Performances of statistical methods for the detection of seasonal influenza epidemics using a consensus-based gold standard

C. Souty, R. Jreich, Y. Le Strat, C. Pelat, P Y Boëlle, C. Guerrisi, S. Masse, T. Blanchon, T. Hanslik, C. Turbelin

Performances of statistical methods for the detection of seasonal influenza epidemics using a consensus based gold standard

C. Souty[1*], R. Jreich[1], Y. Le Strat[2], C. Pelat[2], P.Y. Boëlle[1,3], C. Guerrisi[1], S. Masse[1,4], T. Blanchon[1], T. Hanslik[1,5,6], C. Turbelin[1].

1. Sorbonne Universités, UPMC Univ Paris 06, INSERM, Institut Pierre Louis d'épidémiologie et de Santé Publique (IPLESP UMRS 1136), F-75012, Paris, France

2. Santé publique France, French national public health agency, F-94415, Saint-Maurice, France

3. AP-HP, Hôpital Saint-Antoine, Département de santé publique, F-75012, Paris, France

4. EA7310, Laboratoire de Virologie, Université de Corse-Inserm, FR- 20250, Corte, France

5. Université Versailles Saint Quentin en Yvelines, UFR de Médecine, F-78000, Versailles, France

6. Hôpital universitaire Ambroise Paré AP-HP, Service de médecine interne, F-92100, Boulogne, France

* Corresponding author (Cécile Souty)

IPLESP UMRS 1136 INSERM UPMC

Faculté de médecine Pierre et Marie Curie – Paris 6

27 rue Chaligny,

75571 Paris Cedex 12 France

cecile.souty@upmc.fr ; +33 1 44 73 84 43

Short running head : Evaluation of methods for influenza epidemics detection

Summary

Influenza epidemics are monitored using influenza-like illness (ILI) data reported by health-care professionals. Timely detection of the onset of epidemics is often performed by applying a statistical method on weekly ILI incidence estimates with a large range of methods used worldwide. However, performance evaluation and comparison of these algorithms is hindered by: 1) the absence of a gold standard regarding influenza epidemic periods and 2) the absence of consensual evaluation criteria. As of now, performance evaluations metrics are based only on sensitivity, specificity and timeliness of detection, which definitions are not clear for time repeated measurements such as weekly epidemic detection. We aimed to evaluate several epidemic detection methods by comparing their alerts to a gold standard determined by international expert consensus. We introduced new performance metrics that meet important objective of influenza surveillance in temperate countries: to detect accurately the start of the single epidemic period each year. Evaluations are presented using ILI incidence in France between 1995 and 2011. We found that the two performance metrics defined allowed discrimination between epidemic detection methods. In the context of performance detection evaluation, others metrics than the standard commonly used could better achieve the needs of real-time influenza surveillance.

Keywords

Epidemics; Influenza; Outbreaks; Surveillance system; Surveillance

Background

The yearly global impact of seasonal influenza epidemics has been estimated about 1 billion symptomatic cases, 3 to 5 million severe cases and 250 to 500 thousands of deaths [1]. The duration, severity and geographical spread of influenza activity vary widely from one season to another depending on several factors such as rapid mutating viral strains, sensitivity of the population or climatic factors [2, 3]. Early detection of the start of seasonal epidemics is needed to inform public health authorities in order to implement necessary control measures. Moreover, monitoring influenza epidemics allows analysis about changes in trends, estimation of the global impact on populations and year-to-year comparisons.

The dynamics of influenza epidemics in the general population is monitored using primary care data collected by surveillance networks of health-care professionals who report the number of influenza-like illness (ILI) cases seen among their patients following a specific definition [4]. However, only a portion of ILI cases are due to influenza virus infection [5], thus statistical methods have to be used to determine the influenza epidemic onset from this non-specific data.

A wide variety of statistical methods have been proposed to detect seasonal influenza epidemics based on ILI incidence time series [6] such as regression models [7, 8], hidden Markov models (HMM) [9] and more recently the moving epidemic method (MEM) [10]. However, the evaluation of these methods is hindered by the absence of a gold standard regarding true influenza epidemics periods [6]. Performances of these methods have often been evaluated based from the results of other detection methods [6], using standard epidemiological metrics as sensitivity, specificity, positive predictive value, etc. [6, 11-13] with different definitions [14].

An accurate detection method would be able to detect precisely, *i.e.* with the smallest detection time, each season the whole single epidemic period, and particularly the start, which allowed alerting public health authorities and population.

In France, gold standard for seasonal influenza epidemics periods has been previously determined based on an international expert's consensus using the Delphi method [15]. This allowed identifying the start and end of epidemics using estimated ILI incidences and virological data in primary care.

We propose here to evaluate some common epidemic statistical detection methods by comparing their results to the gold standard determined by this expert consensus [15]. We defined performance metrics according to the monitoring objectives to seek for a global view of the detection methods properties.

Methods

Influenza surveillance data

ILI incidence rates were obtained from the *Sentinelles* network, a nationwide epidemiological surveillance system based on voluntary general practitioners (GP) in France [16, 17]. Sentinel GPs reported on a weekly basis the number of ILI cases seen among their patients, using the following definition: "sudden onset of fever >39°C (102°F) with respiratory signs and myalgia", allowing estimation of weekly ILI incidence rates [18, 19].

Gold standard for influenza epidemic periods

The gold standard for influenza epidemic periods in France were determined by a Delphi method described in Debin et al. [15]. More precisely, 57 experts determined yearly influenza epidemics periods from 1985/86 to 2010/11 using a web interface. For each season, virological results and estimated ILI incidence rates (from *Sentinelles* network) were presented, the experts were asked to determine the beginning and ending dates of each epidemic. In a second round, the same data were presented; adding histograms with the distribution of responses for start and end dates given by all experts on the previous round. A third final round was proposed for seasons when at least 25% of experts changed their responses between the first and the second round. The consensus of start and end dates for each season was then determined by the mode of the response, after removal of 5% of extreme responses on each side. Results for seasons between 1995/96 and 2010/11 are presented in Figure 1 along with estimates of ILI incidence rates from the *Sentinelles* network.

Epidemic detection methods

Four detection methods were evaluated: a periodic regression [7, 20], a robust regression [8], the MEM [10] and a HMM [9]. For each method, several values for the tuning parameters were chosen for calibration (Table 1). The common parameter of these four methods is the length of the learning period involved: the number of past observations (or past seasons for the MEM) provided to perform detection at a given point of time (called "learning size" further). For this parameter, we tested 4 values: 3 years, 5 years, 10 years and the whole available historical data at each time point.

The periodic regression for epidemic detection is a widely used approach from Serfling's work on influenza [7]. To sum up, it is based on a regression model which fits non-epidemic data to predict non-epidemic baseline. The epidemic threshold is defined by an upper percentile of the prediction distribution (here the 90^{th} percentile [21]).

In our evaluation, to prune the data, we removed values in the learning period over a given value (cut-off) that was either a fixed value or one determined from the learning data using a given percentile. To fit the model, we used the following regression equation:

$$I_t = \alpha_0 + \alpha_1 t + \alpha_2 \cos\left(\frac{2\pi t}{52.17}\right) + \alpha_3 \cos\left(\frac{2\pi t}{52.17}\right) + \varepsilon_t$$

where I_t is the incidence on week t, t being the week index and ε_t the residuals.

The robust regression is an alternative to periodic regression described above where all time serie is considered. Data pruning is done by assigning less weight to outliers, computed by a dedicated estimator [22, 23]. We used the same regression equation and the same definition of epidemic threshold as for periodic regression described above.

The MEM is implemented by steps: epidemic periods are first determined using historical time series, then epidemic thresholds are calculated using epidemic periods defined [10]. An extra parameter δ must be specified, corresponding to the minimum increment percentage used to find the optimum epidemic duration [10].

Hidden Markov Models were also used for monitoring such time series. A two-state HMM is applied on incidence time series, assuming that these observations are generated from a mixture of Gaussian distributions [9].

In what follows, we will call “detector” a method used with a given set of fixed parameters. Each detector was applied on ILI incidence rate time series to detect epidemic period in a prospective way (*e.g.* as it would be applied in real-time). Each week, the detector is run only on the data available up to this week.

Epidemic periods generated by detectors are triggered as soon as two consecutive weekly ILI incidence were above the threshold [20, 24, 25] (or classified in “alarm state” for HMM). Moreover, two consecutive ILI observations below the threshold (or classified as “no alarm state” for HMM) were required to determine the end of an epidemic. All weeks inside an epidemic period are classified “on alert” for the detector.

Detection performance

The performance evaluation of each detector was carried out on the period 1995-2011, which included 15 seasonal influenza epidemics and the 2009/10 pandemic. Evaluating the performance of a detector required calculating a number of measures that are in keeping with the objectives of detection. We computed two sets of metrics: 1) “Weekly detection” metrics, which are based on weekly alerts determined by the detector and 2) “Epidemic period detection” metrics, which are focused on detecting the epidemic period as a whole.

For both approaches, we assumed the true state (epidemic or non-epidemic) of a given week was informed by our gold standard - states were called *True* and *False*. States in the evaluated detector ("on alert" or "without alert") were called *Positive* and *Negative*.

Weekly detection metrics

We defined the number of weeks correctly classified by the detector as "true positive" (TP), respectively incorrectly classified as "false positive" (FP); the correctly classified as non-epidemic as "true negative" (TN), respectively the incorrectly as the "false negative" (FN).

Evaluation measures were then defined as:

Sensitivity = TP / (TP + FN)

Specificity = TN / (TN+FP)

We also defined positive predictive value as PPV = TP / (TP + FP) and negative predictive value as NPV = TN / (TN + FN).

These metrics were computed for the whole evaluated period: from ISO week 26 of 1995 to ISO week 25 of 2011, being 835 weeks.

Epidemic period detection metrics

This second evaluation approach focused on the ability of the detector in identifying the start week of each epidemic, and gives less importance to the correct detection of subsequent epidemic weeks. It stems from the reality that, for the management of seasonal influenza epidemics, public health authorities need accurate and timely information about the epidemic start, less so about the epidemic state of each subsequent week as the epidemic unfolds time period detection [26].

As proposed by Tsui et al. [26], we defined for each epidemic a "target" window that consist of the epidemic starting week and its two adjacent weeks (one before and one after) from the gold standard. Then, we considered that a detector correctly detected the start of the epidemic if the start of the first epidemic period detected during the season is in this target window. The associated evaluation metric, called $\mathbf{Detected_{start}}$, was defined as the proportion of epidemic starts correctly detected.

We also defined the **Timeliness** as the mean number of weeks between the first epidemic week in the gold standard and the beginning of the first epidemic period identified by the detector for all the epidemics studied.

Finally, we defined **Multipledetect** which is the number of seasons where the detector identified more than one epidemic period.

Metrics comparisons between detectors

The most desirable detector might detect only one epidemic period per season [24], have maximum $Detected_{Start}$, Sensitivity and Specificity values, and Timeliness close to zero [6, 8, 27]. We prioritized metrics in the evaluation, whenever possible; we selected 1) Multipledetect equal to zero, 2) $Detected_{Start}$ maximal and 3) compromise between high Sensitivity and high Specificity.

For periodic regression and MEM, the impact of the parameters values on the detection performance was studied using linear regression.

Uncertainty about the metrics point estimates was assessed by bootstrapping. The 16 influenza seasons included in the evaluation were resampled with replacement (N=1000). The bootstrap distributions obtained for each metrics allowed estimation of 95% confidence intervals using the 2.5% and 97.5% percentiles. Then, we used paired Student's t-test to compare bootstrap metrics values between detectors.

Results

All the 304 detectors studied (184 periodic regression models, 112 MEM, 4 HMM and 4 robust regression models) detected at least one epidemic period during each of the 16 studied influenza seasons (from 1995/96 to 2010/11).

Link between metrics

Link between $Detected_{Start}$ and Specificity, Sensitivity or Timeliness is a bell-shaped trajectory, with maximal values of $Detected_{Start}$ for Sensitivity between 0.80 and 0.94, Specificity between 0.96 and 0.99 and Timeliness between -1.3 and +0.3 weeks (Figure 2). $Detected_{Start}$ was maximal when Multipledetect was minimal. Multipledetect was equal to zero for a high Specificity and a moderate to high Sensitivity (under 0.954). When Timeliness was close to zero (between -0.5 and 0.5), Sensitivity was between 0.69 and 0.92 and Specificity between 0.97 and 0.99.

Intra-evaluation - by method

Periodic Regression

Over the 184 detectors evaluated using periodic regression method, the prune parameter was the most influential on detection metrics. Increasing the pruning level made the Sensitivity and Multipledetect decreased; the Specificity and Timeliness increased. Regarding the $Detected_{start}$ metric, the relation with the prune parameter was no linear. Indeed, $Detected_{start}$ was minimal (0.375) for extreme values of prune parameters and maximal (0.875) for 36 detectors (cut-off between 160 and 250, percentile between 0.84 and 0.87).

Among the 184 detectors studied, 8 achieved the highest $Detected_{Start}$ value (0.875) and did not detect several epidemics within the same season. Among them, the detector with the highest values for Sensitivity and Specificity was parametrized with a percentile of 0.86 and a maximum learning size period (Sensitivity = 0.874, Specificity = 0.985, Timeliness = -0.2 weeks, PPV = 0.958 and NPV = 0.962).

Robust regression

For robust regression method, among the 4 detectors compared, metrics were often better when the learning size included all available historic data, except for Sensitivity for which a 10 years learning size lead to a slightly higher value (0.80 vs. 0.79). All detectors have at least one multiple epidemic detection within a season (seasons 1995/96 and 2000/01).

Robust regression method parameterised with the largest learning size achieved a $Detected_{Start}$ equal to 0.750, Timeliness to 0.1 weeks, Sensitivity to 0.791 and Specificity to 0.985. During the 2000/01 epidemic, the detector identified two epidemic periods: a first of two weeks between weeks 50 and 52 year 2000 and a second between weeks 03 and 07 year 2001. For this detector, PPV was 0.959 and NPV was 0.941.

Moving Epidemic Method

With the MEM, both the delta parameter and the learning size affected metrics values, excepted Multipledetect. Increasing the delta value led to lower $Detected_{start}$, Sensitivity and higher Timeliness and Specificity. Conversely, higher learning size lead to higher Specificity, Timeliness and lower Sensitivity and $Detected_{start}$.

Twelve detectors achieved a maximal $Detected_{Start}$ (0.875) with no multiple epidemics detected within the same season. These detectors were parameterised with a maximal learning size and a delta value between 1.5 and 1.9, or a learning size equal to 10 years and a delta value between 2.2 and 2.8. Among these detectors, Timeliness was close to zero (between -0.3 and 0 weeks). Sensitivity was more variable (0.83 to 0.92) than Specificity (0.98 to

0.99). The best comprise was MEM parameterized with delta value of 1.5 and whole learning period (Sensitivity = 0.919, Specificity = 0.976, Timeliness = -0.3 weeks, PPV = 0.926, NPV = 0.976).

Hidden Markov model

Among the four detectors parameterized with HMM method, larger learning size led to higher Sensitivity, lower Specificity and $Detected_{start}$. Only one detector had the best $Detected_{Start}$ value (0.500) with no multiple season detections. It was parameterized with a learning size is fixed to 3 years (Sensitivity = 0.946, Specificity = 0.914, Timeliness = -1.1 weeks, PPV = 0.791, NPV = 0.983)

Inter-evaluation: comparison between methods

Among the four detectors identified in the intra-evaluation (Table 2), only robust regression had a multiple epidemic detection (two epidemic periods detected during influenza season 2000/01). Compared with HMM, $Detected_{Start}$ values were higher for MEM and periodic regression method ($p<1.10^{-6}$). Considering these two detectors, we did not highlight differences for $Detected_{Start}$ ($p=0.77$), but the periodic regression led to higher Specificity and lower Sensitivity than the MEM ($p<1.10^{-6}$).

Discussion

We compared performances of several epidemic detection methods and parameterizations for real-time influenza surveillance based on a gold standard determined by an expert consensus [15]. Performance metrics defined here allowed identification of methods able to detect accurately the start of the single epidemic period for each influenza season. The final choice of exact statistical method parameterization depends on the wishes of public health authorities in terms of sensitivity and specificity especially.

Although statistical measures of performances of a classification function are enough consensual - such as sensitivity and specificity, in the case of the detection method evaluation based on a repeated classification function over time - the definition of these measures is less clear [14]. Cowling et al. [6] proposed a definition of sensitivity "*whether there was at least one alarm during the peak season*", allowing a sensitivity equal to 1 for methods which were able to detect for example only the peak of the epidemic. Moreover, the specificity defined by Cowling et al. [6] involved values which are dependant of the epidemic duration. ROC curves, combining sensitivity and specificity, were sometimes used to compare detection methods [28], but they ignored the detection timeliness, which is of paramount importance in practice. We feel that a metric such as $Detected_{Start}$, addresses best what is

expected in practice from an epidemic detection method: identifying the epidemic start "not too early and not too late", as the detection of the epidemic ending is a lesser issue for real-time surveillance. Moreover, as influenza epidemics occur once a year in temperate areas [24], a second metric was used (Multipledetect) to ensure ability of the method to detect only one epidemic period during the season (from September year *n* to August year *n+1*).

In addition to new metrics defined here, standard metrics commonly used such as sensitivity and specificity [6, 20], were also computed. The link between $Detected_{Start}$ and these two metrics was non-linear allowing the selection of detectors achieving a compromise between high values of sensitivity and high values of specificity. In addition, by definition, high $Detected_{Start}$ values lead to Timeliness close to zero. Indeed, $Detected_{Start}$ allowed in one metric, identification of the most desirable method, with high sensitivity and specificity, and timeliness close to zero [8, 27].

The choice of the best method depends on the details of the application, implementation and context of surveillance [11]. Among all method studied here, we observed that HMM is more sensitive and less specific; conversely, robust periodic regression is more specific and less sensitive in comparison to other detectors studied. MEM and periodic regression are more able to be parametrized (delta, cut-off, learning size) involving a more difficult choice for implementation which requires us to test a large number of detectors compared to the two others methods. Overall, we observed that the consideration of all the historical data led to better metrics values.

Epidemic detection methods were applied to ILI incidence time series. According to the chosen ILI definition, specificity for influenza could vary [5] as others respiratory pathogens which also circulating during autumn and winter can cause very similar illness [29]. Virological confirmation of these ILI cases allows estimating the real number of influenza symptomatic cases and would tend to improve epidemic detection. However, laboratory surveillance is not always part of routine surveillance. When data are available, reporting delay is observed and methods, practices and sample size may vary by country [10]. This suggests that detection methods based on clinical data could be a more practical choice. However, when proper virological data is collected along with clinical cases, it should be taken into account to confirm that increasing incidence is largely due to influenza viruses.

Our study was limited by the statistical methods for influenza epidemic detection here compared.

All methods are based only on ILI incidence time series. Assimilation of laboratory-confirmed influenza surveillance data and ILI time series in a same detection method may improve performance. However, definition of ILI used by the French Sentinelles network is very specific [5], allowing estimation of ILI incidence close to

influenza confirmed incidence. Moreover, the methods did not consider spatial information that is often available in influenza surveillance, such as ILI incidence by region. The incorporation of spatial data in statistical models holds the promise of improved sensitivity, timeliness of detection, and possibly specificity [14]. Finally, we did not explore voting algorithm which could combine several detectors.

The metrics presented here allowed to measure ability of statistical epidemic detection methods to detect precisely the beginning of the single epidemic period by year with the smallest detection time. Their implementation on ILI incidence data from primary care surveillance network could improve influenza surveillance by providing accurate epidemic alerts for public health authorities and population.

Acknowledgements

We thank the general practitioners of the Sentinelles network for their participation.

Financial support

Not applicable.

Declaration of interest

None.

References

1. **Influenza** (http://www.who.int/immunization/topics/influenza/en/). Accessed 4 Oct 2016.
2. **Thompson WW*, et al.*** Mortality associated with influenza and respiratory syncytial virus in the United States. *Journal of the American Medical Association* 2003; **289**: 179-186.
3. **Monto AS.** Epidemiology of influenza. *Vaccine* 2008; **26, Supplement 4**: D45-D48.
4. **Ortiz JR*, et al.*** Strategy to enhance influenza surveillance worldwide. *Emerging Infectious Diseases* 2009; **15**: 1271-1278.
5. **Carrat F*, et al.*** Evaluation of clinical case definitions of influenza: detailed investigation of patients during the 1995–1996 epidemic in France. *Clinical Infectious Diseases* 1999; **28**: 283-290.
6. **Cowling BJ*, et al.*** Methods for monitoring influenza surveillance data. *International journal of epidemiology* 2006; **35**: 1314-1321.
7. **Serfling RE.** Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public health reports* 1963; **78**: 494-506.
8. **Wang X*, et al.*** Using an adjusted Serfling regression model to improve the early warning at the arrival of peak timing of influenza in Beijing. *PloS one* 2015; **10**: e0119923.
9. **Le Strat Y, Carrat F.** Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in medicine* 1999; **18**: 3463-3478.
10. **Vega T*, et al.*** Influenza surveillance in Europe: establishing epidemic thresholds by the moving epidemic method. *Influenza and other respiratory viruses* 2013; **7**: 546-558.
11. **Unkel S*, et al.*** Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2012; **175**: 49-82.
12. **Closas P, Coma E, Méndez L.** Sequential detection of influenza epidemics by the Kolmogorov-Smirnov test. *BMC medical informatics and decision making* 2012; **12**: 112.
13. **Choi BY*, et al.*** Comparison of various statistical methods for detecting disease outbreaks. *Computational Statistics* 2010; **25**: 603-617.
14. **Kleinman KP, Abrams AM.** Assessing surveillance using sensitivity, specificity and timeliness. *Statistical methods in medical research* 2006; **15**: 445-464.
15. **Debin M*, et al.*** Determination of French influenza outbreaks periods between 1985 and 2011 through a web-based Delphi method. *BMC medical informatics and decision making* 2013; **13**: 138.
16. **Flahault A*, et al.*** Virtual surveillance of communicable diseases: a 20-year experience in France. *Statistical methods in medical research* 2006; **15**: 413-421.
17. **Sentinelles network database** (http://www.sentiweb.fr/?page=database). Accessed 4 September 2017.
18. **Turbelin C*, et al.*** Age distribution of influenza like illness cases during post-pandemic A(H3N2): comparison with the twelve previous seasons, in France. *PloS one* 2013; **8**: e65919.
19. **Souty C*, et al.*** Improving disease incidence estimates in primary care surveillance systems. *Population Health Metrics* 2014; **12**: 1-9.
20. **Pelat C*, et al.*** Online detection and quantification of epidemics. *BMC medical informatics and decision making* 2007; **7**: 29.
21. **Costagliola D*, et al.*** A routine tool for detection and assessment of epidemics of influenza-like syndromes in France. *American journal of public health* 1991; **81**: 97-99.
22. **Huber PJ.** Robust estimation of a location parameter. *Annals of Mathematical Statistics* 1964; **35**: 73-101.
23. **Fox J**. *An R and S-Plus companion to applied regression*. CA: Sage, 2002.
24. **Viboud C*, et al.*** Influenza epidemics in the United States, France, and Australia, 1972–1997. *Emerging Infectious Diseases* 2004; **10**: 32-39.
25. **Costagliola D.** When is the epidemic warning cut-off point exceeded? *European Journal of Epidemiology* 1994; **10**: 475-476.

26. **Tsui F-C*, et al.*** Value of ICD-9-Coded chief complaints for detection of epidemics. *Journal of the American Medical Informatics Association* 2002; **9**: s41-s47.
27. **Martinez-Beneito MA*, et al.*** Bayesian Markov switching models for the early detection of influenza epidemics. *Statistics in medicine* 2008; **27**: 4455-4468.
28. **Spreco A, Timpka T.** Algorithms for detecting and predicting influenza outbreaks: metanarrative review of prospective evaluations. *BMJ Open* 2016; **6**.
29. **Fleming DM, Cross KW.** Respiratory syncytial virus or influenza? *Lancet (London)* 1993; **342**: 1507-1510.

Tables

Table 1. Methods and parameter combinations used for detectors parameterisation

Method	Parameter	Description	Values
Periodic regression	Learning size	Size of the learning period	3y, 5y, 10y, all [a]
	Prune	Cut-off value	80 to 320 by 10
		Percentile	0.7 to 0.9 by 0.01
Robust regression	Learning size	Size of the learning period	3y, 5y, 10y, all [a]
Moving epidemic method	Learning size	Size of the learning period	3y, 5y, 10y, all [a]
	Delta	Minimum increment percentage to find the optimum epidemic duration	1.3 to 4.0 by 0.1
Hidden Markov model	Learning size	Size of the learning period	3y, 5y, 10y, all [a]

[a] *Considering all available historical data at each point*

Table 2. Metrics values and 95% confidence intervals for the best detector identified for each method tested, influenza epidemics from 1995/96 to 2010/11, France.

Detector	**Multipledetect**	**Detect$_{Start}$** (95%CI)	**Timeliness** in weeks (mean (95%CI))	**Sensitivity** (95%CI)	**Specificity** (95%CI)	**PPV** (95%CI)	**NPV** (95%CI)
Periodic regression (percentile = 0.86; all learning size)	0	0.875 (0.688,1)	-0.2 (-0.2,0.5)	0.874 (0.832,0.912)	0.985 (0.960,1)	0.958 (0.897,1.000)	0.962 (0.948,0.975)
Robust regression (all learning size)	1 [a]	0.750 (0.500,0.938)	0.1 (-1.0,1.0)	0.791 (0.748,0.852)	0.985 (0.960,1.000)	0.959 (0.895,1.000)	0.941 (0.928,0.960)
MEM (delta=1.5; all learning size)	0	0.875 (0.688,1)	-0.3 (-1.4,0.5)	0.919 (0.880,0.952)	0.976 (0.954,0.994)	0.926 (0.871,0.975)	0.976 (0.963,0.986)
HMM (3 years learning size)	0	0.500 (0.250,0.750)	-1.1 (-2.8,0.1)	0.946 (0.899,0.985)	0.914 (0.873,0.953)	0.791 (0.711,0.872)	0.983 (0.969,0.995)

PPV : Positive predictive value ; NPV : negative predictive value

[a] *Two epidemic periods were detected during the 2000/01 season (2000w50 to 2000w52 and 2001w03 to 2001w07)*