A Warning Threshold for Monitoring Tuberculosis Surveillance Data Using a Hidden Markov Model

A thesis submitted to
Tehran University of Medical Sciences
For the degree of Master of Science in Biostatistics

Tehran University
of Medical Sciences

Ali Rafei

Supervisors:
Roohangiz Jamshidi Orak, PhD
Einollah Pasha, PhD

2012

Abstract

Objectives: Early detection of epidemics is one of the most challenging objectives for public health Surveillance worldwide. To date, a surge of research has been carried out in terms of statistical modeling of count data to detect abrupt changes in the disease incidence. However, most of them have ended in a number of complex formulas that are not easily usable except by a limited number of skilled programmers. Therefore, exploring methods that employ conceptually simple frameworks and require less advanced computations, still accurate in detecting outbreaks seems to be essential in public health practice. In the present study, we aimed at introducing a warning threshold for detecting the unexpected incidences of Tuberculosis (TB) using a Hidden Markov Model (HMM).

Methods: we extracted the weekly counts of newly diagnosed patients with sputum smear-positive pulmonary TB from April 2005 to March 2011 nationwide. To detect unexpected incidences of the disease, two approaches: Serfling's technique and HMM, were applied in presence/absence of linear, seasonal and autoregressive components. Parameters were estimated through the least squares error and Baum-Welch methods respectively. A Veterbi algorithm was also employed to decode state sequence of the disease in HMM. Models were subsequently evaluated in terms of goodness-of-fit, and their results were compared in detection of the disease phases. Then, multiple hypothetical thresholds were constructed based on the estimated models and the optimal one was revealed via ROC analysis.

Results: Values of both adjusted coefficient of determination ($\tilde{R}^2$) and Bayesian Information criterion (BIC) reflect a better goodness-of-fit for Periodic Auto-regressive Hidden Markov model (PAHMM) ($BIC = -1323.6$ and $\tilde{R}^2 = 0.74$) than other applied models. Furthermore, according to the ROC curve analysis, higher amounts of Youden's index and area under curve (0.96 and 0.98 respectively) were obtained by the warning threshold on the basis of Periodic Autoregressive Model (PARM).

Conclusions: The warning threshold constructed based on the Periodic Autoregressive Model can be regarded as a useful alternative for HMM in detection of the weeks with unexpected incidence of TB. Therefore, it may be suggested for monitoring TB incidence data in the disease surveillance system.

Keywords: Diseases surveillance, Warning threshold, Sputum smear-positive pulmonary Tuberculosis, Hidden Markov model, Serfling's approach, Periodic autoregressive model.

Acknowledgement

My best gratitude goes toward my supervisor, Dr. R. Jamshidi Orak for her generosity and for providing the opportunity to do research in her chair and for continuously directing the research and correcting my writings. She is a very encouraging supervisor both from a humanitarian and a scientific view and was always available to discuss any problems I encountered. I will never be able to thank her in a convenient manner that he truly deserves for the light she shed on my way which will guide me in the life passage forever. I would also like to thank my father, Mahdi, my mother, Razieh, my little sister, Elaheh, for their comprehension and support during my whole life. Their effect was a bit hidden after moving to Tehran but played a major role in settling me in my new life away from home.

I would like to offer my special thanks to Prof. Einollah Pasha, co-supervisor of the present study, who motivated this thesis by various insightful discussions. I would also like to thank Dr. Mohammad Mehdi Gouya, Dr. Mahshid Nasehi and Dr. Saeid Sharafi from Iran's Center for Communicable Diseases control and prevention for providing me with the data. I am also grateful to Ms Azam Majidi form Tehran University of Medical Sciences for reviewing the manuscript and her useful advices as well as Dr. Fereshteh Asgari from the Center for Non-Communicable Diseases Control and Prevention.

Chapter 1
Introduction

Tuberculosis (TB) still poses a serious threat to global health, despite the significant developments in diagnosis and treatment of the disease over the past century (1, 2). TB currently ranks among the top ten mortality causes and undertakes 2.5% of the Global Burden of Disease. Nearly nine million new cases and two million deaths are annually reported due to TB (3-5). Multi-Drug-Resistant (MDR) TB and TB-HIV co-infecting over recent decades caused global control strategies to be failed (6-11).

In Iran, after establishment of the National Tuberculosis Program in 1990, TB began to be well-controlled, and attributed mortality displayed a descending trend (12, 13). However, increasing the number of MDR-TB and being surrounded by high burden TB countries like Afghanistan, Pakistan and Iraq have doubled the obstacle toward the disease control (14-17). In 2012, there were totally 11,471 old and new cases with TB in Iran, of which 0.8% were MDR and 2.4% were HIV positive (18).

TB surveillance and preventing further spread of the disease requires full understanding of the biological factors affecting TB, and also finding mathematical patterns explaining the mechanism of TB transmission through the community (19). Although TB is not recognized as an infection with rapid dynamics (the chance of getting infection per contact is low), the risk of transmission is higher from patients with sputum smear positive (SS+) (20). Therefore, regarding this assumption, it would be acceptable the fact that the number of persons getting infection over the next time period depends on the number of infectious cases at the current time period. Moreover, some studies have illustrated variable periods of peak seasonality of TB time series with various patterns in different countries. In particular, some of them have reported a higher incidence of the disease in the late winter to early spring in the sense that the indoor activities in the cold weather is much more common than in a warm climate (21, 22).

Detection of epidemics in earlier stages is one of the most challenging objectives for public health surveillance worldwide (23, 24). Since two decades ago, traditional surveillance techniques were replaced by biosurveillance system with the purpose of reducing time delay to detect and report outbreaks (25). Biosurveillance provide early warning system of epidemics by monitoring the data typically consist of time series counts of incident cases of disease, gathered monthly, weekly, or more frequently (26).

There has been already a surge of interest and research in using statistical methods for the early detection of outbreaks based on the routinely surveillance data. Regression techniques, time series analysis, statistical process control and Bayesian methods are some examples of the statistical method which have been used for monitoring the epidemiologic surveillance of infectious diseases (27). However, the majority of such models have ended in a number of complex formulas that are not easily usable except by a limited number of skilled programmers (28, 29). Therefore, exploring methods that employ conceptually simpler frameworks and require less advanced computations, still accurate in detecting outbreaks seems to be essential in public health practice.

Whereas infectious diseases mostly lie into one of the two non-epidemic and epidemic phases (30), it seems using the concept of finite mixture model is preferred to fit models based on a unique distribution. The basic idea of using Hidden Markov Model (HMM) for monitoring the epidemiologic surveillance of infectious diseases was proposed by Le Strat and Carret in 1999 (31). They applied the model to the time series of flu-like disease incidence rates and poliomyelitis counts and demonstrated the ability of HMM in modeling the routinely diseases surveillance data. Nevertheless, it has been rarely applied in public health systems for the same purpose (32). Five years later, Toni M. Rath et al. indicated some problems and shortcomings in Strat's approach and presented some modification to their models (33).

To date, the practical use of HMM has been proved in some infections such as Flu-like diseases, Poliomyelitis (31), Malaria, Leprosy (34), nosocomial infections (29, 35), Hepatitis A and B (32, 36). However, we found no literature that uses this model in TB surveillance for the same purpose. In this study, HMM which seems to be an appropriate tool in this issue will be used for monitoring the anomaly states of the weekly numbers of newly diagnosed cases with SS+. Then, we aim to explore an optimal warning threshold that can be used instead of HMM in the TB surveillance. This threshold was expected to be not only accurate in distinguishing between the disease phases, but also simple in concept and application.

Chapter 2
Material and Methods

2.1. Study area and data source

Data required for this study were derived from the national TB surveillance program in Iran. We extracted the weekly time series of the number of new SS+ pulmonary TB cases detected between April 2005 and March 2011 throughout the country using *TB register software* (version 7.0). The first version of the software was released in 2005 in order to improve the quality of data and statistical reports resulting from the TB surveillance system of Iran (37).

The diagnostic criterion of a new SS+ pulmonary TB case was based on the existence of one of the following conditions: (a) At least two initial sputum smear tests positive for Acid-Fast-Bacilli (AFB), or (b) One sputum smear test positive for AFB plus radiographic diagnosis of active pulmonary TB, or (c) One sputum smear positive for AFB plus sputum culture-positive for Mycobacterium TB. Currently, the detailed information of the new patients with any type of the disease is gathered at TB register units located in any districts across the country and reported quarterly to the "Administration of Tuberculosis and Leprosy Control" of Ministry of Health and Medical Education (12). Figure 1 depicts the time series of the weekly counts of SS+ TB patients over the six-year period 2005-2011.

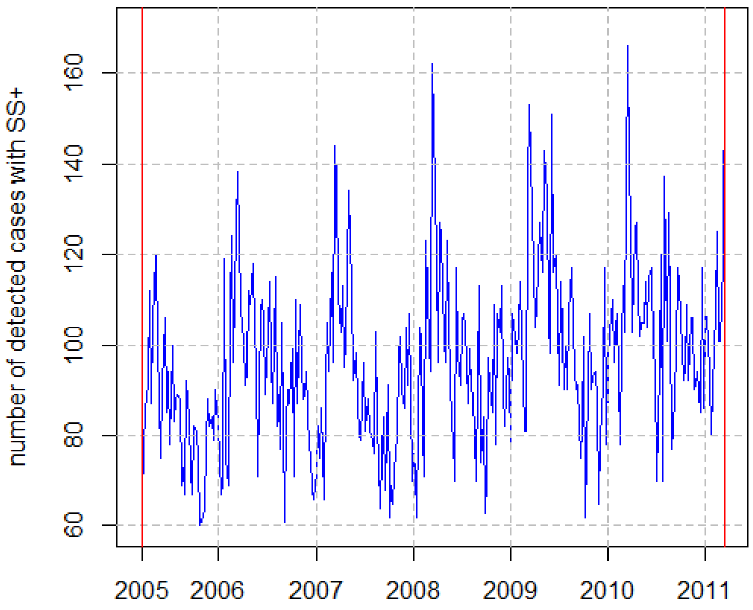

Fig. 1. Time series of the weekly counts of SS+ pulmonary TB in Iran over 2005-2011

2.2. Detecting state sequence of the disease

Two distinct approaches were employed to detect the weeks with unexpected phase of the disease: (a) Serfling's technique and (b) Hidden Markov Model.

2.2.1. Serfling's technique

In the mid-1960s, Surfling developed a Periodic Regression Model (PRM) to monitor and detect the anomaly activities of Pneumonia and Influenza based on the weekly excess mortality data attributed to the disease. The model characterized the historical sequence of the disease time series by combination of a linear term with a trigonometric function describing the seasonal trend (38). This idea originated from the Fourier series and can be formulated with two terms as a multiple linear regression:

$$y_t = \beta_0 + \beta_1 t + \beta_2 \, sin(\frac{2\pi t}{r}) + \beta_3 \, cos(\frac{2\pi t}{r}) + \varepsilon_t$$

Where y_t denotes observed weekly Pneumonia and Influenza deaths at week t for five years period; β_j ; $j=0,1,2,3$ are regression coefficients, as β_0 and β_1 describe the linear part and β_2 and β_3 belong to the seasonal part; r is the time duration of fluctuations; and ε_t is an independent normally distributed error term with a constant variance.

In fact, the Serfling method followed a two-step procedure. The first step was to determine a baseline describing the expected pattern of the historical disease excess mortality. Since the baseline model estimated the non-epidemic phase of the disease, weeks that typically showed high disease incidence were excluded to avoid overestimating the parameters. The main problem in Serfling approach was to determine the epidemic points in this stage. As a criterion, Pelat et al. (2007) proposed excluding the 20% highest values of data to account for past outbreaks in modeling the baseline (39). Then the estimated baseline was used to predict future time series of the expected disease rates. In the second step, an epidemic threshold was obtained by calculating an upper percentile for the prediction distribution according to the baseline. Consequently, an outbreak was detected while an observation exceeds the predefined threshold (27). Moreover, an automated web-based application of Serfling model has been constructed by Pelat et al. for the retrospective and prospective surveillance of diseases (39).

In our study, we initially fitted the mentioned Model (PRM) on the weekly incidence data, assuming $r=52$ due to our weekly data (there are about 52 weeks in a year). In order to capture the effect of the disease incidence in previous weeks, we also added a first-order autoregressive term to our model (PARM). Then, estimation of the parameters was done by use of a least squares error method for both models. Eventually, an upper bound of 95% confidence interval for the prediction ($\hat{y}_t$) was computed and assigned as the alarm threshold for detecting the disease unexpected incidences. It means that any week whose incidence exceeded the threshold an unexpected state was flagged.

2.2.2. Hidden Markov Model

HMM is a statistical tool to fit a mixture distribution on a sequence of dependent data. The application of the models have been recognized in many areas, including automatic speech recognition, electrocardiographic signal analysis, epileptic seizure frequency analysis, DNA sequence analysis, the modeling of neuron firing and meteorology (31). An HMM consists of a bivariate discrete time process like $\{S_t, Y_t\}_{t\geq 1}$, where $\{S_t\}$ is an unobservable Markov chain and, conditional on $\{S_t\}$, $\{Y_t\}$ is a sequence of independent random variables such that the conditional distribution of Y_t only depends on S_t. The sequences $\{S_t\}$ and $\{Y_t\}$ are often called state sequence and observed sequence, respectively (40). The dependence structure of a HMM can be represented by a graphical model as in Figure 2.

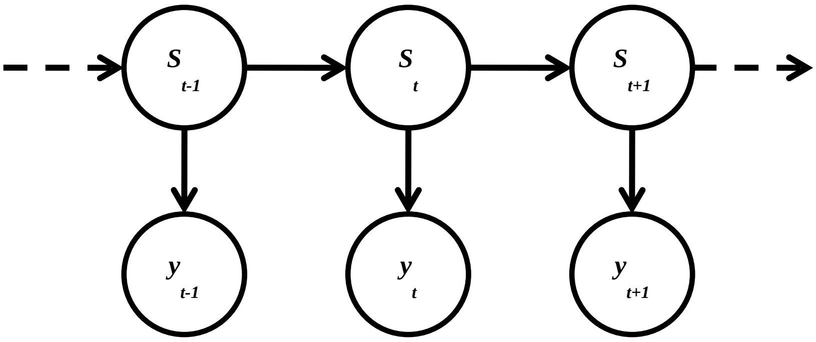

Fig. 2. Graphical representation of the dependence structure of two sequences in a hidden Markov

Let S_t ($t=1,2,\ldots,n$) represents a first order Markov chain which takes on one of the m values $1,2,\ldots,m$ by a transition matrix $\Gamma=(\alpha ij)_{m\times m}$ and initial probability distribution $\pi=(\pi_1,\ldots,\pi_m)^T$, where

$$\alpha_{ij} = P(S_t=j|S_{t-1}=i) \quad i,j = 1,2,\ldots,m; \quad t = 1,2,\ldots,n$$

$$\pi_i = P(S_1=i) \quad i = 1,2,\ldots,m$$

Moreover, the conditional distribution of Y_t given $S_t=i$ follows a parametric form $fi(y_t;\theta_i)$ where θ_i is a vector of unknown parameters. Fitting a HMM to the data requires estimation of the parameters including the initial and transition probabilities and distribution parameters. There are mainly two approaches to estimate the parameters in the HMM literatures. The first can be achieved by a maximum likelihood technique using a modified EM-algorithm, known as the Baum-Welch method. The other is Bayesian framework which assumes the parameters follow a prior distribution and then updates them through a Monte Carlo Markov Chain (MCMC) technique. Consequently, after the estimation of parameters, the most likely sequence of hidden states that produced the data should be decoded by use of the Viterbi algorithm. The following section explains how we used HMMs to detect unusual states of sputum smear- positive pulmonary TB in Iran.

In the current study, a two-state HMM was applied to the incidence data sequence. We initially described $\{Y_t\}$ for each $t=1,2,\ldots,n$ denoting the counts of SS+ TB patients observed over week t. We also supposed $\{S_t\}$ to be a two-state Morkov chain that takes values 1 and 2 corresponding to usual and unusual phases of the disease in week t, respectively. Moreover, $\Pi=(\pi_1, \pi_2)^T$ and $\Gamma=(\alpha_{ij})_{2\times 2}$ were assumed to be the initial probabilities of the state sequence and transition probabilities matrix between the disease states respectively. These elements are defined as below:

$$\pi_i = P\,(S_t=i) \quad i = 1, 2; \quad t = 1, 2,\ldots, n$$
$$\alpha_{ij} = P\,(S_t=j|S_{t-1}=i) \quad i, j = 1, 2; \quad t = 1, 2,\ldots, n$$

Figure 3 illustrates a better representation for the first chain of the model structure and related probabilities applied to the SS+ pulmonary TB data.

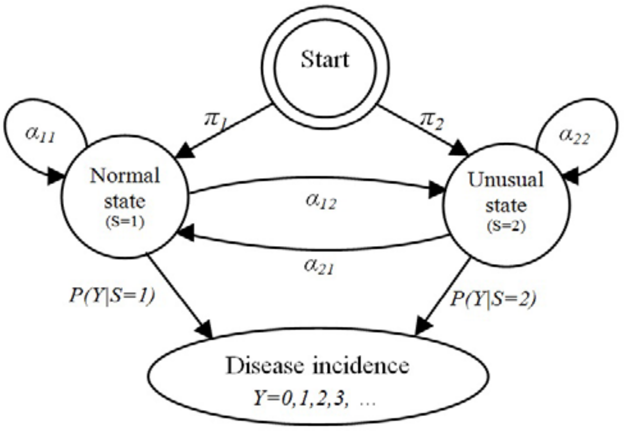

Fig. 3. Graphical representation of the first chain of HMM structure and related probabilities applied to the SS+ pulmonary TB data

Since the data were considered as discrete counts, a mixture of Poisson-Poisson distribution was supposed to fit the data. So, the conditional distribution of $Y_t = y \mid S_t = i \;;\; i = 1, 2$ was defined as below:

$$P(Y_t = y \mid S_t = i) = e^{-\lambda_{it}} \lambda_{it}^{y} / y! \qquad i = 1, 2$$

To determine the expected case load of the disease in each phase ($\lambda_{it} \quad i = 1, 2$), several patterns were considered. Firstly, λ_{it} was assumed to be consonant over time through. So we applied a Simple Hidden Markov Model (SHMM). Next, to control linear and seasonal effects of the disease incidence, a Periodic Regression (PHMM) was utilized to describe the disease case load as below:

$$\lambda_{1t} = E(Y_t \mid S_t = 1) = \beta_0 + \beta_1 t + \beta_2 \cos(\frac{2\pi t}{r}) + \beta_3 \sin(\frac{2\pi t}{r})$$
$$\lambda_{2t} = E(Y_t \mid S_t = 2) = (\beta_0 + \beta_U) + \beta_1 t + \beta_2 \cos(\frac{2\pi t}{r}) + \beta_3 \sin(\frac{2\pi t}{r})$$

Where β_0 ، β_1 ، β_2 and β_3 are the model coefficients, and β_U controls abrupt changes in the disease incidence when moves from usual state to unusual state. Eventually, in order to capture the effect of the disease incidence in previous weeks, again, a first-order autoregressive term was taken into account, with an exception that all parameters were estimated independently in each phases. We defined them as follows:

$$\lambda_{1t} = E(Y_t \mid S_t = 1) = \beta_{10} + \beta_{11} t + \beta_{12} \cos(\frac{2\pi t}{52}) + \beta_{13} \sin(\frac{2\pi t}{52}) + \beta_{14} y_{t-1} + \varepsilon_t$$
$$\lambda_{2t} = E(Y_t \mid S_t = 2) = \beta_{20} + \beta_{21} t + \beta_{22} \cos(\frac{2\pi t}{52}) + \beta_{23} \sin(\frac{2\pi t}{52}) + \beta_{24} y_{t-1} + \varepsilon_t$$

In the current study, we exploited a modified EM-algorithm so-called Baum-Weltch to calculate the Maximum likelihood estimation of the parameters in HMM. We also computed adjusted coefficient of determination ($\tilde{R}^2$) and Bayesian Information Criterion (*BIC*) as two quantities for evaluating models' goodness-of-fit. Then, the model with the highest values of both criteria was chosen to determine the appropriate warning threshold in the next section.

2.3. Exploring an optimal warning threshold

To choose an appropriate warning threshold for detecting unexpected states of SS+ TB, first, through a Veterbi algorithm, the most likely state sequence of the disease states was uncovered, according to the best model selected in the former stage. Next, several hypothetical thresholds were built based on the estimated models. To begin, we initially assumed a time-independent warning threshold for the disease counts. In this case, different values within the range of observations were examined as threshold. Then, we calculated sensitivity and specificity measures for each value and plotted the ROC curve considering the pre-determined state sequence as reference. Then, a Youden's Index (YI) (40) was employed to determine the optimal value. Besides, in order to involve the linear, seasonal and autoregressive terms into our threshold, we exploited parameters estimation of both Serfling's model and HMM in the usual disease phase. Then, an optimal value for the intercept of each model was determined by use of YI. For instance, in PARHMM we defined the threshold for each t as below:

$$C + \hat{\beta}_{11} t + \hat{\beta}_{12} \cos(\frac{2\pi t}{52}) + \hat{\beta}_{13} \sin(\frac{2\pi t}{52}) + \hat{\beta}_{14} y_{t-1} + \varepsilon_t$$

Where $\hat{\beta}_{1i}$; $i = 1,2,3,4$ were parameters estimations, and C was determined in a way to maximize YI. At the end, in order to compare the thresholds and evaluate their accuracy in detection of unexpected states of the disease, we used both YI and Area Under Curve (AUC). We also used the splitting technique to assess how precise the threshold is in predicting future events by separating one third of the data sequence as new observations. All computations were implemented in R version 2.14.1 (Free GNU license), and the source code is available on request.

Chapter 3
Results

3.1. Descriptive analysis

Time series of the weekly counts of SS+ TB comprised of 312 weeks from 2005 to 2011. The minimum and maximum number of patients were reported 60 (Oct. 2005) and 168 (Mar. 2011), respectively. Moreover, the mean and standard error were obtained 97 and 20 new cases per week during these six years. Table 1 show the monthly number of new SS+ cases and associated incidence rates during six-year period 2005-2011. The population estimations of Iran were obtained from the two nationwide censuses 1996 and 2006 for each year.

Table 1. the number of new SS+ cases and associated incidence rates for each of the six periods 2005-2011

Period*	Number of SS+ cases	Population size	Incidence rates per 100,000
2005-2006	4,561	69,390,405	6.57
2006-2007	4,811	70,495,782	6.82
2007-2008	4,677	71,532,062	6.54
2008-2009	4,880	72,583,586	6.72
2009-2010	5,086	73,650,566	6.91
2010-2011	5,171	74,733,230	6.92

* Each period starts from the beginning of April to the end of the next March.

3.2. Results of Surfling's technique

Using Serfling's approach, the estimation of PRM and PARM were obtained respectively as below:

$$\hat{y}_t = 85.67 + \ 0.07t + 11.22\cos(\frac{2\pi t}{52}) + 7.84\sin(\frac{2\pi t}{52})$$

$$\hat{y}_t = 66.23 + 0.05t + 9.15cos(\frac{2\pi t}{52}) + 5.74sin(\frac{2\pi t}{52}) + 0.23y_{t-1}$$

Fisher's test indicated that both models were statistically significant (P-value<0.0001).

3.3. Results of Hidden Markov Model

Table 2 also shows parameters estimation, initial and transition probabilities for both usual and unusual phases of the disease in HMM approach.

Table 2. Parameters estimation of both phases of the disease obtained by HMM approach.

Model	Disease phase	Initial probably	Transition probability		Model parameters				
			Usual	Unusual	β_0	β_1	β_2	β_3	β_4
SHMM[a]	Usual	1	0.79	0.21	82.45				
	Unusual	0	0.21	0.79	111.24				
PHMM[b]	Usual	1	0.70	0.30	76.82	0.07	7.15	7.42	
	Unusual	0	0.45	0.55	100.26	0.07	7.15	7.42	
PARHMM[c]	Usual	1	0.68	0.32	51.10	0.04	7.62	3.71	0.32
	Unusual	0	0.69	0.31	90.52	0.08	16.70	4.99	0.09

[a] Simple Hidden Markov Model
[b] Periodic Hidden Markov Model
[c] Periodic Autoregressive Hidden Markov Model

Results for seeking the best model have been summarized in Table 3. Comparison of the goodness-of-fit (*adjusted-R^2 and BIC*) revealed that among the estimated models PARHMM was the best. Since values of both criteria were greater for PARHMM than the other models. Table 3 also contains the number of weeks detected as unexpected phase of the disease which is least for PARHMM (94 weeks). Figures number 4, 5 and 6 depict the fitted SHMM, PHMM and PARHMM respectively on the TB incidence time series in which points represents weeks with unexpected incidences. Table 4 compares the rest of models with PARHMM in detecting unusual disease states by the use of False Alarm Rate (FAR).

Table 3. Model evaluation and goodness-of-fit for all models fitted on the data

Model	No. Unusual states	No. parameters	R-squared	Adjusted R-squared	Log-likelihood	BIC
PRM[a]	124	4	0.33	0.32	-1308.42	-1319.91
PARM[b]	113	5	0.36	0.36	-1294.34	-1308.69
SHMM	152	5	0.56	0.55	-1375.26	-1386.75
PHMM	120	8	0.72	0.72	-1316.56	-1336.66
PARHMM	94	13	0.75	0.74	-1281.56	-1325.58

[a] Periodic Regression Model
[b] Periodic Autoregressive Model
[c] Simple Hidden Markov Model
[d] Periodic Hidden Markov Model
[e] Periodic Autoregressive Hidden Markov Model

Table 4. Comparison of the accuracy of fitted models with PARHMM in detection of unusual phases of the disease

	False Alarm Rate		
Model	Normal phase	Unusual phase	Both phases
PRM	0.30	0.02	0.12
PARM	0.19	0.00	0.07
SHMM	0.37	0.24	0.34
PHMM	0.17	0.10	0.15

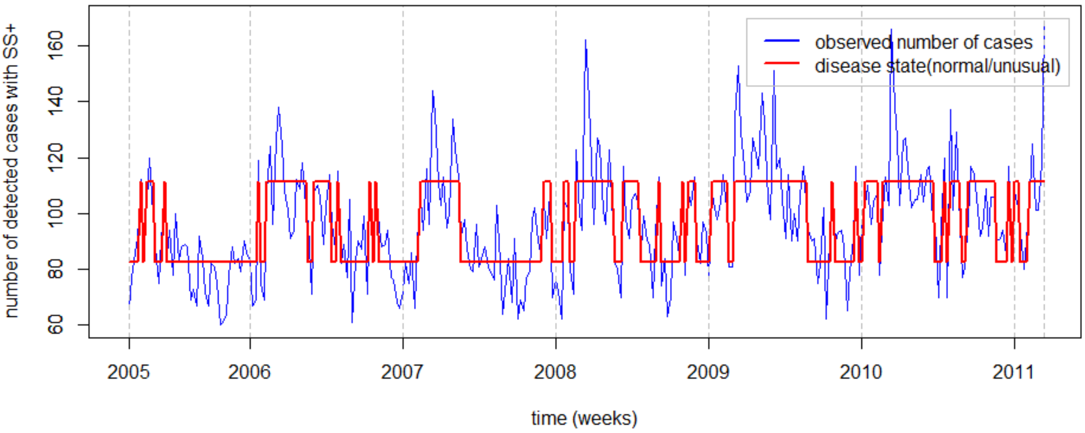

Fig. 4. Unusual states of the weekly SS+ data in Iran over 2005-2011, obtained by applying by Viterbi algorithm in HMM without seasonality

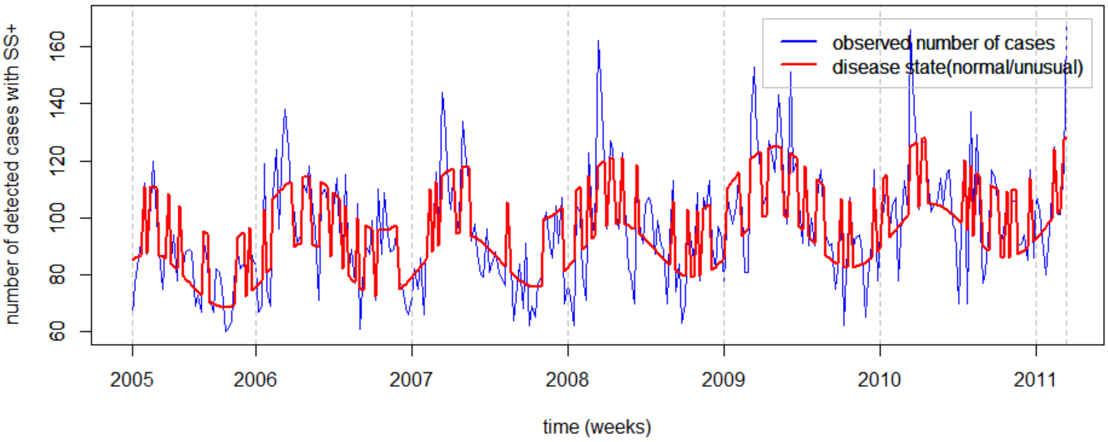

Fig. 5. Unusual states of the weekly SS+ data in Iran over 2005-2011, obtained by applying by Viterbi algorithm in HMM with seasonality

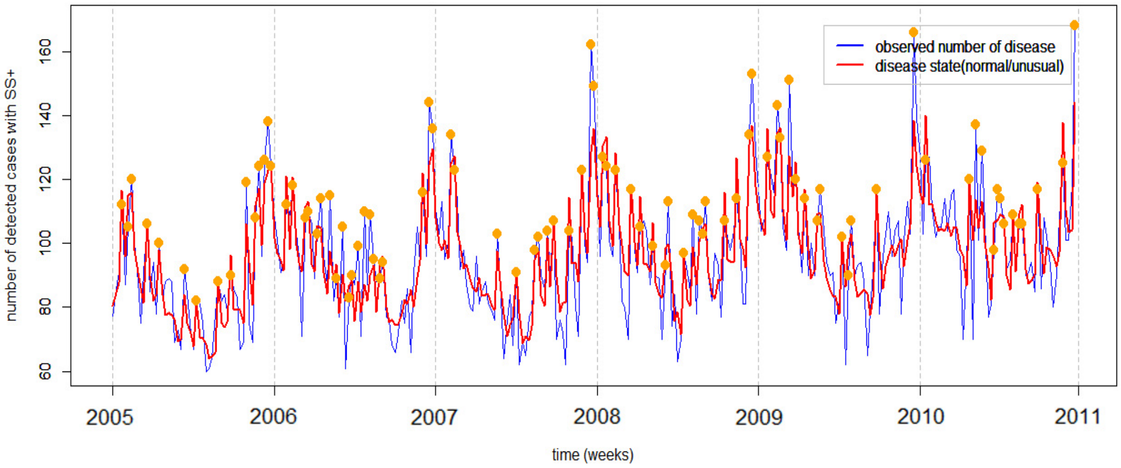

Fig. 6. Estimation of PARHMM and detected weeks with unexpected phase of SS+ Pulmonary TB

3.4. Evaluation of Hypothetical threshold

Figure 7 exhibits the sensitivity and specificity plots as well as ROC curves for each hypothetical threshold. Besides, optimal values of the intercept, according to YI, have been shown in each plot. As ROC curve shows, highest amount of AUC belongs to the threshold that is based on the PARM. Table 5 draws the values of intercept, AUR, sensitivity, specificity and YI to compare the thresholds' accuracy in differentiating between the disease phases. Higher values of both AUC and YI (AUC=0.97; YI=0.96) demonstrate higher degrees of accuracy for the threshold based on PRM.

Table 5. Comparison of the hypothetical threshold in detection of unusual phases of the disease

Threshold	Intercept	AUC	Sensitivity	Specificity	Youden's Index
Time-independent	102	0.77	0.77	0.80	0.57
Based on PRM	93.25	0.89	0.90	0.95	0.85
Based on PARM	72.67	0.97	0.99	0.97	0.96
Based on PHMM	92.61	0.85	0.90	0.92	0.82
Based on PARHMM	64.52	0.96	0.97	0.97	0.94

The result of splitting method has been depicted in Figure 4. The optimal threshold for the first two third of the data sequence has been marked by green color. The red segment shows the threshold's prediction for the next one third of the data. In addition, values of sensitivity and specificity were estimated 1 and 0.96 for the predicted threshold in detecting the disease states.

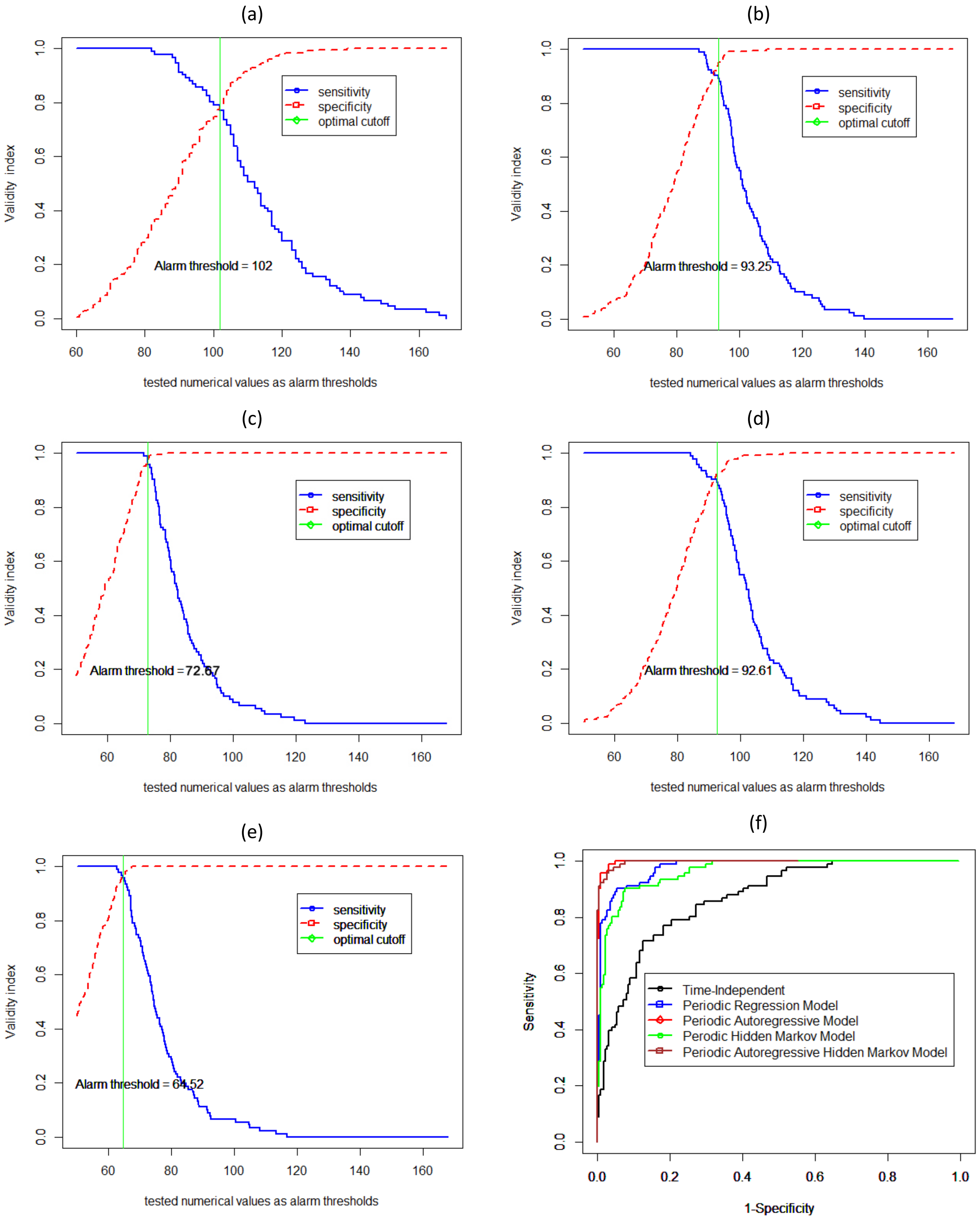

Fig. 7. Sensitivity and Specificity quantities in determining optimal cut-offs based on (a) time-independent, (b) PRM, (c) PARM, (d) PHMM and (e) PARHMM and Figure (f) depicts associated ROC curves

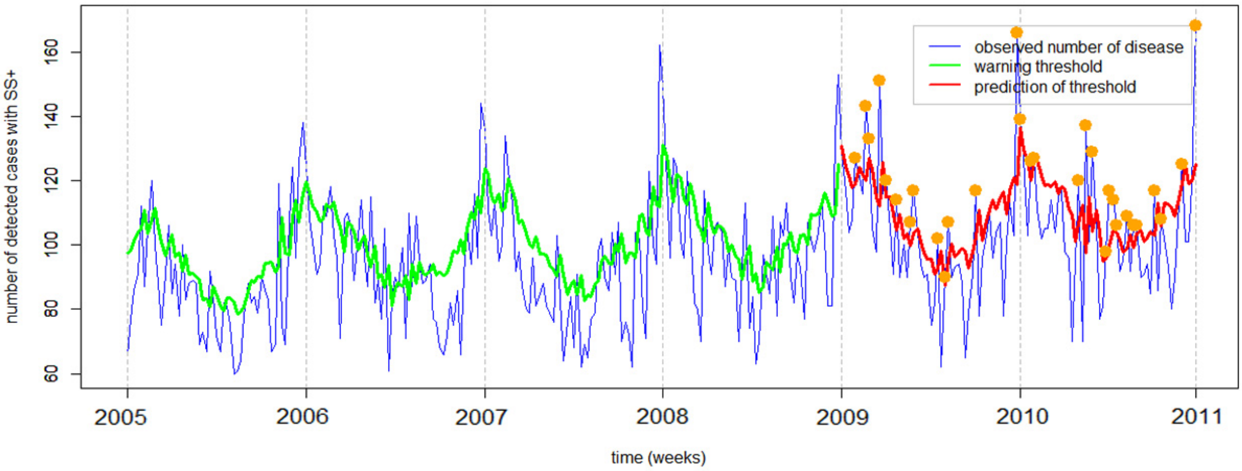

Fig. 8. Evaluating prediction accuracy of the optimal warning threshold using splitting technique

Chapter 4
Discussion

In the present study, we initially made an attempt to develop two different approaches: Serfling's method and HMM, in the presence/absence of linear, seasonal and also autoregressive components, with the aim of monitoring the weekly SS+ TB incidence data in Iran. We then took steps towards exploring an optimal warning threshold in a way to be an appropriate substitution for HMM in discovering unexpected states of the disease. To our knowledge, this is the first effort to find a warning threshold for the TB surveillance data by means of such a model. Our study generally showed that PARHMM had a better goodness of fit ($BIC = -1323.6$ and $\tilde{R}^2 = 0.74$) compared to rest of the models. Furthermore, among the examined hypothetical thresholds, that was constructed based on PARM had more accuracy in detection of the disease unexpected incidences compared to PARHMM. This optimal threshold was obtained as below:

$$\hat{y}_t = 72.67 + 0.05t + 9.15\cos\left(\frac{2\pi t}{52}\right) + 5.74\sin\left(\frac{2\pi t}{52}\right) + 0.23y_{t-1}$$

As shown in Table 5, values of both YI (0.96) and AUC (0.97) reflect a higher performance for the above threshold in monitoring the data at hand. According to this threshold, only 7 weeks out of 312 observations were incorrectly recognized in the unexpected phase of the disease compared to PARHMM. Thus, it can be a satisfactory alternative to HMM in the surveillance of TB.

One brilliant advantage of using such a threshold is that, unlike HMM, we are able to forecast it for the future time periods, i.e. we do not need to wait for the incidence to be observed and then apply a model to the data. To make sure of the predictive capabilities of this threshold, we utilized the splitting technique. As depicted in Figure

8, surprisingly, the predicted threshold for the last one third of the data sequence was able to determine the unexpected states of the disease perfectly (Sen=1). Therefore, we can safely apply the recommended threshold in the surveillance of TB not only for the current time, but for the future as well.

In the present research, we applied a two-state HMM with linear and seasonal trends on the weekly TB incidence data and showed that this model could effectively differentiate between usual and unusual phases of the disease. In addition, an autoregressive term of order one was also added to the model in both Serfling's method and HHM to control the influence of the disease incidence in previous weeks (31). There was also another change in the HMM that we estimated the model parameters independently in each phases of the disease. Although the number of parameters increased from 8 to 13, values of both adjusted-R^2 and BIC increased considerably relative to PHMM. Additionally, the number of unusual states of the disease detected by PARHMM substantially reduced compared to PHMM. This also witnessed the higher flexibility of PARHMM in controlling regular fluctuations.

In addition, our findings demonstrated that adding an autoregressive term to the periodic regression formula in the Serfling's approach significantly increases the goodness-of-fit of the model. In Table 3 the accuracy of the examined models of this study were compared to PARHMM state sequence through the False Alarm Rate. As illustrated, a higher degree of accuracy was obtained for PARM in detection of the disease states (especially unusual phase) compared to other models (FAR=0.07). As a matter of fact, it is expected that this model shows a better performance as a warning threshold in monitoring the disease counts in the second part of the analysis. PHMM, PHM and SHMM took also other places respectively in accuracy.

In the present study, we took the classical approach to estimate the parameters in HMM using a maximum likelihood method (29). This makes us unable to draw inference about parameters or build confidence intervals. Bayesian approach is the other alternative which initially assumes a prior distribution for the model's parameter and then tries to update it given the observations. As a privilege, Bayesian technique enables us to draw any inference or make confidence intervals (32, 42, 43). We also examined higher orders of autoregressive term in modeling the case load of the disease in HMM, but not more significant result was achieved.

The use of count data in this research put restriction on choice of the probability distribution (36). Since Poisson distribution is often used for the rare events, the likelihood in HMM approach presented very small values due to amounts of observations. One way to overcome this problem is to use a Poisson regression framework by applying a log link function to the normal regression formula in both PARM and PARHMM (44, 45). Although this might contribute to a better fitness, it adds to complexity of the models and subsequent thresholds built on their basis. Utilizing weekly incidence rates instead of disease counts offers another way to address this issue. This provides us with a wider range of continuous distribution, for instance Gaussian or exponential distributions (33). This can be defined as a new project for prospective researchers to apply similar approach presented in this study on the time series of the weekly or even monthly TB incidence rates and then, find an optimal threshold for monitoring the disease rates using ROC curve analysis.

By establishment of the national TB registry program, a huge database containing invaluable information on the disease and relevant risk factors has been created in the country. Undoubtedly, analyzing and interpreting such data can effectively help health providers and policy-makers with taking timely actions and strategies to control this deadly epidemic. Authors of the current article strongly believe that statisticians should work more than ever, on modeling these kinds of data in exploring earlier signals for coming outbreaks. This provides unique opportunities for them to put theories into practice and support governors in their critical decisions.

References

1. Migliori G, Loddenkemper R, Blasi F, Raviglione M. 125 years after Robert Koch's discovery of the tubercle bacillus: the new XDR-TB threat. Is “science” enough to tackle the epidemic? European Respiratory Journal. 2007;29(3):423-7.

2. Daniel TM. The history of tuberculosis. Respiratory Medicine. 2006;100(11):1862-70.

3. WHO. Global tuberculosis report 2012: World Health Organization; 2012.

4. Mathers C, Fat DM, Boerma J. The global burden of disease: 2004 update: World Health Organization; 2008.

5. Pinheiro P, Mathers CD, Krämer A. The Global Burden of Infectious Diseases. Modern Infectious Disease Epidemiology. 2010:3-21.

6. Lönnroth K, Castro KG, Chakaya JM, Chauhan LS, Floyd K, Glaziou P, et al. Tuberculosis control and elimination 2010–50: cure, care, and social development. The Lancet. 2010;375(9728):1814-29.

7. WHO. Multidrug and extensively drug-resistant TB (M/XDR-TB): 2010 global report on surveillance and response: World Health Organization; 2010.

8. Zignol M, Hosseini MS, Wright A, Lambregts–van Weezenbeek C, Nunn P, Watt CJ, et al. Global incidence of multidrug-resistant tuberculosis. Journal of Infectious Diseases. 2006;194(4):479-85.

9. Karim SSA, Churchyard GJ, Karim QA, Lawn SD. HIV infection and tuberculosis in South Africa: an urgent need to escalate the public health response. The Lancet. 2009;374(9693):921-33.

10. Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC, et al. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. Archives of internal medicine. 2003;163(9):1009.

11. Dye C, Lönnroth K, Jaramillo E, Williams B, Raviglione M. Trends in tuberculosis incidence and their determinants in 134 countries. Bulletin of the World Health Organization. 2009;87(9):683-91.

12. Nasehi M, Mirhaghani L. National guidelines for TB control. Iranian Ministry of Health. Center for Disease Control. 2009:19-20.

13. WHO. Tuberculosis profile of Iran (Islamic Republic of). 2012.

14. Hadizadeh Tasbiti HT, Yari S, Karimi A, Fateh A, Bahrmand A, Saifi M, et al. Survey of extensively drug-resistant tuberculosis (XDR-TB) in Iran-Tehran: A retrospective study. African Journal of Microbiology Research. 2011;5(22):3795-800.

15. Mirsaeidi MS, Tabarsi P, Farnia P, Ebrahimi G, Morris MW, Masjedi MR, et al. Trends of drug resistant Mycobacterium tuberculosis in a tertiary tuberculosis center in Iran. Saudi medical journal. 2007;28(4):544.

16. Khazaei HA, Rezaei N, Bagheri GR, Dankoub MA, Shahryari K, Tahai A, et al. Epidemiology of tuberculosis in the southeastern Iran. European journal of epidemiology. 2005;20(10):879-83.

17. Masjedi MR, Farnia P, Sorooch S, Pooramiri MV, Mansoori SD, Zarifi AZ, et al. Extensively drug-resistant tuberculosis: 2 years of surveillance in Iran. Clinical infectious diseases. 2006;43(7):841.

18. Administration of Tuberculosis and Leprosy Control. The status of TB/HIV Co-infection. Tehran: Center for Communicable Diseases Control; 2011 [cited 2012]. Available from: www.cdc.hbi.ir.

19. Aparicio JP, Castillo-Chavez C (2009). Mathematical modelling of tuberculosis epidemics. Math Biosci Eng, 6: 209-37.

20. Hertzberg G (1957). The infectiousness of human tuberculosis; an epidemiological investigation. Acta Tuberculosea Scandinavica Supplementum, 38:1.

21. Liu L, Zhao XQ, Zhou Y (2010). A tuberculosis model with seasonality. Bull Math Biol, 72(4): 931-52.

22. Rios M, Garcia J, Sanchez J, Perez D (2000). A statistical analysis of the seasonality in pulmonary tuberculosis. Eur J Epidemiol, 16(5): 483-8.

23. Dawood FS, Iuliano AD, Reed C, Meltzer MI, Shay DK, Cheng P-Y, et al. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. The Lancet infectious diseases. 2012;12(9):687-95.

24. Christian KA, Ijaz K, Dowell SF, Chow CC, Chitale RA, Bresee JS, et al. What we are watching—five top global infectious disease threats, 2012: a perspective from CDC's Global Disease Detection Operations Center. Emerging health threats journal. 2013;6.

25. Shmueli G, Burkom H. Statistical challenges facing early outbreak detection in biosurveillance. Technometrics. 2010;52(1):39-51.

26. Castillo-Chavez C (2010). Infectious Disease Informatics and Biosurveillance. Springer Verlag, pp. 3-8.

27. Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: A review. 2010.

28. McBryde E, Pettitt A, Cooper B, McElwain D. Characterizing an outbreak of vancomycin-resistant enterococci using hidden Markov models. Journal of The Royal Society Interface. 2007;4(15):745-54.

29. Cooper B, Lipsitch M. The analysis of hospital infection data using hidden Markov models. Biostatistics. 2004;5(2):223-37.

30. Lu HM, Zeng D, Chen H (2009). Prospective infectious disease outbreak detection using Markov switching models. IEEE T Knowl Data En, 565-77.

31. Le Strat Y, Carrat F. Monitoring epidemiologic surveillance data using hidden Markov models. Statistics in medicine. 1999;18(24):3463-78.

32. Watkins R, Eagleson S, Veenendaal B, Wright G, Plant A (2009). Disease surveillance using a hidden Markov model. BMC Med Inform Decis Mak, 9(1): 39.

33. Rath T, Carreras M, Sebastiani P (2003). Automated detection of influenza epidemics with hidden Markov models. CHES, 521-32.

34. Jamshidi Orak R, Mohammad K, Pasha E, Sun W, Nori Jalyani K, Rasolinejad M, et al. Modeling the spread of infectious diseases based the Bayesian approach. Journal of School of Public Health and Institute of Public Health Research. 2007;5(1):7-15.

35. McBryde E, Pettitt A, Cooper B, McElwain D. Characterizing an outbreak of vancomycin-resistant enterococci using hidden Markov models. Journal of The Royal Society Interface. 2007;4(15):745-54.

36. Held L, Hofmann M, Höhle M, Schmid V. A two-component model for counts of infectious diseases. Biostatistics. 2006;7(3):422-37.

37. Administration of Tuberculosis and Leprosy Control. TB-register software. Tehran: Center for Communicable Diseases Control; 2011 [cited 2012]. Available from: www.cdc.hbi.ir.

38. Serfling RE. Methods for current statistical analysis of excess pneumonia-influenza deaths. Public health reports. 1963;78(6):494.

39. Pelat C, Boëlle PY, Cowling B, Carrat F, Flahault A, Ansart S, et al. (2007). Online detection and quantification of epidemics. BMC Med Inform Decis Mak, 7(1): 29.

40. Cappé O, Moulines E, Rydén T (2005). Inference in hidden Markov models. Springer Verlag, 1-3.

41. Pepe MS. The statistical evaluation of medical tests for classification and prediction: Oxford University Press; 2003.

42. Conesa D, Martínez-Beneito M, Amorós R, López-Quılez A. Bayesian hierarchical Poisson models with a hidden Markov structure for the detection of influenza epidemic outbreaks. Statistical Methods in Medical Research. 2011.

43. Martínez Beneito MA, Conesa D, López Quílez A, López Maside A. Bayesian Markov switching models for the early detection of influenza epidemics. Statistics in medicine. 2008;27(22):4455-68.

44. Zhu F, Wang D. Estimation and testing for a Poisson autoregressive model. Metrika. 2011;73(2):211-30.

45. Thompson WW, Weintraub E, Dhankhar P, Cheng PY, Brammer L, Meltzer MI, et al. Estimates of US influenza-associated deaths made using four different methods. Influenza and other respiratory viruses. 2009;3(1):37-49.