WATER RESEARCH 53 (2014) 26–34

Available online at www.sciencedirect.com

ScienceDirect

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

Public health and pipe breaks in water distribution systems: Analysis with internet search volume as a proxy

*Julie E. Shortridge**, *Seth D. Guikema*

Department of Geography & Environmental Engineering, Johns Hopkins University, USA

ARTICLE INFO

Article history:
Received 19 September 2013
Received in revised form
1 December 2013
Accepted 9 January 2014
Available online 21 January 2014

Keywords:
Distribution network
Pipe breaks
Gastrointestinal illness
Non-linear regression

ABSTRACT

Drinking water distribution infrastructure has been identified as a factor in waterborne disease outbreaks and improved understanding of the public health risks associated with distribution system failures has been identified as a priority area for research. Pipe breaks may pose a risk, as their occurrence and repair can result in low or negative pressure, potentially allowing contamination of drinking water from adjacent soils. However, measuring this phenomenon is challenging because the most likely health impact is mild gastrointestinal (GI) illness, which is unlikely to result in a doctor or hospital visit. Here we present a novel method that uses data mining techniques and internet search volume to assess the relationship between pipe breaks and symptoms of GI illness in two U.S. cities. Weekly search volume for the terms diarrhea and vomiting was used as the response variable with the number of pipe breaks in each city as a covariate as well as additional covariates to control for seasonal patterns, search volume persistence, and other sources of GI illness. The fit and predictive accuracy of multiple regression and data mining techniques were compared, with the best performance obtained using random forest and bagged regression tree models. Pipe breaks were found to be an important and positively correlated predictor of internet search volume in multiple models in both cities, supporting previous investigations that indicated an increased risk of GI illness from distribution system disturbances.

1. Introduction

While drinking water in developed countries is consistently treated to be compliant with health guidelines, the aging condition of drinking water distribution system infrastructure presents a risk of contaminant intrusion and negative impacts on public health. Breaks and leaks in distribution pipelines can allow pathogens present in surrounding soil or water to enter the distribution system during low or negative pressure events. It is estimated that anywhere from 10 to 50% of waterborne disease outbreaks associated with treated drinking water are attributable to distribution system deficiencies (CDC, 2006; CDC, 2008; CDC, 2011). It is reasonable to assume that the outbreaks analyzed in CDC (2006, 2008, 2011) represent only a small percentage of the overall disease attributable to drinking water, as they require that multiple cases of illness be reported and linked to drinking water exposure (CDC, 2011),

* *Corresponding author*. Tel.: +1 2026796535.
E-mail address: jshortr1@jhu.edu (J.E. Shortridge).

0043-1354/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.watres.2014.01.013

and few people seek medical care for mild to moderate gastrointestinal (GI) illness (Wheeler et al., 1999). Messner et al. (2006) estimate that community water systems are responsible for 16.4 million cases of acute GI illness per year in the United States. While most of these are mild to moderate cases that do not require a doctor or hospital visit, they can still result in high societal costs; for example, in 1988 it was estimated that mild GI illness resulted in $19.5 billion in lost productivity annually (Garthright et al., 1988). Because of these issues, improved understanding of the incidence and severity of health impacts from water distribution systems has been identified as a high priority research area (USEPA and Water Research Foundation, 2010).

Despite the need for additional research on the public health impacts of distribution system deficiencies, a number of issues make detection and measurement of these impacts a challenge. Exposure to pathogens through water distribution systems requires a complex chain of events to occur. While a conceptual model to support microbial risk assessment for low-pressure events is presented by Besner et al. (2011), existing data to support such an analysis is limited and subject to numerous uncertainties and assumptions. While sustained low-pressure events caused by main breaks and maintenance activities are relatively easy to identify, short term pressure transients caused by changes in pump operation, power outages, and sudden changes in demand are unlikely to be identified without high-speed pressure monitoring that is generally not in use in existing systems (Friedman et al., 2004). Population exposure depends on the quantity of pathogens able to enter the distribution system, their transport and dilution throughout the system, and the number of users who eventually consume the contaminated water. Individuals may be exposed to contaminated water at many places other than their homes, such as offices, schools, or restaurants, confounding efforts to monitor illness at fine spatial scales. Furthermore, only a small percentage of GI illness results in a doctor or hospital visit, making health outcomes difficult to track.

Because of these challenges, much existing research on this topic has either focused on the occurrence of pressure transients and external contamination in water distribution systems or survey-based monitoring and intervention trials that aim to estimate the incidence of GI illness attributable to treated drinking water. Sampling studies have indicated that pathogenic microorganisms are frequently present in soil and water adjacent to drinking water pipes (Karim et al., 2003), while low- and negative-pressure transients have been documented in multiple systems (LeChevallier et al., 2003; Karim et al., 2003). Water sampling studies have indicated that distribution systems can allow introduction of viruses into non-disinfected systems (Lambertini et al., 2012). Evaluations of whether low-pressure events lead to measurable increases in GI illness have been mixed. Rates of self-reported GI illness were found to increase following distribution system disruptions in a study conducted in Norway (Nygård et al., 2007) and following self-reported losses in water pressure in the UK (Hunter et al., 2005). However, Malm et al. (2013) monitored calls to a health care hotline system in Sweden and found no statistically significant change in call volume related to GI illness following distribution system disruptions. While not specifically focused on low pressure events, decreased incidence of GI illness have also been observed in water systems with less pipe length per person (Nygård et al., 2004; Tinker et al., 2009) and amongst study participants who drank water that had been bottled at a treatment plant rather than untreated tap water (Payment et al., 1997), indicating that increased rates of illness could be a result of distribution system deficiencies more generally.

The intervention trials and survey-based monitoring evaluations above can provide important insights into health risks associated with the studied distribution systems. However, extrapolating these insights to distribution systems more generally is difficult due to the tremendous variability in water system characteristics. The likelihood of contamination in a given system is likely to depend heavily on factors such as water source, treatment procedures, and distribution system condition and characteristics. For example, the distribution system evaluated by Payment et al. (1991, 1997) was found to be highly susceptible to negative pressure events (LeChevallier et al., 2003). Furthermore, the chance that a contamination event results in observable illness depends on the population served by the system, as certain demographic groups, such as children, the elderly, and immunocompromised individuals, are more likely to become sick after a given exposure. Differences in system characteristics could partly explain the higher rates of illness attributed to drinking water from those studies, as well as differences in research design. While conducting a similar monitoring effort on a larger scale could provide valuable insights into how risks differ amongst different water systems, survey-based monitoring and intervention trials tend to be very resource-intensive and practical only over relatively small scales. Therefore, new methods are needed to support broader studies that can evaluate widescale risks, as well as relative risks in different types of systems.

The use of internet search query data has the potential to prove useful in this regard. It is estimated that 37–52% of Americans search for health information on the internet (Brownstein et al., 2009). Internet search volume has already been shown to be strongly correlated with traditional disease monitoring data in a number of cases. Search volume for influenza-related search terms is capable of providing early detection of influenza epidemics (Ginsberg et al., 2008; Polgreen et al., 2008). This ability has also been demonstrated for a number of GI illnesses, with strong correlations between search volume and confirmed infections of rotavirus (Desai et al., 2012), salmonella (Brownstein et al., 2009), and gastroenteritis (Pelat et al., 2009). These results show that internet search volume has the potential to be an easily and rapidly accessible source of information regarding disease incidence over large areas and long-time scales where traditional monitoring may be infeasible. Surveillance data can also easily be collected through time to support longitudinal evaluations, and thus avoid the difficulties associated with cross-sectional comparisons between or within water service areas.

The objective of this paper is to assess whether a statistical relationship exists between pipe breaks in municipal drinking water distribution systems and GI illness at the metropolitan scale as estimated by internet search volume. We use a novel