Epidemiol. Infect. (2017), **145**, 2166–2175. © Cambridge University Press 2017
doi:10.1017/S0950268817001005

Influenza detection and prediction algorithms: comparative accuracy trial in Östergötland county, Sweden, 2008–2012

A. SPRECO[1]*, O. ERIKSSON[2], Ö. DAHLSTRÖM[3] AND T. TIMPKA[1,4]

[1] *Department of Medical and Health Sciences, Linköping University, Linköping, Sweden*
[2] *Department of Computer and Information Science, Linköping University, Linköping, Sweden*
[3] *Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden*
[4] *Center for Health Services Development, Region Östergötland, Linköping, Sweden*

Received 22 October 2016; Final revision 24 April 2017; Accepted 24 April 2017; first published online 17 May 2017

SUMMARY

Methods for the detection of influenza epidemics and prediction of their progress have seldom been comparatively evaluated using prospective designs. This study aimed to perform a prospective comparative trial of algorithms for the detection and prediction of increased local influenza activity. Data on clinical influenza diagnoses recorded by physicians and syndromic data from a telenursing service were used. Five detection and three prediction algorithms previously evaluated in public health settings were calibrated and then evaluated over 3 years. When applied on diagnostic data, only detection using the Serfling regression method and prediction using the non-adaptive log-linear regression method showed acceptable performances during winter influenza seasons. For the syndromic data, none of the detection algorithms displayed a satisfactory performance, while non-adaptive log-linear regression was the best performing prediction method. We conclude that evidence was found for that available algorithms for influenza detection and prediction display satisfactory performance when applied on local diagnostic data during winter influenza seasons. When applied on local syndromic data, the evaluated algorithms did not display consistent performance. Further evaluations and research on combination of methods of these types in public health information infrastructures for 'nowcasting' (integrated detection and prediction) of influenza activity are warranted.

Key words: Algorithms, epidemiological methods, evaluation research, human influenza, signal detection analysis.

INTRODUCTION

Recent technical developments in the area of public health information infrastructure make it realistic to collect, structure, and statistically analyse infectious disease data in close to real time and in local public health contexts [1]. Early knowledge of influenza epidemics in the community allows local epidemic alerts in primary care and hospital settings before the publication of regional data and could accelerate the implementation of preventive transmission-based precautions both within the local health care services and the community [2]. In the past few years, a considerable amount of research has focused on developing statistical methods to identify aberrations in disease incidence data accurately and quickly [3–5] as well as predicting epidemic progress [6–8]. Other

* Author for correspondence: A. Spreco, Division of Social Medicine, Department of Medical and Health Sciences, Faculty of Health Sciences, Linköping University, SE-581 83 Linköping, Sweden.
(Email: armin.spreco@liu.se)

researchers have focused on the use of alternative data sources, such as internet search engines [9,10], miniblogs [11,12], and records from over-the-counter drug sales [13,14] with the goal of enhancing detection and prediction outcomes. However, few studies have compared the performance of different algorithms in routine practice using prospective designs. In a meta-narrative review of influenza detection and prediction algorithm evaluations in public health settings [15], only three studies covering seven detection algorithms and five studies covering nine prediction algorithms were found to have been performed using prospective designs. We inferred that further research is needed where algorithms are comparatively evaluated in parallel in the same setting using identical data.

The aim of this study was to perform a comparative trial of algorithms for the detection and prediction of influenza activity using local data from a county-wide public health information system.

METHODS

The study applied an accuracy trial design [16] based on two streams of data used for routine influenza surveillance in a Swedish county (population 445 000): data on clinical diagnoses recorded by physicians and syndromic chief complaint data from a national telenursing service. The latter source has been found to provide indications of increased influenza activity up to 2 weeks ahead of the former [10,17]. The primary criteria for inclusion of an influenza detection or prediction algorithm were that it had been evaluated using authentic prospective data and the report had been published in a peer-reviewed scientific journal before 1 February 2016. The secondary criteria were that the algorithm (1) was to be applicable in county-level influenza surveillance, i.e. on unidimensional influenza data from a population of approximately 500 000 inhabitants, (2) it was sufficiently documented to be reproduced, (3) it could be calibrated using a maximum of one season of learning data, and (4) the detailed assumptions about data characteristics were compatible with the county-level data used for the evaluation. The study design was approved by the Regional Research Ethics Board in Linköping (dnr. 2012/104-31).

Data sources

The study data were collected from an electronic health data repository maintained by a Swedish county council [1]. The repository collects data from all patient visits at health care facilities in the county and from calls made by county residents to the nationwide telenursing service. For the study, influenza diagnosis codes (International Classification of Diseases, 10th Revision (ICD-10)/International Conference on Drugs and Pharmacological Classification (ICDPC) J10.0, J10.1, J10.8, J11.0, J11.1, J11.8, and J11.0-P) and telenursing chief complaints potentially associated with influenza, i.e. fever (adult, child), cough (adult, child), headache (adult, child), dyspnea, sore throat, vertigo, lethargy, and syncope were used. Collection of learning data used to calibrate the algorithms covered the winter influenza season of 2008–2009, starting from the end of the previous winter influenza season (4 May 2008 to 25 April 2009). Immediately after the end of the learning period, the evaluation period started, covering one pandemic outbreak and two winter influenza seasons (26 April 2009 to 19 May 2012) (Fig. 1). Because the evaluation period included both the pandemic outbreak and winter influenza seasons, it was divided into two parts; one part covered the pandemic outbreak and the other part covered the winter influenza seasons. The epidemic threshold was defined as two incident influenza diagnosis cases per 100 000 population recorded during a 7-day period.

Evaluation procedure

For diagnostic data, the learning dataset was used to retrospectively decide parameter settings for the different detection and prediction methods. These parameters were then formatively applied in retrospective analyses using the learning dataset. For telenursing data, the learning set was first used to determine which time lag and grouping of chief complaints had the largest strength correlation with the diagnostic data. The chief complaint grouping with the largest correlation strength and best time lag was chosen for the following analyses. Thereafter the learning set was used to determine parameter settings for the different detection and prediction methods.

Detection performance was evaluated using measurements of sensitivity (the proportion of correctly identified weeks with increased influenza activity), specificity (the proportion of correctly identified weeks with no increased influenza activity), and timeliness (the time-difference between the observed and the predicted start of a period with increased influenza activity), whereas prediction performance was compared

https://doi.org/10.1017/S0950268817001005 Published online by Cambridge University Press

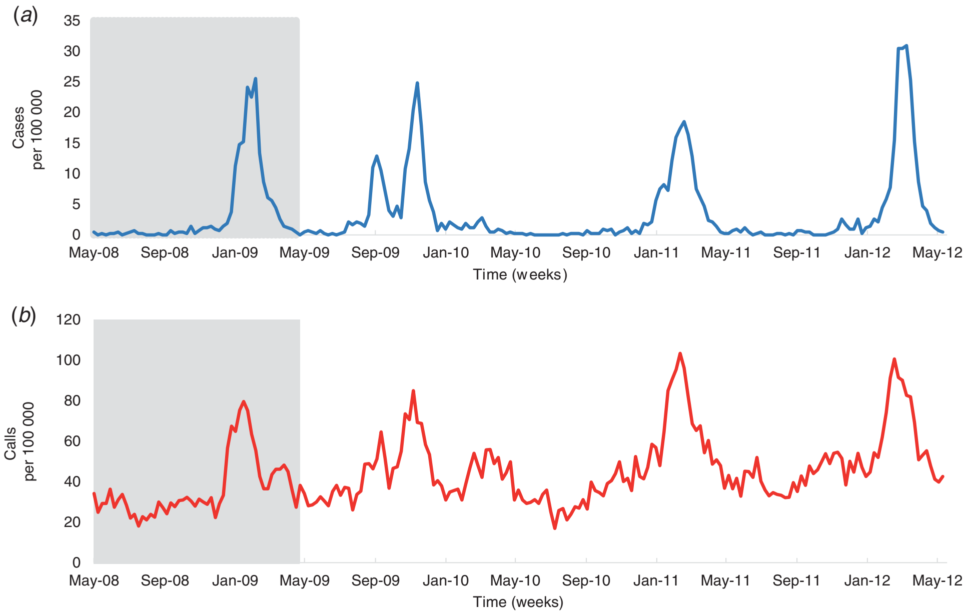

Fig. 1. Weekly rates of influenza diagnosis cases (*a*) and telenursing calls for fever (child, adult) (*b*) in Östergötland County, Sweden, during the retrospective learning period from May 2008 to April 2009 (the gray marked area) and the prospective evaluation period from April 2009 to May 2012.

using Pearson correlation (*r*) and median absolute percentage error (MedAPE), both representing the association between predicted and observed time series of influenza activity.

In ranking the detection algorithms, specificity was given priority over sensitivity because a high level of false alarms is unacceptable in public health practice. If several algorithms performed similarly with regard to specificity and sensitivity; timeliness was used to decide which algorithm was superior. The calculation for specificity was based on the 10 weeks immediately before an epidemic and the calculation for sensitivity was based on the first 10 weeks of an epidemic. The reason why these measures were not based on entire datasets was that detection methods are primarily optimized to detect epidemics. The performance of an algorithm was considered acceptable if the specificity was at least 0·85 and the sensitivity was at least 0·80.

In the evaluation of prediction algorithms, the Pearson correlation coefficient (*r*) was used as the primary measurement of the association between observed and predicted values. The limits used to interpret the observed values were modified from the Cohen scale [18], in which the limits 0·10, 0·30, and 0·50 were defined as small, medium, and large effect sizes. In this study, the limits were set at 0·70, 0·80, and 0·90 denoting acceptable, excellent, and outstanding predictive performance. The secondary evaluation measurement, MedAPE, is a measure of the accuracy of statistical methods for constructing fitted time series values [19]. For a perfect fit, MedAPE is 0; it has no restriction with regard to its upper level. MedAPE gives an idea of the typical percentage error and allows comparisons across different series. The combination of correlation and absolute percentage error (MedAPE before the median is calculated) has been used previously [20].

Detection algorithms

Influenza detection was defined as indicating the initiation of a prolonged period on increased influenza activity in the population under surveillance. At the time of algorithm selection, seven influenza detection algorithms were found to have been evaluated using authentic prospective data. Four of these methods did not meet the secondary inclusion criteria. The algorithm based on the Kolmogorov–Smirnov test evaluated by Closas *et al.* [21] was excluded because it was not applicable on streams of county-level influenza diagnosis data. This test assumes that the rate of influenza diagnosis cases in non-epidemic periods can be represented by a random variable (*y*) that is exponentially distributed. However, influenza diagnosis case rate data in local settings are in general represented by small integers during non-epidemic

https://doi.org/10.1017/S0950268817001005 Published online by Cambridge University Press

periods. Therefore, it is more reasonable to assume that the observations are Poisson distributed. The time series method based on a dynamic model [22], was excluded for similar reasons, i.e. it requires that data follow a normal distribution. Finally, the two hidden Markov models evaluated by Martínez–Beneito *et al.* [23] were excluded because the algorithms partly relied on simulated data.

The three detection algorithms meeting the study criteria were Serfling regression [23,24], 'simple regression' [22,25], and cumulative sum (CUSUM) [22,23]. Serfling regression [23,24] monitors the period when there is no increased influenza activity to determine a baseline defined by a fixed threshold. Defining $\widehat{Y}_t$ as the number of influenza diagnosis cases (or telenursing calls), the model is defined as

$$\widehat{Y}_t = b_0 + b_1 t + b_{2(t)}\cos(kt) + b_{3(t)}\sin(kt) ,$$

where b_0 is a constant intercept, b_1 is the slope of a long-term trend, and $b_{2(t)}$, $b_{3(t)}$ are coefficients for continuous harmonic terms representing seasonal trends, with $k = (2\pi/(365 \cdot 25/7))$ to give a 1-year sinusoidal period of these terms. Using only the non-epidemic phases of the learning set, we first determined the coefficients mentioned above. Using these on the learning set, we searched for the optimal threshold α (the threshold that generates the highest sensitivity and specificity) which is based on the normal distribution, investigating $\alpha = 0\cdot005, 0\cdot010, \ldots, 0\cdot500$.

Simple regression [22,25] raises an alarm if data from the current week fall outside a $100(1-\alpha)\%$ forecast interval from a normal distribution with running mean $\tilde{y}_{(m)}$ and running sample variance $\tilde{s}_{(m)}{}^2$ calculated from the preceding m weeks. The forecast interval is calculated as $\tilde{y}_{(m)} \pm t_{m-1,\,1-\alpha/2}\tilde{s}_{(m)}\sqrt{1+(1/m)}$, where $t_{m-1,1-\alpha/2}$ is the $100(1-\alpha)$th percentile of the Student t-distribution with $m-1$ degrees of freedom. Using the learning set, we searched for the parameter combination that resulted in the highest sensitivity and specificity for this algorithm, investigating all possible combinations of $\alpha = 0\cdot005, 0\cdot010, \ldots, 0\cdot500$ and $m = 3, 4, \ldots, 10$.

Using the CUSUM method [22,23], an alarm is raised if the upper CUSUM C_t^+ exceeds a pre-specified threshold g. For the series of observations y_t, $t = 1, 2, \ldots$, the d-week upper CUSUM at time t, C_t^+ is defined as

$$C_t^+ = \max\left(0,\ \frac{y_t - \tilde{y}_{(7)}}{\tilde{s}_{(7)}} - k + C_{t-1}^+\right),$$

with $C_{t-d}^+ = 0$ [26]. The running mean $\tilde{y}_{(7)}$ and running variance $\tilde{s}_{(7)}{}^2$ are calculated from the series of 7 weeks, $y_{i-d-7}, \ldots, y_{i-d-1}$ preceding the most recent d weeks. The d denotes the number of weeks excluded from the running mean and variance immediately before the index week. This is done in order to avoid contamination with the upswing of an epidemic [27]. The parameter k represents the minimum standardized difference from the running mean, which is not ignored by the CUSUM calculation.

The CUSUM algorithm was evaluated in its original form and in two modified versions based on that the observations y follow a Poisson distribution. In the first modified version the variance is estimated by the sample mean, since the variance equals the expected value in a Poisson distribution. In the second modified version CUSUM at time t (C_t) is expressed in terms of accumulated probability, namely $C_t = \max(0;\ C_{t-1} + P(Y \le y|E = \tilde{y}_{(7)}) - k)$ where P is the Poisson probability function. The second suggested modification is an adaption to Poisson distributed data with so low expected values that normal approximation is inappropriate. In this case, the pre-specified alarm threshold (g) cannot be based on the normal distribution. Using the learning set, we searched for the parameter combination that generated the highest sensitivity and specificity for all three CUSUM methods, investigating all possible combinations of $g = 0\cdot00, 0\cdot01, \ldots, 20\cdot00$, $k = 0\cdot00, 0\cdot01, \ldots, 3\cdot00$ ($k = 0\cdot00 - 1\cdot00$ for the third method) and $d = 0, 1, \ldots, 4$ for diagnostic data; and $g = 0\cdot00, 0\cdot01, \ldots, 100\cdot00$, $k = 0\cdot00, 0\cdot01, \ldots, 4\cdot00$ ($k = 0\cdot00 - 1\cdot00$ for the third method) and $d = 0, 1, \ldots, 4$ for telenursing data.

Prediction algorithms

Influenza prediction was defined as foretelling the amplitude and time span of a detected increase in influenza activity in a specified population. Nine influenza prediction algorithms were found to have been evaluated using authentic prospective data. Six of these did not meet the secondary inclusion criteria for this comparative trial. The Holt–Winters method (generalized exponential smoothing) [19] and the method of analogs [28] were excluded because they required collection of learning data from more than one influenza season. The autoregressive model [28] was excluded because the evaluation data did not comply with its detailed assumptions. A Bayesian network model [20], a Shewhart-type algorithm [17], and a multiple linear regression algorithm [29] were excluded due to that they required access to

https://doi.org/10.1017/S0950268817001005 Published online by Cambridge University Press

multidimensional data, i.e. a syndromic data source to predict influenza case rates.

The first algorithm meeting the study criteria, non-adaptive log-linear regression, fits an ordinary least squares, log-linear model to a learning set to obtain regression coefficients [30]. These coefficients are then used to forecast values beyond the learning data without adjusting for subsequent changes in time series behavior. Modified for weekly data, the model reads as follows:

$$\log(Y_\mathrm{t}+1)=b_0+b_1t+b_2\cos(kt)+b_3\sin(kt),$$

where Y_t is the number of influenza diagnosis cases or telenursing calls on week t, b_0 is a constant intercept, b_1 is the slope of a long-term trend and b_2 and b_3 are coefficients for continuous harmonic terms representing seasonal trends, with

$$k=\frac{2\pi}{(365\cdot25/7)}$$

to give a 1-year sinusoidal period of these terms. The reason for transforming the original weekly counts into log scale is to capture a multiplicative nature of the effects of the trend and seasonal components.

The second algorithm, adaptive log-linear regression with a sliding 8-week baseline interval, recomputes the regression coefficients for each forecast using only the series values from the 8 weeks before the forecast week [31]. The short baseline is intended to capture recent seasonal and trend patterns [19]. Modified for weekly data the model is described as follows:

$$\log(Y_\mathrm{t}+1)=b_0+b_1t,$$

where Y_t is the number of influenza diagnosis cases or telenursing calls in week t, b_0 is a constant intercept, and b_1 is the slope of a long-term trend. In Burkom *et al.* [19] a holiday indicator was added to avoid exaggerated holidays occurring in the short baseline interval; however, because weekly counts are used, the holiday effect are considered to be low or non-existing. Adjustments to the suggested models were made to fit weekly counts, due to that daily counts were used in Burkom *et al.* [19].

The final algorithm, the so-called naive method, predicts that a future incidence F is equal to a current incidence I, hence $F(T+h)=I(T)$, where $h\geqslant 1$ and t is the current week number.

RESULTS

Retrospective algorithm calibration

When re-applied on the retrospective diagnostic data, four of the detections algorithms (Serfling regression and the three CUSUM methods) displayed a perfect performance (i.e. specificity and sensitivity 1·00 and timeliness 0 week) (Table 1), whereas the simple model performed weakly (specificity 0·90, sensitivity 0·70, and timeliness 23 weeks). Regarding the telenursing data, the grouping of telenursing chief complaints with the largest correlation strength on a weekly basis ($r=0{\cdot}91$; $P<0{\cdot}001$) and longest lead time (2 weeks) to diagnostic data in the retrospective dataset was fever (child, adult). Based on these observations, fever (child, adult) was chosen as the complaint grouping for use in the evaluations. The detection algorithms performing best on telenursing data were Serfling regression and the three CUSUM methods, although all displayed less accurate performance than for diagnostic data (specificity 1·00, sensitivity 0·80, and timeliness −2 weeks) (Table 1). The simple model also performed poorly for the telenursing data (specificity 0·90, sensitivity 0·50, and timeliness 12 weeks).

The prediction algorithm that performed best on the retrospective diagnostic dataset was non-adaptive log-linear regression ($r=0{\cdot}72$ for 2-week-ahead predictions and $r=0{\cdot}57$ for 4-week-ahead predictions) followed by the naive method ($r=0{\cdot}62$ for 2-week-ahead predictions and $r=0{\cdot}29$ for 4-week-ahead predictions) (Table 2). However, MedAPE for the non-adaptive log-linear regression was the poorest of the methods (0·92 for 2-week predictions and 1·00 for 4-week predictions). For the telenursing data, the algorithm with the notably best predictive performance was non-adaptive log-linear regression ($r=0{\cdot}72$ for 2-week predictions to $r=0{\cdot}66$ for 4-week predictions and MedAPE 0·15 for 2-week predictions and 0·16 for 4-week predictions) (Table 2).

Prospective evaluations of detection algorithms

In the comparative evaluation using diagnostic data, the best performing detection algorithms were Serfling regression for the winter influenza seasons (specificity 1·00, sensitivity 0·80, and timeliness −2 weeks) and the CUSUM Poisson (alternative 2) for the pandemic outbreak in 2009 (specificity 0·80, sensitivity 1·00, and timeliness 1 week) (Table 3).

In the corresponding evaluation using the syndromic telenursing data, none of algorithms displayed

https://doi.org/10.1017/S0950268817001005 Published online by Cambridge University Press

Table 1. *Performance of influenza detection algorithms when retrospectively applied on the learning set of influenza diagnosis data and syndromic telenursing data*

Algorithm	Parameter combination	Specificity	Sensitivity	Timeliness[a]
Influenza diagnosis data				
Serfling regression	$\alpha = 0·035$	1·00	1·00	0
Simple regression	$m = 9$ weeks, $\alpha = 0·075$	0·90	0·70	23[b]
CUSUM alt1	$d = 2$, $k = 0·25$, $g = 18·49$	1·00	1·00	0
CUSUM alt2	$d = 2$, $k = 0·25$, $g = 20·73$	1·00	1·00	0
CUSUM alt3	$d = 2$, $k = 0·60$, $g = 4·55$	1·00	1·00	0
Telenursing data				
Serfling regression	$\alpha = 0·135$	1·00	0·80	−2
Simple regression	$m = 9$ weeks, $\alpha = 0·075$	1·00	0·50	12
CUSUM alt1	$d = 2$, $k = 0·13$, $g = 15·16$	1·00	0·80	−2
CUSUM alt2	$d = 2$, $k = 0·13$, $g = 25·07$	1·00	0·80	−2
CUSUM alt3	$d = 2$, $k = 0·55$, $g = 3·93$	1·00	0·80	−2

[a] Positive timeliness means that the alarm is raised before the epidemic has started (i.e. the alarm is raised too early) and negative timeliness means that the alarm is raised after the epidemic has started (i.e. the alarm is raised too late).
[b] One value stands out, otherwise the timelines would have been 0.

Table 2. *Performance of influenza prediction algorithms when retrospectively applied on the learning set of influenza diagnosis data and syndromic telenursing data*

	Prediction k weeks ahead	2 weeks	3 weeks	4 weeks
Influenza diagnosis data				
Correlation (r)[a]	Non-adaptive log-linear regression	0·72	0·65	0·57
	Adaptive log-linear regression[b]	0·47	0·26	0·12
	The naive method[c]	0·62	0·45	0·29
MedAPE[a]	Non-adaptive log-linear regression	0·92	1·00	1·00
	Adaptive log-linear regression[b]	0·76	0·86	0·98
	The naive method[c]	0·74	0·77	1·00
Telenursing data				
Correlation (r)[a]	Non-adaptive log-linear regression	0·72	0·69	0·66
	Adaptive log-linear regression[b]	0·45	0·27	0·14
	The naive method[c]	0·64	0·49	0·34
MedAPE[a]	Non-adaptive log-linear regression	0·15	0·16	0·16
	Adaptive log-linear regression[b]	0·28	0·34	0·40
	The naive method[c]	0·14	0·20	0·20

[a] The correlation coefficient (r) ranks the algorithms from highest (best) to lowest (worst) values, while MedAPE ranks them from lowest (best) to highest (worst) values.
[b] The method does not have the same learning set because it is adaptive, which means that the parameters are updated every week.
[c] The method has no learning set. The predicted value k weeks ahead is the same as the true value k weeks before.

a satisfactory performance. The algorithm performing best was Serfling regression with specificity 1·00, sensitivity 0·55, and timeliness −4 weeks for the winter influenza seasons and specificity 1·00, sensitivity 0·70, and timeliness −2 weeks for the pandemic outbreak (Table 3).

Prospective evaluations of prediction algorithms

The best performing prediction algorithm based on diagnostic data for the winter influenza seasons was non-adaptive log-linear regression ($r = 0·77$ for 2-week-ahead predictions, $r = 0·76$ for 4-week-ahead predictions, and MedAPE varying from 0·75 for

Table 3. *Performance of influenza detection algorithms when applied prospectively on influenza diagnosis data and syndromic telenursing data, respectively*

		Winter influenza seasons			Pandemic		
Algorithm	Parameter combination	Specificity	Sensitivity	Timeliness[a]	Specificity	Sensitivity	Timeliness
Influenza diagnosis data							
Serfling regression	$\alpha = 0·035$	1·00	0·80	−2	0·60	0·90	4
Simple regression	$m = 9$ weeks, $\alpha = 0·075$	0·70	0·75	20	0·57	0·30	4
CUSUM alt1	$d = 2$, $k = 0·25$, $g = 18·49$	0·65	1·00	3	0·60	1·00	2
CUSUM alt2	$d = 2$, $k = 0·25$, $g = 20·73$	0·65	0·95	4[b]	0·80	1·00	1
CUSUM alt3	$d = 2$, $k = 0·60$, $g = 4·55$	0·85	0·70	5[b]	1·00	0	–[c]
Telenursing data							
Serfling regression	$\alpha = 0·135$	1·00	0·55	−4	1·00	0·70	−2
Simple regression	$m = 9$ weeks, $\alpha = 0·075$	0·80	0·45	19	0·86	0·30	5
CUSUM alt1	$d = 2$, $k = 0·13$, $g = 15·16$	0·20	1·00	9	1·00	0·50	−5
CUSUM alt2	$d = 2$, $k = 0·13$, $g = 25·07$	0·25	0·95	7	1·00	0·40	−6
CUSUM alt3	$d = 2$, $k = 0·55$, $g = 3·93$	0·40	0·80	7	1·00	0	–[c]

[a] Positive timeliness means that the alarm is raised before the epidemic has started (i.e. the alarm is raised too early) and negative timeliness means that the alarm is raised after the epidemic has started (i.e. the alarm is raised too late).
[b] The mean of the absolute values.
[c] An alarm is never raised.

Table 4. *Performance of influenza prediction algorithms when applied prospectively on influenza diagnosis data and syndromic telenursing data, respectively*

		Winter influenza seasons			Pandemic		
	Prediction k weeks ahead	2 weeks	3 weeks	4 weeks	2 weeks	3 weeks	4 weeks
Influenza diagnosis data							
Correlation (r)[a]	Non-adaptive log-linear regression	0·77	0·77	0·76	−0·27	−0·30	−0·33
	Adaptive log-linear regression[b]	0·52	0·35	0·22	0·09	0·01	0·02
	The naive method[c]	0·62	0·44	0·28	0·38	0·18	0·07
MedAPE[a]	Non-adaptive log-linear regression	0·75	0·77	0·77	1·00	0·97	0·97
	Adaptive log-linear regression[b]	1·00	1·00	1·00	0·96	1·00	1·21
	The naive method[c]	0·80	0·90	0·98	0·70	0·77	0·81
Telenursing data							
Correlation (r)[a]	Non-adaptive log-linear regression	0·77	0·73	0·69	−0·20	−0·27	−0·34
	Adaptive log-linear regression[b]	0·67	0·55	0·42	0·37	0·24	0·18
	The naive method[c]	0·74	0·64	0·51	0·52	0·37	0·22
MedAPE[a]	Non-adaptive log-linear regression	0·17	0·19	0·20	0·31	0·29	0·28
	Adaptive log-linear regression[b]	0·19	0·23	0·33	0·34	0·37	0·39
	The naive method[c]	0·17	0·19	0·23	0·19	0·21	0·26

[a] The correlation coefficient (r) ranks the algorithms from highest (best) to lowest (worst) values, while MedAPE ranks them from lowest (best) to highest (worst) values.
[b] The method does not have the same learning set because it is adaptive, which means that the parameters are updated every week.
[c] The method has no learning set. The predicted value k weeks ahead is the same as the true value k weeks before.

2-week predictions to 0·77 for 4-week predictions) (Table 4). For the pandemic outbreak in 2009, none of the algorithms provided satisfactory evaluation outcomes.

For the syndromic telenursing data, the best prediction algorithm for the winter influenza seasons was non-adaptive log-linear regression ($r = 0·77$ for 2-week predictions, $r = 0·69$ for 4-week predictions, and MedAPE varying from 0·17 for 2-week-ahead predictions to 0·20 for 4-week-ahead predictions) (Table 4). Acceptable performance was also displayed for 2-week-ahead predictions for the naive method ($r = 0{\cdot}74$ and MedAPE = 0·17). For the pandemic outbreak in 2009, the algorithms produced lower correlations; the highest correlation was $r = 0{\cdot}52$ for the naive method for 2-week-ahead predictions.

DISCUSSION

The aim of this study was to perform a comparative trial of algorithms for the detection and prediction of increases in local influenza activity using data streams from a county-wide health information system. Among the detection algorithms evaluated, we found that only the Serfling regression displayed satisfactory performance when applied to influenza diagnosis data during winter influenza seasons. Concerning the evaluated prediction algorithms, the non-adaptive log-linear regression showed acceptable performance when applied both to influenza diagnosis data as well as to syndromic telenursing data. Among the remaining two algorithms, acceptable performance was only displayed for 2-week-ahead predictions for the naive method when applied to syndromic data. It has been pointed out that parametric methods are not suitable when the parameters describing the incidence curve vary considerably from year to year [32–34], as is the case with winter influenza seasons [35]. However, the results of this study show that Serfling regression, a parametric detection method, and non-adaptive log-linear regression, a parametric prediction method, displayed satisfying performance when applied to local influenza diagnosis data. These observations suggest that parametric methods may be considered, although carefully, when developing methods for use in influenza epidemic detection and prediction at local level.

Several factors could explain some of the poor performances of the algorithms observed in this study. One reason for the poor performance of the detection algorithms when applied to the syndromic telenursing data is that most of the evaluated algorithms were threshold based, whereas the baseline of the telenursing data had an increasing trend. For instance, the average number of calls to the telenursing service during the intermittent period before the winter influenza season in 2011–2012 had increased by nearly 18% compared with the corresponding number of calls during the intermittent period before the winter influenza season in 2010–2011 (Fig. 1). When using the learning set of telenursing data to calibrate the algorithms, the thresholds were set lower than would have been optimal for the algorithms to perform well during the evaluation period. In particular, this was reflected in the sensitivity outcomes because low thresholds that lack empirical foundation make the calibrated algorithms excessively sensitive for raising alarms. The need for regular pre-processing of syndromic data, correcting for issues such as daily autocorrelations, seasonal trends and day-of-the-week effects, has previously been emphasized [36]. For instance, in Timpka *et al.* [17], a baseline temporal trend of telenursing calls was estimated in the retrospective data using linear regression ($y = b_0 + b_1 t$ (where y is the incidence, b_0 is the intercept, b_1 is the slope, and t the time unit)) and corrected for. It has also been shown that algorithms applied to syndromic data demonstrate the best performance in specific settings, for example, depending on the shapes of the epidemic signal [37]. These experiences indicate that application of detection algorithms prospectively on syndromic data is a complex enterprise that requires consideration of pre-processing the data streams and combining detection approaches, rather than aspiring to apply one best algorithm on unprocessed unidimensional data [38]. However, it should be taken into account that combined detection approaches may lead to a decreased specificity for the system as a whole [39].

This study has both strengths and limitations that need to be taken into account when interpreting the results. It is one of the first studies to use real empirical data for side-by-side evaluation of influenza detection and prediction algorithms in a prospective setting. Use of real data for algorithm evaluations has been recommended, because the characteristics of baseline conditions in public health practice, such as temporal patterns and noise, are likely to have an influence on algorithm performance [40]. Regarding the limitations of the study, it should be taken into consideration that although some consensus regarding measurements to be used in the evaluation of detection algorithms exists [41], this is not the case for the evaluation of prediction algorithms. For the evaluation of prediction algorithms, we chose the combination of the Pearson correlation coefficient (r) and MedAPE, where the correlation measure estimated how well the predicted time series followed the observed time series and the MedAPE measure estimated the deviance of the level of the predicted time series on the y-axis from the observed

https://doi.org/10.1017/S0950268817001005 Published online by Cambridge University Press

level. Alternative or additional measurements include the median absolute deviation and root-mean-squared error [19], but we believe that the two measurements used in this study provide a sufficiently valid and accurate description of prediction performance. Moreover, because long-term data series of constant quality are seldom available at present at local public health levels, only data from one season were used for algorithm learning. Several interesting algorithms were therefore excluded, and some of the algorithms would probably show a better performance with a longer time period for collection of learning data. This scarcity of longitudinal data is a recognized problem in the evaluation of influenza detection systems [42]. Data simulation has been used to solve this data shortage problem; the main remaining challenge is replicating the complexity of both baseline and epidemic data streams [43,44]. Similarly, algorithms reported only from theoretical settings were excluded, with corresponding implications. In addition, the calculation of specificity was based on the 10 weeks immediately before an observed epidemic and the calculation of sensitivity was based on the first 10 weeks of an observed epidemic. The reason why these measures were not based on entire datasets was that detection methods are primarily optimized to detect epidemics. Evidently, including longer time periods before and during influenza epidemics would have generated higher evaluation values. Although other selections and computation methods could have generated dissimilar results, we believe that the definitions using fixed time periods are valid for the comparative analyses performed in the present study.

We conclude that among the algorithms evaluated, Serfling regression as detection algorithm and non-adaptive log-linear regression as prediction algorithm displayed satisfactory performance when applied on diagnostic data during winter influenza seasons in a local public health setting. When applied on syndromic data, satisfactory performance was shown only for the non-adaptive log-linear regression method among the prediction algorithms evaluated, while all of the detection algorithms evaluated showed poor performance. During the 2009 pandemic outbreak, the evaluated algorithms generally displayed poor performance. Both further evaluation research and research on combination of methods of these types in public health information infrastructures for ‘nowcasting’ (integrated detection and prediction) of influenza activity is warranted.

AUTHOR CONTRIBUTIONS

A.S., O.E., Ö.D., and T.T. conceived and designed the study. A.S., O.E., and Ö.D. analyzed the data. A.S., O.E., Ö.D., and T.T. contributed materials/analysis tools. A.S. and T.T. wrote the paper. Ö.D. and O.E. revised the manuscript and provided intellectual content. A.S., O.E., Ö.D., and T.T. gave final approval of the version to be published. T.T. is the guarantor of the content.

ACKNOWLEDGEMENTS

This study was supported by grants from the Swedish Civil Contingencies Agency (2010-2788) and the Swedish Research Council (2008-5252). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Timpka T, *et al.*** Requirements and design of the PROSPER protocol for implementation of information infrastructures supporting pandemic response: a Nominal Group study. *PLoS One* 2011; **6**(3): e17941.
2. **Gerbier-Colomban S, Potinet-Pagliaroli V, Metzger MH.** Can epidemic detection systems at the hospital level complement regional surveillance networks: case study with the influenza epidemic? *BMC Infectious Diseases* 2014; **14**: 381.
3. **Singh BK, *et al.*** Rapid detection of pandemic influenza in the presence of seasonal influenza. *BMC Public Health* 2010; **10**: 726.
4. **Dórea FC, *et al.*** Syndromic surveillance using veterinary laboratory data: data pre-processing and algorithm performance evaluation. *Journal of The Royal Society Interface* 2013; **10**(83): 20130114.
5. **Dórea FC, *et al.*** Syndromic surveillance using veterinary laboratory data: algorithm combination and customization of alerts. *PLoS One* 2013; **8**(12): e82183.
6. **Ohkusa Y, *et al.*** Real-time estimation and prediction for pandemic A/H1N1(2009) in Japan. *Journal of Infection and Chemotherapy* 2011; **17**(4): 468–472.
7. **Shaman J, Karspeck A.** Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences USA* 2012; **109**: 20425–20430.
8. **Shaman J, *et al.*** Real-time influenza forecasts during the 2012–2013 season. *Nature Communications* 2013; **4**: 2837.
9. **Kim EK, *et al.*** Use of hangeul twitter to track and predict human influenza infection. *PLoS One* 2013; **8**(7): e69305.

https://doi.org/10.1017/S0950268817001005 Published online by Cambridge University Press

10. **Timpka T, *et al.*** Performance of eHealth data sources in local influenza surveillance: a 5-year open cohort study. *Journal of Medical Internet Research* 2014; **16**: e116.
11. **Nagel AC, *et al.*** The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets. *Journal of Medical Internet Research* 2013; **15**(10): e237.
12. **Yom-Tov E, *et al.*** Detecting disease outbreaks in mass gatherings using internet data. *Journal of Medical Internet Research* 2014; **16**(6): e154.
13. **Kirian ML, Weintraub JM.** Prediction of gastrointestinal disease with over-the-counter diarrheal remedy sales records in the San Francisco Bay Area. *BMC Medical Informatics Decision Making* 2010; **10**: 39.
14. **Socan M, Erculj V, Lajovic J.** Early detection of influenza like illness through medication sales. *Central European Journal of Public Health* 2012; **20**(2): 156–162.
15. **Spreco A, Timpka T.** Algorithms for detecting and predicting influenza outbreaks: meta-narrative review of prospective evaluations. *BMJ Open* 2016; **6**: e010683.
16. **Bossuyt PM, *et al.*** Beyond diagnostic accuracy: the clinical utility of diagnostic tests. *Clinical Chemistry* 2012; **58**(12): 1636–1643.
17. **Timpka T, *et al.*** Predictive performance of telenursing complaints in influenza surveillance: a prospective cohort study in Sweden. *Eurosurveillance* 2014; **19**(46). pii: 20966.
18. **Cohen J.** *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum, 1988.
19. **Burkom HS, Murphy SP, Shmueli G.** Automated time series forecasting for biosurveillance. *Statistics in Medicine* 2007; **26**(22): 4202–4218.
20. **Jiang X, *et al.*** Bayesian prediction of an epidemic curve. *Journal of Biomedical Informatics* 2009; **42**(1): 90–99.
21. **Closas P, Coma E, Méndez L.** Sequential detection of influenza epidemics by the Kolmogorov–Smirnov test. *BMC Medical Informatics Decision Making* 2012; **12**: 112.
22. **Cowling BJ, *et al.*** Methods for monitoring influenza surveillance data. *International Journal of Epidemiology* 2006; **35**(5): 1314–1321.
23. **Martínez-Beneito MA, *et al.*** Bayesian Markov switching models for the early detection of influenza epidemics. *Statistics in Medicine* 2008; **27**(22): 4455–4468.
24. **Serfling RE.** Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Reports* 1963; **78**: 494–506.
25. **Stroup DF, *et al.*** Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Statistics in Medicine* 1989; **8**: 323–329; discussion 31–2.
26. **Montgomery DC.** *Introduction to Statistical Quality Control*. Hoboken, NJ: John Wiley & Sons, 2005.
27. **Tokars JI, *et al.*** Enhancing time-series detection algorithms for automated biosurveillance. *Emerging Infectious Diseases* 2009; **15**(4): 533–539.
28. **Viboud C, *et al.*** Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology* 2003; **158**(10): 996–1006.
29. **Yuan Q, *et al.*** Monitoring influenza epidemics in China with search query from Baidu. *PLoS One* 2013; **8**(5): e64323.
30. **Brillman JC, *et al.*** Modeling emergency department visit patterns for infectious disease complaints: results and application to disease surveillance. *BMC Medical Informatics Decision Making* 2005; **5**(4): 1–14.
31. **Burkom HS.** Development, adaptation, and assessment of alerting algorithms for biosurveillance. *Johns Hopkins APL Technical Digest* 2003; **24**(4): 335–342.
32. **Andersson E, Bock D, Frisén M.** Modeling influenza incidence for the purpose of on-line monitoring. *Statistical Methods in Medical Research* 2008; **17**: 421–438.
33. **Bock D, Andersson E, Frisén M.** Statistical surveillance of epidemics: peak detection of influenza in Sweden. *Biometrical Journal* 2008; **50**(1): 71–85.
34. **Schiöler L.** *Modelling the Spatial Patterns of Influenza Incidence in Sweden*. Gothenburg, Sweden: Statistical Research Unit, Department of Economics, University of Gothenburg; 2010. Report no. 2010:1.
35. **Timpka T, *et al.*** Age as a determinant for dissemination of seasonal and pandemic influenza: an open cohort study of influenza outbreaks in Östergötland County, Sweden. *PLoS One* 2012; **7**(2): e31746.
36. **Lotze T, Murphy S, Shmueli G.** Implementation and comparison of preprocessing methods for biosurveillance data. *Advances in Disease Surveillance* 2008; **6**(1): 1–20.
37. **Buckeridge DL, *et al.*** Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics* 2005; **38**(2): 99–113.
38. **Dórea FC, *et al.*** Syndromic surveillance using laboratory test requests: a practical guide informed by experience with two systems. *Preventive Veterinary Medicine* 2014; **116**(3): 313–324.
39. **Dohoo I, Martin W, Stryhn H.** *Veterinary Epidemiologic Research*, 2nd edn. Charlottetown, Canada: Atlantic Veterinary College, 2010.
40. **Buckeridge DL, *et al.*** Predicting outbreak detection in public health surveillance: quantitative analysis to enable evidence-based method selection. *AMIA Annual Symposium Proceedings* 2008; **2008**: 76–80.
41. **Unkel S, *et al.*** Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A* 2011; **175**: 49–82.
42. **Lotze TH, *et al.*** Simulating and evaluating biosurveillance datasets. In: Kass-Hout T, Zhang X, eds. *Biosurveillance: Methods and Case Studies*. New York: CRC Press, 2011, pp. 23–51.
43. **Buckeridge DL, *et al.*** An evaluation model for syndromic surveillance: assessing the performance of a temporal algorithm. *Morbidity and Mortality Weekly Report* 2005; **54**: 109–115.
44. **Buckeridge DL.** Outbreak detection through automated surveillance: a review of the determinants of detection. *Journal of Biomedical Informatics* 2007; **40**: 370–379.

https://doi.org/10.1017/S0950268817001005 Published online by Cambridge University Press