Published in final edited form as:
Prev Med. 2017 August ; 101: 18–22. doi:10.1016/j.ypmed.2017.05.009.

Disparities in digital reporting of illness: A demographic and socioeconomic assessment

Samuel Henly[a], **Gaurav Tuli**[b], **Sheryl A. Kluberg**[b], **Jared B. Hawkins**[b,c], **Quynh C. Nguyen**[d], **Aranka Anema**[e], **Adyasha Maharana**[f], **John S. Brownstein**[b,c], and **Elaine O. Nsoesie**[g,*]

[a]Department of Economics, University of Washington, Seattle, WA, United States

[b]Computational Epidemiology Group, Boston Children's Hospital, Boston, MA, United States

[c]Department of Pediatrics, Harvard Medical School, Boston, MA, United States

[d]Department of Health, Kinesiology, and Recreation, University of Utah College of Health, Salt Lake City, UT, United States

[e]Epidemico, Booz Allen Hamilton, Boston, MA, United States

[f]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States

[g]Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, United States

Abstract

Although digital reports of disease are currently used by public health officials for disease surveillance and decision making, little is known about environmental factors and compositional characteristics that may influence reporting patterns. The objective of this study is to quantify the association between climate, demographic and socio-economic factors on digital reporting of disease at the US county level. We reference approximately 1.5 million foodservice business reviews between 2004 and 2014, and use census data, machine learning methods and regression models to assess whether digital reporting of disease is associated with climate, socio-economic and demographic factors. The results show that reviews of foodservice businesses and digital reports of foodborne illness follow a clear seasonal pattern with higher reporting observed in the summer, when most foodborne outbreaks are reported and to a lesser extent in the winter months. Additionally, factors typically associated with affluence (such as, higher median income and fraction of the population with a bachelor's degrees) were positively correlated with foodborne illness reports. However, restaurants per capita and education were the most significant predictors of illness reporting at the US county level. These results suggest that well-known health disparities might also be reflected in the online environment. Although this is an observational study, it is an important step in understanding disparities in the online public health environment.

[*]Corresponding author at: Institute for Health Metrics and Evaluation, University of Washington, 2301 5th Avenue, Suite 600, Seattle, WA 98121, United States. onelaine@vt.edu (E.O. Nsoesie).

Financial disclosure: No financial disclosures were reported by the authors of this paper.

Appendix A. Supplementary data: Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ypmed.2017.05.009.

Keywords

Digital disease surveillance; Socioeconomic disparities; Foodborne illness surveillance; Foodborne diseases

1. Introduction

Food safety is "the assurance that food will not cause harm to the consumer when it is prepared and consumed according to its intended use" (Codex Alimentarius Commission, 1997). While the number of people affected by poor food safety is unknown, the World Health Organization (WHO) estimated that in 2010, 600 million people were affected by and 420,000 died from foodborne illnesses related to 31 hazards (World Health Organization (WHO), 2015). The United States (U.S.) Centers for Disease Control and Prevention (CDC) attributes 48 million illnesses, 128 thousand hospitalizations and three thousand deaths annually to food-based pathogens and unspecified agents (Scallan et al., 2011a).

Whole genome sequencing and other novel technologies that enable timely detection, investigation and monitoring during outbreaks have the potential to help decrease foodborne illness and deaths (Whole Genome Sequencing (WGS) Program, 2016). Over the past ten years, the design and application of automated, informatics-based disease technologies have bridged gaps in our ability to perform global surveillance of foodborne diseases. In the United States, local departments of health have shown that crowdsourced reports of suspected foodborne illness on social media and business review sites can aid in targeted restaurant inspections and outbreak investigations (Harris et al., 2014; Harrison et al., 2014). Digital monitoring of online news sources is now an integral strategy used by the U.S. Food and Drug Administration (FDA) to enhance early warning signal detection of food contamination in global food supply chains (Bao et al., 2015). Additionally, foods implicated in foodborne illness reports submitted on the business review site, Yelp.com, were shown to correlate with foods implicated in outbreak reports from the US Centers for Disease Control and Prevention FOOD program (Nsoesie et al., 2014).

However, little is known about environmental factors and compositional characteristics that may influence reporting patterns. Such data is important for drawing conclusions about the representativeness of crowdsourced reports. Communities of affluence may be more aggressive about submitting incident reports, which may reflect differential norms around political advocacy and expectations of government services. Also, contextual factors may influence the level of community engagement, reciprocal exchange and social interactions among residents (Jensen et al., 2010), which may boost voluntary reporting of foodborne illness for the benefit of other community members. Areas may also differ with regard to computer literacy and access, food culture, and dining out patterns which can also influence reporting patterns (Jensen et al., 2010; Knight et al., 2009).

1.1. Study aims and hypotheses

We hypothesize that environmental variables and socio-economic factors, such as income and education are significant predictors of the online reporting of foodborne illness at the US county level. To explore this hypothesis, we apply statistical modeling and machine learning techniques to characterize trends in foodservice business reviews, explore associations with demographic and economic variables, and discuss potential biases introduced by reporting disparities based on data from three states.

2. Results

The available data consisted of an estimated 1.5 million reviews submitted for food service businesses in Oregon, Georgia, and Massachusetts between 2004 and 2014 on Yelp.com. Of the reviews submitted, 21,143 (1.41%) and 3900 reviews (0.25%), contained the relevant set of foodborne disease related terms (e.g. diarrhea, vomiting, puking, stomach ache) and were categorized as describing foodborne illness, respectively.

2.1. Seasonality of reviews

A fixed-effects panel regression model indicated that the volume of foodservice business reviews peaked in July and August for all states and also in January, February, and March (see Table S1). July to August peaks were most marked in Oregon and Massachusetts. In addition, review submission was highest on Sunday and Monday, and lowest on Thursday through Saturday.

Similarly, suspected foodborne illness reports peaked in late winter and summer months across all states, with the exception of an October peak in 2013 for Georgia. Regression models were poorly fit to these reviews due to data sparseness (see Table S2).

2.2. Socio-demographic analysis

2.2.1. Demographic variables—With the exception of the race variable, percent black population in the county, all considered variables were positively correlated with the volume of suspected foodborne illness reports per capita. However, these correlations were not statistically significant. The highest correlation was found for logged population size at the county level, r=0.32 (Fig. 1).

2.3. Economic and industry variables

We noted significant positive correlations between reports of suspected foodborne illness on Yelp.com with county-level economic variables associated with affluence, such as higher median income and fraction of the population with a bachelor's degrees. Symmetrically, variables indicating poverty or lower socio-economic status (such as, the fraction of households receiving food stamps) were negatively correlated with reports of foodborne illness.

Additionally, a high concentration of food production/consumption establishments (e.g., hotels, schools, hospitals, amusements, and grocery wholesales) in each county was positively associated with the reporting of foodborne illness on Yelp.com. The strongest associations were observed between the volume of suspected foodborne illness reviews per capita and restaurant establishments per capita and grocery retail establishments per capita (Fig. 1).

2.3.1. County health rankings—Although not statistically significant, negative associations were observed between suspected foodborne illness reports on Yelp.com and limited food access (*r* = −0.18) and food insecurity (*r* = −0.16); factors typically associated with lower socio-economic status. In contrast, water violations (*r* = 0.237) and food index (0.19) were positively correlated with reporting of foodborne illness on Yelp.com.

2.3.2. Regression analysis – all reviews—The following compositional characteristics: restaurants per capita, schools per capita, percent population with bachelor's degrees and high school degrees were selected by a regularization procedure aimed at selecting the most predictive variables from the 25 variables considered (see, SI Table 3 for variables considered and SI Table 4 for pre-regularization regression model). The most significant predictors of review volume at the county level were restaurants per capita, population size, and percent population with a high school degree (Table 1).

2.3.3. Regression analysis – sick-labeled reviews—The regularization procedure selected restaurants per capita, percent of population with a bachelor's degree, and entertainment/amusement venues per capita (e.g., casinos) as the most significant predictors. Restaurants and entertainment venues were highly correlated (*r* = 0.70), so we excluded entertainment venues from the final model (Table 2, pre-regularization model presented in Table S4). Both percent population with a bachelor's degree and restaurants per capita were significant predictors of reporting of foodborne illness at the county level.

3. Discussion

We hypothesized that seasonality and socio-economic factors are significant predictors of the availability of online reports of suspected foodborne illness at the US county level. Our results indicated strong seasonality patterns at the state level, and a significant influence of factors such as education, and variables associated with affluence on county-level observations of overall reviews and foodborne illness reviews. These disparities align with demographic data on Yelp users, which indicate that most users are college educated (57.5%) and have an income of over $100K (46.9%) (Yelp, 2016).

We also note negative associations between low socio-economic indicators (e.g., households receiving SNAP, uninsured, etc.) and occurrence of foodborne illness reports at the county-level. This finding is inconsistent with secondary analyses of the U.S. CDC's Foodborne Diseases Active Surveillance Network (FoodNet) and with population-based cohort studies that have found population of low socio-economic status and racial/ethnic minorities have higher incidence of foodborne illnesses (Quinlan, 2013). Differences in reporting levels may be explained by differential access to the internet, health literacy, computer literacy, computer assistance, and time availability to participate (Jensen et al., 2010). Online reporting of illness is arguably a new form of community engagement and social interaction. Our finding that lower socioeconomic status predicted lower Yelp illness reports, align with studies indicating that community affluence was associated with higher levels of reciprocal exchange and social control among community residents (Sampson et al., 1999). Additionally, socioeconomic status has been found to be a significant predictor of how people access and use the Internet (Silver, 2014; Hargittai, 2010). These disparities in reporting are important for identifying bias and quantifying population representation; factors that could impact the robustness of non-traditional disease surveillance systems (Althouse et al., 2015). While in some cases aggregation at higher geographical levels can capture overall trends in illness, systems that rely on event reports might be missing vulnerable and poor populations (Scarpino et al., 2016; Nsoesie et al., 2016). Further research is needed to evaluate the potential association between digital reports of illness and socio-economic status, drawing on individual-level data.

Additionally, we found a positive association between food retail outlets and reports of foodborne illness on Yelp.com. This observation may be confounded by the fact that Yelp users predominantly report on their experiences with retail businesses, but it raises the question about where on the farm-to-fork food supply chain exposure to foodborne pathogens is actually taking place. Digital surveillance tools, such as the FDA's SupplyChainMap monitors online news for signal detection of chemical, microbial and fungal food product contamination along the producer to retail supply chain (Bao et al., 2015), and may offer critical insights into emerging risks for consumer populations.

In the United States, only a small portion of cases of foodborne illness are identified through traditional surveillance methods. Report to the department of health requires a visit to a physician, request and submission of a stool sample, lab testing of that sample, lab confirmation, and finally reporting to the department of health (Arthur et al., 2009). A 2009 study estimated that in Toronto, only 0.4% of cases of foodborne illness are reported to public health authorities (Arthur et al., 2009). Another study estimated the reporting likelihood for specific diseases in the U.S. and found that the estimated number of true cases for Campylobacter, Shiga toxin-producing *Escherichia coli, Listeria,* and non-typhoidal *Salmonella* are 30, 26, 2, and 29 times higher, respectively, than the number reported to surveillance systems (Scallan et al., 2011b). Indeed, the original sampling frame for the US CDC's FoodNet, which quantifies and monitors the incidence of laboratory-confirmed cases of several pathogens, was not designed to be socio-economically representative of the true U.S. population (Hardnett et al., 2004).

Furthermore, of those who experience acute gastrointestinal illness, a 2006 study indicated that characteristics associated with seeking medical care included male sex, age <5 or ≥65, low household income (<$25,000), and health insurance (Scallan et al., 2006). A 2003 study in England found that lower education and lower SES were associated with a higher rate of physician visit for illness (Tam et al., 2003); however, these findings may not be directly applicable to settings in the United States, as the cost of doctor's visits and salary withholdings for sick leave vary between England and the U.S. Thus, there is some evidence in the literature suggesting that while individuals in low socioeconomic groups face higher risk of gastrointestinal illness, they are less likely to utilize incident event reporting systems.

3.1. Study strengths and limitations

Given that traditional surveillance methods capture very few cases of foodborne illness, newly emerging crowdsourcing platforms have the potential to increase coverage of incident events and allow for greater understanding of how foodborne illness occurs and steps to prevent further outbreaks. Nonetheless, although disease reporting online has been shown to have some advantages in the United States, it is important to understand populations captured through these systems and to account for health disparities in the US. Non-internet users are more likely to be ethnic minorities, older, less educated, and less healthy. However, among internet users, when controlling for age, use of social media did not depend on education, race/ethnicity, or healthcare access (Chou et al., 2009). Contrary to these observations, our study suggests that not all populations that use the internet, are equally active on online disease reporting systems.

Our study was limited by the lack of demographic data for Yelp users, leading us to use county-level demographic data. Approximately 78.3% of Yelp users only reviewed businesses in a single county, suggesting that they either live in that county or a neighboring county. Also, approximately 57.7% of Yelp reviewers wrote a single review, and 88.1% of Yelp reviewers wrote fewer than five reviews. This suggest our data was not dominated by super reviewers, which is important for validating reports of foodborne illness reviews. Also, our county-level perspective prevents us from looking on a more granular geographic scale, such as the city or town level. Additionally, our data was limited to three states and a single data source, thereby limiting how much we can generalize our results to the entire United States. Despite these limitations, our findings provide a first step towards understanding the demographical and socio-economic influence on digital reporting of disease.

3.2. Study significance

Given the wide popularity of Yelp with over 100 million unique users in an average month (Yelp, 2016), the online platform has the potential to further enable public health surveillance of foodborne illness outbreaks by supplementing existing clinical data sources with realtime user-driven data; including reports from individuals who may otherwise not report their foodborne illness experiences. Our study found that summer, water violations, and food establishments per capita were associated with more Yelp reviews containing reports of foodborne illness. Thus, greater public health surveillance of foodborne illness may be warranted in areas with more risk factors including higher density of food establishments, health violations, and during summer months.

Nevertheless, this study also found that social and demographic factors were also associated with foodborne illness reports on Yelp. Concerning is the finding that higher socioeconomic status communities had higher foodborne illness reports on Yelp compared to lower socioeconomic status communities–which runs counter to previous studies identifying higher risk for foodborne illness among low SES groups. Residents in more affluent areas may be more aggressive in reporting issues to authorities (Sampson et al., 2002; Kawachi, 1999). Interventions and public health campaigns can address the incongruence between risk of acute gastrointestinal illness and use of online reporting systems by raising awareness in at-risk communities about the various avenues for reporting foodborne illness, by promoting health literacy and internet literacy, and by supporting internet access for disadvantaged communities. Addressing disparities in online reporting of foodborne illness remains critical to ensuring that public health officials are aware of foodborne outbreaks in all communities, regardless of their demographic and social composition.

4. Methods

4.1. Data

The data consisted of approximately 1.5 million reviews submitted for foodservice businesses in Oregon, Georgia, and Massachusetts between 2004 and 2014 on Yelp.com. To extract reviews indicating food poisoning, we developed a list of relevant terms, including symptoms of gastrointestinal illness (e.g. diarrhea, vomiting, stomach ache), and pathogens (e.g., *E. coli, Salmonella,* Norovirus). Reviews indicating that the writer or their dining party had experienced foodborne illness as a consequence of eating at a foodservice establishment were classified as “relevant” and all others were classified as “irrelevant” by two human labelers. Using manually classified data, we trained a Support Vector Machine (Zou and Hastie, 2005) with an accuracy of 91% and precision and recall of 77% and 71% for “relevant” reviews and 94% and 95% for “irrelevant” reviews, respectively. We applied the trained Support Vector Machine classifier to the entire dataset to identify “relevant” and “irrelevant” reviews.

4.2. Seasonality of reports by state

We fit a panel regression to daily review counts by state location. State, month and day were represented using indicator variables, where Georgia, the month of December, and Sunday were reference values. The fixed-effects panel regression model with clustering by state considered both linear and quadratic time trends. We also considered month-state interactions. Explicitly stated, the model is as follows:

$$y_{s,t}=\beta_0+B_s I_s+B_{\text{dow}} I_{s,t,\text{dow}}+B_{\text{month}} I_{s,t,\text{month}}+B_{\text{month}*\text{state}} I_{s,t,\text{month}*\text{state}}+\beta_{\text{series}}(t-t_0)+\beta_{\text{series}-sq}(t-t_0)^2 \tag{1}$$

where s is the state location on date I; y is the response variable (we considered (a) the log count of all reviews, and (b) the log of suspected foodborne illness labeled reviews); β_0 and B vectors are coefficients; and I^* are vectors of indicator variables for state s on a given date t. Since only fixed-effects are included in the model, the estimated model is in deviations from state means: β_0 and $B_i I_i$ drop out, but are recovered after estimation of other coefficients. As presented in Eq. (1), the coefficients are as follows: B_s is the state-specific intercepts for Massachusetts and Oregon, B_{month} is the month effects, $B_{\text{month*state}}$ is the state-specific month effects for Massachusetts and Oregon, B_{dow} is the day-of-week effects, and $\beta_{\text{series*}}$ is the time trend coefficients on days and days squared since the start of the series. Data for years prior to 2008 were dropped due to sparsity. The remainder of the data were divided into a training (2008–2012) and validation (2013–2014) set.

As stated, we considered two response variables–log of all reviews, and log of sick-labeled reviews–to explore the seasonality of overall reviews and those of suspected foodborne illness. We took the log of the response variables for easier interpretability of results, and to more naturally handle variation of a dependent variable with values across several orders of magnitude.

4.3. Socio-demographic analysis

Data for the socio-demographic analyses were obtained from the 2014 American Community Survey, a product of the US Census Bureau. We extracted major demographic and economic time series data at the county level. We also obtained county-level business establishment counts for several food-related industries from the same agency's 2014 County Business Patterns series.

We identified the county where each reviewed restaurant was located. Also, to present our analysis in the context of health outcomes at the county level, we explored associations between county health rankings from the County Health Rankings and Roadmaps project (Hood et al., 2016) and the volume of illness-labeled reviews.

4.4. Analytic approach

We also aggregated all reviews and sick-labeled reviews submitted on Yelp.com from June 2013 through May 2014. Our response variables were defined as log of overall reviews and sick-labeled reviews per 1000 residents based on county population. Due to the number (n = 26) and the degree of correlation in our potential independent variables, regularization techniques were used to clarify results without sacrificing interpretability. Specifically, we implemented an elastic net regression (Zou and Hastie, 2005) to select explanatory variables before using ordinary least squares to obtain final coefficient estimates. The elastic net regression was implemented using the glmnet package in R. Coefficients were estimated using a 10-fold cross-validation process with hyper-parameters selected by minimizing mean cross-validation error. We also considered state variables in the models and found state-level effects were mostly negligible (see SI Table S5). Furthermore, since several counties had no foodborne illness reviews (n = 99), we considered a model without these counties (see SI Figs. S6, S7 and Table S6).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Conflict of interest statement: The study was funded by a grant from the National Library of Medicine of the National Institutes of Health (Award Number R01LM011965). The funders had no role in the study design; collection, analysis, and interpretation of data; writing the report; and the decision to submit the report for publication.

References

Althouse BM, Scarpino SV, Meyers LA, Ayers JW, Bargsten M, Baumbach J, et al. Enhancing disease surveillance with novel data streams: challenges and opportunities. EPJ Data Sci. 2015; 4:17. [PubMed: 27990325]

Arthur, A., Gournis, E., McKeown, D., Yaffe, B. Toronto Public Health; 2009. Toronto Public Health: Foodborne Illness in Toronto. [Internet][cited 2016 Sep 30]. Available from: http://www.toronto.ca/health/dinesafe/pdf/staffreport_april15_2009_appx_a.pdf

Bao, W., Aman, S., Johnston, D., Ning, B., Freifeld, C., Anema, A. Digital Media Surveillance for Early Warning Detection of Food Contamination in China. 3rd Annual Food and Drug Administration (FDA) Scientific Computing; Day Sept 08–09, 2015; Maryland, USA: FDA White Oak Campus in Silver Spring; 2015.

Chou SW, Hunt MY, Beckjord BE, Moser PR, Hesse WB. Social media use in the United States: implications for health communication. J Med Internet Res. 2009 Nov 27.11(4):e48. [PubMed: 19945947]

Codex Alimentarius Commission. Codex Alimentarius Food Hygiene Basic Texts. first ed. World Health Organization and Food and Agriculture Organization of the United Nations; Rome: 1997.

Hardnett FP, Hoekstra RM, Kennedy M, Charles L, Angulo FJ. For the emerging infections program foodnet working group. epidemiologic issues in study design and data analysis related to foodnet activities. Clin Infect Dis. 2004 Apr 15; 38(Suppl. 3):S121–6. [PubMed: 15095180]

Hargittai E. Digital na (t) ives? Variation in internet skills and uses among members of the “net generation”. Sociol Inq. 2010; 80(1):92–113.

Harris JK, Mansour R, Choucair B, Olson J, Nissen C, Bhatt J. Health department use of social media to identify foodborne illness – Chicago, Illinois, 2013–2014. Morb Mortal Wkly Rep (MMWR). 2014; 63(32):681–685. [PubMed: 25121710]

Harrison C, Jorder M, Stern H, Stavinsky F, Reddy Vasudha, Hanson Heather, et al. Using online reviews by restaurant patrons to identify unreported cases of foodborne illness – New York City, 2012–2013. Morb Mortal Wkly Rep (MMWR). 2014; 63(20):441–445. [PubMed: 24848215]

Hood CM, Gennuso KP, Swain GR, Catlin BB. County health rankings: relationships between determinant factors and health outcomes. Am J Prev Med. 2016 Feb; 50(2):129–135. [PubMed: 26526164]

Jensen JD, King AJ, Davis LA, Guntzviller LM. Utilization of internet technology by low-income adults: the role of health literacy, health numeracy, and computer assistance. J Aging Health. 2010 Sep 1; 22(6):804–826. [PubMed: 20495159]

Kawachi I. Social capital and community effects on population and individual health. Ann N Y Acad Sci. 1999; 896(1):120–130. [PubMed: 10681893]

Knight AJ, Worosz MR, Todd ECD. Dining for safety: consumer perceptions of food safety and eating out. J Hosp Tour Res. 2009 Nov 1; 33(4):471–486.

Nsoesie, EO., Flor, L., Hawkins, J., Maharana, A., Skotnes, T., Marinho, F., et al. Social media as a sentinel for disease surveillance: what does sociodemographic status have to do with it?. PLOS Curr Outbreaks. 2016. http://dx.doi.org/10.1371/currents.outbreaks.cc09a42586e16dc7dd62813b7ee5d6b6

Nsoesie EO, Kluberg SA, Brownstein JS. Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports. Prev Med. 2014; 67:264–269. [PubMed: 25124281]

Quinlan JJ. Foodborne illness incidence rates and food safety risks for populations of low socioeconomic status and minority race/ethnicity: a review of the literature. Int J Environ Res Public Health. 2013; 10(8):3634–3652. [PubMed: 23955239]

Sampson RJ, Morenoff JD, Earls F. Beyond social capital: spatial dynamics of collective efficacy for children. Am Sociol Rev. 1999; 64(5):633–660.

Sampson RJ, Morenoff JD, Gannon-Rowley T. Assessing “neighborhood effects”: social processes and new directions in research. Annu Rev Sociol. 2002; 28(1):443–478.

Scallan E, Griffin PM, Angulo FJ, Tauxe RV, Hoekstra RM. Foodborne illness acquired in the United States–unspecified agents. Emerg Infect Dis J. 2011a; 17(1):16.

Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson M, Roy SL, et al. Foodborne illness acquired in the United States–major pathogens. Emerg Infect Dis J. 2011b; 17(1):7.

Scallan E, Jones TF, Cronquist A, Thomas S, Frenzen P, Hoefer D, et al. Factors associated with seeking medical care and submitting a stool sample in estimating the burden of foodborne illness. Foodborne Pathog Dis. 2006 Dec 1; 3(4):432–438. [PubMed: 17199525]

Scarpino SV, Scott JG, Eggo R, Dimitrov NB, Meyers LA. Data blindspots: high-tech disease surveillance misses the poor. Online Journal of Public Health Informatics. 2016; 8(1)

Silver MP. Socio-economic status over the lifecourse and internet use in older adulthood. Ageing Soc. 2014; 34(6):1019–1034.

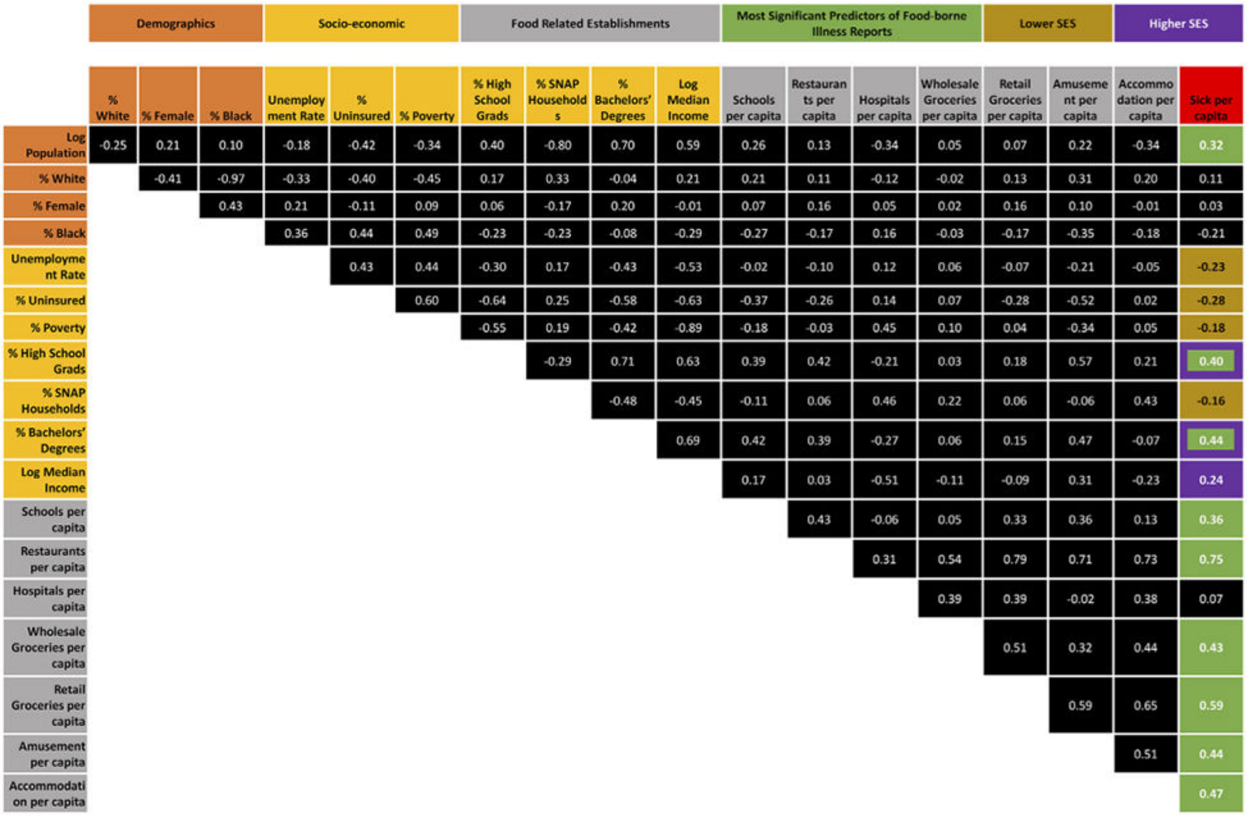
Tam CC, Rodrigues LC, O'Brien SJ. The study of infectious intestinal disease in England: what risk factors for presentation to general practice tell us about potential for selection bias in case-control studies of reported cases of diarrhoea. Int J Epidemiol. 2003 Feb 1; 32(1):99–105. [PubMed: 12690019]

Whole Genome Sequencing (WGS) Program. US Food and Drug Administration. 2016. [Internet] [cited 2016 Sep 29]. Available from: http://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/

World Health Organization (WHO). Geneva: 2015. WHO Estimates the Global Burden of Foodborne Diseases. Foodborne Disease Burden Epidemiology Rference Group 2007. Available at: http://www.who.int/foodsafety/publications/foodborne_disease/fergreport/en/ [Accessed on: Oct 31 2016]

Yelp. An Introduction to Yelp Metrics as of June 30, 2016. 2016. [Internet]. [cited 2016 Oct 13]. Available from: https://www.yelp.com/factsheet

Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B. 2005; 67:301–320.



Demographics | Socio-economic | Food Related Establishments | Most Significant Predictors of Food-borne Illness Reports | Lower SES | Higher SES

	% White	% Female	% Black	Unemployment Rate	% Uninsured	% Poverty	% High School Grads	% SNAP Households	% Bachelors' Degrees	Log Median Income	Schools per capita	Restaurants per capita	Hospitals per capita	Wholesale Groceries per capita	Retail Groceries per capita	Amusement per capita	Accommodation per capita	Sick per capita
Log Population	-0.25	0.21	0.10	-0.18	-0.42	-0.34	0.40	-0.80	0.70	0.59	0.26	0.13	-0.34	0.05	0.07	0.22	-0.34	0.32
% White		-0.41	-0.97	-0.33	-0.40	-0.45	0.17	0.33	-0.04	0.21	0.21	0.11	-0.12	-0.02	0.13	0.31	0.20	0.11
% Female			0.43	0.21	-0.11	0.09	0.06	-0.17	0.20	-0.01	0.07	0.16	0.05	0.02	0.16	0.10	-0.01	0.03
% Black				0.36	0.44	0.49	-0.23	-0.23	-0.08	-0.29	-0.27	-0.17	0.16	-0.03	-0.17	-0.35	-0.18	-0.21
Unemployment Rate					0.43	0.44	-0.30	0.17	-0.43	-0.53	-0.02	-0.10	0.12	0.06	-0.07	-0.21	-0.05	-0.23
% Uninsured						0.60	-0.64	0.25	-0.58	-0.63	-0.37	-0.26	0.14	0.07	-0.28	-0.52	0.02	-0.28
% Poverty							-0.55	0.19	-0.42	-0.89	-0.18	-0.03	0.45	0.10	0.04	-0.34	0.05	-0.18
% High School Grads								-0.29	0.71	0.63	0.39	0.42	-0.21	0.03	0.18	0.57	0.21	0.40
% SNAP Households									-0.48	-0.45	-0.11	0.06	0.46	0.22	0.06	-0.06	0.43	-0.16
% Bachelors' Degrees										0.69	0.42	0.39	-0.27	0.06	0.15	0.47	-0.07	0.44
Log Median Income											0.17	0.03	-0.51	-0.11	-0.09	0.31	-0.23	0.24
Schools per capita												0.43	-0.06	0.05	0.33	0.36	0.13	0.36
Restaurants per capita													0.31	0.54	0.79	0.71	0.73	0.75
Hospitals per capita														0.39	0.39	-0.02	0.38	0.07
Wholesale Groceries per capita															0.51	0.32	0.44	0.43
Retail Groceries per capita																0.59	0.65	0.59
Amusement per capita																	0.51	0.44
Accommodation per capita																		0.47

Fig. 1.

Correlation between variables and count of sick reports per capita (sick per capita).

Table 1

Variables predictive of county-level foodservice business review volume.

Variable	(a) Variables selected in shrinkage procedure Coefficient (SE)	Variables significant in (a) Coefficient (SE)
Log of population	1.257 (0.079) ***	1.295 (0.065) ***
Restaurants per 1000 population	0.812 (0.111) ***	0.881 (0.095) ***
Schools per 1000 population	1.373 (1.702)	–
Percent with bachelor degree	0.012 (0.013)	–
Percent with high school degree	0.042 (0.019) *	0.053 (0.015) ***
Intercept	−13.75 (1.628) ***	−14.889
	$R^2 = 0.85$	$R^2 = 0.849$
	$F_{5,125} = 142.8$ ***	$F_{3,127} = 238.7$ ***

Initial model shown in column two included all variables selected via the regularization procedure. A second regression model comprising only the significant variables is given in column three.

* $p < 0.05$;

** $p < 0.01$;

*** $p < 0.001$.

Table 2

Significant predictors of foodborne illness reviews selected via a regularization procedure described in the methods.

Variable	Coefficient (SE)
Restaurants per 1000 population	0.0315 (0.003) [***]
Percent with bachelor degree	0.0010 (0.0002) [***]
Intercept	−0.045 (0.0051) [***]
	$R^2 = 0.536$ [*]
	$F_{2,206} = 119.1$ [***]

[*] $p < 0.05$;

[**] $p < 0.01$;

[***] $p < 0.001$.