A Bayesian Dynamic Model for Influenza Surveillance

Paola Sebastiani* **Kenneth D Mandl†** **Peter Szolovits‡** **Isaac S Kohane†** **Marco F Ramoni†**

**Department of Biostatistics, Boston University, Boston MA*

†Children's Hospital Informatics Program, Harvard Medical School, Boston MA

‡Computer Science and Artificial Intelligence Laboratory, M.I.T., Cambridge MA

Abstract

The Severe Acute Respiratory Syndrome (SARS) epidemic, the growing fear of an influenza pandemic and the recent shortage of flu vaccine highlight the need for surveillance systems able to provide early, quantitative predictions of epidemic events. We use dynamic Bayesian networks to discover the interplay among four data sources that are monitored for influenza surveillance. By integrating these different data sources into a dynamic model, we identify in children and infants presenting to the pediatric emergency department with respiratory syndromes an early indicator of impending influenza morbidity and mortality. Our findings show the importance of modeling the complex interplay of data collected for influenza surveillance, and suggest that dynamic Bayesian networks could be suitable modeling tools for developing epidemic surveillance systems.

Keywords: Dynamic Bayesian networks, influenza surveillance, syndromic data.

Submitted to: Journal of the American Statistical Association (March 8, 2005).

Address: Paola Sebastiani. Department of Biostatistics, Boston University, 715 Albany Street, T-E417, Boston, MA 02118. PHONE: 617 638 5877, FAX: 617 638-6484, EMAIL: sebas@bu.edu, URL: http://people.bu.edu/sebas.

1. Introduction

The worldwide spread of SARS during the Winter of 2002-2003, the 2001 Anthrax attacks, the current shortage of flu vaccine, and the looming threat of an influenza pandemic motivate the urgent need for surveillance systems able to provide early quantitative predictions of acute respiratory infections. Although less lethal than SARS, influenza is the seventh leading cause of death for people above 65 and below four years of age, and among the top ten in almost all age groups [1]. Every year influenza infects up to 40 million Americans, causes the hospitalization of over 114,000 [31], kills approximately 36,000 [2] and costs the United States between two and five billion dollars in physician visits, lost productivity, and lost wages [29]. This cost could grow to 160 billion dollars in the event of an influenza pandemic: a sudden and widespread outbreak of a new strain of the influenza virus that has significant morbidity and mortality [20]. Historically, the periodicity of pandemics has ranged between nine to 39 years, with the last one in 1968-1969. A new pandemic in the next few years is considered today a looming if not inevitable threat [22], which is reinforced by the increasing number of outbreaks caused by the H5N1 bird-flu strain [10].

Because influenza viruses change continuously by antigenic drift and shift, nobody is immune and surveillance efforts are in place to detect changes in the viruses, to monitor their effects on hospitalization and death rates, and possibly to forecast the resurgence of a new influenza pandemic. The influenza surveillance system in the United States is managed by the Centers for Disease Control and Prevention (CDC) Influenza Branch, which collects and reports weekly information on influenza activity in the United States from September through May. Distribution of influenza activity and identification of viral types is ascertained through about 75 laboratories of the World Health Organization and 50 laboratories of the National Respiratory and Enteric Virus Surveillance System located throughout the United States. About 650 sentinel physicians report the weekly number of patients they have seen and the number of those patients with influenza like illness (ILI) by age group that is defined as fever (temperature above 100^oF) plus either a cough or a sore throat. The effect of in-

fluenza on mortality is monitored through weekly reports filed by the vital statistics offices of 122 cities that contain the total number of death certificates and the number of those for which pneumonia or influenza (P&I) were listed as a contributing cause of death.

The use of poor epidemiology methods and the lack of timeliness of the data limit however the influenza surveillance effort. The current influenza surveillance system monitors all these data streams individually, as shown in Figure 1. For example, ILI data are monitored to detect surges of influenza morbidity nationwide by comparing weekly data to a national baseline defined by a fixed threshold. It is acknowledged by CDC that the national baseline does not provide a useful threshold for local influenza surveillance so that the issue of local monitoring of influenza morbidity is still open [24]. The aggregate mortality data from the 122 cities are monitored to detect influenza epidemics. The detection consists of comparing the aggregated mortality data to an epidemic threshold that is calculated for each week using a baseline defined by a cyclic regression model [30]. To the best of our knowledge, no attempt is made to use and possibly integrate ILI and P&I data to forecast the course of an epidemic.

Furthermore, for both states and physicians, influenza activity reporting is voluntary and data do not become available to health officials until approximately two weeks after the reports have been filed. An evolving approach to more timely detection is syndromic surveillance [19], which identifies infected individuals early in the course of their disease, generally before a confirmed diagnosis is made. Patients are classified by syndrome, such as acute respiratory infection or gastrointestinal illness, based on a variety of data sources ranging from purchase of over the counter medications [9] to primary care physician and emergency department (ED) logs [7, 26]. The primary focus of these efforts has been to provide early detection of bioterrorism, following the Anthrax attacks of 2001, although this data could also provide early indication of naturally occurring disease outbreaks [16].

Despite the increasing amount of data that could potentially inform about outbreaks of disease, biosurveillance research to date has mainly focused on monitoring individual data streams [9, 17, 26]. When multiple signals have been considered, a system architecture based

on a set of parallel, independent surveillance systems has been envisioned [34, 35]. The intuition underlying our approach is that the integration of different data streams should provide complementary information about a disease and improve our ability for early detection of outbreaks. This integration, however, requires a coherent modeling framework to integrate data into a global model.

In this paper, we integrate different data streams into a multivariate model for influenza surveillance. The novelty of our approach is in both the modeling framework that we use and in the type of data. We build a dynamic Bayesian networks that relates pediatric and adult syndromic data in two EDs to the traditional measures of influenza morbidity and mortality, and we show how to use this model for "active" influenza surveillance by forecasting the course of influenza epidemics. Our analysis shows that the use of a dynamic Bayesian network for influenza surveillance has several advantages. By directly modeling measures of influenza morbidity and mortality, our model can be used to forecast the beginning of epidemics, as well as peaks of epidemics. Furthermore, the joint modeling of the four data streams shows that pediatric patients are infected with respiratory viruses well before the general population. In particular, the number of respiratory syndrome cases in a pediatric ED predicts influenza morbidity in the general population as early as two weeks in advance and influenza mortality as early as three weeks in advance. These findings suggest that children with respiratory syndromes seen at the ED act as sentinels for surges in influenza morbidity and mortality and that active surveillance of pediatric populations could become an important component of the influenza surveillance effort.

The next section describes the data that we used to build our dynamic models. Section 3 gives a brief introduction to dynamic Bayesian networks and Section 4 describes how we built the network with the data available and how we use it to forecast the course of an influenza epidemic. Conclusions are reported in Section 5.

2. Data

We relate variations in the weekly frequency of patients with respiratory syndromes presenting to two urban, tertiary care teaching hospitals in Boston, Massachusetts, to ILI data for the influenza seasons 1997-98 through 2001-02, and to deaths for pneumonia and influenza (P&I) in New England from January 1998 to March 2003 published by CDC[1]. Both ILI and P&I data includes all age groups. The hospitals, one pediatric (CH) and the other adult (AH), share the same catchment area and each emergency department has an annual census of approximately 50,000 visits, with an average age of approximately 6 years for CH and 48 years for AH. Syndromic grouping was based on ED chief complaints as described in [25]. The chief complaints were used to select those ED encounters that were related to respiratory illness. At AH, chief complaints are entered as free-text during the triage process. We used two procedures to classify complaints: a text-string search, and a publicly available nave Bayesian classification program [34]. Default probabilities from an ED dataset, supplied with the software, were applied to the AH data. At CH, chief complaint codes were chosen during the triage process, from a pre-defined on-line list of 181 choices. A previously validated subset of the constrained chief complaint set was chosen a priori for inclusion in the respiratory syndromic grouping. The ED dataset dates back to June 1992 at CH and to June 1998 at AH. We will refer to the number of respiratory syndrome cases in either hospitals as CH and AH data.

The four time series are displayed in Figure 2. The plot shows the apparent regularity of the four time series that are characterized by evident winter peaks during the influenza seasons, and more modest peaks in the fall seasons. Typically, the fall peaks correspond to increased activity of respiratory infections that follow the opening of schools. An intriguing feature of these time series is that the peaks do not occur simultaneously but there is an evident order with CH data (red line) peaking first, followed by AH (blue line) and ILI data (green line) and then P&I mortality data (black line).

To ascertain the temporal order of the four time series, we analyzed their pair-wise cross-

[1] http://wonder.cdc.gov/mmwr/mmwrmort.asp

correlations over 52 weeks. We use the notation $\rho_x(A, B)$ to denote the cross-correlation between two time series A and B, with B shifted back by x weeks when $x > 0$, and A shifted back by x weeks when $x < 0$. The cross-correlation plot in the left panel of Figure 3 suggests that CH data are the earliest indicator of influenza-like illness (black-line: ρ_x(CH, ILI)> 0.5 for $1 \leq x =$ week ≤ 5) and deaths in the community (red-line: ρ_x(CH, P&I)> 0.5 for $2 \leq x \leq 8$). The largest correlation between CH and ILI data for a three week lag suggests that the first effects of influenza emerge in CH data well before they become evident in the general population. CH data lead AH data by 1–5 weeks (green line: ρ_x(CH, AH)> 0.5 for $1 \leq x \leq 5$) and the large correlations ρ_1(CH, AH)$= 0.64$ for a one week lag, and $\rho_{2,3}$(CH, AH)$= 0.63$ for a 2–3 week lag show that CH data can be used to predict a substantial proportion of the variability in AH data.

AH data are also able to predict influenza mortality but with a shorter lead time compared to CH data, while they appear to provide only slightly earlier information than ILI data currently monitored through federal surveillance programs. This hypothesis is supported by the cross-correlation between AH data and ILI and P&I mortality data displayed in the right panel of Figure 3: AH data lead P&I data by 1–3 weeks (red-line: ρ_x(AH, P&I)> 0.5 for $1 \leq x \leq 3$). Similarly, ILI data lead P&I mortality data by one week (blue-line: ρ_x(ILI, P&I)> 0.5 for $0 \leq x \leq 3$) thus confirming that AH and ILI data contain similar information to predict P&I mortality. Compared to AH data, CH data show a smaller correlation with P&I deaths, possibly because the elderly are the most at risk of dying from influenza and pneumonia. Nonetheless, this analysis suggests that pediatric patients are the first group to show symptoms of respiratory infections, and that these early effects are seen in the ED.

3. Dynamic Bayesian Networks

To identify the dynamic structure among the four time series, we built a model based on a dynamic Bayesian network [28]. A dynamic Bayesian network is described by a directed acyclic graph in which nodes represent stochastic variables and arrows represent temporal dependen-

cies that are quantified by probability distributions. Following the direction of the arrows, a node Y_1 with an incoming arrow from a node Y_2 is called a child of Y_2, and Y_2 is called a parent of Y_1. We assume that the probability distributions of the temporal dependencies are time invariant, so that the directed acyclic graph of a dynamic Bayesian network represents only the time transitions that are necessary and sufficient to reconstruct the overall temporal process. Figure 4 shows the directed acyclic graph of a dynamic Bayesian network with three variables. The subscript of each node denotes the time lag, so that the arrows from the nodes $Y_{2(t-1)}$ and $Y_{1(t-1)}$ to the node $Y_{1(t)}$ describe the dependency of the probability distribution of the variable Y_1 at time t on the value of Y_1 and Y_2 at time $t-1$. Similarly, the directed acyclic graph shows that the probability distribution of the variable Y_2 at time t is a function of the value of Y_1 and Y_2 at time $t-1$. A dynamic Bayesian network is not restricted to represent temporal dependency of order 1. For example the probability distribution of the variable Y_3 at time t depends on the value of the variable at time $t-1$ as well as the value of the variable Y_2 at time $t-2$.

The topology of the directed acyclic graph describes Markovian properties of the variables that allow us to decompose the network into related modules. The local Markov property asserts that a node Y is independent of its predecessor nodes, given the parent nodes [14], and leads to a direct factorization of the transition distribution of the network variables $Y_1, \ldots, Y_k$ into the product of the conditional distribution of each variable given its parents Π_i:

$$p(y_{1(t)}, \ldots, y_{k(t)}|h_{t-1}) = \prod_i p(y_{i(t)}|\pi_i)$$

where h_{t-1} is the history of the process until time $t-1$, $y_{1(t)}, \ldots, y_{k(t)}$ denotes the value of the network variables at time t, and π_i are the parents of Y_i. For example, we can fully describe the transition distribution of the three variables in Figure 4 given the past history

$$h_{t-1} = y_{1(t-1)}, \ldots, y_{1(0)}, y_{2(t-1)}, \ldots, y_{2(0)}, y_{3(t-1)}, \ldots, y_{3(0)}$$

by using only the three transition distributions:

$$p(y_{1(t)}|h_{t-1}) \quad = \quad p(y_{1(t)}|y_{1(t-1)}, y_{2(t-1)})$$

$$
\begin{aligned}
p(y_{2(t)}|h_{t-1}) &= p(y_{2(t)}|y_{1(t-1)}, y_{2(t-1)}) \\
p(y_{3(t)}|h_{t-1}) &= p(y_{3(t)}|y_{3(t-1)}, y_{2(t-2)})
\end{aligned}
$$

By assuming that these probability distributions are time invariant, they are sufficient to compute the probability that a process that starts from known values $y_{1(0)}$, $y_{2(0)}$, and $y_{3(0)}$ evolves into $y_{1(T)}$, $y_{2(T)}$, $y_{3(T)}$, or that a process with values $y_{1(T)}$, $y_{2(T)}$, $y_{3(T)}$ at time T started from the initial states $y_{1(0)}$, $y_{2(0)}$, $y_{3(0)}$. Exact algorithms exist to perform this inference when the network variables are all discrete, all continuous and modelled with Gaussian distributions, or the network topology is constrained to particular structures [3, 15, 23]. For general network topologies and non standard distributions, we need to resort to stochastic simulation [4]. Among the several stochastic simulation methods currently available, Gibbs sampling [6, 33] is particularly appropriate for Bayesian network reasoning because of its ability to leverage on the decomposition of joint multivariate distributions induced by the directed acyclic graph to improve computational efficiency.

4. Model Building and Validation

We exploited the local Markov property to build our dynamic Bayesian network by modules. For each variable, we identified the best multiple regression model by selecting the significant predictors in the set defined by the variable itself observed at $t-1, \ldots, t-10$ week lags, the other variables observed over the same temporal range, and two auxiliary variables that model the seasonal pattern of the time series. Because of their significant skewness, we modeled the number of P&I deaths by a Poisson log-linear model and the number of patients with ILI symptoms with a log-normal distribution, while we modeled AH and CH data with normal distributions. We used backward step-wise regression and the Akaike information criterion [12] to build initial models for each time series, and then selected only significant predictors. This selection procedure returned the four models in Figure 5, one for each of the four time series. The overall dependency model was built joining the four regression models by their

common predictors using standard path analysis [11], and is reported in Figure 6. Each node in the directed graphs represents one of the four variables at the time point defined by the subscript and the arrows define the temporal dependencies that are quantified by the probability distributions summarized in Table 1.

We validated the model by examining the posterior probability of each selected temporal model versus the others. The posterior probability was derived from the Akaike information criterion as described in [13]. The dependency structure of the selected model is very strong as shown by the plot in Figure 7 that depicts the posterior probability of 400 models describing the dependency of the number of P&I deaths on CH and AH data, for different time lags. The surface peaks when the time lag is three weeks for CH data and one week for AH data and, compared to the other temporal dependencies, there is very strong evidence that this model gives the best fit (posterior probability almost 1).

The temporal dependency structure in Figure 6 confirms some of the results suggested by the cross-correlation analysis. CH data are predictive of AH data with a one week lag, they are predictive of ILI data with a two week lag, and of the number of P&I deaths with a three week lag. This result confirms that children presenting to the ED with respiratory syndromes provide earlier signals of the spread of influenza epidemics compared to ILI data currently monitored through federal surveillance programs. It is worthwhile noting that neither AH data, nor ILI data or P&I deaths are predictive of CH data, thus confirming that pediatric patients presenting to the ED are the first group to show symptoms of influenza. Both AH and ILI data are predictive of P&I deaths with a one week lag. Note that AH data are not predictive of ILI data, once we condition on CH data observed two weeks earlier, thus confirming that CH data contain earlier and stronger signals of influenza morbidity than AH data.

The model in Figure 6 suggests that the spread of respiratory infections, and particularly influenza, in a community begins in children who report to the ED with respiratory syndromes. This observation is confirmed by the absence of predictive signal of ILI and P&I mortality data on CH data. Further support for this conjecture is provided by using the dynamic model for

prediction of influenza morbidity and mortality a few weeks in advance and by quantifying the predictive effect of pediatric patients. For this assessment, we created competitive models by iteratively dropping each of the predictors from the selected model. We trained the competitive models on the data from June 1998 to the end of October 1999 and then computed their two-week ahead predictions of ILI and P&I data from November 1999 through October 2000. We chose this test set as the virus circulating during the influenza season 1999-2000 had significant morbidity. After each prediction, we iteratively updated the model parameters using all data seen. We used the implementation of Gibbs sampling in Winbugs 1.4 to compute the forecasts [32].

Figure 9 reports the BUGS code for this "two-week-ahead" forecast with the model in Figure 6. The model formulation describes the dependency structure between the four data streams assuming that we have initial values at week $t-2$ and we wish to forecast P&I mortality data, and ILI morbidity data at week t. To initialize the dynamic network, we need to provide data for the parent nodes of the four variables at time $t-2$. Therefore, we provide CH data at week $t-4$, and AH data, ILI data and P&I data at week $t-2$ to initialize the conditional distribution of P&I at week $t-1$. Similarly, we provide CH data at week $t-2$ to initialize the conditional distribution of CH at week $t-1$; CH data at week $t-3$ together with AH data at week $t-2$ will initialize the conditional distribution of AH at week $t-1$, while CH data at week $t-3$, and ILI data at week $t-2$ will initialize the conditional distribution of ILI at week $t-1$. The parameters that specify the transition distributions, as well as the time point t and the values of the covariates x_1 and x_2 are provided via the data file, and are iteratively updated for each weekly forecast.

Figure 8 shows the number of patients with ILI symptoms recorded in MA between November 1999 and October 2000 by the sentinel physicians in the CDC influenza surveillance program (black line). The red line depicts the two week ahead prediction computed by the dynamic model, while the green line shows the two week ahead prediction when CH data are removed from the model. The pale blue line is the two week ahead prediction when ILI data

are removed from the model. Besides the very close match between true and predicted values provided by our dynamic model, it is apparent that CH data alone are sufficient to forecast the dynamics of ILI data, while the model in which CH data are removed suffers of a delay. Statistical analysis of the forecasting error of our model showed no significant difference between observed and predicted values, while removing ILI data increases the average forecasting error that is still not significant, and removing CH data makes the average forecasting error significantly different from 0. Similar tests were conducted to assess the predictive effect of both CH and AH data on influenza deaths and confirmed the early signature of syndromic data.

5. Discussion

The increasing emphasis on biosurveillance, motivated by recent epidemic and bioterrorism events, has stimulated the electronic collection of different sources of data that are supposed to contain early signs of disease outbreaks. One of the major challenges that researchers in biosurveillance face at the moment is the integration of these data streams into models that can indeed alert public health officials of impeding disease outbreaks. To achieve these objectives, we need models that can integrate information from different data streams and that can be used for accurate forecasting.

Our analysis shows that dynamic Bayesian networks provide a simple but effective modeling tool for the integration of several time series data into a dynamic model. The modularity of dynamic Bayesian networks makes it possible to learn them from data comprising several variables using standard multiple regression techniques. Furthermore, the availability of software for stochastic computations with directed graphical models makes probabilistic inference with dynamic Bayesian networks very efficient, and removes the need to impose restrictive assumptions on the dependency structure or probability distribution of the network variables such as those imposed on Gaussian networks [5].

From the epidemiology point of view, our analysis identifies in the number of pediatric

patients who present to the ED with respiratory syndromes an early quantitative indicator of influenza epidemics. Compared to hospitalization rates [8] or school absenteeism [18] used for monitoring or detecting ongoing influenza outbreaks, pediatric patients with respiratory syndromes are able to predict upcoming epidemics two to three weeks in advance. The pediatric sentinel signal is early enough to allow ample time to step up vaccination [21, 27] and implement other effective control measures in the community to reduce the burden of illness and mortality.

Acknowledgements

This work was supported by the Alfred P. Sloan Foundation (2002-12-1).

References

[1] E. Arias and B. L. Smith. Deaths: Preliminary data for 2001. *National Vital Statistics Reports*, 51:1–45, 2003.

[2] C. B. Bridges, S. A. Harper, K. Fukuda, T. M. Uyeki, N. J. Cox, J. A. Singleton, and Advisory Committee on Immunization Practices. Prevention and control of influenza. recommendations of the advisory committee on immunization practices (ACIP). *MMWR Recommendation and Reports*, 52:1–34, 2003.

[3] E. Castillo, J. M. Gutierrez, and A. S. Hadi. *Expert Systems and Probabilistic Network Models*. Springer, New York, NY, 1997.

[4] J. Cheng and M. Druzdzel. AIS-BN: An adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. *J Artif Intell Res*, 13:155–188, 2000.

[5] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, New York, NY, 1999.

[6] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE T Pattern Anal*, 6:721–741, 1984.

[7] P. H. Gesteland, M. M. Wagner, W. W. Chapman, J. U. Espino, F. C. Tsui, R. M. Gardner, R. T. Rolfs, V. Dato, B. C. James, and P. J. Haug. Rapid deployment of an electronic disease surveillance system in the state of Utah for the 2002 olympic winter games. In *Proceedings of the Annual AMIA Fall Symposium*, pages 285–289, 2002.

[8] W. P. Glezen, M. Decker, S. W. Joseph, and R. G. Mercready. Acute respiratoty disease associated with influenza epidemics in Houston, 1981–1983. *The Journal of Infectious Diseases*, 155:1119–1126, 1987.

[9] A. Goldenberg, G. Shmueli, R. A. Caruana, and S. E. Fienberg. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proc Natl Acad Sci USA*, 99:5237—5240, 2002.

[10] K. S. Liand Y. Guan, J. Wang, G. J. Smith, K. M. Xu, L. Duan, A. P. Rahardjo, P. Puthavathana, C. Buranathai, T. D. Nguyen, A. T. Estoepangestie, A. Chaisingh, P. Auewarakul, H. T. Long, N. T. Hanh, R. J. Webby, L. L. Poon, H. Chen, K. F. Shortridge, K. Y. Yuen, R. G. Webster, and J. S. Peiris. Genesis of a highly pathogenic and potentially pandemic h5n1 influenza virus in eastern asia. *Nature*, 430:209–213, 2004.

[11] D. J. Hand, H. Mannila H., and P. Smyth. *Principles of data mining*. MIT Press, Cambridge, MA, 2001.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, 2001.

[13] R. E. Kass and A. Raftery. Bayes factors. *J Am Stat Assoc*, 90:773–795, 1995.

[14] S. L. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, UK, 1996.

[15] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J Roy Stat Soc B*, 50:157–224, 1988.

[16] R. Lazarus, K. Kleinman, I. Dashevsky, C. Adams, P. Kludt, A. DeMaria, and R. Platt. Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. *Emerg. Infect. Dis.*, 8:753–760, 2002.

[17] R. Lazarus, D. Vercelli, L. J. Palmer, W. J. Klimecki, E. K. Silverman, B. Richter, A. Riva, M. F. Ramoni, F. D. Martinez, S. T. Weiss, and D. J. Kwiatkowski. SNPs in innate immunity genes: Abundant variation and potential role in complex human disease. *Immunol Rev*, 190:9–25, 2003.

[18] D. D. Lenaway and A. Ambler. Evaluation of a school-based influenza surveillance system. *Public Health Reports*, 110:333–337, 1995.

[19] K. Mandl, J. M. Overhage, M. M. Wagner, W. L. Lober, P. Sebastiani, F. Mostashari, J. A. Pavlin, P. H. Gesteland, T. Treadwell, E. Koski, L. Hutwagner, D. L Buckeridge, R. D. Aller, and S. Grannis. Implementing syndromic surveillance: A practical guide informed by the early experience. *J Am Med Inform Assn*, 2004.

[20] M. I. Meltzer, N. J. Cox, and K. Fukuda. The economic impact of pandemic influenza in the united states: Priorities for intervention. *Emergency Infectious Disease*, 5:659–671, 1999.

[21] K. L. Nichol, A. Lind, K. L. Margolis, M. Murdoch, R. McFadden, M. Hauge, S. Magnan, and M. Drake. The effectiveness of vaccination against influenza in healthy, working adults. *New England Journal of Medicine*, 333:889–893, 1995.

[22] P. A. Patriarca and N. J. Cox. Influenza pandemic preparedness plan for the united states. *Journal of Infectious Disease*, 176:Suppl 1:S4–S7, 1997.

[23] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference*. Morgan Kaufmann, San Francisco, CA, 1988.

[24] T. Rath, M. Carreras, and P. Sebastiani. Automated detection of influenza epidemics with Hidden Markov Models. In M. R. Berthold H. J. Lenz E. Bradley R. Kruse and C. Borgelt, editors, *Advances in Intelligent Data Analysis V. 5th International Symposium on Intelligent Data Analysis, IDA 2003 Berlin, Germany, August 28-30, 2003 Proceedings*, pages 521–531, New York, 2003. Springer.

[25] B. Y. Reis and K. Mandl. Integrating syndromic surveillance data across multiple locations: Effects on outbreak detection performance. In *Proceedings of the Annual AMIA Fall Symposium*, 2003. In press.

[26] B. Y. Reis, M. Pagano, and K. D. Mandl. Using temporal context to improve biosurveillance. *Proc Natl Acad Sci USA*, 100:1961–1965, 2003.

[27] M. B. Rennels, H. C. Meissner, and Committee on Infectious Diseases. Technical report: Reduction of the influenza burden in children. *Pediatrics*, 110:e80, 2002.

[28] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Englewood Cliffs, NJ, 2nd edition, 2003.

[29] S. C. Schoenbaum. Economic impact of influenza. the individual's perspective. *American Journal of Medicine*, 82:26–30, 1987.

[30] R. E. Serfling. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Reports*, 78:494–506, 1963.

[31] L. Simenson, K. Fukuda, L. B. Schonberg, and N. J. Cox. The impact of influenza epidemics on hospitalizations. *The Journal of Infectious Diseases*, 181:831–837, 2000.

[32] D. J. Spiegelhalter, A. Thomas, and N. G. Best. Computation on Bayesian graphical models (with discussion). In J.M. Bernardo, J. Berger, A.P. Dawid, and A.F.M. Smith,

editors, *Bayesian Statistics 5*, pages 4–7–425, Oxford, UK, 1996. Oxford University Press.

[33] A. Thomas, D. J. Spiegelhalter, and W. R. Gilks. Bugs: A program to perform Bayesian inference using Gibbs Sampling. In J. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 837–42. Oxford University Press, Oxford, UK, 1992.

[34] F. C. Tsui, J. U. Espino, V. M. Dato, P. H. Gesteland, J. Hutman, and M. M. Wagner. Technical description of RODS: A real-time public health surveillance system. *J Am Med Inform Assn*, 10:399–408, 2003.

[35] L. Wang, M. F. Ramoni, K. D. Mandl, and P. Sebastiani. Factors affecting the performance of a syndromic surveillance system. *Artif Intell Med*, 2005. In press.

Figures and Tables

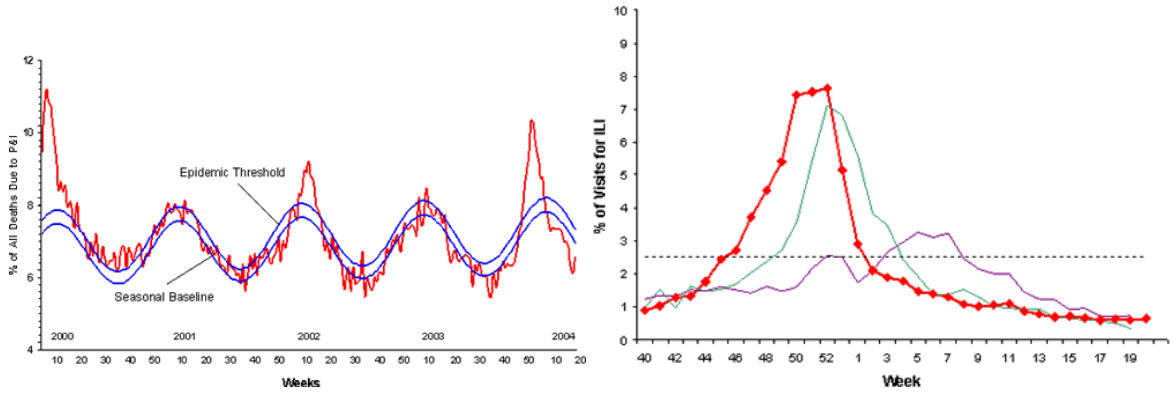


Figure 1: *Baseline model and epidemic threshold of P&I and ILI data used for detection and monitoring of influenza epidemics. The left image reports the P&I data from January 2000 through the end of 2004. The blue lines are the season baseline and epidemic threshold computed using cyclic regression. The red line depicts the percentage of P&I deaths. The right image plots, in red, the percentage of visits for ILI reported by sentinel providers for the influenza season 2003–2004 (week 40 of 2003 through week 20 of 2004). The green and maroon lines are the ILI data for the influenza seasons 1999-2000 and 2002-2003. The 1999-2000 influenza season was characterized by a severe mortality and morbidity, while no influenza epidemic was detected for the season 2002-2003. The dashed line is the national epidemic threshold.*

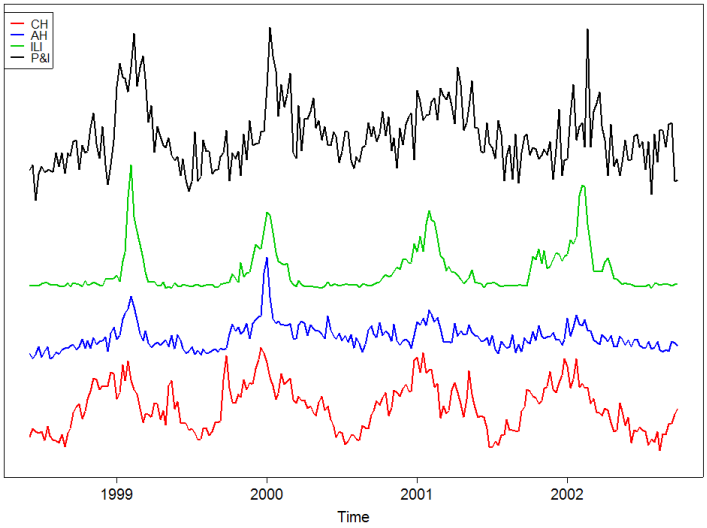


Figure 2: *Plots of the weekly number of respiratory syndrome cases at CH (red) and AH (blue), patients with ILI symptoms (green) and the number of deaths from P&I (black). The x-axis reports time in weeks from June 1998 to the end of September 2002.*

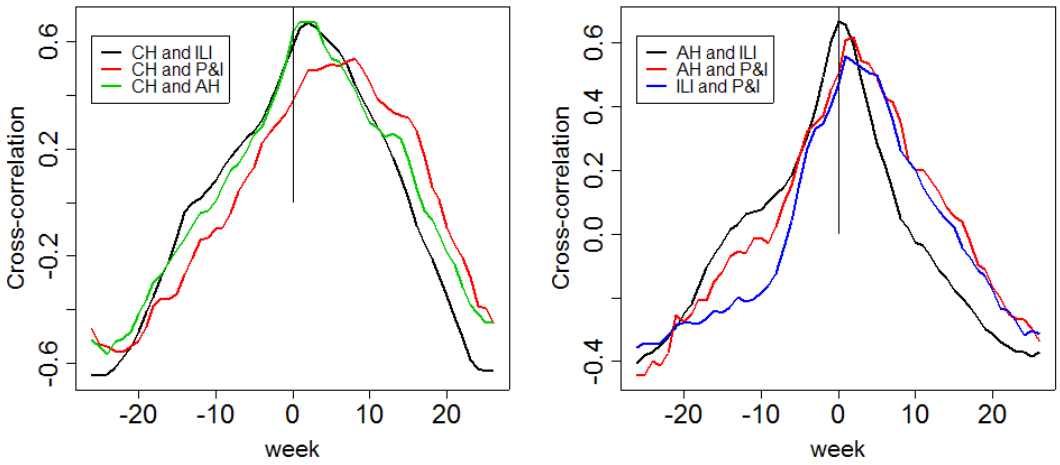
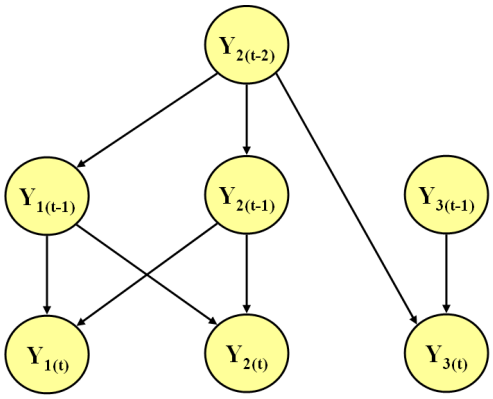


Figure 3: Left: *Cross-correlation (y-axis) between the weekly number of CH and ILI data, deaths from P&I, and the weekly number of AH data for different week lags (x-axis). For each ordered pair of time series in the legend, the cross-correlation for x weeks lag represents the correlation between the first time series shifted back by x weeks when $x > 0$, while the second time series is shifted back by x weeks when $x < 0$.* Right: *Cross-correlation between the weekly number of AH data, ILI cases and deaths from P&I. The line in blue displays the cross-correlation between the series of ILI cases and deaths from P&I.*



		$Y_{3(t)}$	
$Y_{2(t-2)}$	$Y_{3(t-1)}$	-	+
-	-	**0.9**	**0.1**
-	+	**0.7**	**0.3**
+	-	**0.6**	**0.4**
+	+	**0.1**	**0.9**

Figure 4: *A directed acyclic graph that represents the temporal dependency of three categorical variables with states - and +. The conditional probability table shows the transition probabilities that quantify the dependency of Y_3 on its parent nodes.*

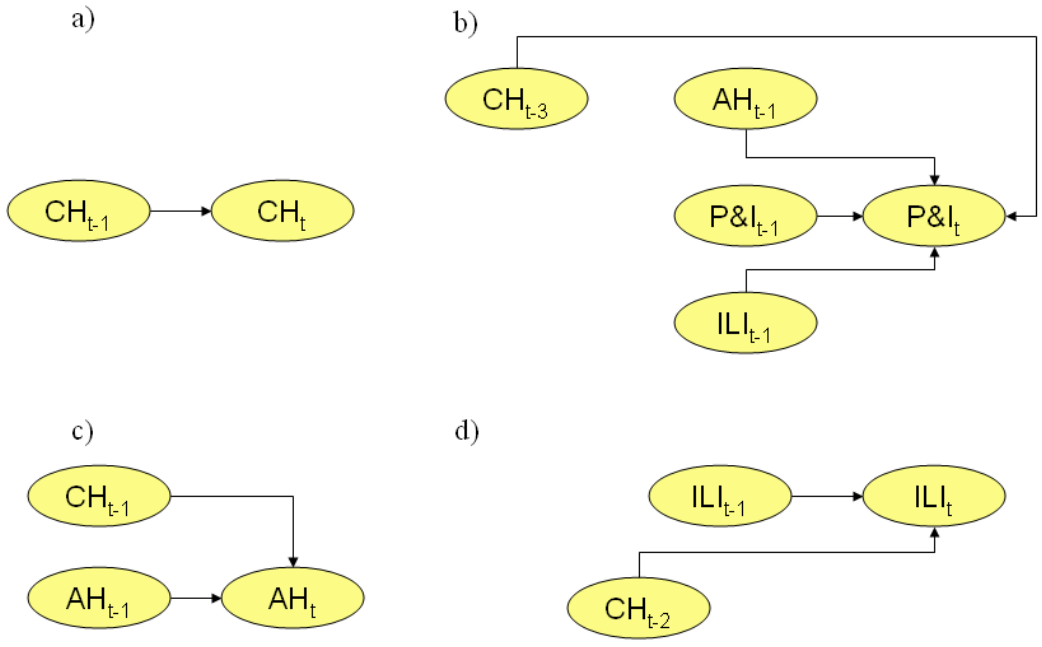


Figure 5: *The regression models for each of the four time series (CH, AH, P&I and ILI). All nodes in the graphs represent stochastic variables and arrows represent temporal dependencies quantified by probability distributions. An arrow from a node A to a node B means that A leads B by the number of weeks identified by the subscript. For example, the graph a) shows that the number of pediatric patients seen at the ED with respiratory syndrome at week* $t-1$ *(*CH_{t-1}*) leads the number of pediatric patients at week* t *(*CH_t*). This is a simple autoregressive model of order 1. The graph b) shows that the model describing the dynamics of P&I at week* t *has an autoregressive component (*$P\&I_{t-1}$*) as well as the three predictors* CH_{t-3}*,* AH_{t-1}*, and* ILI_{t-1}*. Therefore, CH data lead P&I data by three weeks, while both AH and ILI data lead the number of pneumonia and influenza deaths by one week only. Similarly, the graphs c) and d) show that CH data lead AH data by one week, and the number of patients with ILI symptoms by two weeks.*

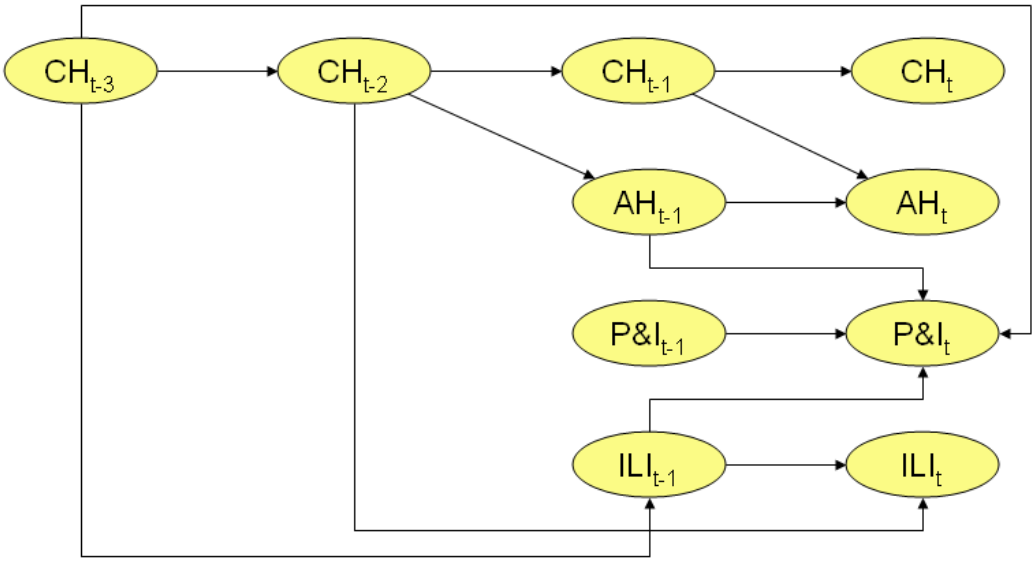


Figure 6: *The overall dynamic Bayesian network that describes the interplay between the four data streams*.

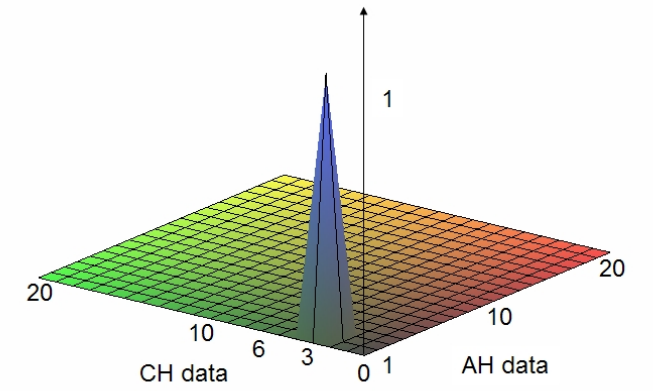


Figure 7: *Posterior probability (z-axis) for the models describing the dependency of the number of deaths for pneumonia and influenza on the number of pediatric and adult patients with respiratory syndromes for varying lags in weeks (x and y axes). The probability is maximum when the number of pediatric patients with respiratory syndrome is measured with a three week lag and the number of adult patients with respiratory syndrome is measured with one week lag and, compared to the other models, there is very strong evidence that this model gives the best fit (posterior probability almost 1).*

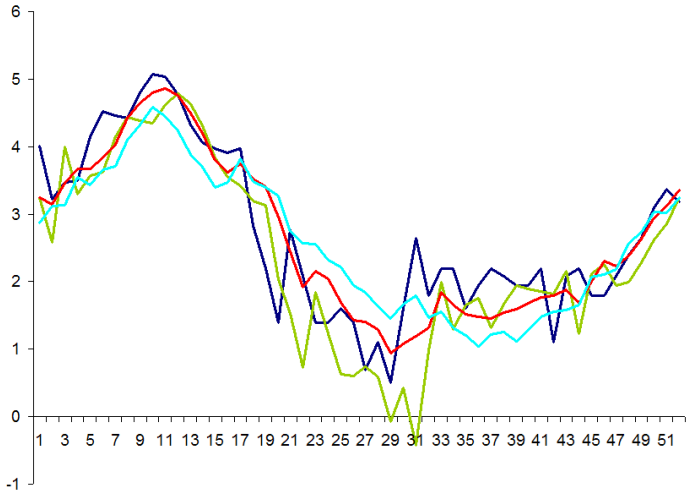


Figure 8: *Two step ahead prediction of ILI cases in Massachusetts between October 1999 and October 2000: the year of a major influenza epidemic. The blue line depicts the observed number of patients with ILI symptom (in logarithmic scale) between November 1999 and the end of October 2000. The red line depicts the values predicted by the model using only the data available until two weeks before. The green line shows the values predicted by the model when the past ILI data are ignored, and the line in pale blue depicts the values predicted by the same model using only the volume of pediatric patients with respiratory syndromes. The synchrony between the lines in dark and pale blue shows that that the signal provided by the pediatric respiratory syndrome data is sufficient to reconstruct the dynamics of influenza and to identify peaks of the epidemic with two weeks anticipation. The x-axis reports the week beginning with November 1999 till the end of October 2000.*

```
model
        {
    ### initial values of the time series
    d.t.2 <- pi[t-2]      ### P&I  at week t-2
    e.i.t.2 <- exp(ili[t-2]); #### ILI at week t-2 ILI are in log-scale
    a.t.2 <- bid[t-2]               #### AH data at week t-2
    c.t.4  <- cc[t-4]; c.t.3<-cc[t-3]; #### CH data at week t-4 and t-3

    #### P&I
    d.t.1 ~ dpois( d.mean.1)    ### P&I at week t-1
    log(d.mean.1)<- int.d+theta.d*d.t.2+ theta.a*a.t.2+ theta.c*c.t.4+
                theta.i*e.i.t.2+theta.x.1*x.1[t-1]+theta.x.2*x.2[t-1]
    d.t ~ dpois( d.mean)          ### P&I at week t
     log(d.mean) <- int.d+theta.d*d.t.1+ theta.a*a.t.1+theta.c*c.t.3+
                theta.i*e.i.t.1+theta.x.1*x.1[t]+theta.x.2*x.2[t]

     #### ILI
      i.t.2<-ili[t-2]
       e.i.t.1 ~ dlnorm( i.mean.1,i.tau)  ### ILI at week t-1
         i.mean.1 <- eta.i*i.t.2+eta.c*c.t.3+eta.x.1*x.1[t-1]
              i.t.1<-log(e.i.t.1)
           e.i.t ~ dlnorm( i.mean,i.tau)  ### ILI at week t
             i.mean <- eta.i*i.t.1+eta.c*c.t.2+eta.x.1*x.1[t]

     #### AH
         a.t.1~ dnorm( a.mean.1, a.tau) ### AH at week t-1
             a.mean.1 <- int.a + beta.a*a.t.2+ beta.c*c.t.2
            a.t ~ dnorm( a.mean, a.tau)  ### AH at week t
             a.mean <- int.a + beta.a*a.t.1+ beta.c*c.t.1

      #### CH
          c.t.1 ~ dnorm( c.mean.1, c.tau)    ### CH at week t-1
          c.mean.1 <- int.c + gamma.c*c.t.2+gamma.x.1*x.1[t-1]
           c.t ~ dnorm( c.mean, c.tau)     ### CH at week t
           c.mean <- int.c + gamma.c*c.t.1+gamma.x.1*x.1[t]
}
```

Figure 9: *BUGS code used for the two-week-ahead forecast. The parameters are updated at each iteration.*

Regression model for P&I			Regression model for $\log$(ILI)		
	Estimate	Std. Error		Estimate	Std. Error
(Intercept)	3.5542	0.0656	$\log(\text{ILI}_{t-1})$	0.6919	0.0443
$\text{P\&I}_{t-1}$	0.0045	0.0007	CH_{t-2}	0.0056	0.0008
AH_{t-1}	0.0031	0.0005	x_1	0.1502	0.0579
ILI_{t-1}	0.0006	0.0003	x_2	-0.1267	0.0542
CH_{t-3}	-0.0014	0.0004			
x_1	0.1153	0.0234			
x_2	0.1370	0.0189			

Regression model for AH			Regression model for CH		
	Estimate	Std. Error		Estimate	Std. Error
(Intercept)	17.6199	4.4835	(Intercept)	64.8846	7.9730
AH_{t-1}	0.5528	0.0530	CH_{t-1}	0.5827	0.0505
CH_{t-1}	0.1755	0.0294	x_1	24.7580	3.5840

Table 1: *Estimates and standard errors of the regression parameters for the four models fitted to CH, AH, ILI and P&I data when all the data are used to induce the network. The model fitted to P&I data is a log-linear model, so that the regression coefficients parameterize the logarithm of the mean. The model fitted to ILI data is in logarithmic scale. The variables* x_1 *and* x_2 *are periodical functions defined as* $x_1 = \sin(2\pi t/52)$ *and* $x_2 = \cos(2\pi t/52)$.