Foodborne illness is prevented by inspection and surveillance conducted by health departments across America. Appropriate restaurant behavior is enforced and monitored via public health inspections. However, surveillance coverage provided by state and local health departments is insufficient in preventing the rising number of foodborne illness outbreaks. To address this need for improved surveillance coverage we conducted a supplementary form of public health surveillance using social media data: Yelp.com restaurant reviews in the city of San Francisco. Yelp is a social media site where users post reviews and rate restaurants they have personally visited. Presence of keywords related to health code regulations and foodborne illness symptoms, number of restaurant reviews, number of Yelp stars, and restaurant price range were included in a model predicting a restaurant’s likelihood of health code violation measured by the assigned San Francisco public health code rating. For a list of major health code violations see (S1 Table). We built the predictive model using 71,360 Yelp reviews of restaurants in the San Francisco Bay Area. The predictive model was able to predict health code violations in 78% of the restaurants receiving serious citations in our pilot study of 440 restaurants. Training and validation data sets each pulled data from 220 restaurants in San Francisco. Keyword analysis of free text within Yelp not only improved detection of high-risk restaurants, but it also served to identify specific risk factors related to health code violation. To further validate our model we applied the model generated in our pilot study to Yelp data from 1,542 restaurants in San Francisco. The model achieved 91% sensitivity 74% specificity, area under the receiver operator curve of 98%, and positive predictive value of 29% (given a substandard health code rating prevalence of 10%). When our model was applied to restaurant reviews in New York City we achieved 74% sensitivity, 54% specificity, area under the receiver operator curve of 77%, and positive predictive value of 25% (given a prevalence of 12%). Model accuracy improved when reviews ranked highest by Yelp were utilized. Our results indicate that public health surveillance can be improved by using social media data to identify restaurants at high risk for health code violation. Additionally, using highly ranked Yelp reviews improves predictive power and limits the number of reviews needed to generate prediction. Use of this approach as an adjunct to current risk ranking of restaurants prior to inspection may enhance detection of those restaurants participating in high risk practices that may have gone previously undetected. This model represents a step forward in the integration of social media into meaningful public health interventions.

Published paper: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152117