Syndromic surveillance for local outbreak detection and awareness: evaluating outbreak signals of acute gastroenteritis in telephone triage, web-based queries and over-the-counter pharmacy sales

T. ANDERSSON[1,2,3]*, P. BJELKMAR[1,4], A. HULTH[1], J. LINDH[1,5], S. STENMARK[6,7] AND M. WIDERSTRÖM[7]

[1] *Swedish Institute for Communicable Disease Control (SMI), Solna, Sweden*
[2] *National Food Agency (SLV), Sweden*
[3] *Division for Mathematical Statistics, Department of Mathematics, Stockholm University, Sweden*
[4] *Inera AB, Sweden*
[5] *Department of Microbiology, Tumour and Cell Biology, Karolinska Institutet, Sweden*
[6] *County Medical Officer, Västerbotten, Sweden*
[7] *Department of Clinical Microbiology, Umeå University, Sweden*

SUMMARY

For the purpose of developing a national system for outbreak surveillance, local outbreak signals were compared in three sources of syndromic data – telephone triage of acute gastroenteritis, web queries about symptoms of gastrointestinal illness, and over-the-counter (OTC) pharmacy sales of antidiarrhoeal medication. The data sources were compared against nine known waterborne and foodborne outbreaks in Sweden in 2007–2011. Outbreak signals were identified for the four largest outbreaks in the telephone triage data and the two largest outbreaks in the data on OTC sales of antidiarrhoeal medication. No signals could be identified in the data on web queries. The signal magnitude for the fourth largest outbreak indicated a tenfold larger outbreak than officially reported, supporting the use of telephone triage data for situational awareness. For the two largest outbreaks, telephone triage data on adult diarrhoea provided outbreak signals at an early stage, weeks and months in advance, respectively, potentially serving the purpose of early event detection. In conclusion, telephone triage data provided the most promising source for surveillance of point-source outbreaks.

Key words: Foodborne infections, outbreaks, statistics, syndromic surveillance, waterborne infections.

INTRODUCTION

In syndromic surveillance, two functions need to be addressed, Early Event Detection (EED) and Situational Awareness (SA) [1, 2]. EED refers to the process of gathering and analysing signals of relevance for timely detection of disease outbreaks. SA represents real-time monitoring and assessment of epidemics: their size, location, and spread. EED is easier to translate into automatic surveillance systems as it involves real-time data collection and analysis. SA is about determining and understanding the situation at hand, and is less well-defined and more difficult to formalize. However, an ideal system for

* Author for correspondence: Dr T. Andersson, Swedish Institute for Communicable Disease Control (SMI), 171 82 Solna, Sweden.
(Email: tom.andersson@msb.se)

syndromic surveillance needs to integrate the SA and EED functions. In practice, there must be a trade-off. EED benefits from preclinical signals, e.g. self-diagnosis, absenteeism, pharmacy sales and patient contact rates. Accurate SA requires evidence-based information, e.g. epidemiological studies, clinical diagnosis and laboratory test results. These conflicting demands raise the question of whether certain data sources are better suited to bridge SA and EED. The main purpose of this study was to evaluate the efficiency of data sources for EED and SA.

To examine the suitability of different data sources, a reasonable strategy is to evaluate signals with respect to outbreaks that are well-defined in time and space, i.e. point-source outbreaks, such as local foodborne and waterborne outbreaks. This allows for easier mapping than propagated or seasonal epidemics with unclear temporal and spatial limits. However, few studies of this type have been conducted to date. The majority of empirical studies target seasonal epidemics, e.g. influenza, winter vomiting disease (norovirus), rotavirus and respiratory syncytial virus (RSV) [3]. The research on point-source and local outbreak surveillance is more limited. The issue is mainly discussed in relation to larger waterborne or foodborne outbreaks [4], or event monitoring at healthcare centres, hospitals and emergency departments [5]. Furthermore, systematic mapping of signal and outbreak characteristics is rare in these studies. Two studies of telephone triage data from NHS Direct (National Health Service, UK) have been published, showing positive and negative results, respectively [6, 7]. Studies of over-the-counter (OTC) pharmacy sales have reported similar conflicting results [8, 9]. To our knowledge, no comparative analysis of syndromic data with respect to multiple point-source outbreaks has previously been published.

In the presented study, we evaluated the potential of different sources of syndromic data for both SA and EED. From Swedish official outbreak reports, we selected point-source outbreaks during 2007–2011 that allowed for comparisons across the data sources. We first validated outbreak signals through testing for significant signal-to-noise (STN) ratios. For validated outbreak signals identified by this procedure, we subsequently explored the potential for SA and EED. This was achieved by analysing the correspondence between signal properties and outbreak sizes. For the strongest outbreak signals, we assessed the potential of different symptoms for EED by applying a simple detection algorithm.

METHOD

Data sources

Swedish Health Care Direct 1177 is a 24-hour nurse-on-call service comprising healthcare advice by telephone (1177) and by a website (www.1177.se). The record created for each call includes a contact cause, i.e. the main symptom [10]. For the purposes of this study, we extracted five data streams on the number of calls per day and municipality: (i) gastrointestinal illness across age groups, grouping the following symptoms: nausea, vomiting, diarrhoea, stomach pain and stomach illness; (ii) adult gastrointestinal illness, grouping the same symptoms, but excluding children (<18 years); (iii) diarrhoea in adults; (iv) nausea and vomiting in adults; and (v) stomach pain in adults. All data used in this study were anonymized.

For the investigated period, 2007–2011, web query data were obtained from 'Vårdguiden', the Stockholm County Council website providing information to the public on illnesses, health and healthcare. The Swedish Institute for Communicable Disease Control (SMI) has direct access to the data from the Vårdguiden website and produces regular analyses on selected queries submitted to the website [11, 12]. For this study, we extracted the number of web queries per day on the following gastrointestinal symptoms: vomiting (kräkningar), diarrhoea (diarré), stomach pain (magont) and gastrointestinal illness (magsjuka). The data represented word stems, allowing for inflections and spelling variations.

Data on OTC sales of antidiarrhoea medication were purchased from Pharmacy Services Ltd (Apotekens Service AB). After consultation with Pharmacy Services Ltd, we included all OTC antidiarrhoea drugs with ATC codes A07B and A07D. The extracted data covered daily unit sales of antidiarrhoea medication in pharmacies per municipality between 2006 and 2011. All Swedish pharmacies report daily OTC sales.

A list of point-source outbreaks was made to establish a basis for comparison of data sources. The point of departure was official reports of waterborne and foodborne outbreaks issued by the National Food Agency in Sweden. We decided on three criteria for inclusion of outbreaks: First, we selected larger outbreaks, excluding outbreaks comprising fewer than 100 cases. Second, the time window was limited to 2007–2011 avoiding the first years of establishment of the telephone triage and web-based healthcare services. Third, the time of the outbreak had to be

within the temporal limits of all data sources. In the following, we refer to an outbreak by the name of the municipality in question.

Methods of analysis

The evaluation of data sources consisted of three parts: (1) validation of outbreak signals, (2) estimation of signal rates and (3) signal detection analysis. The validation part contained visual inspection and statistical analysis of count data per day, i.e. number of 1177 calls, web queries and OTC units sold. The purpose was to identify true outbreak signals (deviations) as distinct from background noise (baseline variation). The estimation part consisted of calculating the magnitude of the outbreak signals to establish signal rates, i.e. the average number of signals per case. Finally, the detection part involved statistical analysis to identify abnormal signals before outbreak peaks, as well as calculations to assess the sensitivity and specificity of the data stream in question.

Signal validation

The validation process began by defining outbreak periods and midpoints. The midpoint was defined as the day when the local or regional authorities first issued public information about the outbreak. If public information was issued in the evening, the midpoint was taken as the date of the following day. Consequently, the outbreak midpoint divided the outbreak period into two phases: low and high public awareness, respectively. For outbreaks without any official public information, the midpoint was defined by the date of the first consumer complaint to the regional or local authorities.

For each outbreak, two outbreak periods were defined with respect to the midpoint, one narrow (±7 days, 15 days in total) and one wide (±14 days, 29 days in total). For each outbreak, two baseline periods were also defined, ±14 days and ±28 days, respectively, minus the corresponding outbreak period, creating baseline periods of 14 and 28 days. Daily count data were plotted and visually inspected for each combination of outbreak and source of data, and also extracted and summed for outbreak and baseline periods. The sums of signal counts for outbreak and baseline periods were compared using Pearson's χ^2:

$$\chi^2 = \frac{(OS - E_{OS})^2}{E_{OS}} + \frac{(BS - E_{BS})^2}{E_{BS}},$$

where OS is sum of outbreak signal counts, E_{OS} is expected outbreak signal counts, BS is the sum of baseline signal counts, and E_{BS} is the expected baseline signal counts.
Furthermore:

$$E_{OS} = \frac{OD}{OD + BD} \times TSC,$$

$$E_{BS} = \frac{BD}{OD + BD} \times TSC,$$

where OD is number of days of the outbreak period, BD is number of days of the baseline period and TSC is total signal count (i.e. OS + BS).

The STN ratio was calculated by dividing the difference in means between signal counts for outbreak baseline periods by the standard deviation of the baseline counts:

$$STN = \frac{mean(OSC) - mean(BSC)}{SD(BSC)},$$

where OSC is daily signal counts during the outbreak period, BSC is the daily signal counts during the baseline period and SD(BSC) is standard deviation of daily signal counts during the baseline period.

Outbreak signals were considered validated if the following criteria were met: (1) positive visual inspection; (2) STN > 1; and (3) $\chi^2 > 6 \cdot 635$ (upper limit for 99% confidence) for at least one time period (2 or 4 weeks).

Signal estimation

For the validated outbreak signals, the signal rates, i.e. the signal-to-case ratios, were estimated. This was done by calculating the deviation of signal counts from their expected values for observed outbreak periods, and then relating the magnitude of deviation to the number of cases in the outbreak:

$$SR = \frac{SC_O - SC_E}{NC},$$

where SR is signal rate (signal-to-case ratio), SC_O is signal count for the observed outbreak period, SC_E is expected signal count for the observed outbreak period and NC is total number of outbreak cases, according to epidemiological studies or official outbreak reports.

The observed outbreak periods should not be confused with the fixed outbreak periods defined for signal validation. The observed periods were defined by pooling existing information on the outbreaks, including epidemiological studies, outbreak

investigations and the syndromic data sources in question. The criterion was to define periods broad enough to cover real outbreak durations, while as narrow as possible to minimize signal noise. Small variations in observed outbreak periods were not critical for the point estimations of signal rates, although they influenced the confidence intervals.

Regression analysis of signal counts on municipality population size, excluding the targeted municipality, was used to estimate the expected (predicted) signal count and the prediction interval for the targeted municipality. Linear regression was used when the mean signal counts for the observed outbreak periods were >25. For mean signal counts <25, Poisson regression analysis was used. The linear regression model was as follows:

$$SC_E = \beta_1 \times \text{population size} + \beta_0,$$

$$SC_O = \beta_1 \times \text{population size} + \beta_0 + \text{residual},$$

$$SR = \frac{SC_O - SC_E}{NC},$$

$$SC_{E,High} = SC_E + 2 \times PE,$$

$$SC_{E,Low} = SC_E - 2 \times PE,$$

$$PI : \left[\frac{SC_O - SC_{E,High}}{NC}, \frac{SC_O - SC_{E,Low}}{NC} \right],$$

where PE is the prediction error in the regression model for the targeted municipality, $SC_{E,Low(High)}$ is the low(high) limit of expected signal count for the outbreak period and PI is the prediction interval.

Signal detection

For the largest outbreaks and the data source with the largest STN and highest SR values, a signal detection analysis was carried out to evaluate the potential of different data streams for EED, i.e. outbreak signal detection before the outbreak midpoint. For the observed outbreak periods before the outbreak midpoints, a binomial distribution was applied and expected values and standard deviations of daily signal counts were calculated. The signal count at day t for municipality i ($C_{t,i}$) was classified as an outbreak signal if it exceeded a threshold $T_{t,i}$:

$$T_{t,i} = \max(L, V),$$

$$L = [0, 1, 2, 3, \ldots],$$

$$V = (E[C_{t,i}] + L \times SD(C_{t,i})),$$

$$L = [0, 1, 2, 3, \ldots] E[C_{t,i}] = p_{t,i} \times N_i,$$

$$SD(C_{t,i}) = \sqrt{N_i \times p_{t,i} \times (1 - p_{t,i})},$$

$$p_{t,i} = \frac{\sum_{j=1, j \neq i}^{n_i} \sum_\tau C_{\tau,j}}{4 \times \sum_{j=1, j \neq i}^{n_i} N_j},$$

$$\tau \in \{t - 14, t - 21, t - 28, t - 35\},$$

where L is the minimum number of signal counts for a positive outbreak signal, V is the threshold for a positive outbreak signal based on binomial distribution, $C_{t,i}$ is the daily signal count of municipality i at time t, N_i is the population size of municipality i, $p_{t,i}$ is the probability of a single 1177 call at time t from municipality i and n_i is the number of municipalities in the county where municipality i is located.

The threshold $T_{t,i}$ was taken as the maximum of the fixed value L and the varying value V, defined by the number L of standard deviations above the expected value. The level L set the minimum number of signal counts that the daily count needed to exceed to qualify as a positive outbreak signal. Daily counts exceeding level 3 (low) were described as weak signals and daily counts exceeding level 5 (high) as strong signals. The probability $p_{t,i}$ was calculated on the basis of the sum of signal counts in a county for four weekdays, 2–5 weeks back in time, divided by four times the population size of the county.

To evaluate the sensitivity and specificity of the signal detection during observed outbreak periods, the target municipality defined the outbreak condition. The control condition was defined by non-neighbouring municipalities in the same county. Daily signal counts $C_{t,i}$ above and below $T_{t,i}$ in the outbreak condition defined hits and misses, respectively, whereas $C_{t,i}$ above and below $T_{t,i}$ in the control condition defined false alarms (FA) and correct rejections (CR), respectively.

$$\text{Sensitivity} = \frac{\text{Hits}}{\text{Hits} + \text{Misses}},$$

$$\text{Specificity} = \frac{\text{CR}}{\text{CR} + \text{FA}}.$$

RESULTS

Signal validation

Nine outbreaks were included in the study (Table 1). The three largest outbreaks were caused by contamination of drinking water, and the others were related to local foodborne contamination, e.g. a bakery,

Table 1. *List of larger waterborne and foodborne outbreaks in Sweden 2007–2011*

Midpoint	Municipality	Population	County	County population	Agent	Vehicle	Point-source	Reported cases	Signal validation
27 Nov. 2010	Östersund	59416	Jämtland	126691	*Cryptosporidium*	Drinking water	Surface source	27000	1/1/0
19 Apr. 2011	Skellefteå	71580	Västerbotten	259667	*Cryptosporidium*	Drinking water	Surface source	20000	1/1/0
11 Sept. 2008	Lilla Edet	12831	Västra Götaland	1558130	Calicivirus	Drinking water	Surface source	2400	1/0/0
7 Mar. 2008	Helsingborg	126754	Skåne	1214758	Calicivirus	Bakery products	Bakery	369	1/0/0
12 July 2009	Karlskrona	63342	Blekinge	152591	Norovirus	Unknown	Elderly care	185	0/0/0
21 Oct. 2009	Uddevalla	51518	Västra Götaland	1569458	Calicivirus	Unknown	Primary school	145	0/0/0
27 Aug. 2009	Alingsås	37515	Västra Götaland	1569458	Calicivirus	Frozen raspberries	High school	130	0/0/0
30 May 2007	Skövde	50610	Västra Götaland	1558130	Norovirus	Mixed foods	Restaurant	100	0/0/0
17 Dec. 2008	Skövde	50610	Västra Götaland	1558130	Norovirus	Mixed foods	Restaurant	100	0/0/0

restaurants, schools and elderly care. For the three largest outbreaks, the number of cases was supported by local cross-sectional surveys carried out by SMI or regional county medical officers. For the remaining six outbreaks, epidemiological data were limited to outbreak investigations conducted by local health protection offices, basing the case numbers on more informal case-by-case interviews and questionnaires.

Outbreak signals were validated for the four largest outbreaks (Table 2). The 1177 telephone triage data captured all four, while the OTC sales data enabled detection of the two largest. No outbreaks could be validated in the web query data. The STN ratios were generally higher for the 1177 triage data ($1 \cdot 41 < \text{STN} < 5 \cdot 6$), about twice as high as for OTC sales data on corresponding outbreaks ($0 \cdot 95 < \text{STN} < 2 \cdot 37$), indicating stronger signals in the 1177 triage data. A visual illustration of the differences in signal quality was obtained by plotting the signal counts in the 1177 and OTC data for the two largest outbreaks (cf. Figs 1 and 2).

The OTC sales peaks lagged 2–4 days behind the 1177 call peaks. Elevated OTC sales were also more short-lived than elevated call intensity. Changing from a fixed outbreak period of 2 weeks to a period of 4 weeks for the two largest outbreaks increased the STN ratios for the triage data, whereas they remained more or less the same for the OTC sales data. This indicates that there were broader peaks in the 1177 triage data than in the OTC sales data. For the two smaller validated outbreaks, changing the period from 2 to 4 weeks resulted in a reduction in STN ratios, supporting the assumption of fast, transient outbreaks.

For the remaining outbreaks, visual inspections of data and criteria for validation revealed no unusual or irregular signal pattern at outbreak time. Furthermore, no association could be established between outbreaks and web query counts, although results were inconclusive for the largest outbreak. Visual inspection and validation criteria showed peaks surrounding the outbreak midpoint.

Signal estimation

For the outbreaks with validated signals, the following observed outbreak periods were defined. For the largest outbreak (Östersund), the starting and end points were set to 1 November 2010 and 31 January 2011 (92 days). For the second largest outbreak (Skellefteå), an epidemiological survey and the 1177 triage data indicated elevated gastrointestinal illness

Table 2. *Signal validation for the four largest outbreaks in Östersund, Skellefteå, Lilla Edet and Helsingborg*

| Source	Statistics		Municipality							
			Östersund		Skellefteå		Lilla Edet		Helsingborg	
1177 calls	Visual confirmation		Yes		Yes		Yes		Yes	
	Time window		2 weeks	4 weeks	2 weeks	4 weeks	2 weeks	4 weeks	2 weeks	4 weeks
	Signal count	Target	995	1316	669	1096	78	99	362	566
		Baseline	321	429	427	486	21	35	204	405
	Signal-to-noise ratio		4·16	5·6	1·41	3·22	2·64	2·15	2·82	0·93
	Test	χ^2	591·6	420·4	75·0	214·4	57·1	28·4	66·8	21·4
		P	**<0·001**	**<0·001**	**<0·001**	**<0·001**	**<0·001**	**<0·001**	**<0·001**	**<0·001**
Over-the-counter sales	Visual confirmation		Yes		Yes		No		No	
	Time window		2 weeks	4 weeks	2 weeks	4 weeks	2 weeks	4 weeks	2 weeks	4 weeks
	Count	Target	1043	1487	751	1290	62	115	823	1537
		Baseline	444	724	539	888	53	83	714	1401
	Signal-to-noise ratio		2·37	2·05	0·95	1·07	0·12	0·35	0·2	0·16
	Test	χ^2	202·0	237·3	21·8	60·8	0·2	4·1	2·0	2·4
		P	**<0·001**	**<0·001**	**<0·001**	**<0·001**	0·639	0·043	0·153	0·119
Web queries	Visual confirmation		Yes		No		No		No	
	Time window		2 weeks	4 weeks	2 weeks	4 weeks	2 weeks	4 weeks	2 weeks	4 weeks
	Count	Target	2160	4867	1279	2525	816	1548	2106	3998
		Baseline	2707	3945	1246	2687	732	1392	1892	3955
	Signal-to-noise ratio		−0·97	0·45	−0·12	−0·19	0·16	0·28	0·13	−0·08
	Test	χ^2	105·1	66·8	1·2	12·3	0·6	3·7	1·4	1·2
		P	**<0·001**	**<0·001**	0·281	**<0·001**	0·435	0·054	0·229	0·279
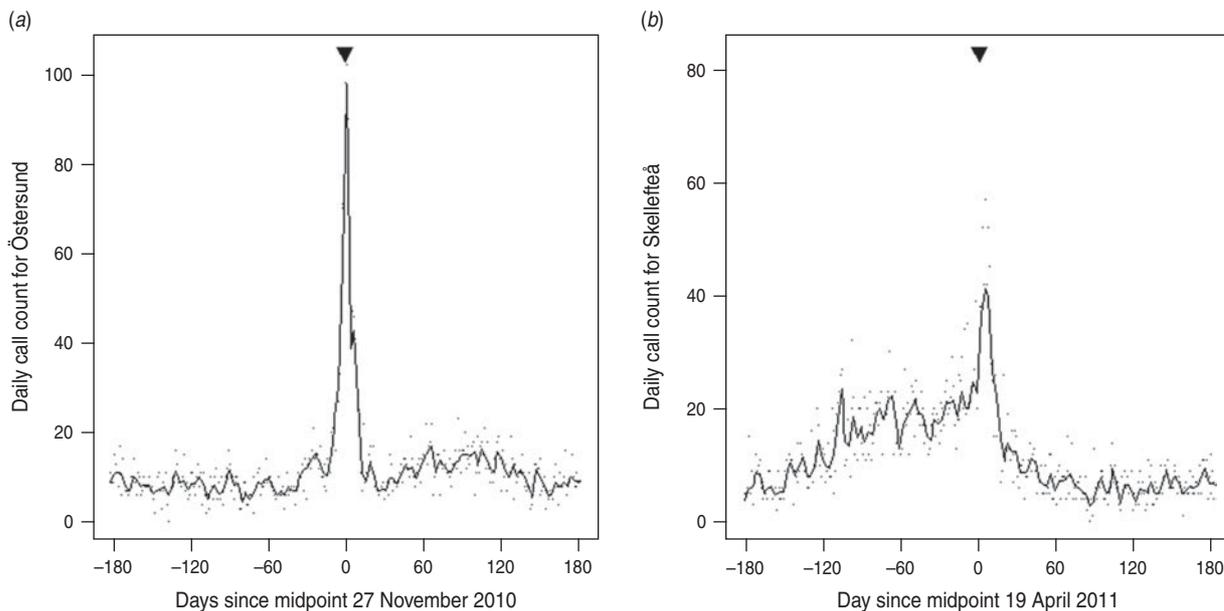
Fig. 1. Number of 1177 calls relating to adult gastrointestinal symptoms during the outbreaks in (*a*) Östersund and (*b*) Skellefteå. The smoothed curve is based on a locally weighted polynomial regression performed with the R function 'lowess', using a smoother span of 14 days. The solid triangles indicate the call count at the outbreak midpoint, i.e. the day when regional and local authorities issued official public information. The vertex indicates the signal count at the midpoint.
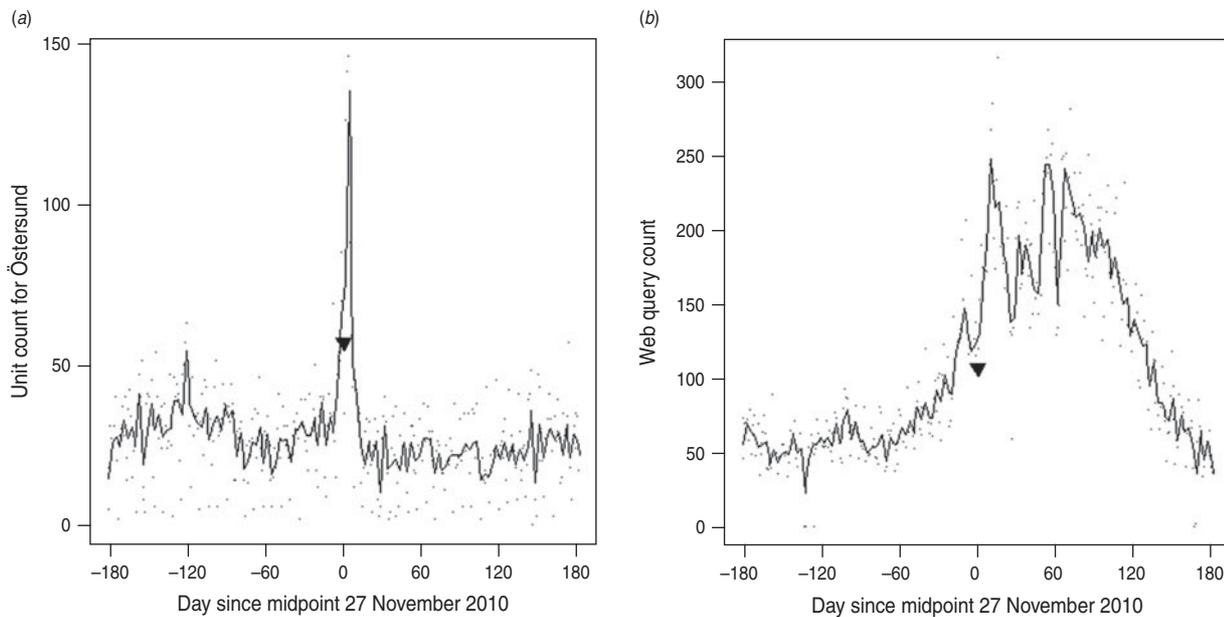
Fig. 2. (*a*) Pharmacy over-the-counter sales of antidiarrhoeals and (*b*) daily sums of web queries on gastrointestinal symptoms during the outbreak in Östersund. The smoothed curve is based on a locally weighted polynomial regression performed with the R function 'lowess', using a smoother span of 14 days. The solid triangles indicate the unit and search counts at the outbreak midpoint, i.e. the day when regional and local authorities issued official public information on the outbreak. The vertex indicates the signal count at the midpoint.

from the beginning of 2011. Therefore a long outbreak period was defined: 15 December 2010 to 30 June 2011 (198 days). The data suggested a more rapid increase in illness from March 2011 onwards. Therefore a short outbreak period was also defined: 1 March to 30 June 2011. For the remaining two outbreaks, the outbreak periods were clearly short and were narrowly set to 7 days, ±3 days around outbreak midpoints.

Details of the calculations can be found in Supplementary Tables S1 and S2 (available online). The regression analysis of calls relating to gastrointestinal illness for all ages on population size resulted in the following predicted signal rates (lower boundary, upper boundary): 0·042 (0·028, 0·055) for Östersund; 0·061 (0·005, 0·116) for Skellefteå; 0·019 (0·016, 0·021) for Lilla Edet; and 0·111 (−0·036, 0·257) for Helsingborg (Supplementary Table S1). When the data were limited to adults (>17 years), similar figures were obtained, but with narrower prediction boundaries: 0·039 (0·030, 0·048) for Östersund; 0·054 (0·030, 0·078) for Skellefteå; 0·013 (0·011, 0·014) for Lilla Edet; and 0·163 (0·053, 0·273) for Helsingborg. Thus, adults represented most of the excess signals due to the outbreaks.

Limiting the analysis to single gastrointestinal symptoms in the 1177 triage data (adult diarrhoea,

vomiting and stomach pain), calls relating to adult diarrhoea represented the majority of the excess signals in the two largest outbreaks: 0·027 (0·025, 0·03) and 0·037 (0·030, 0·045) for Östersund and Skellefteå, respectively. The outbreak in Skellefteå was also marked by an elevated rate of adult vomiting [0·011 (0·005, 0·017)], in particular in the first phase of the outbreak [0·022 (0·009, 0·034)]. For the outbreak in Lilla Edet, the two symptoms were more balanced: 0·0067 (0·0058, 0·0071) and 0·0052 (0·0046, 0·0054) for adult diarrhoea and adult vomiting, respectively. For Helsingborg, the signal rate was highest for adult vomiting [0·070 (0·049, 0·092)], followed by adult diarrhoea [0·046 (0·023, 0·069)]. The signal rate of stomach pain was only significant for the outbreak in Östersund [0·0089 (0·0027, 0·0151)].

The regression analysis of OTC sales resulted in signal rates with wide intervals: 0·032 (−0·001, 0·064) and 0·012 (−0·088, 0·111) for Östersund and Skellefteå, respectively. The wider intervals compared with the triage data indicate weaker specificity of the OTC data. Further visual inspection of the OTC data revealed a marked example of the weaker specificity. One Swedish municipality, Strömstad, demonstrated a clear extreme value during the outbreak period for Östersund, as well as during the
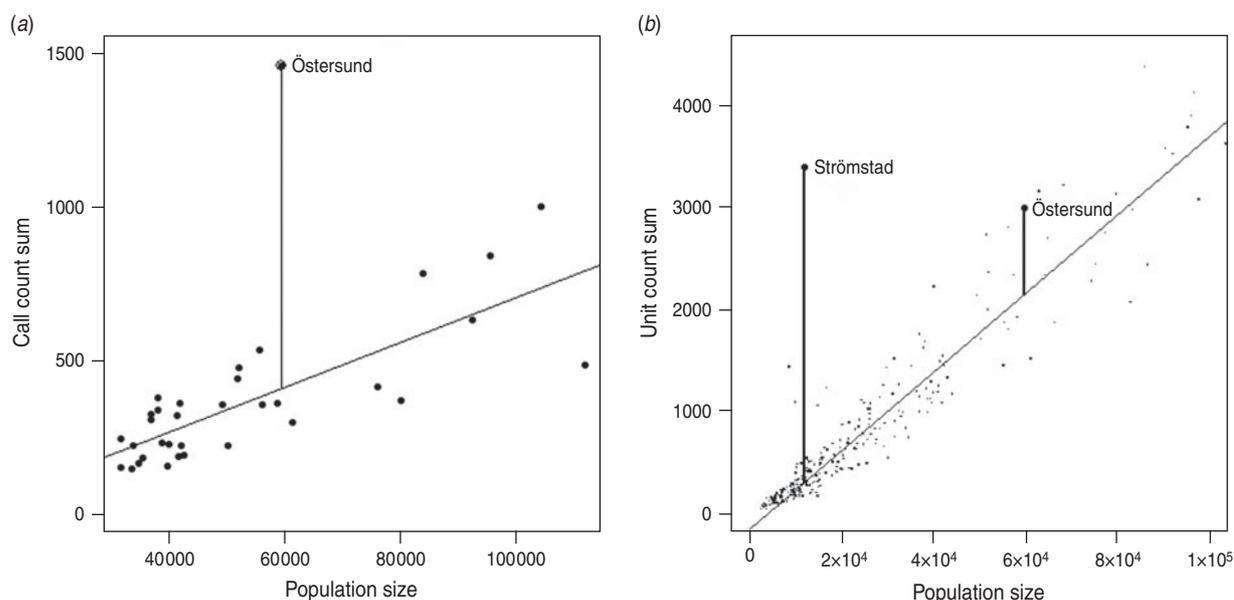
Fig. 3. Signal rates. Regression analysis of count data during the observed outbreak period of Östersund on municipality population size for (a) adult diarrhoea calls and (b) over-the-counter (OTC) sales. The analyses included municipalities from half, to twice the size of the targeted municipality (Östersund), excluding municipalities affected by outbreaks. The OTC plot extends beyond the range of the analysis.

outbreak period for Skellefteå, without corroboration from official outbreak reports (Fig. 3).

Signal detection

Due to the relatively weak signals in the OTC sales data, signal detection analysis was limited to the triage data. Furthermore, the week-long outbreak periods in Lilla Edet and Helsingborg were comparatively short. Visual inspection of data showed that early warnings could at best be issued 1–2 days before the outbreak midpoint, thus not warranting any signal detection analysis. Therefore, the analysis was limited to the larger outbreaks in Östersund and Skellefteå. Since these both involved *Cryptosporidium*, the analysis was further limited to three syndromes: all adult symptoms of gastrointestinal illness, adult diarrhoea and adult stomach pain.

When the thresholds for weak and strong signals were applied to the triage data on calls from Östersund, a cluster of significant outbreak signals appeared for the period 2–9 November 2010. The count data on all adult symptoms of gastrointestinal illness together resulted in three strong and three weak signals during these days. There were one strong and five weak signals for adult diarrhoea; and two strong and three weak signals for stomach pain. A cluster of strong and sustained

outbreak signals appeared on 21 November, 6 days before the outbreak midpoint (Fig. 4).

During the initial phase of the outbreak period, from 1 to 26 November, applying a single threshold of 3 to count data on all adult gastrointestinal symptoms generated 17 outbreak signals, giving a sensitivity of 0·653 (17/26). During the same period, no outbreak signals were observed for controls, giving a specificity of 1. Comparing adult diarrhoea, vomiting and stomach pain, diarrhoea was the most efficient classifier of outbreak signals, as judged by the overall differences between hit rates and false alarm rates (0·577). Detailed information on the effects of different thresholds on the sensitivity and specificity for different syndromes are given in Table 3.

The detection analysis of the outbreak in Skellefteå revealed several strong and weak signals at the end of 2010 and the beginning of 2011 (Fig. 4). After this cluster, strong and weak signals of adult diarrhoea, vomiting and stomach pain reappeared in Skellefteå sporadically until 20 March, after which strong and weak signals of diarrhoea began to increase. Applying a threshold of 3 from 15 December 2010 to 18 April 2011, the sensitivity was 0·416 for adult gastrointestinal symptoms and 0·400 for adult diarrhoea, while the specificity was 0·998 and 0·992, respectively. With a shortened initial outbreak period from 1 March 2010 to 18 April 2011, the sensitivity for adult gastrointestinal symptoms and adult diarrhoea increased

Table 3. *Signal detection analysis for the outbreaks in Östersund and Skellefteå*

Municipality	Threshold*	Signal†	Diarrhoea	Vomiting	Stomach pain	Adult GI‡	All GI§
Östersund	Low	FAR	0%	0%	0%	0%	0%
		HR	57·7%	3·8%	53·8%	65·3%	46·2%
	High	FAR	0%	0%	0%	0%	0%
		HR	30·8%	0%	26·9%	46·2%	46·2%
Skellefteå (long outbreak period)	Low	FAR	0·8%	0·1%	0%	0·2%	1·5%
		HR	40·0%	18·4%	16·0%	41·6%	30·4%
	High	FAR	0%	0%	0%	0%	0·1%
		HR	13·6%	4·8%	4·8%	13·6%	9·6%
Skellefteå (short outbreak period)	Low	FAR	0%	0%	0·1%	0·1%	0·7%
		HR	53·1%	2·4%	16·3%	49·0%	26·5%
	High	FAR	0%	0%	0%	0%	0%
		HR	28·6%	0·8%	8·2%	18·4%	12·2%

* Low/High: +3/+5 standard deviations.
† FAR/HR: False alarm rate/Hit rate.
‡ GI, Gastrointestinal illness (diarrhoea, vomiting, stomach pain).
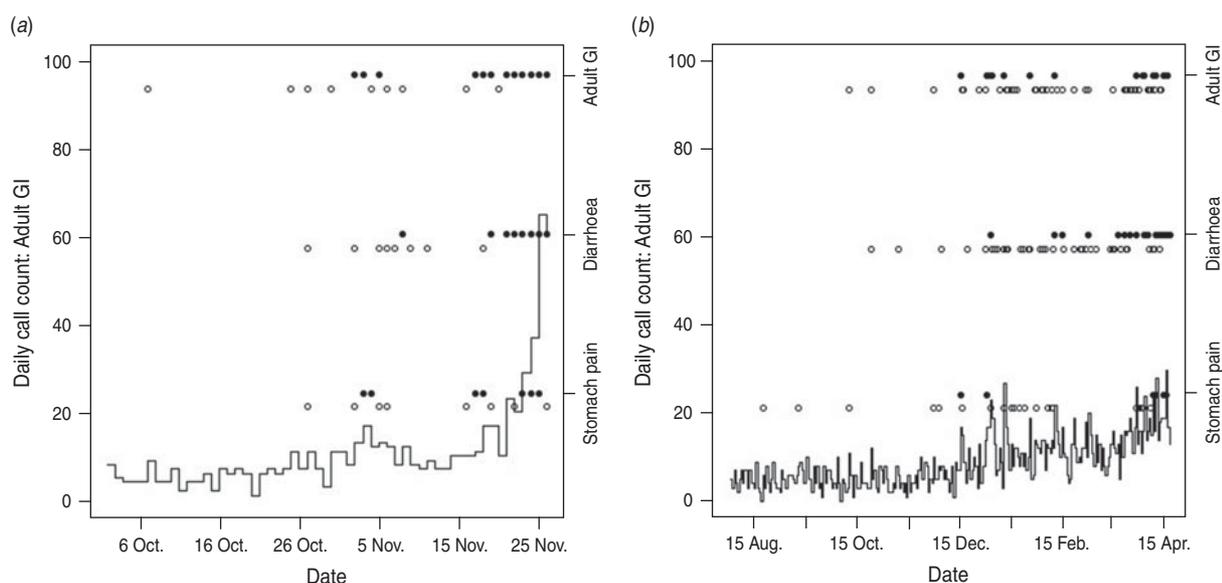§ Including children (<18 years).

Fig. 4. Signal detection analysis. The stepped graphs represent daily counts of adult gastrointestinal (GI) calls during the outbreak periods in (*a*) Östersund and (*b*) Skellefteå, before the outbreak midpoints (27 November 2010 and 19 April 2011, respectively). The solid and open circles indicate strong and weak outbreak signals when the detection algorithm was applied to three streams of 1177 triage data: adult GI calls (upper circles), diarrhoea (middle circles), and stomach pain (lower circles).

to 0·490 and 0·531, respectively, while still maintaining high specificity (1 and 0·999, respectively).

DISCUSSION

To summarize the findings, outbreak signals were validated in syndromic data for four out of nine point-source outbreaks in Sweden between 2007 and 2011. The four largest outbreaks had significant effects on signal counts in the triage data. The two largest outbreaks were also manifested in the OTC sales data. No outbreak signal could be validated for web query data. Several potential factors may have contributed to the comparatively weaker sensitivity and specificity of web query signals. Most importantly, the web query data lack geographical resolution, i.e. there is no geographical marker connected to an

individual query, limiting the analysis to a spatially unspecified population. In addition, the website traffic is concentrated in the county of Stockholm, but all included outbreaks occurred outside this county. An alternative source to use would be Google trends, but this source is associated with similar problems, i.e. limited temporal and spatial resolution and non-transparent data formats. The usage of too wide a geographical area may also explain the previous conflicting findings reported for triage data [6, 7].

The OTC sales data were only sensitive to the two largest outbreaks and revealed extreme values that did not correspond to any known outbreaks. The explanation is straightforward. First, the two largest outbreaks involved diarrhoea as the main symptom, while for the other two outbreaks the symptoms were varied and more transient, making the use of antidiarrhoeal medication less relevant. Second, in May 2009, a pharmacy opened in a shopping centre in Strömstad close to the border with Norway, after which the OTC sales of antidiarrhoeals increased significantly, from 19·8 units per day (S.D. = 10·5), to 33·0 units per day (S.D. = 15·3). These averages were calculated on daily volumes for 365 days preceding and following the opening day, using ± 30 days from the opening day as a dead zone in the calculations. Just-in-case rather than emergency purchases lower the specificity of OTC sales. This behaviour may partly explain discrepancies in previous studies using OTC sales data for syndromic surveillance [8, 9].

The signal rates for 1177 calls varied from 1% to 10% depending on outbreak and syndrome. The rates were higher for the smallest of the four validated outbreaks (Helsingborg), but there are several reasons for questioning the officially reported size of this outbreak. First, the number of outbreak cases ($n = 369$) was based on a local outbreak investigation relying on traditional methods, i.e. case-by-case contacts, and no cross-sectional survey was conducted in the population. Second, since the outbreak involved a common disease agent with well-known symptoms (norovirus) during high season, the expectation is for rather low contact rates. Third, as the outbreak passed without an official Swedish public warning (VMA – Important Message to the Public) the news media coverage was limited. Considering these factors, the signal rates in Helsingborg should be more in line with those for the outbreak in Lilla Edet, which involved the same agent, but in Lilla Edet, the signal rates were about tenfold lower. Thus, an alternative and more reasonable hypothesis for the high rates in

Helsingborg is that the outbreak was in fact larger than the official figure, perhaps as much as tenfold larger. This illustrates an important potential use of syndromic surveillance for SA, i.e. outbreak size estimation.

For several reasons, we decided to apply our own detection algorithm, despite the availability of various outbreak detection algorithms [13]. The objective was to compare different data streams (symptoms), not detection algorithms. Furthermore, dealing with local point-source outbreaks, limited in space and time, we needed to take spatial and temporal variation into account, but exclude large-scale disease trends, e.g. winter vomiting disease (norovirus). Last, we wanted a simple detection algorithm that was sufficiently transparent for non-statisticians. Practitioners ultimately decide on which signals to act, and non-transparent signals can then be a problem.

The two largest outbreaks were extended in time, from one to several months. The detection analysis showed that early warnings could have been issued weeks to months in advance and could potentially have contributed to crisis preparedness and prevention, reducing the burden of disease. However, the identification of outbreak signals does not by itself constitute an efficient system for syndromic surveillance or outbreak detection. Beyond outbreak signals, the system must also include decision-making and operational measures that aim at epidemic control and outbreak management. Thus, it is impossible to say whether the outbreak signals in question would have affected the epidemics in Östersund and Skellefteå.

This study shows that syndromic surveillance of point-source outbreaks of acute gastroenteritis can serve both SA and EED. In particular, telephone triage data, with sufficient temporal and spatial resolution, revealed clear and strong outbreak signals for outbreaks involving more than 1000 cases, assuming that the outbreak in Helsingborg was larger than the official figure ($n = 369$). However, it is still difficult to generalize the findings. First, we lacked data on outbreaks of moderate size (300–1000 cases). Thus, we cannot draw any conclusion regarding outbreak detection limits from this study. Second, technological, medical, psychological and organizational factors influence signal rates. In order to determine the real potential of syndromic surveillance, all these factors need to be addressed and controlled in future research. It is a difficult task, but essential if we are to improve our capacity and capability for SA.

Other important work that remains is to pool our knowledge and experience of syndromic surveillance of local point-source outbreaks across national borders. For obvious reasons, large-scale epidemics motivate international cooperation and research. Only a handful of studies have been published on syndromic surveillance for local point-source outbreaks. Point-source outbreaks are hard to detect, monitor and predict, thereby reducing the power of EED. The problem is to a large extent due to the quality of data, quality being proportional to outbreak size. Small outbreaks do not motivate large investigations. For the purpose of SA, however, we need better data on local point-source outbreaks to map outbreak characteristics and signal properties. By sharing and evaluating local outbreak data across national borders, we will also be better equipped to synchronize national syndromic surveillance systems that are based on national, regional and local solutions to healthcare information and communication.

SUPPLEMENTARY MATERIAL

For supplementary material accompanying this paper visit http://dx.doi.org/10.1017/S0950268813001088.

ACKNOWLEDGEMENTS

DECLARATION OF INTEREST

None.

REFERENCES

1. **Bradley CA, *et al.*** BioSense: implementation of a National Early Event Detection and Situational Awareness System. *Morbidity and Mortality Weekly Report* 2005; **54** (Suppl.): 11–19.
2. **Fricker RD Jr.** Some methodological issues in biosurveillance. *Statistics in Medicine* 2011; **30**: 403–415.
3. **Buckeridge DL.** Outbreak detection through automated surveillance: a review of the determinants of detection. *Journal of Biomedical Informatics* 2007; **40**: 370–379.
4. **Berger M, Shiau R, Weintraub JM.** Review of syndromic surveillance: implications for waterborne disease detection. *Journal of Epidemiology and Community Health* 2006; **60**: 543–550.
5. **Morse SS.** Public health surveillance and infectious disease detection. *Biosecurity and Bioterrorism : Biodefense Strategy, Practice, and Science* 2012; **10**: 6–16.
6. **Cooper DL, *et al.*** Can syndromic surveillance data detect local outbreaks of communicable disease? A model using a historical cryptosporidiosis outbreak. *Epidemiology and Infection* 2006; **134**: 13–20.
7. **Smith S, *et al.*** Value of syndromic surveillance in monitoring a focal waterborne outbreak due to an unusual Cryptosporidium genotype in Northamptonshire, United Kingdom, June–July 2008. *Eurosurveillance* 2010; **15**: 19643.
8. **Edge VL, *et al.*** Syndromic surveillance of gastrointestinal illness using pharmacy over-the-counter sales. A retrospective study of waterborne outbreaks in Saskatchewan and Ontario. *Canadian Journal of Public Health* 2004; **95**: 446–450.
9. **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 and Decision Making* 2010; **10**: 39.
10. **Ernesater A, Holmstrom I, Engstrom M.** Telenurses' experiences of working with computerized decision support: supporting, inhibiting and quality improving. *Journal of Advanced Nursing* 2009; **65**: 1074–1083.
11. **Hulth A, Rydevik G, Linde A.** Web queries as a source for syndromic surveillance. *PLoS One* 2009; **4**: e4378.
12. **Hulth A, Rydevik G.** Web query-based surveillance in Sweden during the influenza A(H1N1)2009 pandemic, April 2009 to February 2010. *Eurosurveillance* 2011; **16**(18).
13. **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.