Title

Gastrointestinal Disease Outbreak Detection Using Multiple Data Streams from Electronic Medical Records

Journal

Foodborne Pathogens and Disease, 9(5)

ISSN

1535-3141

Authors

Greene, Sharon K
Huang, Jie
Abrams, Allyson M
et al.

Publication Date

2012-05-01

DOI

10.1089/fpd.2011.1036

FOODBORNE PATHOGENS AND DISEASE
Volume 9, Number 5, 2012

DOI: 10.1089/fpd.2011.1036

Gastrointestinal Disease Outbreak Detection Using Multiple Data Streams from Electronic Medical Records

Sharon K. Greene,[1] Jie Huang,[2] Allyson M. Abrams,[1] Debra Gilliss,[3] Mary Reed,[2] Richard Platt,[1] Susan S. Huang,[4] and Martin Kulldorff[1]

Abstract

Background: Passive reporting and laboratory testing delays may limit gastrointestinal (GI) disease outbreak detection. Healthcare systems routinely collect clinical data in electronic medical records (EMRs) that could be used for surveillance. This study's primary objective was to identify data streams from EMRs that may perform well for GI outbreak detection. ***Methods:*** Zip code-specific daily episode counts in 2009 were generated for 22 syndromic and laboratory-based data streams from Kaiser Permanente Northern California EMRs, covering 3.3 million members. Data streams included outpatient and inpatient diagnosis codes, antidiarrheal medication dispensings, stool culture orders, and positive microbiology tests for six GI pathogens. Prospective daily surveillance was mimicked using the space-time permutation scan statistic in single and multi-stream analyses, and space-time clusters were identified. Serotype relatedness was assessed for isolates in two *Salmonella* clusters. ***Results:*** Potential outbreaks included a cluster of 18 stool cultures ordered over 5 days in one zip code and a *Salmonella* cluster in three zip codes over 9 days, in which at least five of six cases had the same rare serotype. In all, 28 potential outbreaks were identified using single stream analyses, with signals in outpatient diagnosis codes most common. Multi-stream analyses identified additional potential outbreaks and in one example, improved the timeliness of detection. ***Conclusions:*** GI disease-related data streams can be used to identify potential outbreaks when generated from EMRs with extensive regional coverage. This process can supplement traditional GI outbreak reports to health departments, which frequently consist of outbreaks in well-defined settings (e.g., day care centers and restaurants) with no laboratory-confirmed pathogen. Data streams most promising for surveillance included microbiology test results, stool culture orders, and outpatient diagnoses. In particular, clusters of microbiology tests positive for specific pathogens could be identified in EMRs and used to prioritize further testing at state health departments, potentially improving outbreak detection.

Introduction

In 2007, foodborne disease outbreaks were associated with over 21,000 reported illnesses in the United States (CDC, 2010b). Health departments (HDs) are commonly notified of focal (e.g., restaurant-associated) outbreaks by passive reports from clinicians or patients. These reports may be incomplete, delayed, and/or non-representative. Outbreaks of intermediate scope, such as those caused by contaminated commercial products, are often detected via laboratory testing. HDs can monitor for unusual increases in passive laboratory-based reports of notifiable diseases (CDC, 2009b) or for clusters of isolates with identical pulsed-field gel electrophoresis (PFGE) patterns (Swaminathan *et al.*, 2001; Gerner-Smidt *et al.*, 2006). However, some gastrointestinal (GI) pathogens are not nationally notifiable (e.g., norovirus, campylobacteriosis), and due to resource limitations at HDs, laboratory testing may be delayed. Generalized outbreaks (e.g., seasonal rotavirus increases) are seldom reported to HDs.

In electronic medical records (EMRs), healthcare systems routinely collect GI disease-related clinical and laboratory data. Using these data may improve the timeliness and representativeness of outbreak surveillance. A prior evaluation of 1 year of EMR data in four states for nine syndromes, including upper and lower GI, did not identify any clusters of public health interest (Yih *et al.*, 2010); however, only

[1]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts.
[2]Kaiser Permanente Division of Research, Oakland, California.
[3]Disease Investigations Section, Infectious Diseases Branch, California Department of Public Health, Richmond, California.
[4]Division of Infectious Diseases and Health Policy Research Institute, University of California Irvine School of Medicine, Orange, California.

ambulatory care (AC) diagnoses were used, though other data streams may be more informative. It is also possible that no single clinical data stream is optimal for surveillance, but that multiple syndromic and laboratory data sources could be useful to identify outbreaks.

Space-time clusters in laboratory-based data have been identified for *Escherichia coli* O157 (Pearl *et al.*, 2006) and *Shigella* (Jones *et al.*, 2006; Stelling *et al.*, 2009). Syndromic surveillance for GI outbreaks has been evaluated using AC diagnoses (Yih *et al.*, 2005, 2010), emergency department (ED) chief complaints (Balter *et al.*, 2005; Kulldorff *et al.*, 2005), telephone helplines (Caudle *et al.*, 2009), and medication sales (Kirian and Weintraub, 2010; Pelat *et al.*, 2010). To facilitate capacity planning, one hospital conducted univariate temporal analyses of three streams and determined that GI complaints in free text were more useful for rotavirus outbreak detection than either rotavirus antigen laboratory tests or diarrheal disease discharge diagnoses (Levin and Raman, 2005). To our knowledge, no study has used the same patient population to evaluate multiple EMR data streams for spatiotemporal GI outbreak detection.

Prior work using EMR data to detect localized excess influenza-like illness (ILI) suggested that AC diagnoses, reverse transcription–polymerase chain reaction (RT-PCR) tests ordered, and antiviral dispensings were most useful for surveillance (Greene *et al.*, 2011). However, the analogous data streams may not be useful for GI outbreak detection. ILI is caused by respiratory viruses transmitted person-to-person, peaks each winter in temperate climates, and has widespread illness activity. In contrast, GI illness has heterogeneous viral, bacterial, parasitic, and chemical etiologies; is transmitted person-to-person via the fecal-oral route or by a contaminated vehicle; may or may not have a demonstrable seasonal pattern; and the scale may be attributable to a highly localized point source or an internationally disseminated product. True increased ILI activity should be reflected across all streams around the same time, but true increased GI activity may be detectable in only some streams. For example, lower GI syndrome and positive *Shigella* tests may reflect a shigellosis outbreak, while upper GI syndrome and positive norovirus tests may reflect a norovirus outbreak.

The objectives of this study were to use data from a comprehensive regional health system to (1) create syndromic and laboratory-based data streams for GI disease from EMRs, (2) mimic near real-time prospective surveillance to identify which streams perform well for outbreak detection, and (3) compare results of single stream and multi-stream analyses.

Methods

Study population

Kaiser Permanente Northern California (KPNC) is a large, integrated healthcare delivery system utilizing comprehensive EMRs with inpatient, outpatient, laboratory, and pharmacy data. As of January 2009, KPNC included 18 medical centers and 3.3 million members, representing approximately 22% of the total population residing in 46 counties (951 zip codes) in the Central Valley and San Francisco Bay area. KPNC laboratory-based reports of notifiable diseases are sent to the local HD, which in turn report them to the California Department of Public Health (CDPH) (Backer *et al.*, 2001; California Code of Regulations, 2009).

Data streams

Twenty-two data streams were analyzed, based on syndromic definitions ($n=14$), prescription drug dispensings ($n=1$), microbiology tests ordered ($n=1$), and microbiology test results ($n=6$) (Table 1). Syndromic definitions were adapted from lists previously developed by a Centers for Disease Control and Prevention/Department of Defense working group (CDC, 2003). Upper GI (UGI) captured vomiting, and lower GI (LGI) captured diarrhea and gastroenteritis (Table 2). Each stream consisted of residential zip code-specific daily episode counts in 2009. Within each stream, an "episode" was the first patient encounter after at least 42 days with no encounter (Jung *et al.*, 2009).

Additional streams were generated but ultimately excluded from analyses. Pathogen-specific ICD-9 codes were excluded because of unreliability, e.g., there were fewer ICD-9 code episodes for *Salmonella* (003.0, 003.20, 003.29, 003.8, 003.9) than laboratory tests positive for *Salmonella*, and ICD-9 codes are assigned before test results become available. Microbiology test results for specific pathogens (e.g., rotavirus antigen tests) were excluded due to low usage.

Univariate single stream analyses

Near real-time prospective surveillance was mimicked by analyzing data "each day." The prospective space-time permutation scan statistic (Kulldorff *et al.*, 2005) was used to detect and evaluate the strength of potential outbreaks. Using a variable-sized cylinder, where the circular base represents space and the height represents time, the method scans the geographic area for potential outbreaks at different locations, with different radii and lengths of time. For each location and cylinder size, a likelihood ratio-based test statistic compares the observed number of cases within the cylinders with what would be expected if the spatial and temporal locations of all cases were independent of each other so that there is no space-time interaction. As such, it adjusts for any purely spatial and any purely temporal clusters. The cylinder with the maximum likelihood ratio is the most likely cluster, the least likely to have occurred by chance. Calculations were done using SaTScan™ (www.satscan.org).

The maximum geographical size of the cylinder was set to contain at most 50% of the observed episodes, and the maximum temporal size was set to 14 days. Since the weekly pattern of health-seeking behavior may vary geographically, we adjusted for space by day-of-week interaction, with holidays treated as Sundays and the day after holidays treated as Mondays. The surveillance period was January 1 to December 31, 2009. A 365-day rolling control period established local baselines for each zip code.

To determine statistical significance, 9,999 Monte Carlo simulations (Dwass, 1957) were performed "each day." The recurrence interval (RI) for each cluster represents the length of follow-up required to expect one cluster at least as unusual as the observed cluster by chance (Kleinman *et al.*, 2004). The single stream RIs were statistically adjusted for multiple testing in terms of the thousands of cylinders considered for each data stream, but not for the fact that 22 different single streams were analyzed. In Table 7 below, where only five streams were considered, we also present RIs that were not only adjusted for the many cylinders, but also for the five streams analyzed. This was done by dividing each RI by five.

Table 1. Data Stream Definitions

Category	#	*Data stream*	*Notes*
UGI and LGI, ICD-9 code based	1	UGI in AC	ICD-9 codes in Table 2
	2	LGI in AC	
	3	UGI in ED	
	4	LGI in ED	
	5	UGI hospital discharge	Primary discharge diagnosis
	6	LGI hospital discharge	
	7	GI in AC, <5 year-olds	UGI and LGI diagnosis codes combined.
	8	GI in ED, <5 year-olds	<5 year-olds are the age category at highest
	9	GI hospital discharge, <5 year-olds	risk for rotavirus gastroenteritis (Peck and Bresee 2006), and a population at higher risk for outbreaks in daycare centers and petting zoos.
	10	UGI in AC with Rx	Antibiotics (Rx) were identified using National
	11	LGI in AC with Rx	Drug Codes for macrolides, quinolones,
	12	UGI in ED with Rx	metronidazole, and trimethoprim-
	13	LGI in ED with Rx	sulfamethoxazole
	14	GI hospital admission	Text strings for vomiting, diarrhea, and gastroenteritis, with exclusions for chronic diarrhea, pregnancy, chemotherapy, and intoxication (Chapman *et al.*, 2010; Chapman 2011)
Prescription antidiarrheals	15	Antidiarrheal dispensing	NDCs for diphenoxylate and loperamide. Note that a portion of these dispensings would reflect prescriptions in advance of international travel, rather than acute illness treatment. Over-the-counter anti-diarrheal agents were not considered, as relevant clinician recommendations for symptomatic treatment would be captured only sporadically in EMR text fields.
Microbiology tests ordered in any setting	16	Stool culture tests	
Microbiology tests positive for GI pathogens	17	*Campylobacter*	Positive stool culture test. Analyzed according to
	18	*Salmonella*	date test ordered, not according to lagged date
	19	*Shigella*	when test results became available.
	20	*E. coli* O157:H7	
	21	*Vibrio parahaemolyticus*	
	22	*Cryptosporidium*	Positive stain

AC, ambulatory care; ED, emergency department; GI, gastrointestinal; LGI, lower gastrointestinal; Rx, antibiotic prescription; UGI, upper gastrointestinal.

For streams other than tests positive for GI pathogens, all clusters from prospective analyses with RI of >365 days were identified. Clusters from the same stream on consecutive days and overlapping in space were grouped together into "cluster sequences." These cluster sequences were then compared across streams, and sequences with at least 1-day temporal overlap and spatial overlap (cluster center in other cluster) were further grouped together into "potential outbreaks."

For the six streams of tests positive for GI pathogens, clusters with RI of >60 days were identified. This lower RI threshold was selected because of the sparseness of microbiology data, and because the trigger to begin a cluster investigation may be lower for these streams, which are more specific for acute infections than the syndromic streams.

The epidemiological interpretation of clusters is subjective. Clusters may be prioritized for possible investigation by simultaneously weighing (1) the number of observed cases (the greater, the more urgent the need for a public health intervention), (2) the observed/expected number of cases (the greater, the more excess risk), (3) the RI (the greater, the less likely the observed clustering is due to chance), and (4) the degree of localization (a smaller radius more strongly suggests a common source or localized person-to-person transmission).

Multi-stream analyses

Multivariate analyses use multiple streams in the same statistical analysis (Burkom *et al.*, 2005; Kulldorff *et al.*, 2007; Rolka *et al.*, 2007). Five streams were included in multivariate analyses: three microbiology-based streams were selected because cluster detection could prompt case interviews and PFGE testing of isolates, and two syndromic streams were selected based on frequency of episodes and contribution to potential outbreak detection in single stream analyses. To avoid multicollinearity across streams, a patient appearing in more than one stream within 14 days was retained in only one stream. A priority order of increasing frequency was used so

Table 2. Syndromic Definitions for Upper and Lower Gastrointestinal Illness

Syndrome	*Definition*	*International Classification of Diseases, Ninth Revision codes*
UGI illness	Epidemic vomiting syndrome	078.82
	Gastritis and duodenitis	535.0, 535.4, 535.5, 535.6
	Persistent vomiting	536.2
	Nausea and vomiting	787.0
LGI illness	Other bacterial food poisoning	005.89, 005.9
	Intestinal infection due to other organisms	008.49
	Bacterial enteritis unspecified	008.5
	Enteritis due to other viral enteritis	008.69
	Intestinal infection due to other organism not elsewhere classified	008.8
	Ill-defined intestinal infections	009
	Regional enteritis	555.0, 555.1, 555.2
	Other and unspecified noninfectious gastroenteritis and colitis	558.2, 558.9
	Unspecified disorder of intestine	569.9
	Visible peristalsis	787.4
	Diarrhea	787.91

LGI, lower gastrointestinal; UGI, upper gastrointestinal.

that patients would be retained in the stream in which they were most proportionally informative: *E. coli* O157:H7, *Salmonella, Campylobacter,* LGI in ED, and LGI in AC.

With these five streams, the multivariate scan statistic simultaneously searched for clusters in all 31 possible combinations of one or more streams, adjusting for the multiple testing inherent in both the thousands of cylinders evaluated for each combination and in the 31 data stream combinations evaluated. On each surveillance day, the combined log-likelihood was defined as the sum of the individual log-likelihoods for those streams with more observed events than expected (Kulldorff *et al.,* 2007). The maximum cylinder size and other settings were the same as for the single stream analyses.

State HD data

For context, but not for direct comparison with potential outbreaks identified using KPNC data, we compiled a list of GI disease outbreaks known to CDPH occurring in non-institutional settings affecting any of the 16 counties for which KPNC had ≥10% population coverage. One author (D.G.) provided preliminary GI illness outbreak reports, which had been sent to CDPH by local HDs soon after outbreak detection. CDPH's final foodborne disease outbreak reports reflected the best available information after completed outbreak investigations. The preliminary and final outbreak reports were unlinked and represented separate information sources.

The CDPH state laboratory performs *Salmonella* serotyping. Previously collected serotype information was obtained to evaluate the two potential *Salmonella* outbreaks with the most observed cases identified in KPNC EMR data.

Results

Figure 1 shows the frequencies and seasonal patterning of the 22 data streams. The relative frequencies of pathogens (Fig. 1A) were consistent with the most common laboratory-confirmed infections in 2009 at the California FoodNet site (CDC, 2010a).

Single stream analyses

In an illustrative potential outbreak, a signal occurred on November 9, 2009 for stool culture tests ordered, with signals continuing over the subsequent four days (Tables 3 and 4, #19). Ultimately, 18 tests were ordered in one zip code over 5 days, with less than four tests expected. Statistically, this was very unlikely to occur by chance. No corresponding cluster was detected of tests positive for any of the six specific pathogens under surveillance, but the cluster could have reflected an outbreak of a viral pathogen.

In single stream analyses of 16 non-pathogen-specific streams, 24 potential outbreaks were detected using a 365-day RI threshold (Table 4). Of the 16 streams, two had no clusters with RI of >365 days: UGI with Rx, in the AC and ED settings. Three streams each contributed to the identification of >5 potential outbreaks, all diagnoses in outpatient settings: LGI in ED, UGI in ED, and LGI in AC.

In single stream analyses of six pathogen-specific streams, five potential outbreaks (two *Campylobacter* and three *Salmonella*) were detected using a 60-day RI threshold (Table 5). In potential outbreak B, serotype information was available for five of the six isolates, all from patients residing in one zip code. All isolates were *Salmonella enterica* serotype Thompson. Nationwide, *Salmonella* serotype Thompson represents only 1% of all serotyped *Salmonella* isolates (CDC, 2008), which suggests this may have been an outbreak with a common source or person-to-person transmission. This event was only detected using microbiology data, not with any of the less specific and noisier syndromic streams.

In potential outbreak C, serotype information was available for all 10 isolates: four Enteritidis (the most common serotype (CDC 2010a)), three Montevideo, one Heidelberg, one Infantis, and one Paratyphi B L(+) tartrate +. This mix of serotypes over 10 days across 71 zip codes is unlikely to represent a common source outbreak.

Multi-stream analyses

Multivariate analyses detected six potential multi-stream outbreaks (Table 6), two of which were not detected in single stream analyses. Some potential outbreaks were detected by both single and multi-stream analyses, but at different RI strengths. Table 7 shows examples where multi-stream

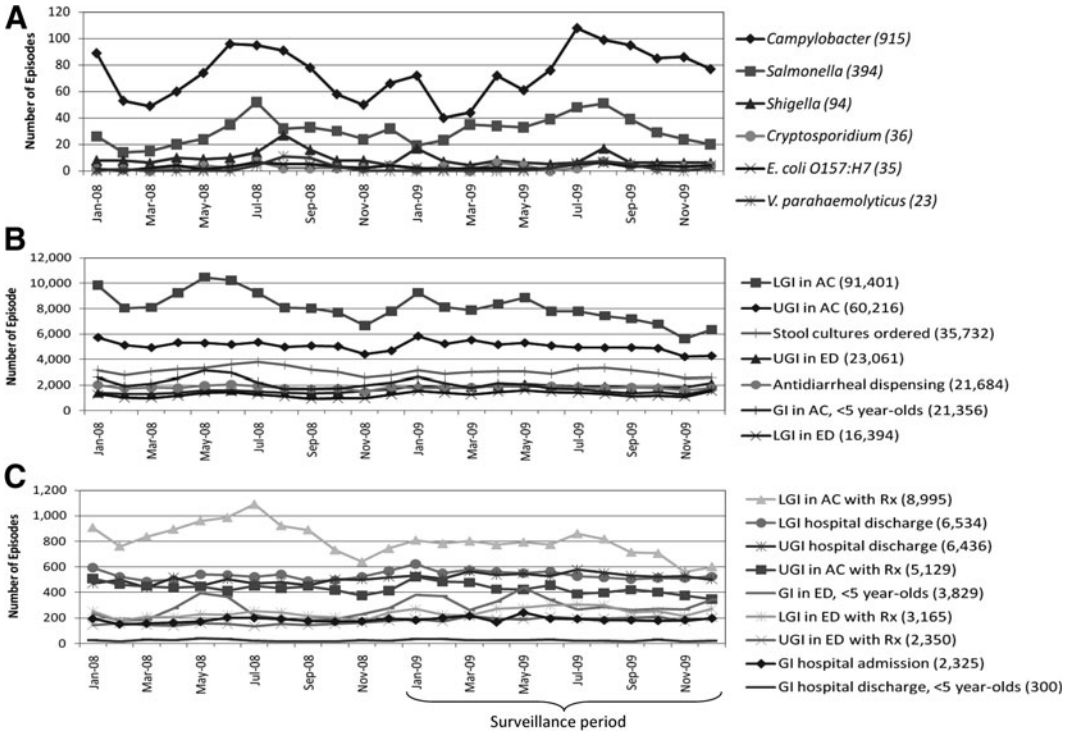

FIG. 1. Monthly number of episodes in data streams related to gastrointestinal illness in Kaiser Permanente Northern California, 2008–2009. **(A)** Data streams for microbiology tests positive for specific pathogens ($n=6$). **(B)** Non-laboratory-based data streams with ≥10,000 episodes in 2009 ($n=7$). **(C)** Non-laboratory-based data streams with <10,000 episodes in 2009 ($n=9$). Legend indicates data stream and number of episodes in 2009.

analyses detected a cluster with a higher RI than single stream analyses, and vice versa. In one example centered in Firebaugh over 6 days, multi-stream analysis resulted in more timely detection of a potential outbreak than single stream analysis (Fig. 2). Maps of clusters on all days in 2009 with signals in either single or multi-stream analyses are available (see Supplementary Material available online at www.liebertonline.com/fpd).

Outbreaks reported to CDPH

Twenty-four GI outbreaks affecting the study area in 2009 were reported to CDPH. Three were multi-state outbreaks associated with contaminated commercial food products (Nielsen *et al.*, 2010; CDC, 2009a; FDA, 2009). Two additional outbreaks were laboratory-confirmed: a restaurant-based norovirus outbreak ($n=8$) and a home-based *E. coli* O157:H7 outbreak ($n=5$). The remaining 19 reported outbreaks had no laboratory-confirmed etiology, were mostly in restaurants and child day care centers, and had few reported cases (median, 10; range, 3–35).

Discussion

GI disease-related data streams can be generated from EMRs and prospectively used to identify potential outbreaks in settings with extensive regional coverage, such as KPNC. Different surveillance systems will have different strengths and weaknesses, and one should not expect EMR data streams to detect all outbreaks reported to CDPH and vice versa. Not all true outbreaks are recognized, and not all recognized outbreaks are reported (CDC, 2010b). In addition, the three multi-state outbreaks known to CDPH may have clustered only in time but not in space within the catchment area, so that they could not be detected by spatial methods. Also, with a median number of nine cases, the other 21 outbreaks reported to CDPH would have been difficult to detect. Given

Table 3. Illustrative Potential Outbreak: Stool Culture Tests Ordered over Five Consecutive Days in One Zip Code

Start date	*Signal date*	*Number of zip codes*	*Observed (O)*	*Expected (E)*	*O/E*	*Recurrence interval (days)*
11/9/09	11/9/09	1	9	0.6	14.4	3,333
11/9/09	11/10/09	1	11	1.4	8.1	769
11/9/09	11/11/09	1	14	2.1	6.8	2,000
11/9/09	11/12/09	1	17	2.7	6.4	5,000
11/9/09	11/13/09	1	18	3.4	5.4	10,000

Table 4. Clusters from Single Stream Analysis with Recurrence Interval of >365 Days, Grouped into "Potential Outbreaks" and Listed in Chronological Order as of the Date with the Maximum Recurrence Interval Within Each "Cluster Sequence"

Potential outbreak	*Central location*	*Radius (km)*	*Number of zip codes*	*Percentage of KPNC population in geographic area*	*Data stream*	*Details as of date with maximum recurrence interval*					
						Signal date (2009)	*Number of days in cluster*	*Observed (O)*	*Expected (E)*	*O/E*	*Recurrence interval*
1	San Francisco	14	64	7.8	GI in AC, <5 year-olds	1-Jan	14	88	46.5	1.9	5,000
	Sausalito	18	70	6.5	LGI in AC	1-Jan	14	246	161.5	1.5	10,000
2	San Jose	12	49	6.2	UGI in ED	4-Jan	14	48	20.8	2.3	3,333
3	Westley	36	34	10.1	LGI in ED	23-Jan	8	39	15.4	2.5	769
4	Half Moon Bay	29	67	6.9	LGI in AC	27-Jan	13	563	435.9	1.3	10,000
	Sausalito	24	102	11.3	GI in AC, <5 year-olds	14-Feb	14	159	103.8	1.5	500
	San Francisco	1	6	0.3	LGI in AC with Rx	16-Feb	4	7	0.4	15.7	625
5	Hilmar	32	31	0.9	UGI in ED	15-Feb	11	36	12.3	2.9	5,000
	Crows Landing	35	31	2.5	LGI in ED	19-Feb	14	34	12.5	2.7	526
	Huron	195	210	12.5	UGI hospital discharge	24-Feb	5	26	7.9	3.3	2,500
6	Sacramento	7	19	2.0	LGI in ED	30-Mar	12	31	10.2	3.0	1,111
7	Sacramento	13	27	5.9	GI hospital discharge, <5 year-olds	29-Apr	13	8	1.2	6.9	625
8	Merced	41	24	17.0	LGI in ED	5-May	11	12	1.9	6.5	435
	Firebaugh	101	134	0.4	LGI in AC	14-May	14	413	309.9	1.3	10,000
	Chowchilla	106	151	6.8	LGI in ED	14-May	14	93	46.9	2.0	10,000
	Linden	53	89	9.9	UGI hospital discharge	23-May	11	31	11.2	2.8	417
	Merced	97	138	2.9	GI in AC, <5 year-olds	26-May	7	93	48.7	1.9	3,333
	Dos Palos	80	72	13.6	UGI in ED	28-May	14	77	36.0	2.1	10,000
	Fowler	8	2	0.1	LGI in AC	31-May	1	5	0.1	38.4	769
	Valley Springs	55	81	6.6	LGI in AC	1-Jun	14	324	240.5	1.4	370
	Lodi	30	39	11.2	UGI in ED	2-Jun	3	21	5.0	4.2	1,667
9	Berkeley	7	18	2.3	LGI in ED with Rx	21-Jun	13	14	2.6	5.4	714
10	Winters	16	3	0.5	LGI in ED	21-Jun	13	15	2.9	5.1	400
11	San Jose	0	1	4.2	LGI in ED	24-Jun	4	8	0.6	14.3	1,667
	Santa Clara	12	54	3.5	UGI in ED	1-Jul	12	65	28.0	2.3	10,000
	Mountain View	9	23	2.5	GI hospital discharge, <5 year-olds	3-Jul	5	5	0.4	12.7	1,429
	Santa Clara	12	57	7.0	LGI in ED	21-Jul	11	57	24.0	2.4	10,000
	San Jose	7	32	4.1	GI in ED, <5 year-olds	16-Aug	13	16	3.2	5.0	3,333
	Mountain View	19	76	5.6	LGI in ED	30-Aug	14	78	38.1	2.1	10,000
12	Montague	366	257	26.0	LGI in AC	11-Jul	5	318	225.9	1.4	10,000
	Kenwood	16	21	2.3	LGI hospital discharge	17-Jul	4	14	2.3	6.0	1,111
13	Campo Seco	52	74	5.0	GI in AC, <5 year-olds	25-Jul	13	72	35.9	2.0	2,500
14	Vallejo	0	1	0.5	GI in ED, <5 year-olds	29-Jul	2	4	0.1	46.7	417
15	Stockton	42	57	7.3	UGI in AC	11-Oct	3	57	26.1	2.2	909
16	Hidden Valley Lake	42	31	0.8	LGI in AC	17-Oct	13	120	72.1	1.7	556
	Guinda	58	54	1.7	UGI in ED	26-Oct	12	53	22.8	2.3	3,333
17	San Francisco	2	2	0.5	Antidiarrheal dispensing	29-Oct	1	8	0.6	14.4	1,111

(*continued*)

Table 4. Continued

Potential outbreak	Central location	Radius (km)	Number of zip codes	Percentage of KPNC population in geographic area	Data stream	Details as of date with maximum recurrence interval					
						Signal date (2009)	Number of days in cluster	Observed (O)	Expected (E)	O/E	Recurrence interval
18	Rough and Ready	46	43	1.9	LGI in AC with Rx	2-Nov	14	22	5.8	3.8	714
19	Elk Grove	0	1	0.8	Stool cultures ordered	13-Nov	5	18	3.4	5.4	10,000
20	San Rafael	19	42	4.1	GI hospital admission	18-Nov	13	15	2.6	5.7	10,000
21	San Jose	1	2	0.8	LGI in AC with Rx	25-Nov	6	9	0.9	9.9	909
22	San Francisco	5	32	2.7	Antidiarrheal dispensing	2-Dec	4	30	10.2	3.0	500
23	Boulder Creek	27	60	5.3	LGI in ED	8-Dec	6	32	10.9	3.0	1,667
24	Fairfield	24	11	5.9	UGI in ED	23-Dec	13	57	24.8	2.3	10,000
	Capay	34	16	0.7	GI in ED, <5 year-olds	25-Dec	3	5	0.2	27.6	588

Microbiology tests positive for GI pathogens are excluded.

AC, ambulatory care; ED, emergency department; GI, gastrointestinal illness; LGI, lower gastrointestinal illness; Rx, antibiotic prescribed; UGI, upper gastrointestinal illness; KPNC, Kaiser Permanente Northern California.

Table 5. Clusters from Single Stream Analysis with Recurrence Interval of >60 Days in Data Streams for Microbiology Tests Positive for Gastrointestinal Pathogens

Potential outbreak	Central location	Radius (km)	Number of zip codes	Percentage of KPNC population in geographic area	Data stream	Details as of date with maximum recurrence interval					
						Signal date	Number of days in cluster	Observed	Expected	Observed / expected	Recurrence interval
A	San Francisco	3	17	1.0	*Campylobacter*	10-Jan	3	4	0.1	28.4	169
B	Union City	5	3	2.1	*Salmonella*	4-Jul	9	6	0.7	8.6	68
C	San Jose	27	71	10.7	*Salmonella*	23-Aug	10	10	2.0	5.1	147
D	Moraga	21	24	5.9	*Salmonella*	13-Sep	5	5	0.4	11.8	71
E	Suisun City	23	13	4.8	*Campylobacter*	2-Oct	8	9	1.2	7.6	909

KPNC, Kaiser Permanente Northern California.

Table 6. Clusters from Multi-Stream Analysis with Recurrence Interval (RI) of >365 Days, as of the Date with the Maximum RI

Potential outbreak from Table 4	Central location	Radius (km)	Number of zip codes	LGI in AC			LGI in ED			Campylobacter			Salmonella			Multi-stream		
				O	E	O/E	O	E	O/E	O	E	O/E	O	E	O/E	Signal date (2009)	Number of days in cluster	RI
1	Sausalito	19	83	327	229.3	1.4	81	64.2	1.3	–	–	–	2	0.9	2.3	1-Jan	14	10,000
8	Firebaugh	101	137	404	332.7	1.2	82	42.1	2.0	4	1.9	2.1	1	0.6	1.6	12-May	14	10,000
None	Arnold[a]	103	190	693	569.7	1.2	166	124.8	1.3	6	5.6	1.1	–	–	–	15-Jul	13	10,000
11[c]	Santa Clara	12	57	343	331.9	1.0	66	27.7	2.4	–	–	–	–	–	–	24-Jul	14	10,000
11[c]	Sunnyvale	14	62	274	244.6	1.1	47	18.1	2.6	7	3.8	1.9	6	2.1	2.9	30-Aug	13	10,000
None	Discovery Bay[b]	0	1	7	0.5	15.0	–	–	–	–	–	–	1	0.03	36.5	26-Oct	2	417

[a]Weakest candidate for possible investigation: despite maximum RI, low O/E across all constituent data streams and widest geographic coverage.
[b]Strongest candidate for possible investigation: high O/E, high geographic localization, with RI well over 365 days.
[c]Both of these correspond to the same potential outbreak from single stream analysis.
AC, ambulatory care; ED, emergency department; LGI, lower gastrointestinal illness; UGI, upper gastrointestinal illness; O, observed; E, expected.

the 22% KPNC population coverage, on average, only two cases in each outbreak would be KPNC members. Thus, monitoring EMR streams would complement rather than replace traditional outbreak detection systems, which typically detect outbreaks in well-defined settings with no laboratory-confirmed pathogen.

Several EMR streams emerged as most promising for GI outbreak detection. First, microbiology test results are very specific for acute enteric illness, and even with a low RI, a cluster suggestive of a true outbreak was identified of genetically related isolates (Table 5, potential outbreak B). Also, stool culture orders reflect clinician suspicion of acute illness and disproportionately represent patients with bloody diarrhea and diarrhea duration of ≥3 days (Scallan *et al.*, 2006); it is unknown whether the intriguing potential outbreak identified (Table 3) reflected a true outbreak or something else, but it represents the type of event that public health officials may be interested in prospectively detecting. Finally, outpatient diagnoses contributed to the most potential outbreaks (Table 4). Data are rapidly available in the KPNC EMR: diagnoses and stool culture orders typically within one day, and positive stool culture test results in a median of three days. Multi-stream analyses could identify potential outbreaks too faint for detection in single stream analyses (Table 7) and improve the timeliness of detection (Fig. 2), but were not consistently superior or inferior to single stream analyses.

Several limitations should be noted. First, analyses were by zip code of patient residence, so point source outbreaks where people congregate, then disperse, may be missed. Second, only a portion of the total population are KPNC members, and only a fraction of members with GI illness will seek care or submit a stool specimen and thus appear in an EMR; generally, about 20% of patients with an acute diarrheal illness seek medical care, and 4% submit a stool specimen (Jones *et al.*, 2007). Hence, an EMR-based surveillance system cannot be expected to detect very small outbreaks. Third, we evaluated only one geographical area during 1 year. The apparent relative strengths across streams could be different in other places or years.

Microbiology test results in EMRs seem to be especially promising for outbreak detection. Given limited resources and competing priorities within HDs, there are delays in serotyping and an inability to perform PFGE testing on all isolates. Currently at CDPH, *Salmonella* isolates are prioritized for investigation (e.g., patient interview and/or PFGE testing) based on the presence of an unusual serotype, an increase in a common serotype, or recognized clusters. An alert of a cluster of *Campylobacter* isolates might trigger CDPH to collect case report details or conduct PFGE testing, which are non-routine activities for campylobacteriosis. HDs could strengthen cooperative partnerships with healthcare systems like KPNC, such that in addition to routinely submitting their isolates to the HD for possible further testing, laboratories in these healthcare systems could provide HDs with counts of microbiology tests ordered and positive, by zip code, for automated daily analyses at the HD, in order to identify unusual space-time clustering. In concert with other strategies for near real-time laboratory-based surveillance (Nielsen *et al.*, 2006; Miller *et al.*, 2010), this could more efficiently prioritize testing and patient interviews, potentially improving outbreak detection.

Table 7. Six Illustrative Examples in Which Single and Multi-Stream Analyses Detect the Same Cluster with Different Strength

Did single or multi-stream analysis detect cluster with higher recurrence interval?	*Signal date (2009)*	*Recurrence interval from single stream analysis, not adjusted for the five data streams evaluated*		*Recurrence interval from single stream analysis, adjusted for the five data streams evaluated*		*Recurrence interval from multi-stream analysis, adjusted for the 31 combinations of the five data streams evaluated*
		LGI in AC	*LGI in ED*	*LGI in AC*	*LGI in ED*	
Multi-stream	7-May	33	118	7	24	435
	15-Jul	2,500	189	500	38	10,000
	26-Oct	227	–	45	–	417
Single stream	30-Mar	–	1,111	–	222	84
	10-Jul	5,000	–	1,000	–	313
	21-Jul	–	10,000	–	2,000	2,000

AC, ambulatory care; ED, emergency department; LGI, lower gastrointestinal illness.

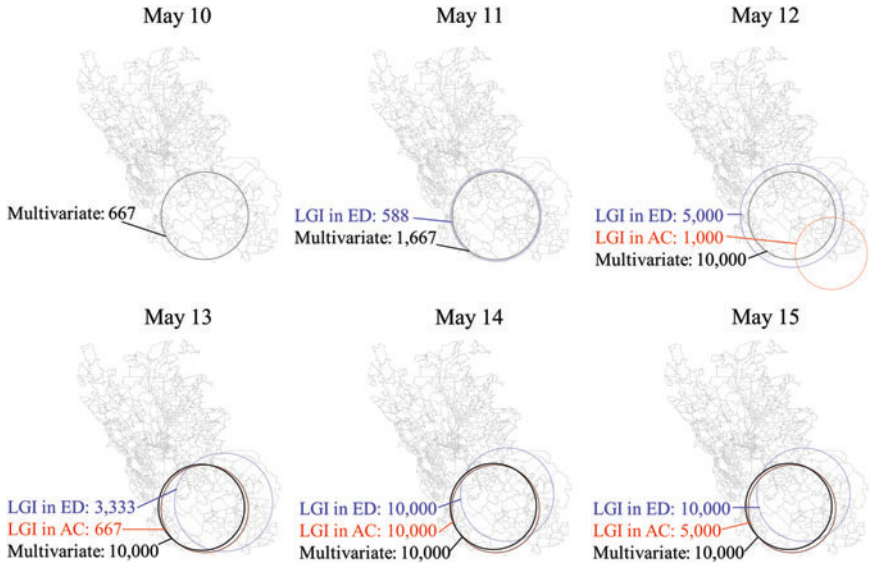

FIG. 2. Illustrative example where multi-stream analysis had more timely detection of a potential outbreak than single stream analysis. On May 10, multivariate analysis first detected a cluster, with a recurrence interval of 667. The next day on May 11, the single stream analysis of lower gastrointestinal illness (LGI) in the emergency department (ED) first detected the same cluster. On May 12, the single stream analysis of lower gastrointestinal illness (LGI) in ambulatory care (AC) detected a cluster, but it was geographically offset. On May 13, the single stream analysis of LGI in AC detected the same cluster that the multivariate analysis first detected 3 days earlier. Color images available online at www.liebertonline.com/fpd

Acknowledgments

We thank James M. LePan (Kaiser Permanente Northern California Regional Laboratory) for assistance in determining serotype relatedness of isolates in clusters; Shu C. Sebesta (California Department of Public Health) for assembling the final foodborne disease outbreak reports; and W. Katherine Yih, Ph.D., M.P.H., and Julie D. Lankiewicz, M.P.H. (Harvard Pilgrim Health Care Institute) and John Hsu, M.D., M.B.A., and Richard Brand, Ph.D. (Kaiser Permanente Medical Care Program) for helpful discussions. Research support was provided by the National Institute of General Medical Sciences (cooperative agreement U01GM076672 under the Models of Infectious Disease Agent Study program). Preliminary results were presented in oral presentations at the International Society for Disease Surveillance Ninth Annual Conference (Park City, UT, December 2, 2010) and at the American Public Health Association 139th Annual Meeting (abstract 236721, Washington, DC, October 31, 2011).

Disclosure Statement

No competing financial interests exist.

References

Backer HD, Bissell SR, Vugia DJ. Disease reporting from an automated laboratory-based reporting system to a state health department via local county health departments. Public Health Rep 2001;116:257–265.

Balter S, Weiss D, Hanson H, Reddy V, Das D, Heffernan R. Three years of emergency department gastrointestinal syndromic surveillance in New York City: What have we found? MMWR Morb Mortal Wkly Rep 2005;54(Suppl):175–180.

Burkom HS, Murphy S, Coberly J, Hurt-Mullen K. Public health monitoring tools for multiple data streams. MMWR Morb Mortal Wkly Rep 2005;54(Suppl):55–62.

California Code of Regulations. Reportable infectious diseases: Reporting by laboratories (California Code of Regulations. Title 17, Section 2505). 2009. Available at: http://www.cdph.ca.gov/programs/Documents/Title%2017%20Section%202502%2010-28-11.pdf, accessed February 1, 2012.

Caudle JM, van Dijk A, Rolland E, Moore KM. Telehealth Ontario detection of gastrointestinal illness outbreaks. Can J Public Health 2009;100:253–257.

[CDC] Centers for Disease Control and Prevention. Syndrome definitions for diseases associated with critical bioterrorism-associated agents. Atlanta: U.S. Department of Health and Human Services, CDC. 2003. Available at: www.bt.cdc.gov/surveillance/syndromedef/, accessed February 1, 2012.

[CDC] Centers for Disease Control and Prevention. *Salmonella* surveillance: Annual summary, 2006. Atlanta: U.S. Department of Health and Human Services, CDC. 2008. Available at: http://www.cdc.gov/ncidod/dbmd/phlisdata/salmonella.htm, accessed February 1, 2012.

[CDC] Centers for Disease Control and Prevention. Multistate outbreak of *E. coli* O157:H7 infections linked to eating raw refrigerated, prepackaged cookie dough. Atlanta: U.S. Department of Health and Human Services, CDC. 2009a. Available at: http://www.cdc.gov/ecoli/2009/0807.html, accessed February 1, 2012.

[CDC] Centers for Disease Control and Prevention. Nationally notifiable infectious diseases, United States, 2009. Atlanta: U.S. Department of Health and Human Services, CDC. 2009b. Available at: http://www.cdc.gov/osels/ph_surveillance/nndss/phs/infdis2009.htm, accessed February 1, 2012.

[CDC] Centers for Disease Control and Prevention. Preliminary FoodNet data on the incidence of infection with pathogens transmitted commonly through food—10 states, 2009. MMWR Morb Mortal Wkly Rep 2010a;59:418–422.

[CDC] Centers for Disease Control and Prevention. Surveillance for foodborne disease outbreaks—United States, 2007. MMWR Morb Mortal Wkly Rep 2010b;59:973–979.

Chapman WW. Syndromic surveillance from chief complaints: Consensus syndrome definitions. 2011. Available at: http://isds.wikispaces.com/Syndromic+Surveillance+from+Chief+Complaints, accessed February 1, 2012.

Chapman WW, Dowling JN, Baer A, Buckeridge DL, Cochrane D, Conway MA, Elkin P, Espino J, Gunn JE, Hales CM, *et al.* Developing syndrome definitions based on consensus and current use. J Am Med Inform Assoc 2010;17:595–601.

Dwass M. Modified randomization tests for nonparametric hypotheses. Ann Math Stat 1957;28:181–187.

[FDA] Food and Drug Administration. FDA alerts the public to Uncle Chen and Lian How brand dry spice product recall. Silver Spring, MD: U.S. Department of Health and Human Services, FDA. 2009. Available at: http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2009/ucm149555.htm, accessed February 1, 2012.

Gerner-Smidt P, Hise K, Kincaid J, Hunter S, Rolando S, Hyytia-Trees E, Ribot EM, Swaminathan B. PulseNet USA: A five-year update. Foodborne Pathog Dis 2006;3:9–19.

Greene SK, Kulldorff M, Huang J, Brand RJ, Kleinman KP, Hsu J, Platt R. Timely detection of localized excess influenza activity in Northern California across patient care, prescription, and laboratory data. Stat Med 2011;30:549–559.

Jones RC, Liberatore M, Fernandez JR, Gerber SI. Use of a prospective space-time scan statistic to prioritize shigellosis case investigations in an urban jurisdiction. Public Health Rep 2006;121:133–139.

Jones TF, McMillian MB, Scallan E, Frenzen PD, Cronquist AB, Thomas S, Angulo FJ. A population-based estimate of the substantial burden of diarrhoeal disease in the United States; FoodNet, 1996–2003. Epidemiol Infect 2007;135:293–301.

Jung I, Kulldorff M, Kleinman KP, Yih WK, Platt R. Using encounters versus episodes in syndromic surveillance. J Public Health (Oxf) 2009;31:566–572.

Kirian ML, Weintraub JM. Prediction of gastrointestinal disease with over-the-counter diarrheal remedy sales records in the San Francisco Bay Area. BMC Med Inform Decis Mak 2010;10:39.

Kleinman K, Lazarus R, Platt R. A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. Am J Epidemiol 2004;159:217–224.

Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. PLoS Med 2005;2:e59.

Kulldorff M, Mostashari F, Duczmal L, Yih WK, Kleinman K, Platt R. Multivariate scan statistics for disease surveillance. Stat Med 2007;26:1824–1833.

Levin JE, Raman S. Early detection of rotavirus gastrointestinal illness outbreaks by multiple data sources and detection algorithms at a pediatric health system. AMIA Annu Symp Proc 2005;2005:445–449.

Miller ND, Draughon FA, D'Souza DH. Real-time reverse-transcriptase–polymerase chain reaction for *Salmonella enterica* detection from jalapeno and serrano peppers. Foodborne Pathog Dis 2010;7:367–373.

Nielsen CF, Langer A, Pringle J, Heffernan R, Klos R, Monson T, Rauch M, Ball J, Hoekstra M, Archer J, *et al.* First documented multistate outbreak of *Salmonella* Carrau infections—United States, 2009. Presented at the 59th Annual Epidemic Intelligence Service Conference, Atlanta, 2010.

Nielsen EM, Scheutz F, Torpdahl M. Continuous surveillance of Shiga toxin-producing *Escherichia coli* infections by pulsed-field gel electrophoresis shows that most infections are sporadic. Foodborne Pathog Dis 2006;3:81–87.

Pearl DL, Louie M, Chui L, Dore K, Grimsrud KM, Leedell D, Martin SW, Michel P, Svenson LW, McEwen SA. The use of outbreak information in the interpretation of clustering of reported cases of *Escherichia coli* O157 in space and time in Alberta, Canada, 2000–2002. Epidemiol Infect 2006;134:699–711.

Peck AJ, Bresee JS. Viral gastroenteritis. In: *Oski's Pediatrics: Principles and Practice.* McMillan JA, Feigin RD, DeAngelis C, Jones MD (eds.) Philadelphia: Lippincott Williams & Wilkins, 2006, pp. 1288–1293.

Pelat C, Boelle PY, Turbelin C, Lambert B, Valleron AJ. A method for selecting and monitoring medication sales for surveillance of gastroenteritis. Pharmacoepidemiol Drug Saf 2010;19:1009–1018.

Rolka H, Burkom H, Cooper GF, Kulldorff M, Madigan D, Wong WK. Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: Research needs. Stat Med 2007;26:1834–1856.

Scallan E, Jones TF, Cronquist A, Thomas S, Frenzen P, Hoefer D, Medus C, Angulo FJ. Factors associated with seeking medical care and submitting a stool sample in estimating the

burden of foodborne illness. Foodborne Pathog Dis 2006;3:432–438.

Stelling J, Yih WK, Galas M, Kulldorff M, Pichel M, Terragno R, Tuduri E, Espetxe S, Binsztein N, O'Brien TF, *et al.* Automated use of WHONET and SaTScan to detect outbreaks of *Shigella* spp. using antimicrobial resistance phenotypes. Epidemiol Infect 2010;138:873–883.

Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV. PulseNet: The molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis 2001;7:382–389.

Yih WK, Abrams A, Danila R, Green K, Kleinman K, Kulldorff M, Miller B, Nordin J, Platt R. Ambulatory-care diagnoses as potential indicators of outbreaks of gastrointestinal illness—Minnesota. MMWR Morb Mortal Wkly Rep 2005;54(Suppl): 157–162.

Yih WK, Deshpande S, Fuller C, Heisey-Grove D, Hsu J, Kruskal BA, Kulldorff M, Leach M, Nordin J, Patton-Levine J, *et al.* Evaluating real-time syndromic surveillance signals from ambulatory care data in four states. Public Health Rep 2010;125:111–120.

Address correspondence to:
Sharon K. Greene, Ph.D., M.P.H.
Department of Population Medicine
Harvard Medical School
and Harvard Pilgrim Health Care Institute
133 Brookline Avenue, 6th Floor
Boston, MA 02215-3920

E-mail: Sharon_Greene@harvardpilgrim.org