Mining Big Data to Improve Food Safety

It's not uncommon for people to use the internet to research their health concerns by entering their symptoms in a search engine like Google. Now, that data may help identify potentially unsafe restaurants faster and with more accuracy than traditional methods. Developed by Google, the model, called FINDER, uses machine learning on large-scale, de-identified data to more proactively identify restaurants that might have gaps in food safety.

"Our model works by classifying web search queries that suggest experiencing foodborne illness, such as 'diarrhea' or 'vomiting,'" explains Evgeniy Gabrilovich, senior staff research scientist at Google. "We then use de-identified location history—from users who have chosen to allow us to record their locations—to identify restaurants whose visitors eventually search for symptoms of foodborne illness."

Today, U.S. municipal health departments inspect every restaurant once or twice a year, but in most cases, they are performed randomly. Health departments also rely on consumer complaints to point them in the direction of potentially hazardous food safety gaps. According to Gabrilovich, the FINDER model has benefits over traditional random inspections and complaint-based inspections.

In collaboration with researchers at Harvard T.H. Chan School of Public Health and the health departments of Las Vegas and Chicago, the Google team tested the model. Health departments in both cities were given a list of restaurants that were identified by the model as being potential sources of foodborne illness and health inspectors were dispatched to these restaurants. "Our findings indicate that over half of the venues that our system suggested for inspection had critical or severe health violations," says Gabrilovich. The study results are published in *npj Digital Medicine*.

"Interestingly, we also found our method to be more accurate than consumer complaints," says Gabrilovich. "When consumers call health departments to complain about a restaurant (or post their experiences on social media), they will most likely complain about the last restaurant they visited … However, the incubation period of various kinds of foodborne illness can be up to 72 hours." In fact, the study revealed that in 38% of the cases, it was the penultimate restaurant that was the source of the food safety issue.

"While individual query classification and location attribution might be imperfect, the power of our method comes from the combination of the two, as well as from using large-scale aggregated data that allows us to identify patterns," explains Gabrilovich.

"To date, we have just scratched the surface of what is possible in the realm of machine-learned epidemiology," says Gabrilovich. "In our future work, we plan to study methods for finding other types of illness that spread from other types of locations. One example scenario would be identifying food poisoning outbreaks that start at grocery stores, such as the recent nationwide recall of romaine lettuce."

Predicting animal sources of *Salmonella*

Machine learning may soon play a key role in identifying the animal source of certain foodborne pathogens, thanks to the work of scientists at the University of Georgia Center for Food Safety. The research, detailed in a study published in *Emerging Infectious Diseases*, used more than 1,000 genomes to predict the animal sources of *Salmonella Typhimurium*.

About 95% of the more than 9 million annual episodes of foodborne illness in the United States are sporadic and remain uninvestigated, says study coauthor Xiangyu Deng. Although routine use of whole-genome sequencing (WGS) has created a wealth of genomes, much of the data has been untapped beyond routine investigation of foodborne illness outbreaks. The researchers decided to take advantage of the large volume of WGS data for *Salmonella* in public databases.

"We trained a machine-learning algorithm called Random Forest with many *Salmonella Typhimurium* genomes from livestock origins," explains Deng. "Through training, the algorithm 'learned' how to classify genomes of the pathogen by their livestock sources, including poultry, pigs, and cattle." Altogether, the researchers were able to identify a set of about 50 genetic markers that were sufficient for robust livestock source prediction of the pathogen, a finding that may lead to a scalable source identification tool. In a retrospective analysis of eight major zoonotic outbreaks in the United States between 1998 and 2013, the classifier attributed seven to the correct source.

Because the vast majority of foodborne illnesses are uninvestigated, their contamination sources often remain unknown. The researchers' methodology provides a potential tool to fill the information and knowledge vacuum as WGS of foodborne pathogens becomes increasingly routine. "It may also speed up source tracking analysis for outbreak investigations," notes Deng, who points to the need for quick identification of contaminated food sources to ensure timely recall and prevent more consumers from contracting the pathogen. "Our study indicates that isolates from previous fresh produce outbreaks could be traced back to animal sources," he explains, "therefore providing clues to identify root causes of some *Salmonella* contaminations of fresh produce." **FT**

IFT*NEXT content is made possible through the generous support of Ingredion, the* IFT*NEXT Platinum Content Sponsor.*