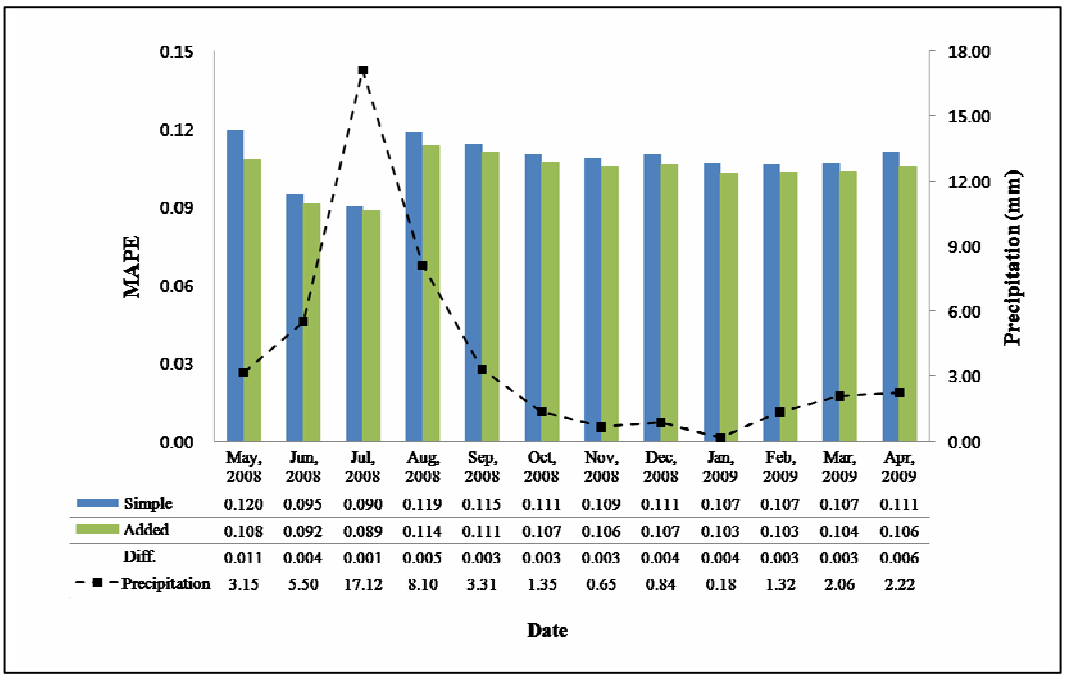

Fig. 6 Cumulative mean graphs for MAPEs of actual and predicted visits during the prediction period (May 1, 2008 to April 30, 2009). The result and estimation of predictions of the numbers of visiting patients from May, 2008 to April, 2009 were shown with cumulative mean of MAPE. The environmental factor-added model predicted the number of visiting patients more accurately than the simple models. MAPE = Mean Absolute Percentage Error; AD = Acute Diarrhea.

Table 1 List of ICD-10 codes used for selecting patients with acute diarrhea. The diagnostic codes for acute diarrhea were classified by emergency department doctors on the basis of the research conducted by the Centers for Disease Control and Prevention of America and ESSENCE (based on ICD-9).

ICD-10 code	Diagnosis	Sub-codes
A00	Cholera	A00.0, A00.1, A00.9
A01	Typhoid and paratyphoid fevers	A01.0, A01.4
A02	Other salmonella infections	A02.0 - A02.2
A03	Shigellosis	A03.0 - A03.2, A03.9
A04	Other bacterial intestinal infections	A04.0, A04.3 - A04.9
A05	Other bacterial foodborne intoxications, NEC	A05.0 - A05.3, A05.8, A05.9
A06	Amoebiasis	A06.0 - A06.6, A06.8, A06.9
A07	Other protozoan intestinal diseases	A07.0
A08	Viral and other specified intestinal infections	A08.0 - A08.5
A09	Other gastroenteritis and colitis of infectious and unspecified origin	A09.0, A09.9
K58	Irritable bowel syndrome	K58.0, K58.9
K59.1	Functional diarrhea	K59.1

Table 2 List Descriptive daily statistics for variables. The characteristics of the NEDIS data and other selected weather variables are provided (for the first training/prediction with 12 months). Variables of the day of the week and holidays are not shown

Variable	Training period (May 1, 2007~April 30, 2008)	Prediction period (May 1, 2008~April 30, 2009)	*P* value*
Number of institutes	27.16(±1.01)	29.43(±0.99)	<0.001
Number of patients	1659.69(±378.36)	1976.00(±413.79)	<0.001
Number of patients with acute diarrhea	110.95(±45.08)	130.29(±49.45)	<0.001
Average temperature (°C)	13.01(±10.12)	13.07(±9.95)	Ns
Highest temperature (°C)	17.26(±10.30)	17.46(±10.20)	Ns
Lowest temperature (°C)	9.40(±10.20)	9.24(±9.99)	Ns
Temperature difference (°C)	7.86(±2.75)	8.22(±2.58)	Ns
Precipitation (mm)	3.14(±9.07)	3.85(±14.38)	Ns
Average wind speed (m/s)	2.45(±0.75)	2.47(±0.79)	Ns
Relative humidity (%)	60.55(±15.20)	60.34(±14.70)	Ns
Sunshine duration (Hours)	5.30(±3.86)	5.76(±3.93)	Ns

* Independent-Samples T test; Ns = Non-significant (p>0.05).

References

1. Mostahari F, Hartman J. Syndromic surveillance: a local perspective. J Urban Health. 2003; 80(2 Suppl. 1): i1-i7.
2. Henning KJ. What is syndromic surveillance? MMWR Morb Mortal Wkly Rep. 2004; 53 (Suppl.): 5-11.
3. Green MS, Kaufman Z. Surveillance for early detection and monitoring of infectious disease outbreaks associated with bioterrorism. Isr Med Assoc J. 2002; 4(7): 503-506.
4. Buehler JW, Berkelman RL, Hartley DM, Peters CJ. Syndromic surveillance and bioterrorism-related epidemics. Emerg Infect Dis. 2003; 9(10): 1197-1204.
5. Cho JP, Min YG, Choi SC. Syndromic surveillances based on the emergency department. J Prev Med Public Health. 2008; 41(4): 219-224.
6. Buckeridge DL. Outbreak detection through automated surveillance: a review of the determinants of detection. J Biomed Inform. 2007; 40(4): 370-379.
7. Bravata DM, McDonald KM, Smith WM, Rydzak C, Szeto H, Buckeridge DL, et al. Systematic review: surveillance systems for early detection of bioterrorism-related diseases. Ann Intern Med. 2004; 140(11): 910-922.
8. Larkin M. Technology and public health. Lancet Infect Dis. 2002; 2(7) :448-449.
9. Bravata D, Rahman M, Luong N, Divan H, Cody S. A comparison of syndromic incidence data collected by triage nurses in Sata Clara county with regional infectious disease data.] Journal of Urban Health. 2003; 80: i112.
10. Doroshenko A, Cooper D, Smith G, Gerard E, Chinemana F, Verlander N, et al. Evaluation of syndromic surveillance based on National Health Service Direct derived data-England and Wales. MMWR Morb Mortal Wkly Rep. 2005; 54 (Suppl.): 117-122.
11. Yuan CM, Love S, Wilson M. Syndromic surveillance at hospital emergency departments-southeastern Virginia. MMWR Morb Mortal Wkly Rep. 2004; 53 (Suppl): 56-58.
12. Wu TS, Shih FY, Yen MY, Wu JS, Lu SW, Chang KC, et al. Establishing a nationwide emergency department-based syndromic surveillance system for better public health responses in Taiwan. BMC Public Health. 2008; 8: 18.
13. Muscatello DJ CT, Kaldor J, Zheng W, Chiu C, Correll P, Jorm L. An automated, broad-based, near real-time public health surveillance system using presentations to hospital Emergency Departments in New South Wales, Australia. BMC Public Health. 2005; 5: 141.
14. Miller B, Kassenborg H, Dunsmuir W, Griffith J, Hadidi M, Nordin JD, et al. Syndromic surveillance for influenzalike illness in ambulatory care network. Emerg Infect Dis. 2004; 10(10): 1806-1811.
15. Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice, New York City. Emerg Infect Dis. 2004; 10(5): 858-864.
16. Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE) [Internet]. [cited 2010 March 25]. Available from: http://www.dhss.mo.gov/ESSENCE/index.html.
17. Ansaldi F, Orsi A, Altomonte F, Bertone G, Parodi V, Carloni R, et al. Emergency department syndromic surveillance system for early detection of 5 syndromes: a pilot project in a reference teaching hospital in Genoa, Italy. J Prev Med Hyg. 2008; 49(4): 131-135.
18. Reis BY, Mandl KD. Time series modeling for syndromic surveillance. BMC Med Inform Decis Mak. 2003; 3: 2.
19. Doyle TJ, Bryan RT. Infectious disease morbidity in the US region bordering Mexico, 1990-1998. J Infect Dis 2000; 182(5): 1503-1510.
20. Balter S, Weiss D, Hanson H, Reddy V, Das D, Heffernan R. Three years of emergency department gastrointestinal syndromic surveillance in New York City: what have we found? MMWR Morb Mortal Wkly Rep. 2005; 54 (Suppl.): 175-180.
21. Hripcsak G, Soulakis ND, Li L, Morrison FP, Lai AM, Friedman C, et al. Syndromic surveillance using ambulatory electronic health records. J Am Med Inform Assoc. 2009; 16(3): 354-361.
22. Quenel P, Dab W. Influenza A and B epidemic criteria based on time-series analysis of health services surveillance data. Eur J Epidemiol. 1998; 14(3): 275-285.
23. Stoto MA SM, Mariono LT. Syndromic Surveillance: Is it Worth the Effort? Chance. 2004; 17(1): 19-24.
24. Kleinman KP, Abrams A, Mandl K, Platt R. Simulation for assessing statistical methods of biologic terrorism surveillance. MMWR Morb Mortal Wkly Rep. 2005; 54 (Suppl.): 101-108.
25. Wang L, Ramoni MF, Mandl KD, Sebastiani P. Factors affecting automated syndromic surveillance. Artif Intell Med. 2005; 34(3): 269-278.
26. Magruder SF. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. In: Johns Hopkins APL technical digest Johns Hopkins University Applied Physics Laboratory; 2003: P349-P353.

27. Davies GR, Finch RG. Sales of over-the-counter remedies as an early warning system for winter bed crises. Clin Microbiol Infect. 2003; 9(8): 858-863.
28. Hogan WR, Tsui FC, Ivanov O, Gesteland PH, Grannis S, Overhage JM, et al. Detection of pediatric respiratory and diarrheal outbreaks from sales of over-the-counter electrolyte products. J Am Med Inform Assoc. 2003; 10(6): 555-562.
29. Levy K, Hubbard AE, Eisenberg JN. Seasonality of rotavirus disease in the tropics: a systematic review and meta-analysis. Int J Epidemiol. 2009; 38(6): 1487-1496.
30. Callejas D, Estevez J, Porto-Espinoza L, Monsalve F, Costa-Leon L, Blitz L, et al. Effect of climatic factors on the epidemiology of rotavirus infection in children under 5 years of age in the city of Maracaibo, Venezuela. Invest Clin. 1999; 40(2): 81-94.
31. Ghazi HO, Khan MA, Telmesani AM, Idress B, Mahomed MF. Rotavirus infection in infants and young children in Makkah, Saudi Arabia. J Pak Med Assoc. 2005; 55(6): 231-234.
32. Chung JB, Ahn ME, Ahn HC, You KC, Kim H, Cho JW, et al. Early Aberration Reporting System Modelling of Korean Emergency Syndromic Surveillance System for Bioterrism. Journal of the Korean Society of Emergency Medicine. 2003; 14(5): 638-645.
33. Jones SS, Thomas A, Evans RS, Welch SJ, Haug PJ, Snow GL. Forecasting daily patient volumes in the emergency department. Acad Emerg Med. 2008; 15(2): 159-170.
34. Ivanov O, Gesteland PH, Hogan W, Mundorff MB, Wagner MM. Detection of pediatric respiratory and gastrointestinal outbreaks from free-text chief complaints. AMIA Annu Symp Proc. 2003: 318-322.
35. Irvin CB, Nouhan PP, Rice K. Syndromic analysis of computerized emergency department patients' chief complaints: an opportunity for bioterrorism and influenza surveillance. Ann Emerg Med. 2003; 41(4): 447-452.
36. Marsden-Haug N, Foster VB, Gould PL, Elbert E, Wang H, Pavlin JA. Code-based syndromic surveillance for influenzalike illness by International Classification of Diseases, Ninth Revision. Emerg Infect Dis. 2007; 13(2): 207-216.
37. Campbell J, Francesconi S, Boyd J, Worth L, Moshier T. Environmental air sampling to detect biological warfare agents. Mil Med. 1999; 164(8): 541-542.
38. Fricker RD, Jr., Hegler BL, Dunfee DA. Comparing syndromic surveillance detection methods: EARS' versus a CUSUM-based methodology. Stat Med. 2008; 27(17): 3407-3429.
39. The Korea center for disease control and prevention (KCDC). [Internet]. [cited 2010 March 25]. Available from: http://www.cdc.go.kr.
40. National Emergency Medical Center (NEMC). [Internet]. [cited 2010 March 25]. Available from: http://www.nemc.go.kr/eng/major/major_information.jsp.
41. Josseran L, Caillere N, Brun-Ney D, Rottner J, Filleul L, Brucker G, et al. Syndromic surveillance and heat wave morbidity: a pilot study based on emergency departments in France. BMC Med Inform Decis Mak. 2009; 9: 14.
42. Turner RM, Muscatello DJ, Zheng W, Willmore A, Arendts G. An outbreak of cardiovascular syndromes requiring urgent medical treatment and its association with environmental factors: an ecological study. Environ Health. 2007; 6: 37.
43. Cifuentes E, Suarez L, Solano M, Santos R. Diarrheal diseases in children from a water reclamation site in Mexico city. Environ Health Perspect. 2002; 110(10): A619-A624.
44. Simanjuntak CH, Larasati W, Arjoso S, Putri M, Lesmana M, Oyofo BA, et al. Cholera in Indonesia in 1993-1999. Am J Trop Med Hyg. 2001; 65(6): 788-797.
45. Rowland MG. The Gambia and Bangladesh: the seasons and diarrhoea. Dialogue Diarrhoea. 1986: 26: 3.
46. Sutra S, Srisontrisuk S, Panpurk W, Sutra P, Chirawatkul A, Snongchart N, et al. The pattern of diarrhea in children in Khon Kaen, northeastern Thailand: I. The incidence and seasonal variation of diarrhea. Southeast Asian J Trop Med Public Health. 1990; 21(4): 586-593.
47. Hashizume M, Armstrong B, Hajat S, Wagatsuma Y, Faruque AS, Hayashi T, et al. Association between climate variability and hospital visits for non-cholera diarrhoea in Bangladesh: effects and vulnerable groups. Int J Epidemiol. 2007; 36(5): 1030-1037.
48. Curriero FC, Patz JA, Rose JB, Lele S. The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948-1994. Am J Public Health. 2001; 91(8): 1194-1199.
49. Korea Meteorological Administration (KMA). [Internet]. [cited 2010 March 25]. Available from: http://www.kma.go.kr/sfc/sfc_03_02.jsp.
50. Schweigler LM, Desmond JS, McCarthy ML, Bukowski KJ, Ionides EL, Younger JG. Forecasting models of emergency department crowding. Acad Emerg Med. 2009; 16(4): 301-308.
51. Box GEP, Jenkins GM, Reinsel GC. Time Series Analysis: Forecasting and Control. San Francisco, CA: Holeden day; 1970.

52. Eberhard ML, Nace EK, Freeman AR, Streit TG, da Silva AJ, Lammie PJ. Cyclospora cayetanensis infections in Haiti: a common occurrence in the absence of watery diarrhea. Am J Trop Med Hyg. 1999; 60(4): 584-586.
53. Begier EM, Sockwell D, Branch LM, Davies-Cole JO, Jones LH, Edwards L, et al. The National Capitol Region's Emergency Department syndromic surveillance system: do chief complaint and discharge diagnosis yield different results? Emerg Infect Dis. 2003; 9(3): 393-396.