Preventive Medicine 67 (2014) 264–269

Contents lists available at ScienceDirect

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports

Elaine O. Nsoesie [a,b,*], Sheryl A. Kluberg [a], John S. Brownstein [a,b,c]

[a] Children's Hospital Informatics Program, Boston Children's Hospital, Boston, MA, 02215, USA
[b] Department of Pediatrics, Harvard Medical School, Boston, MA 02215, USA
[c] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec H3A 1A2, Canada

ARTICLE INFO

Available online 11 August 2014

Keywords:
Foodborne illness
Foodborne diseases
Disease surveillance
Social media
Gastroenteritis
Population surveillance

ABSTRACT

Objective. Traditional surveillance systems capture only a fraction of the estimated 48 million yearly cases of foodborne illness in the United States. We assessed whether foodservice reviews on Yelp.com (a business review site) can be used to support foodborne illness surveillance efforts.

Methods. We obtained reviews from 2005 to 2012 of 5824 foodservice businesses closest to 29 colleges. After extracting recent reviews describing episodes of foodborne illness, we compared implicated foods to foods in outbreak reports from the U.S. Centers for Disease Control and Prevention (CDC).

Results. Broadly, the distribution of implicated foods across five categories was as follows: aquatic (16% Yelp, 12% CDC), dairy–eggs (23% Yelp, 23% CDC), fruits–nuts (7% Yelp, 7% CDC), meat–poultry (32% Yelp, 33% CDC), and vegetables (22% Yelp, 25% CDC). The distribution of foods across 19 more specific food categories was also similar, with Spearman correlations ranging from 0.60 to 0.85 for 2006–2011. The most implicated food categories in both Yelp and CDC were beef, dairy, grains—beans, poultry and vine-stalk.

Conclusions. Based on observations in this study and the increased usage of social media, we posit that online illness reports could complement traditional surveillance systems by providing near real-time information on foodborne illnesses, implicated foods and locations.

Introduction

An estimated 48 million people experience foodborne illness in the United States each year (CDC Estimates of Foodborne Illness in the United States). Most foodborne illnesses are associated with acute gastroenteritis (defined as diarrhea and vomiting) (Lucado et al., 2013), but affected individuals can also experience abdominal cramps, fever and bloody stool (Daniels et al., 2002; McCabe-Sellers and Beattie, 2004). Although there are several surveillance systems for foodborne illnesses at the local, state and territorial levels, these systems capture only a fraction of the foodborne illness burden in the United States mainly due to few affected individuals seeking medical care and lack of reporting to appropriate authorities (McCabe-Sellers and Beattie, 2004). One way to improve surveillance of foodborne illnesses is to utilize nontraditional approaches to disease surveillance (Brownstein et al., 2009).

Nontraditional approaches have been proposed to supplement traditional systems for monitoring infectious diseases such as influenza (Aramaki et al., 2011; Yuan et al., 2013) and dengue (Chan et al., 2011). Examples of nontraditional data sources for disease surveillance include social media, online reports and micro-blogs (such as Twitter) (Aramaki et al., 2011; Chan et al., 2011; Madoff, 2004; Yuan et al., 2013). These approaches have been recently examined for monitoring reports of food poisoning and disease outbreaks (Brownstein et al., 2009; Wilson and Brownstein, 2009). However, only one recent study by New York City Department of Health and Mental Hygiene in collaboration with researchers at Columbia University (Harrison et al., 2014) has examined foodservice review sites as a potential tool for monitoring foodborne disease outbreaks.

Online reviews of foodservice businesses offer a unique resource for disease surveillance. Similar to notification or complaint systems, reports of foodborne illness on review sites could serve as early indicators of foodborne disease outbreaks and spur investigation by proper authorities. If successful, information gleaned from such novel data streams could aid traditional surveillance systems in near real-time monitoring of foodborne related illnesses.

The aim of this study is to assess whether crowdsourcing via foodservice reviews can be used as a surveillance tool with the potential to support efforts by local public health departments. Our first aim is to summarize key features of the review dataset from Yelp.com. We study reviewer–restaurant networks to identify and eliminate reviewers whose extensive reviewing might have a strong impact on the data.

* Corresponding author at: Boston Children's Hospital, 1 Autumn St. #14, Boston, MA 02215, USA.
E-mail address: onelaine@vt.edu (E.O. Nsoesie).

http://dx.doi.org/10.1016/j.ypmed.2014.08.003

Furthermore, we identify and further investigate report clusters (greater than two reports in the same year). Our second aim is to compare foods implicated in outbreaks reported to the U.S. Centers for Disease Control and Prevention (CDC) Foodborne Outbreak Online Database (FOOD) to those reported on Yelp.com. Attribution of foodborne illness and disease to specific food vehicles and locations is important for the monitoring and estimation of the extent of foodborne illness, which is necessary for public policy and regulatory decisions (Kuchenmuller et al., 2009; Nyachuba, 2010; Scallan et al., 2013; Woteki and Kineman, 2003).

Methods

Data sources

Yelp

Yelp.com is a business review site created in 2004. Data from Yelp has been used to evaluate the correlation between traditional hospital performance measures and commercial website ratings (Bardach et al., 2013), and the value of forecasting government restaurant inspection results based on the volume and sentiment of online reviews (Kang et al., 2013). We obtained data from Yelp containing de-identified reviews from 2005 to 2012 of 13,262 businesses closest to 29 colleges in fifteen states (Table A.1). 5824 (43.9%) of the businesses were categorized as Food or Restaurant businesses.

CDC

We also obtained data from CDC's Foodborne Outbreak Online Database (FOOD) (CDC Foodborne Outbreak Online Database) to use as a comparator. FOOD contains national outbreak data voluntarily submitted to the CDC's foodborne disease outbreak surveillance system by public health departments in all states and U.S. territories. The data comprises information on the numbers of illnesses, hospitalizations, and deaths, reported food vehicle, species and serotype of the pathogen, and whether the etiology was suspected or confirmed. Note, outbreaks not identified, reported, or investigated might be missing or incomplete in the system. For each of the fifteen states represented in the Yelp data, we extracted data from FOOD in which reported illness was observed between January 2005 and December 2012.

Analysis

Keyword matching

We constructed a keyword list based on a list of foodborne diseases from the CDC and common terms associated with foodborne illnesses (such as diarrhea, vomiting, and puking) (Table A.2). Each review of a business listed under Yelp's food or restaurant category (Table A.5) was processed to locate mentions of any of the keywords. 4088 reviews contained at least one of the selected keywords. We carefully read and selected reviews meeting the classification criteria (discussed in the next section) for further analysis.

Classification criteria

We focused on personal reports and reports of alleged eyewitness accounts of illness occurring after food consumption (see Table 1 for examples). We concentrated on recent accounts of foodborne illness and eliminated episodes in the distant past, such as childhood experiences. For each relevant review, we documented the following information, if reported: date of illness, foods consumed, business reviewed, and number of ill individuals.

Bias and cluster analysis

Data bias could be introduced by false reviews from disgruntled former employees and competitors. Yelp has a process for eliminating such reviews. We therefore focused on identifying bias introduced by individuals with a large number of negative reviews compared to the median in the dataset using network analysis and visualization. If a reviewer had significantly more reports than the median, we would investigate the impact of including and excluding this individual from the analysis. We also identified and investigated restaurants with more than two foodborne illness reports in the same year, since most restaurants appeared to have one or two reports, and because the CDC defines a foodborne disease outbreak as more than one case of a similar illness due to consumption of a common food (Daniels et al., 2002; Jones et al., 2013).

Table 1
Sample reports of alleged foodborne illness. Keywords are in bold.

business_id = rblZR9xtCUgwjE19AU2y8w user_id = −1rqMSXzoQ7iYTRipDNhPA stars = 1 date = 2009-01-10 I got HORRIBLE **food poisoning** from this place. If I could give it negative stars I would. And no, I'm not a lightweight: I eat Indian food all the time (and have even been to India) without getting **sick**. I know it was from the place because it was the only thing I had eaten that was out of the ordinary for the entire week. As it turns out, one of my coworkers had the same experience the same week from the same place, although unfortunately he only told me later and thus I was not able to avoid it. So, in summary, yikes! If you value your health, stay away!
business_id = 279Aj_4Hd7EhoAZOiip42g user_id = j0FOcXf6WQeVqlQVdAEt4w stars = 1 date = 2005-07-11 I went here on a Thursday during their free taco day … I dont know about you all but I don't find getting cold and hot flashes up my spine while **puking** at 4 in the morning very exciting. Me and my friend got f*king **food poisoning** there! I didn't feel better until 5 pm the next day. Tch … I am still mad at that … Damn taco meat must've been rat meat. I guess free food means free sh*t at their restaurant. I'll still go there for the two dollar beer special but the bartenders' attitude could be a little less b..chy, I dunno just a thought.
business_id = x52nVXRLWAwf3Rw76jcKMg user_id = MGL6GNXBjchbHx2D70MFbg stars = 1 date = 2010-01-02 Epic fail. Yesterday, I looked at the reviews and decided to post a four-star review, as I headed over to Zorba's to meet a few out-of-town friends. “Why such a bad rap?” I thought — and figured I'd help boost the reviews of this place that I'd been to twice before, and enjoyed. Well, I went there yesterday for lunch. Today, I woke up **deathly ill**, and proceeded to kick off 2010 by **vomiting**. Nice. I'm still **sick** but my family is taking care of me. The three of us had different items – not sure what took us all down – but we suspect Zorba's as we all went our separate ways and are all **deathly ill** today. Now I will add that I'm sure they run a good business and are decent people. But **food poisoning** is the one thing that cannot happen when you run a restaurant. I will touch base with them to see if they will do anything for us.

Comparison of food vehicles

We extracted food vehicles mentioned in the FOOD outbreak reports and the Yelp data according to the CDC convention of categorizing and grouping implicated foods (Painter et al., 2009, 2013). Broadly, the taxonomy consisted of three major categories: aquatic animals, land animals and plants. These categories were hierarchically distributed into subcategories as shown in Fig. 2. Initially, we grouped the data into five major categories: aquatic, dairy–eggs, fruits–nuts, meat–poultry, and vegetables. Based on observations from this grouping, we further analyzed nineteen more specific categories, capturing all the major food groups. The nineteen categories consisted of fish, crustaceans, mollusks, dairy, eggs, beef, game, pork, poultry, grains–beans, fruits–nuts, fungi, leafy, root, sprout, vine-stalk, shellfish, vegetables, and meat. The aquatic, shellfish, vegetables and meat categories consisted of all foods that belonged to these categories but could not be assigned to the more specific categories such as leafy, crustaceans, poultry, etc. We excluded the oils–sugars category since most meals include natural or processed oils and/or sugars.

Foods implicated in foodborne illness were either categorized as *simple* or *complex*. Simple foods consisted of a single ingredient (e.g., lettuce) or could be classified into a single category (e.g., fruit salad). Complex foods consisted of multiple ingredients that could be classified into more than one commodity (e.g., pizza). For example, if pizza were implicated in an alleged foodborne illness report, we documented three food categories: grains–beans (crust), vine-stalk (tomato sauce), and dairy (cheese). If a report included a food item not easily identifiable (such as a traditional dish), we used Google search engine to locate the main ingredients in a typical recipe (e.g., meat, vegetable, aquatic, etc.) and categorized the food accordingly.

To compare foods implicated by Yelp and the CDC, we focused on reports from 2006 to 2011, because the 2012 Yelp data were incomplete. We ranked the nineteen food categories separately for Yelp and FOOD, according to the frequency with which each food category was implicated per year. Food categories with the same frequency were assigned the average of their rankings. Correlations of the ranked food categories were assessed using Spearman's rank correlation coefficient, ρ. Analyses were performed in SAS 9.1.3 (SAS Institute, Inc., Cary, NC).