Chapter 5

A Data-Driven Approach to Food Safety Surveillance and Response

N.P. Greis and M.L. Nogueira
University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

5.1 INTRODUCTION

Maintaining a safe and secure food supply is critical to the well-being of millions around the world. An increasingly global food chain—in which products are sourced from locales far from the end consumer—has increased the potential for contamination. These pressures have only increased the sense of urgency in addressing gaps in the food safety system. In particular, early detection and rapid response are challenges that must be met to minimize the impact of a contamination event—whether due to unintentional failure of the food chain or due to an intentional terrorist act. This chapter explores the potential of data-driven informatics tools to provide situational awareness and decision-making intelligence for an intrinsically complex and dynamic process—the detection of and response to a foodborne illness outbreak. A data-driven approach is introduced that builds situational awareness by coalescing real-time data fusion of both traditional and nontraditional sources, analytics based on tools of data science, visualization using a Common Operating Picture (COP), and real-time collaboration across stakeholders of the system to reduce the latency in detecting an emerging contamination event. By reducing the latency of detection, responses such as medical alerts and product recalls can be accelerated, thereby saving lives and cost. These principles of situational awareness were used to develop a prototype software tool for the State of North Carolina, the North Carolina Foodborne Events Data Analysis Tool or NCFEDA. Latencies reductions in surveillance and response are illustrated using a typical example—a cluster of unspecified illness cases reported with symptoms of gastrointestinal distress that may (or may not) indicate a possible foodborne disease outbreak.

Food Protection and Security. DOI: http://dx.doi.org/10.1016/B978-1-78242-251-8.00005-9

5.2 CHALLENGES OF FOOD SAFETY

Recent high-profile contamination events have elevated the need to adopt a data-driven approach to assure the safety of the country's food system. One of the most widely reported contamination events was the recent closure of more than 40 restaurants belonging to Chipotle, a US-based fast-food restaurant chain, in Washington and Oregon in October 2015 due to an *Escherichia coli* contamination. The Centers for Disease Control and Prevention (CDC) reported that 45 people were sickened by the *E. coli O26* outbreak strain and, of those, 43 reported eating at Chipotle. Sixteen people were hospitalized although no deaths were reported. In February 2016, the CDC concluded their investigation, unable to find the source of the *E. coli* contaminations.

The CDC has also linked the Chipotle outbreak in the US Pacific Northwest with other reported *E. coli* cases in California, Ohio, New York, and Minnesota. And only a few months before, Chipotle had been linked to two other cases of foodborne contamination and resulting illness—a norovirus outbreak in California in August and cases of *Salmonella* in Minnesota that have been traced to tomatoes from out-of-state farms. Then, in December 2015, 80 individuals were sickened after eating at a Chipotle restaurant in Massachusetts. And before the chain of outbreaks, Chipotle had taken the step of removing pork from its restaurant menus when one of the company's suppliers failed to follow animal welfare standards.

Despite recent efforts and the passage of the Food Safety Modernization Act (FSMA) in 2012, foodborne infections continue to be an important public health problem in the United States. Federal data released by the Foodborne Diseases Active Surveillance Network (FoodNet) in 2015 showed little improvement in terms of foodborne illnesses when compared with data collected between 2006 and 2008, and between 2011 and 2013. The data indicated that illness due to *Campylobacter*—usually caused by consuming undercooked poultry—has risen by 13%. In addition, illnesses from two strains of *Salmonella*, javiana and infantis, typically found in undercooked eggs, milk, and meat, have more than doubled. And *Listeria*, the likely culprit in this year's massive Blue Bell Creameries outbreak in the United States, was responsible for the most deaths of any strain last year. Of the 118 people who were diagnosed with listeriosis, 18 of them died.

The problems experienced by Chipotle and other food purveyors are emblematic of the challenges faced by today's food industry. Our food supply chains are dynamic and complex—with an array of governmental agencies at different jurisdictional levels charged with regulating and supervising the safety of millions of food products produced by thousands of companies across the globe. Assuring safe food depends critically on our ability to collect, interpret, and disseminate electronic and other information across organizational and jurisdictional boundaries. The lack of visibility due to interoperability across the stakeholders of the food chain makes it difficult to quickly determine when a contamination has taken place. And once a contamination has been confirmed, the lack of visibility makes it difficult to trace contaminated food products back to the farm or country where they were produced—and also forward to locations where similar products may be waiting to be sold.

In response to these challenges, the food industry has looked to data science and "big data" for insight and a way forward. Significant efforts are being made to marshal big data tools to the cause of food safety. The definition of big data remains in flux depending on industry and application, but it typically involves the digital generation of data, often passively produced and automatically collected and stored, but also actively generated through events that serve as a trigger for marshaling response to an emerging contamination event. The premise of data science and big data for improved food safety is that, when fusing multiple types and formats of data including new and nontraditional sources, new analytics will make it possible to enhance our visibility of the food system to better monitor and respond in (near) real time to contamination threats as they occur.

5.3 MOVING TO DATA-DRIVEN FOOD SAFETY

Major advances in many industries can be attributed to the convergence of multiple technological advances whose synergistic effects enable major transformation within that industry. Defined as the coming together of two or more disparate disciplines or technologies, convergence has been associated with advances from early in the industrial age—from firearms and sewing machines at the beginning of the 20th century to jet engines today. The fax revolution was produced by a convergence of telecommunications technology, optical scanning technology, and printing technology. Today fundamental shifts in our basic industries are emerging from a confluence of the internet and related communication technologies with technical advances in specific domains—including the food industry.

A fortuitous and simultaneous convergence of internet and communications technologies along with a new generation of sensors and analytical tools is reshaping the food industry—and its ability to reduce the risk of food contamination and resulting foodborne illness. The availability of low-cost sensors, scanners, and various mobile devices, along with new communications technologies linked to the internet, offers visibility across the food chain. When combined with data analytical tools capable of fusing extremely large quantities of data of different formats and extracting relevant information, these technologies are opening the door to real-time, end-to-end monitoring, and control of the movement of food products across the chain. And, as we will see later in this chapter, this convergence can be marshaled to make it possible to reduce the latencies in both detecting a food contamination event and responding to it.

The food supply chain starts at the farm and encompasses food transportation companies, processing facilities, distributors, retailers, brokers, importers, and governmental agencies responsible for overseeing and regulating the system—and ends at the consumer's table. Given the large-scale and distributed nature of the food system, it can be viewed as a "system of systems" whose components are complex, heterogeneous, self-organizing networks of systems that operate independently but are ultimately integrated into a dynamic, evolving "organism" that expertly manages the continuous production, distribution, and sale of food. Bringing these stakeholder systems together into an efficient and effective food safety network has been the signature challenge of regulatory agencies such as the US Food and Drug Administration (FDA).

Across many of these food chains today, sensors and other hardware are able to record a wide range of parameters—from location of a pallet or even item of food to its temperature while in transit from farm to fork. These sensors provide a level of granularity that was not available previously. A sensor attached to a carton of New Zealand milk will record the swings in temperature that accompany that carton as it moves from the New Zealand dairy farm by truck to airplane hold and by truck to retailer in China or elsewhere in Asia. This information, alone, can assist in identifying milk that might have spoiled before it is placed on the grocer's shelf. Temperature traces in route when combined with weather data, as well as shelf-life curves for that product, can also let retailers know what the remaining shelf-life is for that product.

In addition to preventing food spoilage and contamination, these new technologies are enabling better surveillance to determine the onset of foodborne illness. Although the specific authority varies from country to country, surveillance has typically been the purview of public health departments. Public health officials engage in surveillance activities to determine whether reported cases of foodborne illness are part of a large outbreak. Local public health departments are usually the first to pick up the signals of foodborne illness. These signals may correspond to isolated reports of illness or they may be causally linked and part of a larger outbreak. Or they may be uncorrelated and isolated cases that are not precursors of an emerging event.

When public health officials suspect a set of causally related cases, samples are sent to official laboratories such as the CDC for DNA "fingerprinting" to confirm that the illness is due to the same pathogen. Confirmation of the pathogenic source becomes the starting point for investigations by response teams to determine the specific food types that are responsible for the illness. Numerous delays occur in the surveillance and response processes. The promise of data science and big data is that these latencies can be reduced by timely fusion and interpretation of information on potential cases of foodborne illness.

Large amounts of data are already collected during the surveillance and response processes. What separates "big data" from "small data" in the food chain? Big data is distinguished by five characteristics referred to as the "5 Vs"—the volume, velocity, variety, veracity, and value of the data generation process. More data is being collected faster and in many different formats. The highly structured data that is typical of processing histories, shipment records, and lab reports is being augmented by data generated and/or transmitted from many nontraditional sources including wireless sensors such as RFID, temperature, and chemical sensors that monitor ambient conditions during transport, and mobile technologies—as well as satellite images, real-time data collected by drones, text data from telephone hotline calls, electronic medical data, and even social media.

Except for highly sensored food chains, the volume of data currently collected across a food chain is not extremely large when compared with other industrial processes such as aerospace where voluminous data is reported by aircraft in flight to ground stations for analysis. Similarly, the velocity with which the data is gathered is not extremely high compared with other domains such as financial systems. However, in both the food and agriculture industries, there is a proliferation of data variety with different levels of value. To build the capabilities necessary for improved surveillance and response, data must be collected and combined from the multiple and heterogeneous sources listed previously. And with increasing numbers of sources, there is inevitably a data quality and confidence problem—so veracity is an issue as well.

The proliferation of multiple data systems and tools that lack interoperability hinders effective information gathering and timely response to emerging but yet unconfirmed foodborne illness. As already noted, most of the public health and food safety informatics work in the United States—from early detection of food-related outbreaks by local and state health departments to confirmation by the CDC through "fingerprinting" of pathogenic contaminants—takes place at different local, state, and federal jurisdictional levels causing significant delays that have significant cost in terms of lives and dollars. A data-driven approach to food safety would reduce these latencies by bringing together: (1) traditional and new nontraditional data sources across all stakeholders in the food safety network; (2) new information and communication technologies for fusing and interpreting this data; and (3) new informatics and visualization tools capable of extracting knowledge that establishes "evidence" that can be used effectively by *all* the stakeholders across the food chain.

5.4 NEW FOOD SAFETY STAKEHOLDER MODEL

In the United States, the passage of the FMSA of 2012 was an attempt to bridge gaps in food safety surveillance and response activities by mandating the implementation of new information processes and informatics tools that reduce both the scale and scope of a food contamination event—whether unintentional or intentional. Not only did the passage of FSMA serve as a milestone in food safety law in the United States that sets the stage for a "big data" approach to assuring food safety, it also fundamentally changed the landscape of stakeholders that play a role in assuring safe food.

The FSMA signaled the emergence of a new food safety stakeholder model in which the private and public sectors, as well as the consumer, assume new roles in meeting the challenge of safe food. While the public sector has traditionally been the guardian of food safety, increasingly private sector enterprises and the consumer are playing an important role. The private sector is being given more responsibility for recording and providing information about its processes, suppliers, and customers (when requested by the FDA). And consumers have more opportunity to provide information to regulatory agencies and private sector enterprises about the quality and safety of their food.

Under FSMA new responsibilities fall on private sector companies. Food manufacturers are required to register and to examine their processing systems to identify possible ways that food products can become contaminated and to develop detailed plans to keep that from occurring. Companies must share those plans with the FDA, and provide the agency with records, including product test results, showing how effectively they can carry them out. The FDA was mandated to work with private sector companies on pilot projects to develop traceability systems that strike a balance between protecting public health and preventing any undue burden to businesses.

Increasingly the consumer is also a key stakeholder in the system. Previously, the consumer has had limited direct input into the food safety system. Official laboratory reports of cases of foodborne illness typically take many days, or even weeks, to find their way into the food safety system. Increasingly, however, consumer input into the surveillance and response processes is occurring through new channels. "Complaint" hotlines to food retailers and to public agencies provide real-time signals of possible foodborne illness. Consumers also "blog" information related to food using social media and other emerging technologies. Harnessing these sources of real-time consumer information can be critical in reducing delays in detecting foodborne illness.

Fig. 5.1 presents the new food safety stakeholder model comprised of the food safety system's four major stakeholders. They are: (1) a private sector that controls the production and commercialization of food products, the sale and distribution of potentially contaminated products, and participates in the recall of tainted products; (2) a public health system in charge of surveillance and management of outbreaks of disease caused by food contamination; (3) the governmental agencies pertaining to agricultural activities and the protection of the environment and natural resources, which regulate the production of food for human consumption by the agricultural and food

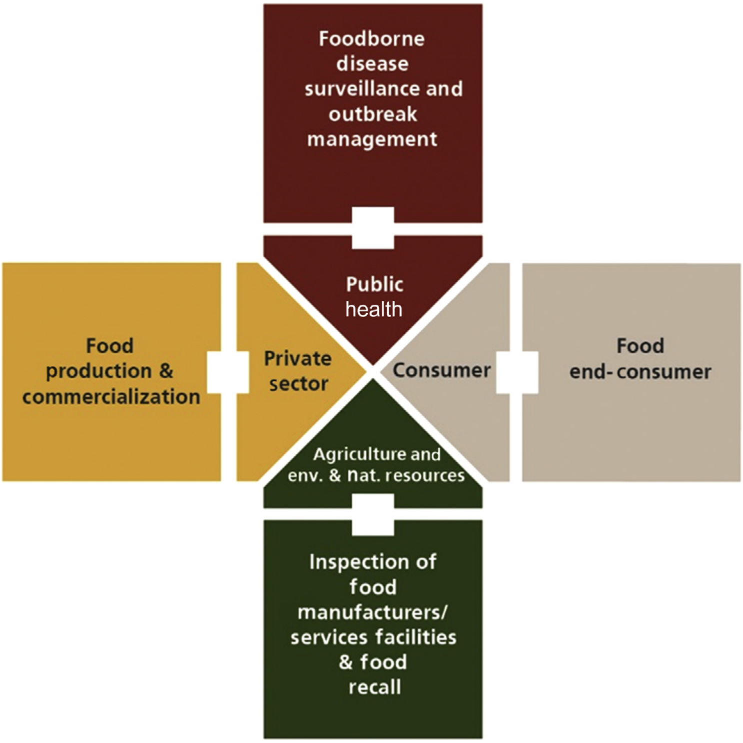

Foodborne disease surveillance and outbreak management
Public health
Food production & commercialization
Private sector
Consumer
Food end-consumer
Agriculture and env. & nat. resources
Inspection of food manufacturers/ services facilities & food recall

FIGURE 5.1 Food safety stakeholder model.

manufacturing sectors, oversee the safe use of natural resources and the environmental and sanitary conditions of establishments offering food services, as well as monitor and assist in food recall efforts; and (4) consumers of food products.

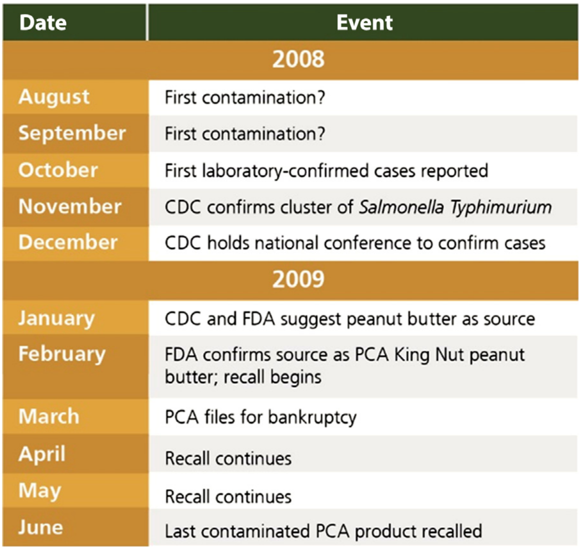
5.5 REDUCING LATENCY IN SURVEILLANCE AND RESPONSE

Continuous surveillance for early detection of foodborne outbreaks and rapid response to reduce the scale and scope of outbreaks are essential components of timely response for safeguarding our food supply. As noted earlier, our current ability to detect and respond to foodborne illness outbreaks is hampered by a number of gaps in the food safety system that create latencies in these processes. FSMA provides increased authority and resources for FDA to address many of the existing gaps in our food safety system. The law seeks to bridge some of the biggest gaps by mandating the implementation of new information processes and informatics tools that reduce both the scale and scope of a food contamination event.

An illustrative example of the degree of latency in responding to food contamination outbreaks was the 2008–09 *Salmonella Typhimurium* contamination of peanut butter produced by the now-defunct Peanut Corporation of America (PCA). The outbreak sickened 714 people in 46 states and may have contributed to nine deaths, according to the CDC. The illnesses began in January 2009 and ultimately prompted one of the largest food recalls in US history. This contamination triggered the most extensive food recall in US history up to that time, involving 46 states, more than 360 companies, and more than 3900 different products manufactured using PCA ingredients. The cost to food companies and the government was estimated to be more than $1 billion.

The timeline for the PCA outbreak is shown in Fig. 5.2. As shown in the figure, the first suspected contamination occurred in August 2008. It took almost 6 months to confirm that a foodborne outbreak had occurred, and another 6 months to locate all the contaminated products and remove them from retail shelves across the country. Nearly 6 years later, on September 21, 2015, the owner of the now-defunct PCA was sentenced to 28 years in prison for knowingly shipping out salmonella-contaminated peanut butter and hiding the evidence. This was the toughest punishment in US history to date for a producer in a foodborne illness case.

The key tasks associated with the surveillance and response processes are represented by four phases of the food safety wheel shown in Fig. 5.3. The right-hand side of the food safety wheel represents the surveillance phase. The left-hand side represents the response phase. During the first phase public health officials engage in detection activities to determine whether



Date	Event
2008	
August	First contamination?
September	First contamination?
October	First laboratory-confirmed cases reported
November	CDC confirms cluster of *Salmonella Typhimurium*
December	CDC holds national conference to confirm cases
2009	
January	CDC and FDA suggest peanut butter as source
February	FDA confirms source as PCA King Nut peanut butter; recall begins
March	PCA files for bankruptcy
April	Recall continues
May	Recall continues
June	Last contaminated PCA product recalled

FIGURE 5.2 Timeline of Peanut Corporation of America outbreak.

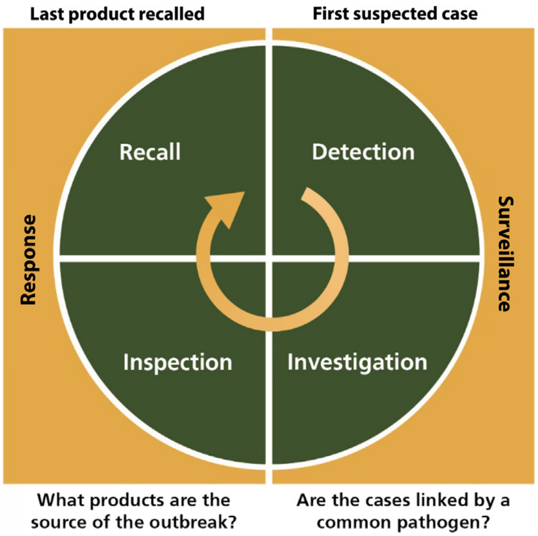


FIGURE 5.3 Key tasks in surveillance and response.

individual cases of foodborne illness are part of a larger outbreak. Laboratory testing to look for common pathogens (e.g., *Salmonella*) is used to confirm that an outbreak has occurred and that a cluster of cases has a common pathogen. Once a common pathogen has been identified and an outbreak has been confirmed, epidemiologists conduct interviews to discover the offending food types (e.g., tomatoes).

During the recall phase, the specific food products (e.g., Red Ripe Tomatoes) and facilities (e.g., Best Produce Company) are tested and inspected to identify specific product brands and/or production facilities. Once a source has been located, the difficult task of recalling all contaminated products in the food chain begins. The scale and scope of a food contamination event is directly related to the speed with which these tasks can be performed. Reducing the latencies associated with these events is crucial to saving lives and reducing costs.

Thus, we can reduce the costs of contamination significantly by reducing the surveillance period during which evidence is gathered and used to confirm the outbreak. In this section, we apply big data principles and techniques to the problem of reducing these food safety gaps.

5.6 BUILDING SITUATIONAL AWARENESS ACROSS THE FOOD CHAIN

The example of the PCA demonstrates the overarching need for capabilities that enhance situational awareness across the stakeholders in the food safety system. Theoretical frameworks for situational awareness are based on an understanding of cognitive processes of the human mind for decision making. The most common theoretical framework is provided by Endsley who defined situational awareness as the "perception of the elements in the environment within a volume of time and space, comprehension of their meaning and the projection of their status in the near future." We target four essential capabilities that contribute to enhanced situational awareness in food safety: (1) data integration; (2) visualization; (3) analytical tools; and (4) real-time collaboration.

Collectively, these four capabilities support an operational environment necessary for understanding, evaluating, and responding to foodborne outbreak events in an effective and timely manner. In continuously and rapidly changing environments such as public health, capabilities that support situational awareness maximize results of operational procedures, improve team collaboration, and enable better-informed decision making. The relevance of each capability to our end goal of reducing latencies in surveillance and response to foodborne illness outbreaks is described briefly in the following paragraphs.

Data Integration. Fusing data from all major food safety stakeholders can offer a more complete and clear picture of an emerging or ongoing (i.e., near real-time) event. In order to create situational awareness an informatics tool must provide a coherent representation of those data elements that are relevant to respective food safety stakeholders and that are essential to perceiving the status, attributes, and dynamics of any emerging or ongoing event. Currently, each major food safety stakeholder (c.f., public health official or private company) has only partial knowledge of what is happening based on that stakeholder's limits of responsibility and authority. Combining relevant information across all relevant food safety stakeholders into a single shared view, i.e., common operational picture, will create a more complete representation of present conditions that may allow faster recognition of existing problems and generate new knowledge that will contribute to latency reductions.

Visualization. A visualization tool not only provides a graphical representation of data that is more easily interpreted, but can also be used as a problem-solving tool. Trying to answer questions by examining large numerical tables or spreadsheets is typically more difficult and time-consuming than allowing a user to process the same data presented in graphs or maps or charts. Exploring different visual views of the same data facilitates analytical reasoning by taking advantage of human capabilities to process images. Benefits obtained from fusing diverse data sources can be augmented by adding visual analysis capabilities to the food safety system.

Analytical Tools. Analytics are broadly defined as a set of tools based on logic, statistics, or data science that are used to support decision making. In food safety, analytical tools can discover disease or exposure patterns that require further epidemiological investigation and will, as a result, speed up the process of identifying possible sources of contamination. For example, analytical tools can generate clusters based on similar foods consumed, places visited, or other common elements among data records that may help point out the source of contamination or uncover a totally new, still unreported, existing problem. Such tools can also assist in reducing latencies in the recall process by making the recall and effectiveness checks more efficient. Analytical tools can also be used to assess the likelihood of the emergence of a food safety event from fused data that can then be used to guide response.

Real-Time Collaboration. The need for better mechanisms for informal and formal communication among stakeholders is multifold and: (1) calls for a communication vehicle that enables exchange of information between participants; (2) offers 24/7 access; and (3) entails keeping a comprehensive roster of responders and public health officials at the state level including direct contact information and location, and an analogous roster of local healthcare providers' representatives and physicians at the local level. Such capability enables anytime, anywhere collaboration and exchange of ideas and information.

5.7 BUILDING A DATA ANALYTICS ENGINE FOR SURVEILLANCE

Savings lives and reducing the costs of a foodborne illness outbreak depend directly on the ability to reduce the latency with which a contamination event can be confirmed and the speed with which the offending products can be removed from the shelves of retail stores, as illustrated in Fig. 5.3. There are many cases of illness due to food consumption every day. Not all of them signal an impending food safety crisis. Individuals may react poorly to certain types of food. And, in other cases, food safety problems may be attributed to an individual's malfunctioning refrigerator. Distinguishing between these two cases is essential in responding effectively and efficiently to potential food contamination problems.

In making this assessment, the human decision-making process takes into account what is known by the decision maker. This includes *facts* describing the situation at hand and preestablished procedures/regulations, or *processes*, that dictate how that particular situation must be handled. The human processes this information through an activity known as *logical reasoning*, which allows the human to identify relationships among seemingly independent elements of a problem in the search for a solution. When it is not possible to apply any known processes to the known facts, humans can resort to using logical reasoning to link apparently unrelated facts to get a better understanding of the problem and to find an answer, or to delay any decision until more information is available or a new method is devised.

Connecting information, or finding the relationships among isolated facts, and selecting what is relevant to the task at hand is key to enabling humans to make better decisions in a timely and efficient manner. Today, representing facts and processes in a format so that they fit traditional execution models for computers is an ordinary task which makes it possible to automatically control many operations with these machines. Facts are well-suited to be represented in databases and processes as sequences of instructions for computer programs. In the case of food safety surveillance and response, these instructions are analogous to the thought processes that assist the decision-making process of the human.

In translating the cognitive processes by which we assess an emerging food safety event, we think of the food safety surveillance process as one in which many different bits of (big) data are being received in sequence. These data points contain information such as an admission to the emergency room with presenting gastroenteritis, a physician's report of a suspected foodborne illness case to public health authorities, personal blogs on social media that report illness after eating at a particular restaurant, FDA food product recalls, or even calls to government poison hotlines. These bits of data can be thought of as "events" that contain information that can help determine when a food contamination event has occurred and to distinguish that contamination event from an isolated case of food poisoning.

An illustrated example of such an events sequence is shown in Fig. 5.4. In the figure, a couple enjoys a meal at *MyFoodChain* restaurant and blogs

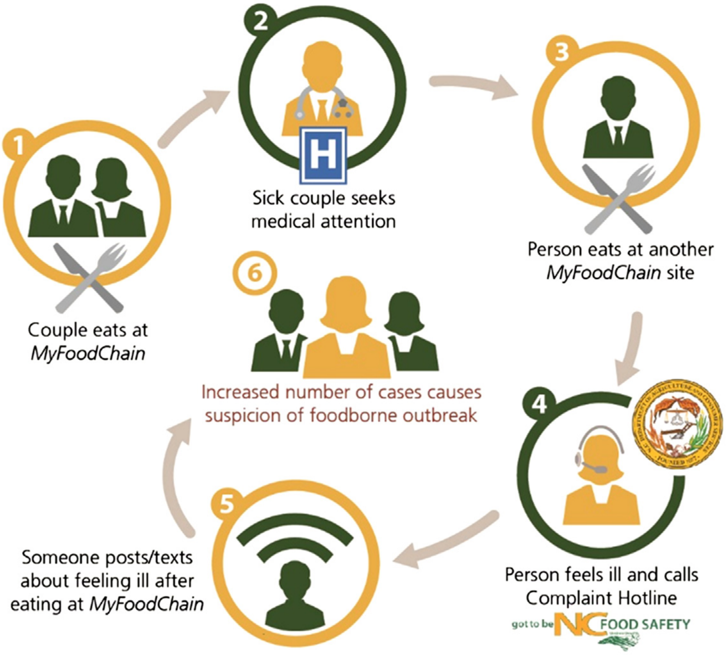

1
Couple eats at *MyFoodChain*
2
Sick couple seeks medical attention
3
Person eats at another *MyFoodChain* site
4
Person feels ill and calls Complaint Hotline
got to be NC FOOD SAFETY
5
Someone posts/texts about feeling ill after eating at *MyFoodChain*
6
Increased number of cases causes suspicion of foodborne outbreak

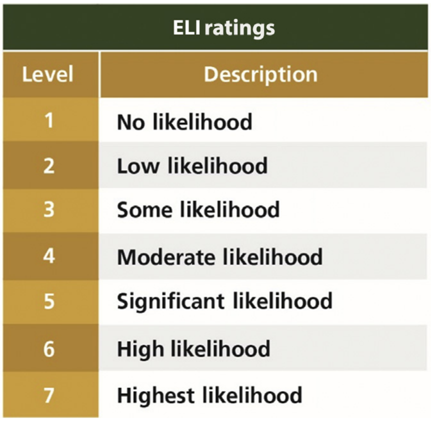
FIGURE 5.4 Events sequence for foodborne illness analysis.

about it to friends. They fall ill shortly thereafter and visit the local emergency room for treatment. A single report of gastroenteritis by a single couple does not, by itself, confirm a food safety event. However, soon thereafter another person eats at *MyFoodChain* and falls ill. Not sick enough to report to the hospital, the individual calls a state complaint hotline to report the problem. The likelihood that the illness is due to a contaminated food product at *MyFoodChain* has increased with the second report. The likelihood increases even more when a third person posts a blog about feeling ill after also eating at *MyFoodChain.*

The food safety challenge, then, is to develop analytics that interpret this sequence of events in real time and assess the likelihood that a foodborne illness outbreak is emerging. A number of data science methods can be adapted to look for clues in the various information events that are being received to determine the strength of the supporting evidence. Conceptually, the task is to "connect-the-dots" between possibly related pieces of information. As a new piece of evidence is observed (c.f., another hotline report of illness), it is compared to the available set of events to determine whether or not the newly received event increases the likelihood that the current situation signals an emerging foodborne illness outbreak.

Representing the necessary logical reasoning in such a way that it can be performed by computers, on the other hand, is not an easy endeavor because the relationships to be represented may require a complex set of rules that cannot be easily encoded in a database or a program with a well-defined flow of control. Usually logical reasoning is encoded as *inference rules* using some computer programming language and these rules are processed, together with facts, by another software application called a *reasoning engine* to produce answers. The analytics engine described herein performs rule-based predictive analytics and "reasons" about an existing situation as described by the known facts and encoded rules. The analytics engine deduces relationships among events to generate an *evidence set* of relevant events and information, which is shared with food safety stakeholders to improve their situational awareness and help in determining the likelihood that a food contamination event is emerging.

In addition to identifying relevant information concerning possible emerging events that can be "pushed" to users, an analytics engine can also rate the strength of the relationships among the events for users and compute a measure of the likelihood that the event under consideration is indeed an emerging event. In NCFEDA, the strength of the relationship among events in the evidence set is captured by the computation of the Event Likelihood Index (ELI) metric. This metric is based on the number and "connectedness" of the events that comprise the evidence set. The ELI metric is captured in an ordinal scale as shown in Fig. 5.5. In the example, seven possible levels of the ELI ratings scale range from "no relationship" at ELI = 1 and "highest likelihood" at ELI = 7.



ELI ratings	
Level	Description
1	No likelihood
2	Low likelihood
3	Some likelihood
4	Moderate likelihood
5	Significant likelihood
6	High likelihood
7	Highest likelihood

FIGURE 5.5 Event Likelihood Index (ELI).

5.8 NCFEDA—NORTH CAROLINA FOODBORNE EVENTS DATA INTEGRATION AND ANALYSIS TOOL

The North Carolina Foodborne Events Data Integration and Analysis (NCFEDA) prototype tool demonstrates the potential of improved situational awareness—created through real-time data fusion, analytics, visualization, and real-time communication—to reduce latency of response to foodborne illness outbreaks by North Carolina public health personnel. Data integration occurs across responding agencies—the North Carolina Department of Public Health (NCDPH), the North Carolina Department of Agriculture and Consumer Services (NCDA&CS), and the North Carolina Department of Environmental and Natural Resources (NCDENR)—as necessary for situational awareness. NCFEDA also includes new data sources from the private sector and the consumer. For example, on the private sector side, FDA recall alerts and enforcement reports provide information about contaminated food products as reported by manufacturing companies to the FDA and USDA. On the consumer side, consumer complaints collected by agencies' complaint hotlines are used as triggers for the NCFEDA system.

At its present state, the NCFEDA Analytics Engine processes triggering event data against other food safety data already stored in its databases and generates one or more possible "models" of the situation being evaluated. By definition, a model is a consistent set of knowledge assertions that the engine infers from the given inputs and the concepts it knows. The Analytics Engine data usage flow is illustrated by the diagram in Fig. 5.6 and indicates the types of results expected to be produced by the engine for two different use case scenarios.

Fig. 5.6 presents a high-level view of the major components that comprise NCFEDA's modular architecture including input data from stakeholders, analytical tools such as the Analytics Engine, and stakeholder dashboards. Input

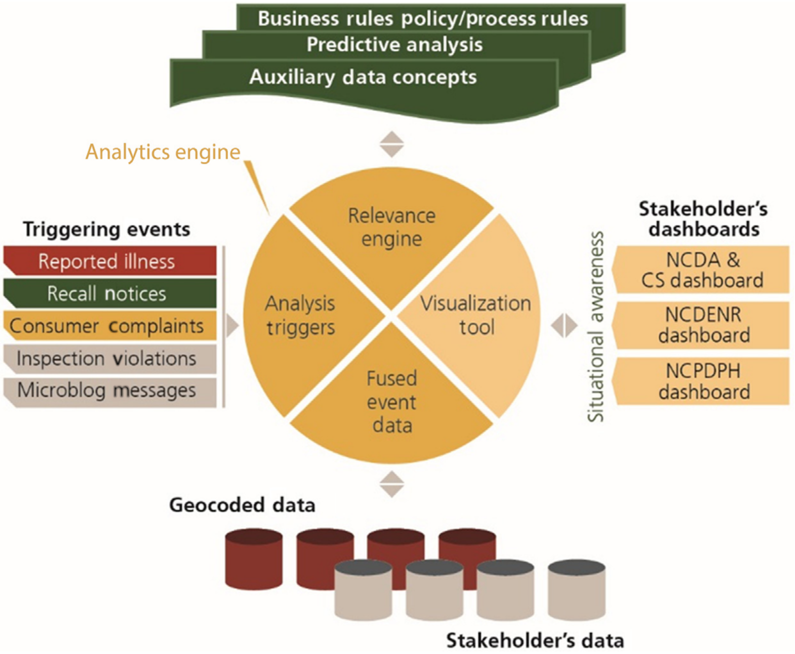


FIGURE 5.6 High-level view of NCFEDA's modular architecture.

data are observations associated with a food-related event that triggers NCFEDA analysis. These triggering events could be illness cases reported to public health officials, recall notices issued by FDA and USDA, or consumer complaints reported to NCDA&CS. These events are then provided to NCFEDA's Analytics Engine to determine whether they are relevant to the stakeholders' decision-making process and support the likelihood of an emerging foodborne illness. These new events can be thought of as signals that may indicate an emerging event or confirm an existing event.

When new events—or signals—arrive, they are interpreted by NCFEDA to determine whether they are relevant to other previously received data. Every new arrival may or may not activate one or more NCFEDA logical rules which are the basis for NCFEDA's Analytics Engine. If arriving event information is determined by the NCFEDA Analytics Engine to be relevant to a suspected emerging event, NCFEDA "pushes" that information to the appropriate stakeholder dashboard. When NCFEDA determines that the event may be relevant, the Analytics Engine computes the ELI—a measure of the likelihood that the suspected outbreak is real, which can assist public health officials in planning a response.

The major components of NCFEDA's modular architecture are described in the following paragraphs.

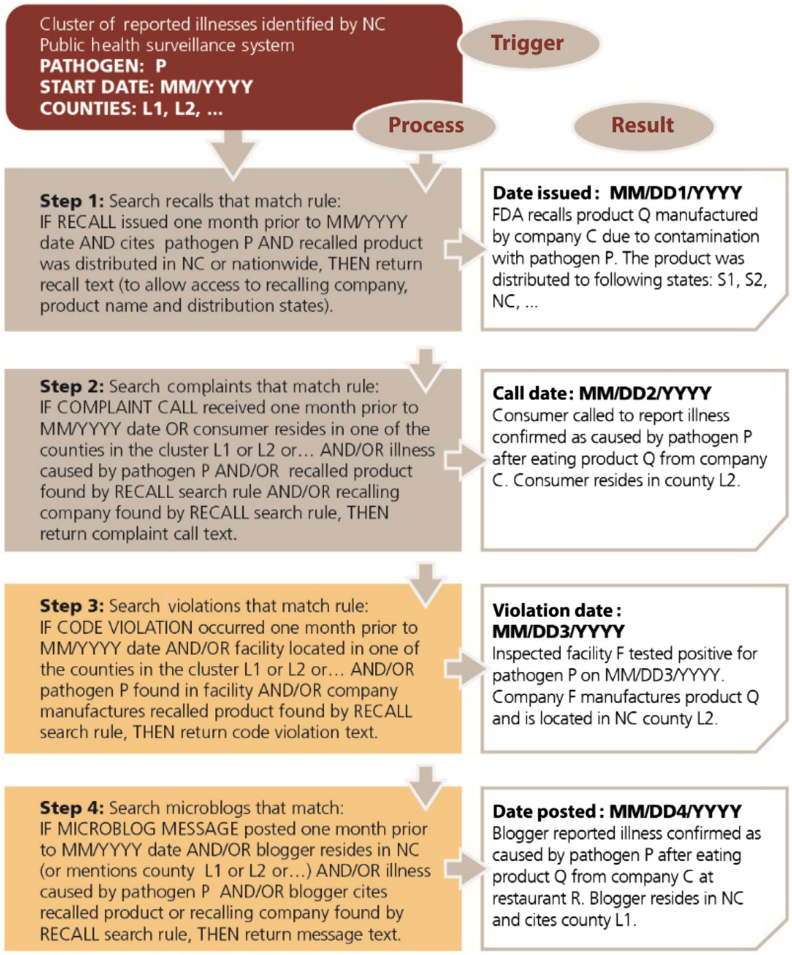
Analytics Engine. The Analytics Engine is the intelligent component of the system responsible for drawing conclusions about a given food safety situation. The engine is the core, domain-independent inference module and includes a set of inference rules that were created to define the food safety domain problem for North Carolina. NCFEDA reasoning capabilities are powered by formal logic. This means that the Analytics Engine “reasons” about food safety events by applying deductive reasoning, i.e., inference rules, to facts informed to the engine in order to infer (new) knowledge. The Analytics Engine’s modules execute the following main functions: (1) analysis of the incoming triggering information events; (2) fusion of known event data previously acquired directly or indirectly from various stakeholder’s surveillance and reporting systems; and (3) processing of new trigger information against the known data by using the relevance engine’s deductive mechanisms together with various sets of rules, i.e., predictive analytics. A sample of the logical rule set that reasons to build the evidence set in NCFEDA is shown in Fig. 5.7.

Auxiliary Data Concepts. The auxiliary data concepts are a set of seven factual databases which store concepts of interest necessary for the task of reasoning about food safety events, and which are represented as logical knowledge for easy processing by the Analytics Engine. These concepts include four (simplified) ontologies for food, foodborne illness, and geographical information, as well as three databases which contain FDA’s Food Code and the medical and consumer complaints codes utilized by the NCDA&CS to process consumer complaints about food products.

Stakeholder Databases. NCFEDA’s databases store all event data obtained from both private and public sources and are the source of all information analyzed by the Analytics Engine and displayed on stakeholders’ dashboards. All received event information is recorded in NCFEDA databases so that the databases are kept up to date. These event data constitute the history of food safety in North Carolina and are used by the Analytics Engine to support or refute possible conclusions regarding emerging and other food events.

The following data are provided to NCFEDA and stored in the NCFEDA databases:

- **Public Health Illness Data**. Records of patient illness reported to the North Carolina Division of Public Health containing, among other fields, the office visit date, probable diagnosis, and patient’s county of residence.
- **Food Recall Notifications**. Recall notices of food products issued by FDA containing the recall issuing date, the product recalled, the company recalling the product, the cause for the recall (i.e., pathogen causing the contamination when available), and areas (states) where the product has been distributed.



Cluster of reported illnesses identified by NC Public health surveillance system
PATHOGEN: P
START DATE: MM/YYYY
COUNTIES: L1, L2, ...

Step 1: Search recalls that match rule:
IF RECALL issued one month prior to MM/YYYY date AND cites pathogen P AND recalled product was distributed in NC or nationwide, THEN return recall text (to allow access to recalling company, product name and distribution states).

Date issued: MM/DD1/YYYY
FDA recalls product Q manufactured by company C due to contamination with pathogen P. The product was distributed to following states: S1, S2, NC, ...

Step 2: Search complaints that match rule:
IF COMPLAINT CALL received one month prior to MM/YYYY date OR consumer resides in one of the counties in the cluster L1 or L2 or… AND/OR illness caused by pathogen P AND/OR recalled product found by RECALL search rule AND/OR recalling company found by RECALL search rule, THEN return complaint call text.

Call date: MM/DD2/YYYY
Consumer called to report illness confirmed as caused by pathogen P after eating product Q from company C. Consumer resides in county L2.

Step 3: Search violations that match rule:
IF CODE VIOLATION occurred one month prior to MM/YYYY date AND/OR facility located in one of the counties in the cluster L1 or L2 or… AND/OR pathogen P found in facility AND/OR company manufactures recalled product found by RECALL search rule, THEN return code violation text.

Violation date: MM/DD3/YYYY
Inspected facility F tested positive for pathogen P on MM/DD3/YYYY. Company F manufactures product Q and is located in NC county L2.

Step 4: Search microblogs that match:
IF MICROBLOG MESSAGE posted one month prior to MM/YYYY date AND/OR blogger resides in NC (or mentions county L1 or L2 or…) AND/OR illness caused by pathogen P AND/OR blogger cites recalled product or recalling company found by RECALL search rule, THEN return message text.

Date posted: MM/DD4/YYYY
Blogger reported illness confirmed as caused by pathogen P after eating product Q from company C at restaurant R. Blogger resides in NC and cites county L1.

FIGURE 5.7 Building the evidence set by reasoning.

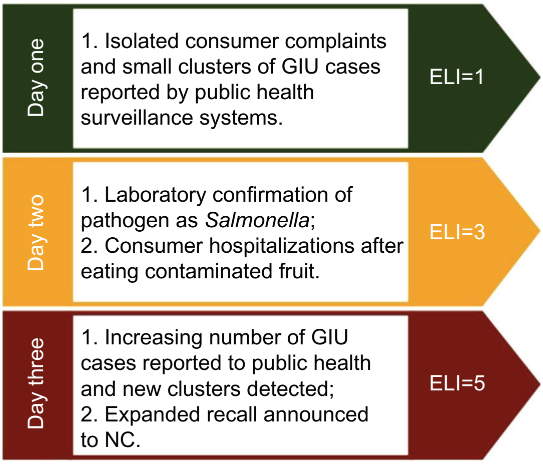
- **Consumer Complaints**. Consumer complaint calls to the NCDA&CS implicating a possible contaminated food product including date of the call, complainant county of residence, product implicated, retailer/manufacturer/food service provider implicated, complainant medical status (i.e., illness, hospitalization), diagnosis, and description of the complaint.

Visualization Dashboards. The visualization tolls in NCFEDA create visual representations of the results produced by the Analytics Engine for display on users' dashboards, increasing users' situational awareness by presenting information in a user-friendly interface. These dashboards are a set of dynamic graphical user interfaces that provide each agency-user and stakeholder with a COP and additional customized screens that, together, convey situational awareness to the various stakeholders.

5.9 PULLING IT ALL TOGETHER

In the following section we illustrate how NCFEDA works using a typical example of the progression of a foodborne illness event. The emerging event occurs over a 3-day period during which time a cluster of unspecified illness with symptoms of gastrointestinal problems is recorded by the system. This cluster may be an indication of an ongoing foodborne illness outbreak. Over the 3-day period, new information from various sources is provided daily to NCFEDA's Analytics Engine. As each new "event" is received, NCFEDA continuously evaluates this newly acquired information against knowledge previously acquired by the system to determine what information is "connected" and whether it belongs to the evidence set. NCFEDA also provides a measure of the likelihood that a foodborne illness or threat is occurring based on the strength of evidence contained in the evidence set, referred to as the ELI, or Evidence Likelihood Index. The 3-day simulation is summarized in Fig. 5.8.

Users connect to NCFEDA by accessing a login page, shown in Fig. 5.9, and then entering the name of the agency for whom they work, their user identification number, and a personal password to be verified by the system before any further access can be granted. The login page can also provide users with links to sites hosting relevant news related to food safety. For example, the login page provides a direct link to the latest recall issued by FDA, to the latest recall issued by USDA, and to an additional link to a site hosting recent food safety news.

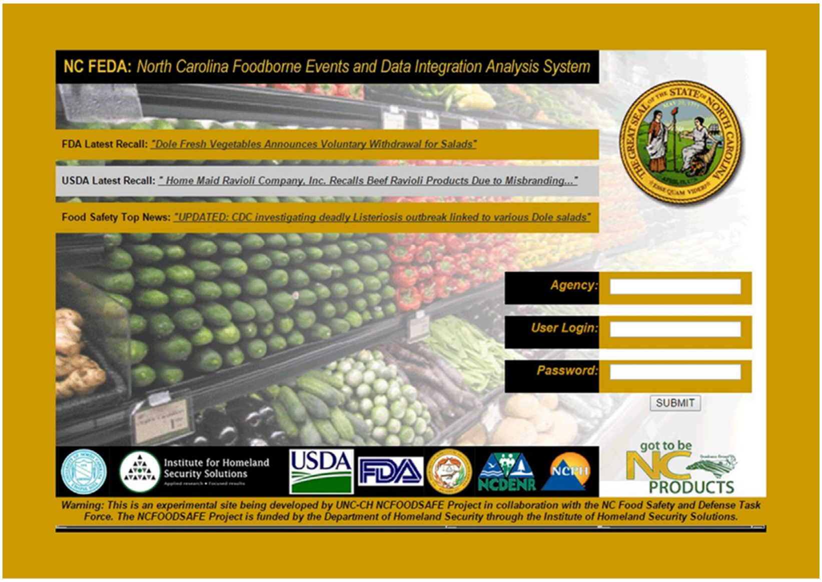


1. Isolated consumer complaints and small clusters of GIU cases reported by public health surveillance systems.

1. Laboratory confirmation of pathogen as *Salmonella*;
2. Consumer hospitalizations after eating contaminated fruit.

1. Increasing number of GIU cases reported to public health and new clusters detected;
2. Expanded recall announced to NC.

FIGURE 5.8 NCFEDA simulation timeline.



NC FEDA: *North Carolina Foodborne Events and Data Integration Analysis System*

FDA Latest Recall: *"Dole Fresh Vegetables Announces Voluntary Withdrawal for Salads"*

USDA Latest Recall: *" Home Maid Ravioli Company, Inc. Recalls Beef Ravioli Products Due to Misbranding..."*

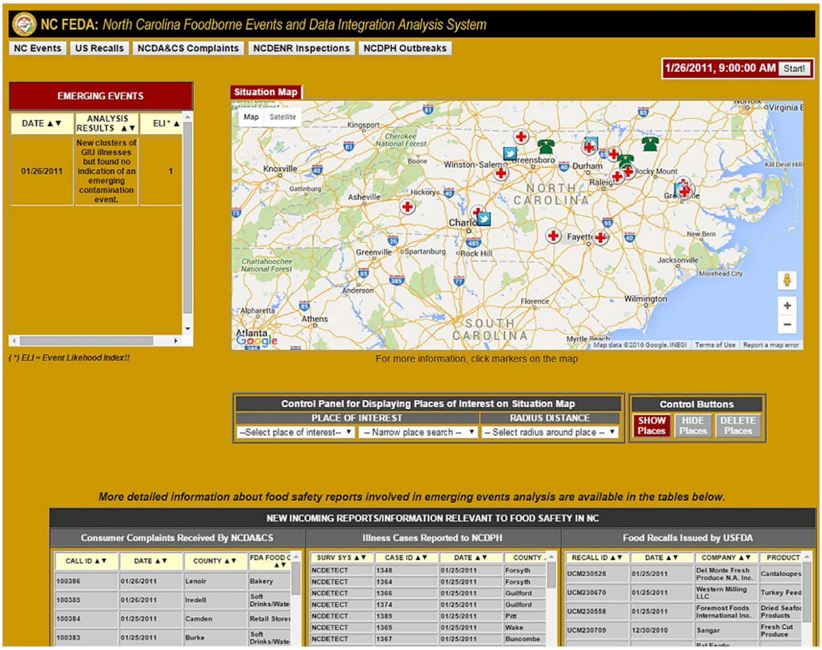
Food Safety Top News: *"UPDATED: CDC investigating deadly Listeriosis outbreak linked to various Dole salads"*

Warning: This is an experimental site being developed by UNC-CH NCFOODSAFE Project in collaboration with the NC Food Safety and Defense Task Force. The NCFOODSAFE Project is funded by the Department of Homeland Security through the Institute of Homeland Security Solutions.

FIGURE 5.9 NCFEDA user login page.

The NCFEDA screen shown as Fig. 5.10 corresponds to a COP of all food-related events occurring in North Carolina and is intended to be used by all agencies as their primary NCFEDA work screen. Its goal is to increase user situational awareness across all users of ongoing events. The key areas of the COP are as follows:

1. The *North Carolina Map* provides the primary view of the COP. The map offers visual cues as to where "events" are occurring to help users assimilate the spatial distribution of possible food contamination threats.
2. The *Emerging Events Table* keeps a continuous record of any possible emerging event identified by the NCFEDA engine. The *Emerging Events Table* is "pushed" to users via separate pop-up windows. The pop-up window contains a short description of the Analytics Engine result and the corresponding ELI rating at any point in time.
3. The *New Incoming Reports/Information Relevant to Food Safety in NC* area in the middle of the screen displays three tables that contain key data fields from the three primary sources—consumer complaints received by NCDA&CS, illness cases reported to NCDPH, and food recalls issued by USFDA.
4. Finally, the *NCFEDA Searchable Database of Food Safety Reports* table at the bottom of the screen provides an easy mechanism for users to query NCFEDA databases by typing words of interest on dedicated search areas attached to each field.



NC FEDA: *North Carolina Foodborne Events and Data Integration Analysis System*

EMERGING EVENTS

DATE ▲▼	ANALYSIS RESULTS ▲▼	ELI* ▲
01/26/2011	New clusters of GIU illnesses but found no indication of an emerging contamination event.	1

*(*) ELI = Event Likehood Index!!*

For more information, click markers on the map

Control Panel for Displaying Places of Interest on Situation Map

PLACE OF INTEREST		RADIUS DISTANCE
--Select place of interest-- ▾	-- Narrow place search -- ▾	-- Select radius around place -- ▾

More detailed information about food safety reports involved in emerging events analysis are available in the tables below.

NEW INCOMING REPORTS/INFORMATION RELEVANT TO FOOD SAFETY IN NC

Consumer Complaints Received By NCDA&CS

CALL ID ▲▼	DATE ▲▼	COUNTY ▲▼	FDA FOOD C ▲▼
100386	01/26/2011	Lenoir	Bakery
100385	01/26/2011	Iredell	Soft Drinks/Wate
100384	01/25/2011	Camden	Retail Stores
100383	01/25/2011	Burke	Soft Drinks/Wate

Illness Cases Reported to NCDPH

SURV SYS ▲▼	CASE ID ▲▼	DATE ▲▼	COUNTY .
NCDETECT	1348	01/25/2011	Forsyth
NCDETECT	1364	01/25/2011	Forsyth
NCDETECT	1366	01/25/2011	Guilford
NCDETECT	1374	01/25/2011	Guilford
NCDETECT	1389	01/25/2011	Pitt
NCDETECT	1369	01/25/2011	Wake
NCDETECT	1367	01/25/2011	Buncombe

Food Recalls Issued by USFDA

RECALL ID ▲▼	DATE ▲▼	COMPANY ▲▼	PRODUCT
UCM230528	01/25/2011	Del Monte Fresh Produce N.A. Inc.	Cantaloupes
UCM230670	01/25/2011	Western Milling LLC	Turkey Feed
UCM230558	01/25/2011	Foremost Foods International Inc.	Dried Seafoo Products
UCM230709	12/30/2010	Sangar	Fresh Cut Produce
		Pat Exotic	

FIGURE 5.10 NCFEDA Common Operating Picture with ELI = 1.

5.9.1 Day One: Reports of Gastrointestinal Illness

On Day One a small cluster of illnesses with general symptoms of gastrointestinal ulceration (GIU) is reported by the public health department to NCFEDA. These cases are represented as icons on the North Carolina map. Without confirmatory test results, or a more precise diagnosis, no pathogen can be identified. These GIU records are also displayed in both the NCDPH records table and the Searchable Database table appearing on the COP.

The receipt of information about this cluster can be viewed as a trigger in NCFEDA. When this cluster is reported to NCFEDA, it is compared against other data "events" by the system, such as recent food product recalls and consumer complaint calls, to determine whether these events can be linked to these cluster cases. The locations of hospital visits (marked by an iconic red cross) and complainant counties of residence (marked by the icon of a green telephone) are plotted in NCFEDA's North Carolina map.

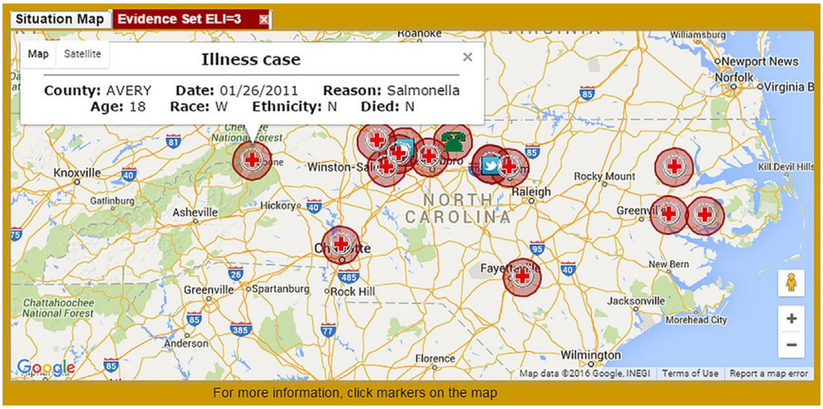
Using all information that is determined to be relevant and part of the evidence set, NCFEDA's engine computes the ELI rating for this situation and displays a short message in the Emerging Events Table to inform the user of its findings. As shown previously in Fig. 5.6, ELI ranges from 1 to 7 where a score of 7 indicates the highest likelihood. Without any confirmatory information to indicate an emerging foodborne illness outbreak, the ELI rating is computed to be 1 (ELI = 1). The emerging events for Day One are shown in Fig. 5.10.

5.9.2 Day Two: Lab Results and Consumer Complaint Calls

On Day Two a new cluster of illness cases is detected by public health officials. This new information is analyzed by NCFEDA to determine whether it is part of the evidence set, and thus increases the likelihood of an emerging event. Laboratory tests confirm that these cases are associated with the pathogen *Salmonella*. Given that a pathogen has now been positively identified, NCFEDA searches among both incoming and previously active recall notices to verify if any of those are also a result of contamination by *Salmonella*. NCFEDA also looks to see whether any products associated with these recalls are known to have been shipped to North Carolina. But no results are found.

NCFEDA also searches among incoming consumer complaint calls, and any complaints currently under investigation by NCDA&CS, for any illnesses confirmed to have been caused by *Salmonella*, or for implicated good products susceptible to this pathogen. NCFEDA searches its databases and locates a complaint call in which the caller reported being hospitalized because of possible consumption of contaminated fruit. Because fruit is susceptible to *Salmonella*—as documented by existing recall data—the NCFEDA relevance engine deduces that there is a possible emerging *Salmonella* contamination event and issues a warning to responsible agencies.

An emerging events map pops up in a separate window, as shown in Fig. 5.11, displaying the location of all events in the evidence set linked to this threat. When the user hovers the computer mouse over the map icons, detailed information about each reported case/complaint is displayed. In light



Illness case

County: AVERY Date: 01/26/2011 Reason: Salmonella
Age: 18 Race: W Ethnicity: N Died: N

For more information, click markers on the map

FIGURE 5.11 Evidence set for ELI = 3 on day two.

of the confirmatory evidence, the ELI for the event is computed to be 3 (ELI = 3) by the Analytics Engine and appears in the corner of this pop-up window. A screen shot of NCFEDA with ELI = 3 is shown in Fig. 5.11.

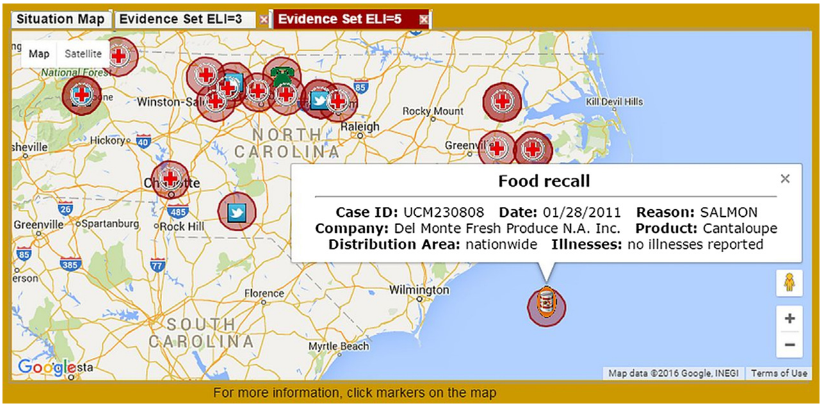
5.9.3 Day Three: Food Recall Issued

On Day Three, an increasing number of new illness cases are reported to NCFEDA from the public health system and new clusters are detected. Because we do not have personalized information about patient identity due to government HIPAA regulations and other privacy concerns, NCFEDA cannot deduce an exact relationship among patient cases beyond same county of residence.

However, the arrival of a new recall notice from FDA expanding the area of distribution of recalled cantaloupe to the state of North Carolina is thought to be linked to the cluster of *Salmonella* cases. The cantaloupe recall had previously been restricted to three states on the west coast of the United States. NCFEDA recognizes that cantaloupe is a fruit and that the *Salmonella* pathogen causing this recall is also the same pathogen causing a reported illness and hospitalization as reported by a consumer complaint call.

The emerging events map appears in the COP displaying all events linked to this threat, which now includes a recall notice (shown in the gray box on the bottom left corner of the pop-up window viewed in Fig. 5.12). The new ELI rating has been elevated to a score of 5 (ELI = 5) because more connections among the data have been discovered and confirmed by the relevance engine in the Analytics Engine.

Now that the threat level has been elevated to ELI = 5, indicating a high likelihood of an emerging threat, a warning message is pushed to users that



Food recall

Case ID: UCM230808 **Date:** 01/28/2011 **Reason:** SALMON
Company: Del Monte Fresh Produce N.A. Inc. **Product:** Cantaloupe
Distribution Area: nationwide **Illnesses:** no illnesses reported

For more information, click markers on the map

FIGURE 5.12 Evidence set for ELI = 5 on day three.

includes a complete set of information about the relevant events and possible threat including the suspect food product (cantaloupe) and the pathogen (*Salmonella*). The evidence set with ELI = 5 is shown in Fig. 5.12.

5.10 FUTURE TRENDS

Public agencies and private companies alike are working hard to adopt methods of data science in the interest of food safety. In 2013 the US FDA awarded a $50 million federal contract to Dynamics Research Corporation (now part of Engility, Inc.) to help move the agency into the big data era. And, given the prevalence of mobile apps and smart phones, it is not surprising that a number of efforts are mining social media data, much as Google did for influenza. The City of Chicago Department of Public Health is working with the Smart Chicago Collaborative to develop mobile applications that monitor Twitter for possible food poisoning references. The New York City Department of Health and Mental Hygiene is working with Columbia University to review restaurant-goer comments on Yelp for possible clues to a food contamination event or outbreak.

Many private sector companies are also contributing data-driven technologies and analytics to support the push to an integrated, data-driven approach to food safety. IBM recently announced a new predictive analytics technology that the company claims is capable of identifying contaminated products "within as few as 10 outbreak case reports." Like NCFEDA, the goal of IBM's technology is to reduce the time required to identify the likely contamination sources by days or even weeks. Predictive analytics and other algorithms look through petabytes of grocery store food sales data from retailers and distributors in search of patterns and relationships that may indicate contamination. Visualization techniques link the data to geographical information to connect suspected contaminations with clinical and lab reports, as well as other data. A pilot is being conducted with the Department of Biological Safety at the German Federal Institute. The project will process information from 1.7 billion supermarket items sold in each country.

These developments are the first steps toward the integration of the food chain within an Internet-of-Things (IoT) environment. Situational awareness of complex and lengthy food chains is currently constrained by difficulties associated with the timeliness of data collection and fusion—in fact, much data is still manually entered into systems. In an IoT environment, end-to-end data needed for both surveillance and response can be autonomously and automatically collected using the sensor-enabled network environment of the IoT. In this, hopefully, near-term future, all stakeholders in the supply chain from the farm to the consumer will have sensors and systems in place to monitor both the food as it moves through the food chain and the health data and laboratory data needed to identify and confirm foodborne illness—and most importantly the connections between them that enable the incidence of illness to be linked immediately with the offending products in the food chain and with its source.

5.11 FURTHER INFORMATION

Further information about the application of data science in food safety can be found in several disciplines. *Food Safety Magazine* (http://www.foodsafetymagazine.com/) offers many articles about the critical challenges of food safety and the application of new data-driven and digital tools to address those challenges. The CDC website offers up-to-date information about the current state of food safety and capabilities of surveillance and response (http://www.cdc.gov/foodsafety/fsma/index.html). Their website provides basic information about the current responsibilities and procedures for managing a foodborne disease outbreak. Two studies published by the CDC address the state of food safety in the United States. The CDC's annual food safety progress report measures foodborne illnesses from nine key germs and is produced from data compiled by the FoodNet. The National Outbreak Reporting System (NORS) publishes an annual summary of foodborne outbreaks reported to CDC by state and local health departments.

ACKNOWLEDGMENTS

This research was funded by the US Department of Homeland Security through the Research Triangle Institute—Institute for Homeland Security Solutions (IHSS) under Contract HSHQDC-08-C-00100.

REFERENCES

FoodNet Report Shows Mixed Bag of Foodborne Illness Trends, May 15, 2015, <http://www.foodsafetymagazine.com/news/foodnet-report-shows-mixed-bag-of-foodborne-illness-trends/>.

Beach, C., CDC declares Chipotle *E. coli* outbreaks over; cause unknown, Food Safety News, February 1, 2016, <http://www.foodsafetynews.com/2016/02/cdc-declares-chipotle-e-coli-outbreaks-over-cause-unknown/#.VrjH4v7bKig>.

Dube, L., LAbban, A., Moubarac, J.-C., Heslop, G., Ma, Y., 2014. A nutrition/health mindset on commercial big data and drivers of food demand in modern and traditional systems. Ann. N. Y. Acad. Sci. USA 1331, 278–295.

Doinea, M., Boja, C., Batagan, L., Toma, C., Popa, M., 2015. Internet of things based system for food safety management. Inf. Econ. 19 (1), 87–97.

Greis, NP, Nogueira, M, MacDonald, P, Wilfert, R., NCFEDA North Carolina Foodborne Events Data Integration and Analysis Tool: A New Informatics Tool for Food Safety in North Carolina, 2011. Prepared by RTI International–Institute for Homeland Security Solutions under contract HSHQDC-08-C-00100.

IBM. The four V's of big data, 2014. <www.ibmbigdatahub.com/infographic/four-vs-big-data>.

Keller, M., Blench, M., Tolentino, H., et al., 2009. Use of event-based unstructured reports or global infectious disease surveillance. Emerg. Infect. Dis. 15 (5), 689−695.

Manyika, J., Chui, M., Brown, B., et al., Big data: the next frontier for innovation competition, and productivity, 2011, <http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_in_innovation>.

Piramuthu, S., Zhou, W., 2016. RFID and Sensor Network in the Food Industry. Wiley Blackwell.

Strawn, L. K., E. W. Brown, J.R.D. David, H.C. Den Barker, P. Vangay, F. Yiannas, and M. Wiedmann, Big Data in Food, Food Technol., 2015, 69(2), pp. 42−49, Retrieved at: <https://www.researchgate.net/publication/272238175>.

Wang, Y., Yang, B., Luo, Y., He, J., Tan, H., 2015. The Application of Big Data Mining in Risk Warning for Food Safety. Asian Agric. Res. 7, 83−86.