Accelerated warning of a bioterrorist event or endemic disease outbreak

Luke Marsh
Dstl, Porton Down, Salisbury, Wiltshire, SP4 0JQ
Email: lmarsh@dstl.gov.uk

Abstract - This paper discusses an alarm algorithm that utilises the Exponentially Weighted Moving Average (EWMA) statistic in order to provide early warning of a disease outbreak in a military environment. This is followed by the examination of implementing a Bayesian network to complement this capability, so that a probabilistic assessment of the outbreak can be obtained once an alarm has been triggered.

Keywords - Syndromic surveillance, EWMA, alarm algorithm, Bayesian networks

I. Introduction

A fast response to either a bioterrorist or natural biological incident saves lives. Syndromic data, which is the grouping of medical symptoms e.g. vomiting, can be used to indicate the presence of a disease. Therefore, the Defence Science and Technology Laboratory (Dstl) is developing a surveillance tool that exploits syndromic data in order to reduce the time period between exposure of individuals to the incident and a response being initiated. A response could include the timely administration of medical countermeasures or the implementation of control strategies to limit the number of newly infected individuals. To improve the effectiveness of a response, a diagnostic aid that aims to narrow the field of suspected outbreaks is also being investigated. These two capabilities will provide warning and potential identification of an outbreak, so that an effective response can be initiated.

II. Tool Overview

The tool exploits syndromic information from the Defence Medical Information Capability Programme (DMICP), which is the UK in-service electronic medical database for military personnel [1]. DMICP is based around a fully interoperable electronic clinical record system, provided by Egton Medical Information Systems (EMIS) [2].

Syndromic data is employed by the tool in order to detect unusual clusters of key syndromic indicators within a particular area and time window. Key syndromic indicators include symptoms such as diarrhoea, vomiting or flu-like symptoms. The tool employs an algorithm that utilises the EWMA statistic [3] in order to estimate a quasi-stationary distribution of background symptom reports; any deviation from which initiates an alarm. This alarm can be used by a decision maker in conjunction with any other information available to make an informed decision on the best course of action.

The tool then exploits the data in order to provide a probabilistic assessment of the likelihood of a particular biological agent. This type of information can be used, for example, to direct medical screening. Due to the uncertainty in the probabilistic relationship between disease and symptoms, a Bayesian approach is being taken.

III. Alarm Algorithm

A univariate alarm algorithm that detects unexpected increases in the number of presentations per day of a particular symptom was first considered. This algorithm is ideal for detecting unusual increases in the number of people who are displaying a particular symptom, for example, suffering from diarrhoea, since unexpected increases in this symptom alone is a good indicator of a potential outbreak.

Let X_i be the symptom count for day i with $i = 1, .., n$. If $X_1 = 0$, this is reset to $X_1 = 1$. The EWMA statistic Z_i, which averages the data by giving more weight to recent data, is defined as:

$$Z_i = \lambda X_i + (1 - \lambda) Z_{i-1},$$

where $0 < \lambda \leq 1$ is the smoothing constant. Typically $0.05 \leq \lambda \leq 0.25$. In order to limit false alarms whilst the algorithm establishes itself, the first EWMA statistic Z_0 is taken as $Z_0 = 2X_1$. The variance $\sigma^2_{Z_i}$ of Z_i is calculated by:

$$\sigma^2_{Z_i} = \lambda(X_i - Z_{i-1})^2 + (1 - \lambda)\sigma^2_{Z_{i-1}}.$$

The initial variance is taken as $\sigma^2_{Z_0} = 2X_1$. The evolving threshold alarm T_i at time i is given as:

$$T_i = Z_{i-1} + L\sigma_{Z_{i-1}},$$

where L is a control parameter and $\sigma_{Z_{i-1}}$ is taken as the positive square root of the variance. If $X_i > T_i$, then an alarm is initiated. Typically, the control parameter used is $L = 3$ [4]. It should be noted that the algorithm will not work before $i = -1/(4\ln|1-\lambda|)$, rounded up to the nearest integer.

The univariate alarm algorithm is extended to a multivariate alarm algorithm in order to detect increases in the number of presentations per day of correlated symptoms. For example, increases of a cough symptom alone might not be a good representation of the background activity, but monitored with other symptoms e.g. flu like symptoms, would present a more realistic portrayal of the background syndromic activity. This multivariate algorithm would be more sensitive to unusual syndromic behaviour.

Let $\mathbf{X}_i = (x_1, .., x_p)$ be the symptom count vector for day i for p symptoms with $i = 1, .., n$. Any zero entries of $\mathbf{X}_1$ are set to 1. The multivariate EWMA statistic vector $\mathbf{Z}_i = (z_1, .., z_p)$ is calculated by:

$$\mathbf{Z}_i = \lambda\mathbf{X}_i + (1-\lambda)\mathbf{Z}_{i-1},$$

where $0 < \lambda \leq 1$ is a scalar. The vector $\mathbf{Z}_0 = 2\mathbf{X}_1$ is used as the initial EWMA statistic vector. The covariance matrix Σ_i of $\mathbf{Z}_i$ can be computed as:

$$\Sigma_i = (1-\lambda)\Sigma_{i-1} + \lambda(\mathbf{X}_i - \mathbf{Z}_{i-1})(\mathbf{X}_i - \mathbf{Z}_{i-1})^T.$$

The matrix $\Sigma_0 = diag(2x_1, .., 2x_p)$ is used for the initial covariance matrix. An alarm is produced if

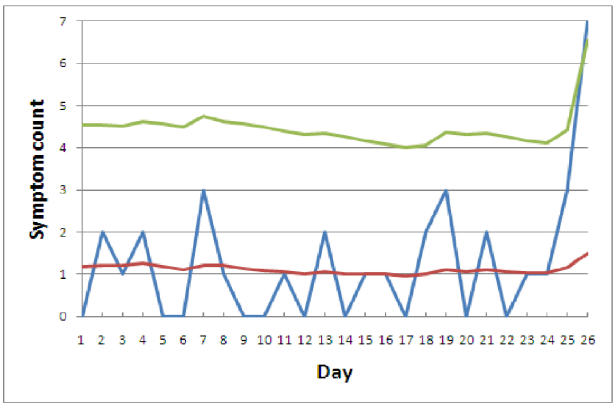

Figure 1: A graph of X_i (blue), Z_i (red) and T_i (green) over time for the univariate alarm algorithm example of Table 1.

i	X_i	Z_i	$\sigma^2_{Z_i}$	T_i	Alarm
1	1	-1	-1	-	
2	1	-1	-1	-	
3	0	-1	-1	-	
4	1	-1	-1	-	
5	2	-1	-1	-	
6	0	1.17859	1.25795	4.54334	
7	2	1.22552	1.22552	4.54661	
8	1	1.21263	1.21263	4.51622	
9	2	1.25761	1.25761	4.62191	
10	0	1.18577	1.26609	4.56139	
11	0	1.11803	1.26517	4.49243	
12	3	1.22554	1.37277	4.7405	
13	1	1.21266	1.29693	4.62915	
14	0	1.14338	1.29753	4.56065	
15	0	1.07806	1.2898	4.48514	
16	1	1.0736	1.21643	4.38235	
17	0	1.01227	1.20547	4.30609	
18	2	1.0687	1.18615	4.33602	
19	0	1.00765	1.1764	4.2615	
20	1	1.00721	1.1092	4.16676	
21	1	1.0068	1.04583	4.07478	
22	0	0.949283	1.03757	4.00512	
23	2	1.00931	1.03436	4.06042	
24	3	1.12303	1.17653	4.37707	
25	0	1.05887	1.17337	4.30854	
26	2	1.11264	1.15132	4.33163	
27	0	1.04908	1.14842	4.26401	
28	1	1.04627	1.08294	4.16821	
29	1	1.04363	1.04363	4.10837	
30	3	1.15539	1.17839	4.412	
31	7	1.48927	2.8459	6.55022	Alarm

Table 1: Simulated univariate alarm algorithm output for a generic symptom count X_i with $\lambda = 0.057$ and $L = 3$. An alarm is produced on day $i = 31$.

$T_i > L$, where L is a control parameter and T_i is the Mahalanobis distance [5] given by:

$$T_i = (\mathbf{X}_i - \mathbf{Z}_{i-1})^T\Sigma^{-1}(\mathbf{X}_i - \mathbf{Z}_{i-1}).$$

IV. Alarm Algorithm Results

The application of the univariate alarm algorithm for a single generic symptom count is demonstrated. Presented in Table 1 are the values of X_i, Z_i, $\sigma^2_{Z_i}$, T_i over time and notification that the algorithm has detected unusual trend behaviour on day 31. Figure 1 shows how X_i, Z_i and T_i evolve over time. Inspection of the X_i counts show that an alarm at this time is appropriate, since X_{31} has distinctly increased above

i	X_{1i}	X_{2i}	T_i	Alarm
1	1	1	1	
2	0	1	1.44874	
3	1	0	1.38396	
4	0	0	1.58972	
5	1	2	0.563959	
6	2	1	0.592393	
7	1	0	1.20759	
8	2	0	1.16631	
9	0	1	1.21686	
10	0	2	1.2363	
11	3	1	1.13329	
12	1	0	1.11919	
13	0	0	1.38128	
14	2	1	0.493951	
15	0	1	1.08954	
16	1	0	0.997542	
17	1	1	0.353246	
18	0	0	1.25996	
19	2	0	1.15726	
20	3	0	1.60584	
21	0	0	1.24194	
22	2	1	0.475486	
23	0	0	1.20397	
24	1	1	0.275083	
25	1	0	0.800666	
26	3	1	1.31289	
27	7	2	3.05344	Alarm
28	3	1	0.761891	
29	2	0	0.948565	
30	1	1	0.47424	
31	0	1	1.04471	
32	2	0	0.977764	

Table 2: Simulated multivariate alarm algorithm output for two symptom counts with $\lambda = 0.05$ and $L = 2$.

the trend. The smoothing parameter λ needs to be chosen to ensure that, whilst the algorithm does not produce false alarms, it does not fail to alarm when a biological incident has occurred. It can be seen in Figure 1 around days 7 and 18 that if λ is chosen well, the alarm threshold adapts robustly to small fluctuations in the symptom counts to avoid unnecessarily alarming. The EWMA statistic can also be seen to be a relatively good representation of the historical data. For this example, the algorithm runs for five days in order for the algorithm to generate a threshold. This unfortunately cannot be avoided, but once this time period has elapsed, the univariate alarm algorithm performs well.

The application of the multivariate alarm algorithm is now demonstrated for two generic symptom counts. Presented in Table 2 are the counts of two symptoms X_{1i} and X_{2i}, with the corresponding value of T_i for day i and notification that the algorithm has alarmed on day 27. Figure 2 shows how the symptom counts X_{1i} and X_{2i} behave over time and how T_i evolves over time compared to the alarm threshold $L = 2$. Inspection of the symptom counts show that an alarm at this time is appropriate since X_1 has noticeably increased above the trend at $i = 27$. The Mahalanobis distance T_i, once established, can be seen to mimic the joint symptom counts fluctuations fairly well, and is a relatively good representation of the overall syndromic behaviour. The algorithm is much more sensitive to unusual symptom fluctuations compared to the univariate algorithm and is therefore more likely to alarm. Like any iterative algorithm, a few initial iterations are required in order for the algorithm to truly establish itself, but once established, the multivariate alarm algorithm performs well.

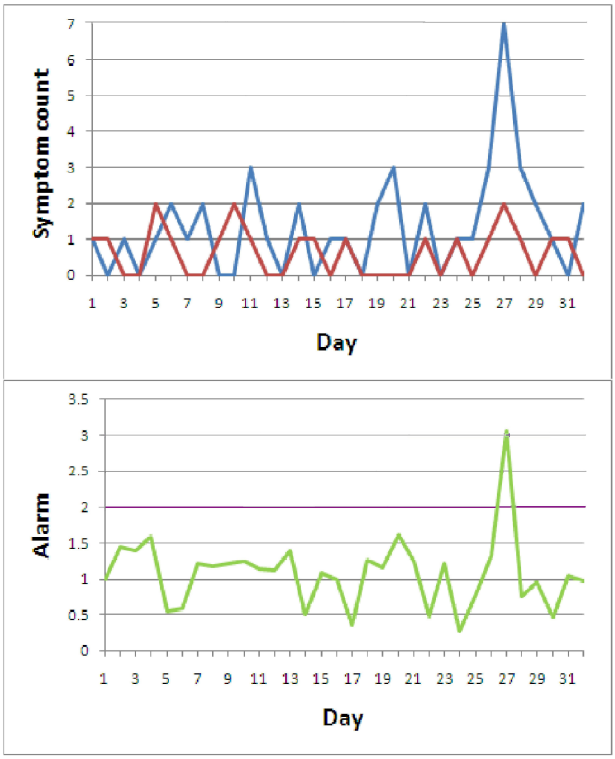

Figure 2: Top: A graph to show how the symptom counts X_{1i} (blue) and X_{2i} (red) evolve over time for the multivariate alarm algorithm example of Table 2. Bottom: A graph to show how the Mahalanobis distance T_i (green) changes over time in relation to the alarm threshold $L = 2$ (purple).

V. Bayesian Networks

A Bayesian network [6] is being considered in order to try and assign a probabilistic assessment of the likelihood of a particular biological agent once the alarm has been triggered. A Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a Directed Acyclic Graph (DAG). Several Bayesian networks have been constructed and explored. The symptom variables are taken to be Boolean since an individual either has the symptom or not. Changing a symptom variable to either true or false to indicate the presence of the symptom in question, updates the probability that the patient has a particular disease.

The first Bayesian network constructed was in order to answer a users specific question such as, "what is the probability that the alarm is due to an anthrax attack?" Since this question is channelled in the sense that the user has supplied the perceived threat, a Bayesian network can be constructed for each biological agent of concern. An example of such a network is shown in Figure 3 for inhalation anthrax. Diagnosis

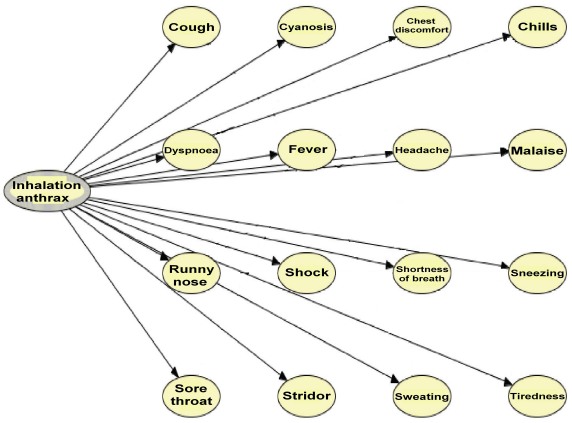

Figure 3: A Bayesian network for inhalation anthrax.

of symptoms for a patient can be used to update the probability that the patient has, in this example case, anthrax. This is shown in Figure 4, whereby a few symptoms have been diagnosed for the patient. This network calculates the probability that a single patient has the disease in question. The progression of the disease probability over time for multiple patients can be introduced into the network by including an extra variable that denotes the previous probability of the disease. This variable is directed into the current disease probability variable and can be continuously updated to gain an overall probability of a biological

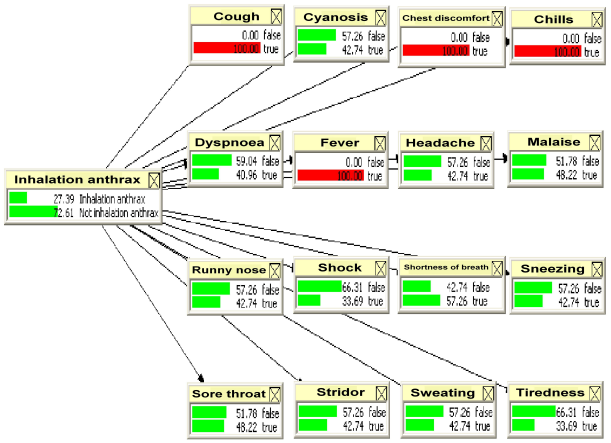

Figure 4: A Bayesian network for anthrax with diagnosed symptoms in red for a single patient.

incident (See Figure 5).

Since only limited data on disease/symptom relationships exist to base probabilities on, the data to populate the probability tables of a Bayesian network will come from Subject Matter Experts (SMEs). It can be difficult for an SME to quantify their knowledge in the form of a probability, and therefore a probability elicitation method will be used [7, 8, 9]. The probability scale method is deemed the most relevant for this application. This method has a scale from 0 to 1 in probability terms, that is represented on a line with verbal or numerical anchors placed at equal distances along the scale. The SME is asked to mark on the scale where they think the required probability lies.

The Bayesian network shown in Figure 3 assumes that each symptom within the network is independent, but in reality, if say a patient has a fever symptom then there is a higher probability that the pa-

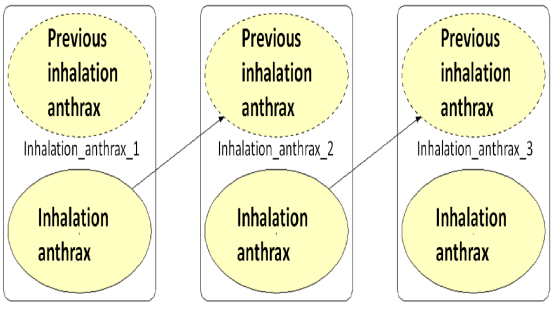

Figure 5: Overall probability of the incident over time.

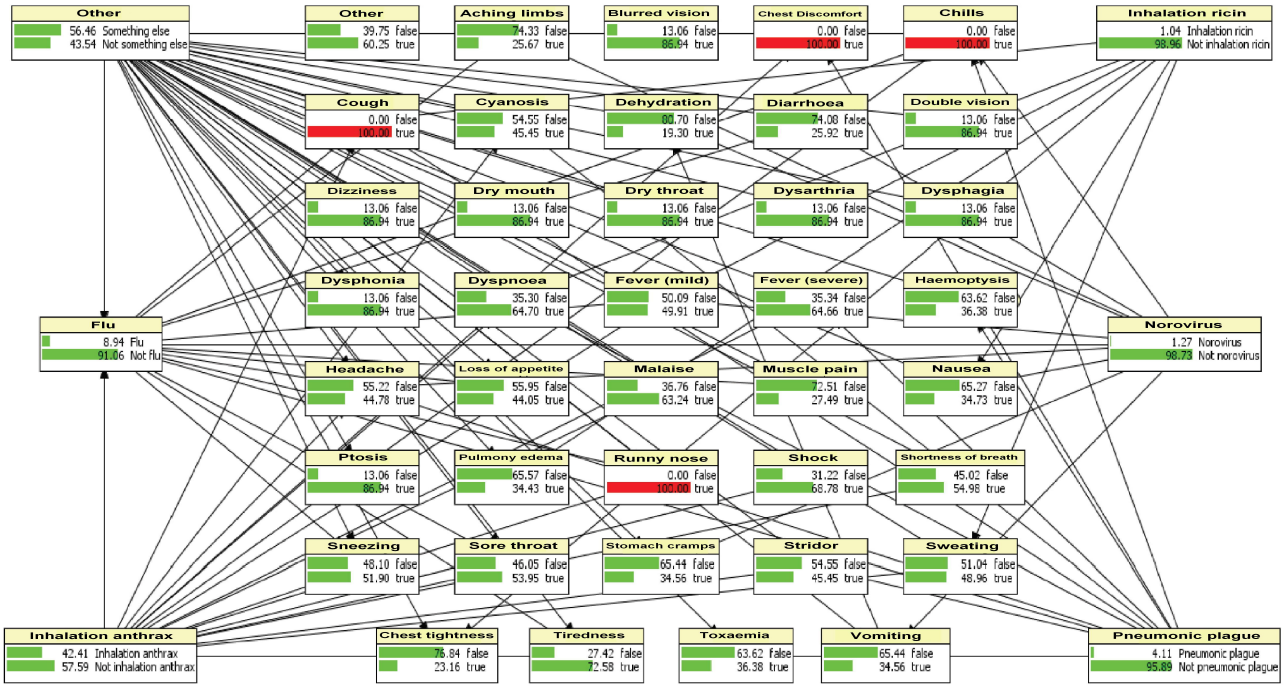

Figure 7: A single Bayesian network for multiple diseases with the symptoms in red diagnosed as being present.

tient might be diagnosed with say a sweating symptom. Generally, due to the complex nature of disease and symptoms, establishing relationships between all symptoms can be difficult. The inclusion of this extra layer will result in the probability table becoming larger, and in some cases, more complex. An example for the inhalation anthrax example of Figure 4 is shown in Figure 6.

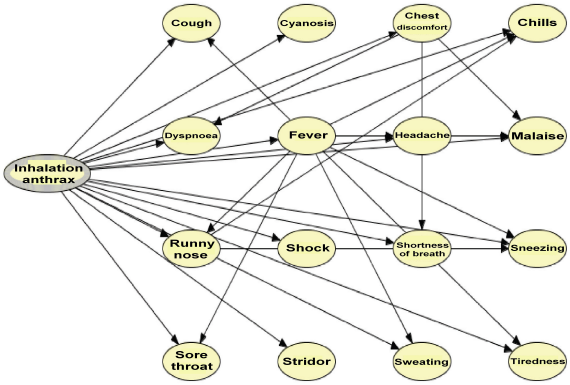

Figure 6: A Bayesian network for inhalation anthrax with an added layer.

Potentially, a Bayesian network can be used to combine multiple diseases into a single network, so that diagnosing a single symptom will update the probabilities of all included diseases. An example is shown in Figure 7, with flu, inhalation anthrax, inhalation ricin, norovirus, pneumonic plague and an option for other diseases being included. This network is more complex than the networks shown so far. Since there is a reasonable possibility that a patient might have a disease outside the diseases included within the network, an "other diseases" option has been included. There is no easy way to quantify this variable. If this variable is omitted, then the results could be misleading, and the user would have to understand that the network is answering the question "if the patient had one of these diseases, what is the more likely?" This suggests that the user has already narrowed the field of suspected causes, and therefore the user has had to supply a "best guess" to the cause of the alarm outbreak. Constructing a universal network that has no user input is extremely difficult, especially since new outbreaks and strains are common. Determining probabilities tables for a universal network approach is near impossible, since some tables would be extremely large and the SME would have to be able to assign a probability to every eventuality.

VI. Conclusion

An alarm algorithm that employs the EWMA statistic and a diagnostic aid that uses a Bayesian network have been investigated in order to develop a powerful decision aid that aims to provide warning and identification of disease outbreaks in a military environment. Examination of the alarm algorithm, both the univariate and multivariate forms, has shown great potential in detecting unusual syndromic activity. An investigation of a Bayesian approach for a diagnostic aid has shown that the user will need to supply a "best guess" of the alarmed outbreak, otherwise populating the probability tables for the network becomes too daunting for an SME. Pursuing this approach will enable a manageable network, whereby the probability tables can be sensibly populated.

References

[1] http://site.logica.com/file/7821

[2] http://www.emis-online.com

[3] Michael S. Saccucci, Raid W. Amin and James M. Lucas, Exponentially weighted moving average control schemes with variable sampling intervals, Communications in statistics - simulation and computation, volume 21, issue 3, 1992.

[4] D. C. Montgomery, Introduction to statistical control quality control, 2001.

[5] P. Mahalanobis, On tests and measures of group divergence I. Theoretical formulae J. and Proc. Asiat. Soc. of Bengal , 26 (1930), pp. 541-588.

[6] F. V. Jensen, Bayesian networks and decision graphs, Springer 2001.

[7] A. Ford, G. Shaw and A. Webb, Elicitation of prior probabilities for a BMD Bayesian network. QinetiQ/S&E/SPI/TR042247, September 2004.

[8] Cilia Witteman and Silja Renooij, Evaluation of a verbalnumerical probability scale, International Journal of Approximate Reasoning 33 (2003), pp. 117-131.

[9] Paul H. Garthwaite, Joseph B. Kadane, and Anthony OHagan, Statistical methods for eliciting probability distributions, Journal of the American Statistical Association June 2005, Vol. 100, No. 470, pp. 680-700.