Algorithms for detecting and predicting influenza outbreaks: meta-narrative analysis of prospective evaluations: Detailed performance of the evaluated algorithms

Evaluation settings and algorithms

The three outbreak detection studies that fulfilled the study criteria used data sets from the United States (n=1), Spain (n=2), and Hong Kong (n=1). The five studies that evaluated outbreak prediction algorithms were set in the United States (n=2), France (n=1), Sweden (n=1) and China (n=1). In all outbreak detection studies, the syndromic data used originated from emergency department records. All three studies reported evaluations of regression-based and temporal modelling (RBTM) techniques. These methods compare observed patterns with those predicted by a model developed using historical data. This approach requires a model of the baseline pattern as well as the selection of a threshold to signal an alarm. In one study, an RBTM-based technique was the only algorithm evaluated; in one study, an RBTM-based technique was the main algorithm (compared with other algorithms); and in one study, an RBTM-based technique was one of the main algorithms. Two studies evaluated a statistical process control (SPC) method. SPC methods rely on cumulative differences between observed and expected data in a time window compared with a threshold. Both studies evaluated a cumulative sum (CUSUM)-based algorithm. The evaluation designs involved application of a decision rule on a weekly basis.

In the five outbreak prediction studies fulfilling the selection criteria, syndromic data from hospital emergency department visits (n=4), laboratory data (n=1), telenursing data (n=1), military clinic visits (n=1) and search query data (n=1) were analysed. One of the studies

evaluated a regression-based algorithm in combination with an exponential smoothing technique/algorithm, one assessed a non-parametric time series forecasting method (the method of analogues), a linear autoregressive model, and the naive method[43], one appraised a Bayesian network-based technique, one evaluated a Shewhart-type algorithm[44] and one study applied a multiple linear regression model. For the evaluations, data were analyzed on a daily (n=3), weekly (n=2) or on a monthly (n=1) basis.

Evaluation outcomes for outbreak detection algorithms

Two studies used sensitivity and specificity as evaluation measurements. One of these studies used only these two measurements while the other used these measurements in combination with timeliness and VUTROC. The third study used AUWROC as evaluation measurement.

From a Spanish study[17], acceptable performance was reported for the RBTM technique using the Kolmogorov-Smirnov test, based on 1.00 sensitivity and 0.88 specificity. From another Spanish study[18], outstanding performances were reported for the RBTM techniques using two hidden Markov models and the Serfling model. These algorithms produced AUWROC values between 0.93 and 0.98. Here, the SPC method using CUSUM and the RBTM technique of simple regression showed poor performance according to AUWROC.

From a study using data from Hong Kong and the United States[19], excellent performances were reported in 2-week evaluations (alarms generated within the first 2 weeks of the peak season) for the RBTM techniques using time series analysis (dynamic linear model) and simple regression and the SPC method using CUSUM when applied on US data, based on estimates of VUTROC between 0.81 and 0.90. The performances on Hong

Kong data for the RBTM techniques were acceptable (VUTROC 0.77 for the time series model and VUTROC 0.75 for the simple regression model), whereas the performance was poor for the SPC method using CUSUM (VUTROC 0.56) for the 2-week evaluation period.

Evaluation outcomes for outbreak prediction algorithms

Two of the five studies used correlations between predicted and observed data as a performance measurement (one of these used percent error in addition to correlation), and one study used area under curve (AUC), positive predictive value (PPV), sensitivity and specificity as performance metrics. The two remaining studies used metrics based on estimates of residuals only allowing rank comparisons to be made, one of these studies using only absolute percent error (APE) and the second using median absolute residual (MAD) and median absolute percent error (MedAPE).

In a French study predicting influenza outbreaks over 18 seasons[24], excellent performance (r=0.81) was observed for a non-parametric time series method in 1-week-ahead predictions, and poor performance (r=0.66) in 10-week-ahead predictions. The performance of an autoregressive algorithm was reported as acceptable (r=0.73) in 1-week-ahead predictions and poor (r=0.07) in 10-week-ahead predictions; the performance of the naive method was reported as poor in both predictions.

A study using county-level data from one single influenza outbreak in the United States[22] reported outstanding performance for a Bayesian network algorithm for predictions of the remaining outbreak made on day 13 (r=0.94) and day 22 (r=0.97) of an ongoing outbreak. Another study used respiratory illness data from the United States[23] to evaluate methods for automatic preconditioning of syndromic data (adjusting for day-of-week and seasonal effects) to enable use of SPC methods for outbreak detection. Two regression models

(adaptive and non-adaptive) and the Holt-Winters approach for data-driven smoothing were applied to ten different series of count data on respiratory illness, computing 7-days-ahead forecasts. Using MAD as the evaluation metric, the Holt-Winters approach was found to be superior to the adaptive and non-adaptive regression models in seven of the ten series, and in eight of the ten series using MedAPE as metric.

A study using telenursing data from a Swedish county to predict influenza outbreaks over 3 seasons including the H1N1 pandemic in 2009, showed outstanding performance for seasonal influenza outbreaks on a daily basis (AUC 0.89; PPV 0.93) and excellent performance on a weekly basis (AUC 0.83; PPV 1.00)[20]. However, the performance for the pandemic outbreak was poor on a daily basis (AUC 0.84; PPV 0.58) and also poor (at most acceptable) on a weekly basis (AUC 0.78; PPV 0.79). For one study from China using monthly search query data (from Baidu) to predict laboratory confirmed data it was difficult to evaluate the absolute performance, however it displayed a mean absolute percent error (MAPE) of 11% over 8 months, ranging from 1.2% to 22.2%[21].

References

43. Stone L. Coloured noise or low-dimensional chaos? *Proc Biol Sci* 1992;250:77–81.

44. Montgomery DC. Introduction to Statistical Quality Control. 6th ed. New York: John Wiley and Sons 2008.