Food Control 59 (2016) 53–58

Contents lists available at ScienceDirect

Food Control

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

A modeling framework to accelerate food-borne outbreak investigations

Kun Hu*, Sondra Renly, Stefan Edlund, Matthew Davis, James Kaufman

IBM Research, Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA

ARTICLE INFO

Article history:
Received 18 April 2014
Received in revised form
4 May 2015
Accepted 12 May 2015
Available online 14 May 2015

Keywords:
Food safety
Food-borne infections
Geospatial data
Geospatial modeling
Likelihood-based method
Epidemiology
Outbreaks
Public health informatics

ABSTRACT

Food safety procedures are critical to reducing pathogen caused food-borne disease (FBD). However there is no way to completely eliminate the risk of consuming contaminated products. When prevention efforts fail, rapid identification of the contaminated product is essential. The medical and economic losses incurred grow with the duration of the outbreak. In this paper we show that before an outbreak occurs, analysis of food sales data, as a proactive intervention, can provide useful product intelligence that we can exploit during an outbreak investigation to accelerate the identification process. Using real grocery retail sales data from Germany, we have implemented a likelihood-based approach to study how such data can be used to accelerate the investigation during the early stages of an outbreak.

1. Introduction

Food-borne disease (FBD) is a global public health problem that affects millions of people every year and is caused by contamination by a variety of pathogens including bacteria, viruses, and parasites. The European Centre for Disease Prevention and Control (ECDC) gathers and reports incidence data on common pathogens that cause food-borne illness across Europe including *Norovirus*, *Campylobacter*, *Salmonella*, *Shigella*, *Listeria*, *Escherichia coli* (VTEC), and Hepatitis A (European Centre for Disease Prevention and Control, 2012). Healthcare clinicians report suspect and confirmed cases of food-borne disease to public health authorities. An outbreak is when two or more confirmed case reports are linked to the same pathogen after digesting a common food or ingredient (Rocourt, Moy, & Schlundt, 2003). Public health officials investigate outbreaks to try to identify the common food or ingredient as quickly as possible to remove the contaminated product from sale and restore consumer trust in the safety of the food supply (Marvin et al., 2009).

There have been many outbreaks where it has been difficult to identify the contaminated product using current best practices (Januszkiewicz et al., 2012; Reingold, 1998; Scavia et al., 2013). Best practices include very thorough questionnaires about food consumption for both individuals who got ill and those that did not. Public health officials can search the home for discarded containers and stored foods. Food histories can be compared to annual product consumption surveys to identify a higher correlation of particular product consumption in the ill population than in the general population (Centers for Disease Control and Prevention (CDC), 2014; Jones & Schaffner, 2003). Also, public health officials can obtain permission to retrieve data on customer loyalty program or warehouse membership card for grocery purchases (Barret et al., 2013; Gieraltowski et al., 2013). Even with these best practices, public health officials face a significant challenge and longtime delays in obtaining critical information to help identify the contaminated product.

Historically, roughly 40% of outbreaks occur from consumption from sources other than food served in restaurants or institutions (e.g., schools, nursing homes, prisons) (Jones & Angulo, 2006). Much of this food is purchased from grocery retailers. These retailers survive on razor thin profit margins and as a result have significant financial incentives to invest heavily in information

* Corresponding author.
E-mail address: khu@us.ibm.com (K. Hu).

http://dx.doi.org/10.1016/j.foodcont.2015.05.017
0956-7135/© 2015 Published by Elsevier Ltd.

technology to help them efficiently manage their inventory. The grocery retailers have thus accumulated large data sets of near real-time information about food sales by stores that can provide accurate and timely information about the population's food consumption patterns. We hypothesize that before an outbreak even occurs, analysis of all retail food sales data, as a proactive intervention, can provide useful food product intelligence we can exploit during an outbreak investigation.

Kaufman et al. (2014) obtained a retail dataset that includes 580 anonymous products in two undefined food product groups from grocery retail stores in Germany. The dataset is comprised of 8176 unique retail stores distributed across 3518 postal code areas. The dataset includes the amount of each product sold per week per store during the years from 2008 to 2010 (i.e., 157 weeks in total). Leveraging this set of empirical data from Germany, we showed that before an outbreak occurs, analyzing food sales data using a proposed likelihood-based method provides useful product intelligence, and the resulting food consumption model enables one to accelerate identification of a contaminated product during the early stage of a food-borne disease outbreak. This work builds upon Kaufman et al. (2014), introducing several new and important measurements for the statistical model as well as a description of how to implement these new algorithms as components of a future system.

2. Methods

Our predictive analytics framework was created with three exchangeable components. The first component is a food distribution model that predicts where product consumption occurs. The second component is an outbreak generator. This generator uses information from the food distribution model and creates a simulated set of linked geo-coded public health case reports. The synthetic case reports capture hypothetical ill persons who consumed the contaminated product. The third component in the framework applies a statistical analysis to rank products based on the probability that each product is the cause of the outbreak. We calculate the probability leveraging the product sales distribution and location of the geo-coded public health case reports. By creating this as a flexible framework, we are able to study the individual effects of each component on performance of the system as a whole. It also enables us to compare models and methodologies, and maintain the ability to quickly load new data sets for study.

2.1. Food distribution model component

Grocery retail sales data provides us with temporal and geospatial information on food sales but not food consumption. A food distribution model is employed to show where the food is consumed in a population. Literature offers general retail shopping models such as the Huff Gravity Model (Huff, 1963), though retailers today could leverage their customer address information from loyalty or membership card programs to create a more precise spatial model.

In this work, we use a simple food distribution model $f_s(n, r)$ that assumes each product n is distributed and consumed only within a postal code region r where the product was originally purchased shown in Eq. (1). We feel this is a reasonable approach in high population density regions where people shop frequently in neighborhood markets. So if $f_c(n,r)$ is the probability that product n is consumed in region r, and $f_s(n,r)$ is the probability that product n is *sold* in region r, in our simplified model we assume:

$$f_c(n, r) = f_s(n, r) \tag{1}$$

Let *sales*(n, r) represent the number of units of food product n sold in region r over this three-year period. We can now define a function $f_s(n, r)$ representing the probability that product n is sold in region r as:

$$f_s(n, r) = \frac{sales(n, r)}{\sum\limits_{\widehat{r} \in R} sales(n, \widehat{r})} \tag{2}$$

The food sales data is aggregated and normalized across all German postal codes in set R such that the sum of each product is one (refer to Eq. (3)). This model simplification allows us to focus our research on examining differences in food sales distributions across the country and isolating unique patterns.

$$\sum_{r \in R} f_c(n, r) = 1 \tag{3}$$

2.2. Outbreak generator component

An outbreak generator creates individual geo-located public health case reports based on a contamination event that can include one or more products. The generator component can be used to create synthetic outbreaks or to re-create historical outbreaks for retrospective study.

In this study, we leverage knowledge about Germany and one contaminated product's distribution to generate public health case reports for a simulated food-borne disease outbreak. The normalized food consumption data (refer to Eq. (3)) from our food distribution model is input to our Monte Carlo outbreak simulation method. We generate synthetic outbreak case reports for a selected "contaminated" product x (*where we use x instead of n to indicated a single contaminated product*). Using A. J. Walker's alias method (Walker, 1977), we draw M random locations by sampling from $f_c(x, r)$ over all locations r in R. In separate trials, synthetic case report data are generated assuming each of the 580 products, in turn, as the source of contamination. We assume the products are independent so $f_c(x, r)$ also defines the probability of a case report at location r due to contaminated product x. It is true that two "products" with different local "brands" or "ids" could in fact be the same food item simply rebranded when repackaged locally. In this case, we enable noise in the generator introducing the ability to relocate small amounts of product consumption events outside the original sales postal code, which does not require the changes to the food distribution model. Conversely, a product sold on a national scale under one single brand could become contaminated at a single point of retail site (e.g., a butcher shop at a grocery store). For the purposes of this study, the simulated case reports were generated self consistently from the retail data using the assumption that the data provided to us by product id were independent. Depending upon the spatial distribution of product x, it is likely that, during one simulated outbreak of 100 cases, multiple case reports will come from a same postal code. In this work, we defined an outbreak to include between 100 and 1000 case reports and generated 100 simulated outbreaks per contaminated product for our statistical analysis.

2.3. Statistical analytical component

The statistical analytical component creates an ordered ranking of all the known products from the most likely contaminated product to the least for each synthetic outbreak generated. This component is designed to be completely independent of the food distribution model in order to create a plug and play environment that best supports fine-tuning to component-specific requirements.