Application of Information Technology ■

Design of a National Retail Data Monitor for Public Health Surveillance

Michael M. Wagner, MD, PhD, J. Michael Robinson, Fu-Chiang Tsui, PhD,
Jeremy U. Espino, MD, William R. Hogan, MD

Abstract The National Retail Data Monitor receives data daily from 10,000 stores, including pharmacies, that sell health care products. These stores belong to national chains that process sales data centrally and utilize Universal Product Codes and scanners to collect sales information at the cash register. The high degree of retail sales data automation enables the monitor to collect information from thousands of store locations in near to real time for use in public health surveillance. The monitor provides user interfaces that display summary sales data on timelines and maps. Algorithms monitor the data automatically on a daily basis to detect unusual patterns of sales. The project provides the resulting data and analyses, free of charge, to health departments nationwide. Future plans include continued enrollment and support of health departments, developing methods to make the service financially self-supporting, and further refinement of the data collection system to reduce the time latency of data receipt and analysis.

■ **J Am Med Inform Assoc.** 2003;10:409–418. DOI 10.1197/jamia.M1357.

The rapid, early detection of disease outbreaks has become a national priority and an emerging field of research.[1] Kaufmann et al.,[2] after analyzing several bioterrorism scenarios, concluded that "delay in starting a prophylaxis program is the single most important factor leading to increased loss of life and health." The urgency of the problem is reflected in an explosion of research on new computer-based disease surveillance systems.[3–12]

Wagner et al.[1] recently discussed four possible ways to improve the earliness of outbreak detection. Prominent among them is the use of new types of surveillance data that track sales of over-the-counter (OTC) health care products such as cough syrup that are purchased early in the course of illness by sick individuals for the symptomatic treatment of illness. For common syndromes such as upper respiratory illness ("flu") and asthma, the sick are more likely to self-treat with OTC health care products than to see a physician. In a random digit dialing survey of 1,505 individuals conducted by the Consumer Healthcare Products Association in 2001, 72% (505 of 701) of those with cough, cold, "flu," or sore throat in the previous six months treated themselves with an OTC health care product. Importantly, in 42%, purchase or use of OTC preparations was their first action, and in 34%, self-observation was the first action.[13] In less than 9% was seeking professional medical care the first action. For the symptom of headache, the findings were even stronger (81% self-medication and 52% self-observation, respectively, with 4% seeking a physician as the first action).[13] In a population-based survey of 42,333 adults in the province of Ontario, Canada, only 14% of adults with upper respiratory tract infections visited a doctor, whereas 76% engaged in self-care with OTC medications.[14] Sales of OTC health care products have attractive characteristics for outbreak detection. In the United States, use of scanners and Universal Product Codes at checkout counters in retail industry stores facilitates routine collection of such data in real time. A small number of national companies own the majority of retail outlets that sell such products, and these corporations integrate their sales data at the national level in near to real time. For these reasons, the technical effort and cost to obtain these data for public health surveillance are comparatively low.

Preliminary studies suggest that sales of OTC health care products can be used for the early detection of outbreaks,[15–17] yet research progress has been slow due to the difficulty in obtaining data to adequately test the hypothesis in a sufficiently large number of sizable outbreaks. The first such study, reported in 1979, showed an association between influenza B activity and purchases of cold remedies for a single outbreak.[17] A recent study of 18 seasonal outbreaks in children showed that sales of pediatric electrolyte solutions correlated strongly with hospitalizations of children for gastrointestinal and respiratory illnesses and usually preceded the hospitalizations by more than two weeks.[15] Unpublished research shows increases in sales of OTC "flu" and cough preparations preceding increases in influenza activity as measured by outpatient billing diagnoses.[16] Other research (conducted by our laboratory and others) simulates

Affiliations of the authors: The RODS Laboratory, Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania (MMW, F-CT, JUE, WRH); Global Strategic Solutions, Cincinnati, Ohio (JMR).

Supported by Pennsylvania Department of Health Award number ME-01-737. The authors are indebted to Judith Hutman, Joel Hersh, William Yasnoff, James Seligman, Julie Gerberding, MD, Information Resources, Inc., ACNielsen, Procter & Gamble, and the participating retailers. The authors thank the anonymous reviewer who provided detailed comments.

Correspondence and reprints: Michael M. Wagner, MD, PhD, The RODS Laboratory, Center for Biomedical Informatics, University of Pittsburgh, Suite 550, 100 Technology Drive, Pittsburgh, PA 15219; e-mail: <mmw@cbmi.pitt.edu>.

Received for publication: 03/06/03; accepted for publication: 05/13/03.

the effects of outbreaks of different magnitudes and time courses on sales of OTC health care products (and other types of data) to understand the smallest increase in sales that would be detectable above background levels of sales.[18,19]

A rationale for not monitoring nonspecific data such as sales of OTC health care products as a means of alerting has been presented by Broome et al.,[20] who raise concerns that data with low specificity will require an increment of hundreds of units over baseline levels before detection would occur, and that the cost of investigating false alarms may be too high.

Because of the threat of bioterrorism, the relative ease and low cost of monitoring sales of OTC health care products, and the accumulating evidence in favor of monitoring sales of OTC health care products, the Real-time Outbreak and Disease Surveillance (RODS) laboratory worked with the retail industry to build a retail data monitor for the Commonwealth of Pennsylvania. The authors quickly realized that the same effort could create a National Retail Data Monitor. This report describes the design and current status of the National Retail Data Monitor.

Background

Public Health Surveillance

The role of public health surveillance in general is to collect, analyze, and interpret data about biological agents, diseases, risk factors, and other health events and to provide timely dissemination of collected information to decision makers.[21–23]

Public health surveillance traditionally relies on manual operations and off-line analysis. Key components of such activities include reporting by clinicians and laboratories on notifiable diseases and dependence on astute clinicians to notice and report suspicious clusters of cases to health departments. The utility of these traditional disease surveillance approaches for the early detection of outbreaks is limited severely by delays in obtaining and analyzing the data, by the reliability of the antiquated "reportable disease" system, and by the delays that result from clinicians waiting to make reports until the diagnosis is confirmed by definitive testing.

Newer automated surveillance systems and other monitoring techniques have been developed to detect epidemics more rapidly, utilizing routinely collected and often prediagnostic data. These efforts have been pioneered by regional projects located in New York City, Washington, DC, Utah, Indiana, Seattle, Pennsylvania, and other locations.[3–11]

Universal Product Codes

Monitoring of retail data would be impossible if not for the existence of a standard coding system for retail products. Retailers and manufacturers in North America use Universal Product Bar codes ("UPC codes") in their data systems. UPC codes are 12-digit numbers used by manufacturers to uniquely identify themselves and their products worldwide. UPC codes consist of black and white bars and a number. Retailers scan products at the cash register to detect these symbols and thereby collect data in real time about their sales. UPC codes originate from the Uniform Code Council, Inc. (UCC), a not-for-profit standards organization. The mission of the UCC is to establish and promote multi-industry standards for product identification and related electronic communication. The UCC administers the UPC codes and provides a range of standards and business solutions for over 250,000 member companies doing business in 25 major industries. A manufacturer pays an annual fee to join the council. In return, the council issues the manufacturer a six-digit *manufacturer identification number* and provides guidelines on how to use it. The manufacturer identification number is the first six digits of the UPC number, and the next five digits are the item number followed by a one-digit field used as a check digit to ensure UPC number accuracy. A person employed by the manufacturer, called the *UPC coordinator*, is responsible for assigning item numbers to products, making sure the same code is not used on more than one product, retiring codes as products are removed from the product line, and other duties. In general, every item the manufacturer sells, as well as every size package and every repackaging of the item, needs a different item code. So, a 4-ounce bottle of *Tylenol Children's Cold Great Grape Flavor—Alcohol Free, Aspirin Free*, is assigned a different item number from an 8-ounce bottle of *Tylenol Children's Cold Great Cherry Flavor—Alcohol Free, Aspirin Free*.

Regions that Monitor OTC Sales

Surveillance systems developed by the Johns Hopkins Applied Physics Laboratory in the National Capital Area [24] and by the New York City Department of Health [25] collect and analyze sales of OTC health care products for public health surveillance. The New York City project obtains data in nine product categories from a single local chain. The National Capital Area project obtains data from two large national chains, but the data are limited to the National Capital Area. These efforts have similarities to the current project in the type of data being collected and the purpose of the systems. A key difference of the National Retail Data Monitor is that it receives nationwide data from the national data warehouses of retailers.

Design Objectives

The purpose of the National Retail Data Monitor is to collect and analyze sales of OTC health care products to detect outbreaks of disease. Because of the relative lack of specificity of OTC health care products as indicators of disease, the types of outbreaks that the current system is intended to detect are those that involve a relatively large proportion of individuals in the geographic region being analyzed (in this case, zip code). Currently, the methods used in the National Retail Data Monitor also are focused on detecting sudden outbreaks, although this current specialization is not a fundamental limitation of the approach. In particular, the current niche for the National Retail Data Monitor is early detection of a mass exposure of a large number of people through contamination of the air, food, or water (a *cohort exposure*). Soon after such an exposure, the cohort will become symptomatic, and, depending on the symptoms, may begin self-treatment and then either recover or seek medical care. If the cohort is large enough, sales of OTC health care products will increase significantly above the normal, background level of sales.

A key requirement for early detection is to minimize the time latency between time of purchase and time when data become available for analysis. Time intervals as small as hours can make a difference when a large cohort is exposed to rapidly progressing diseases such as anthrax.[1,2,26] Therefore, a design objective is to collect and analyze the data in as near as real time as possible, with at most a day's delay from time of sale. A second requirement is completeness of sales data collection, which is important for both early detection and sensitivity to smaller outbreaks. The number of independent stores not participating in centralized data collection limits available data to less than 100% coverage of sales. The project's specific objective, therefore, is to collect sales data sufficient to achieve at least 70% market share nationally and to achieve 70% share in each of the 20 largest urban regions (Fig. 2). A third requirement concerns the need for precise spatial information in outbreak detection. Ideally, one would receive data that support spatial analysis at the level of individual store locations, or at least by the zip codes of stores.

Additional desiderata include collection of supplemental data—for example, indicating when retailers feature promotions or how day of the week affects local sales volumes. Because UPC codes change, a system for maintaining UPC code masters and mappings from UPC codes to analytic categories also is a requirement. Finally, it is essential to create an effective link between the surveillance data, public health review, and response. If the information collected by the National Retail Data Monitor is not reviewed by the intended users of the system (local, state, and federal public health authorities) and does not influence response (e.g., quarantine and medical treatment), the National Retail Data Monitor cannot have an impact on preventing morbidity and mortality and therefore will have no utility beyond supporting retrospective research.

System Description

This section describes methods used in the National Retail Data Monitor to receive data from retailers, analyze and process the data, and define and maintain product categories.

The "Data Utility" Model

The National Retail Data Monitor is a *data utility* for the collection, analysis, and distribution of data on sales of OTC health care products and provision of the data to health departments. This approach reduces the resources required for health departments to monitor sales of OTC health care products. Without a data utility, each health department would have to negotiate data-sharing agreements with many retailers, work with the retailers to understand the data, build systems to collect and analyze the data, and maintain UPC master medication lists as new product codes are assigned. Few health departments have such resources, and retailers have already expressed resistance to this approach. A national data-utility approach also is efficient because most large urban population centers cross jurisdictional boundaries of health departments. Without a centralized approach, each health department would either collect data only for its own jurisdiction (and thus have an incomplete picture of the health of the region) or redundantly collect data for overlapping nearby jurisdictions.

Data Agreements with Retailers

Working with Information Resources, Inc. (IRI) and ACNielsen, the major syndicated data analysts for U.S. retail industries, the authors identified the significant market share leaders for OTC health care products in the country. The analysis showed that the top five national retailers account for approximately 48% of sales of OTC health care products nationwide, the top 10 retailers for 65%, and the top 20 retailers for 76% of sales. Although the effect of incomplete monitoring of OTC sales on sensitivity of outbreak detection is not fully understood, 65% or 76% sampling far exceeds the sampling efficiency of most public health surveillance schemes.[6,27,28] The process of determining which retailers would be ideally suited for participation involved merging and analyzing a significant amount of industry knowledge and market share data. Figure 1 shows the market coverage provided by the top four national drugstore chains for the 20 most populous U.S. cities. In Washington DC, for example, two retailers account for approximately 75% of all sales of OTC health care products. The chart also identifies the market share of the largest nonnational retailer in the region. In New York City, a large local retailer accounts for an estimated 25% of market share. That retailer is already providing OTC data for surveillance to the New York City Department of Health, and the additional data provided by only two national chains could bring retail data monitoring for the New York City area to over 80%.

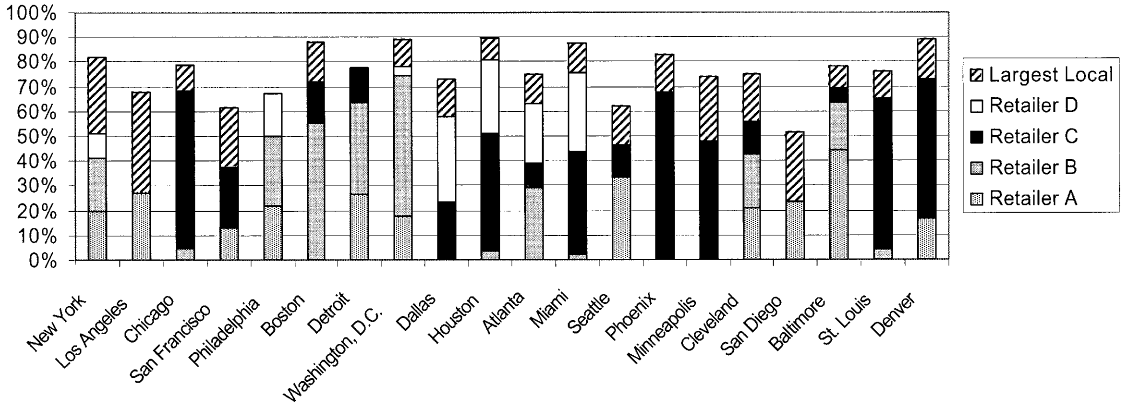


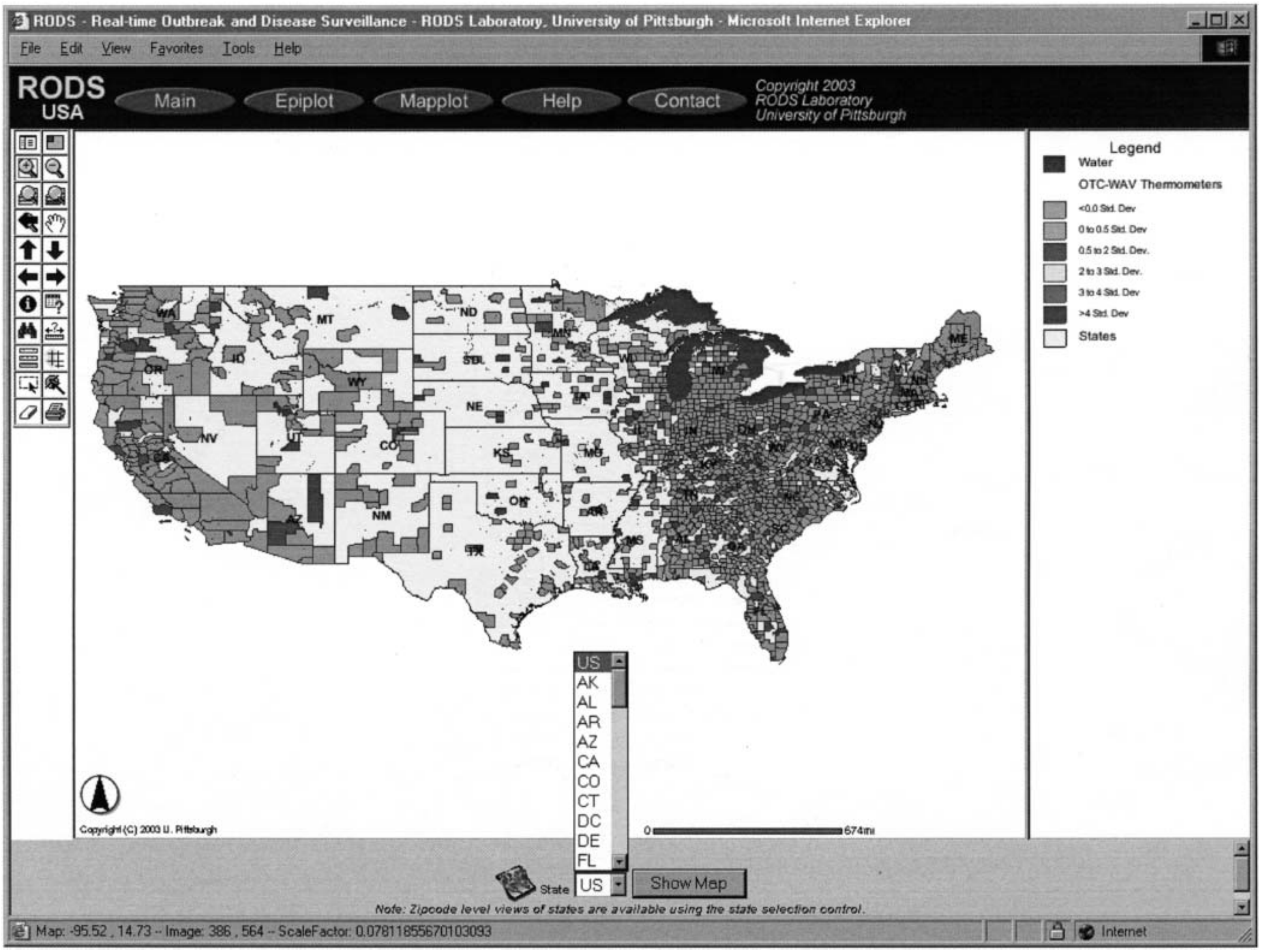
Figure 1. Market share of over-the-counter (OTC) health care products of the four largest national retailers for the 20 most populated metropolitan areas in the United States. Also shown is the market share of the largest local retailer for each metropolitan area. The combination of the four national retailers and the largest local retailer provides 50% to 90% market share coverage for all cities. The four retailers do not necessarily correspond to retailers participating in this project. Source: Estimated from industry statistics using ACNielsen Scantrack, IRI Infoscan, and Racher Press Chain Drug Research.

Once the authors identified the retailers with sufficiently large market share to make a meaningful surveillance impact, they asked the retailers to provide their data. Discussions with the retailers were facilitated by introductions made by the IRI and ACNielsen corporations and by letters provided by the Pennsylvania Department of Health and by the Director of the U.S. Centers for Disease Control and Prevention (CDC), Dr. Julie L. Gerberding. Currently, four retailers are providing OTC sales data. For confidentiality reasons, the authors cannot disclose their names, but they comprise approximately 10,000 individual pharmacies representing 23% market share nationally (Fig. 2). Additionally, three other national retailers have agreed to provide data for their 8,000 pharmacies, which will bring the total to approximately 18,000 pharmacies representing approximately 33% of total U.S. OTC health care product sales.

The retailers and the University of Pittsburgh have executed data sharing agreements that permit the University to redistribute the data to those Departments of Health (DOHs) that are participating in the project as well as to the CDC. The agreements stipulate that the data must be aggregated with similar other data by zip code. Each participating DOH is permitted to access only those portions of the data that are relevant to the jurisdiction of that DOH. The DOHs may also review aggregated data using the RODS monitoring system interfaces (described below and in another article in this issue of *JAMIA*). Several retailers expressed concerns that their data not be shared with competitors. These concerns were satisfied by a clause in the data-sharing agreement that prohibits the aggregated data from being shared back to any retailer. Finally, the companies make no warranties of any kind. Retail data that are being collected by the National Retail Data System are not governed by Health Insurance Portability and Accountability Act (HIPAA) regulations because they are not Personal Health Information —the data describe the quantity of product that the stores are selling per day, not what any individual is buying.

Data Feeds

The requirements for the data feeds from the retailers were minimal time latency and data content with a level of



State US

Note: Zipcode level views of states are available using the state selection control.

Map: -95.52 , 14.73 -- Image: 386 , 564 -- ScaleFactor: 0.07811855670103093

Figure 2. National Retail Data Monitor: Sales of thermometers by county on May 17, 2003. This map summarizes data received from approximately 10,000 stores belonging to four national retail chains. The color *green* (the lighter shade) indicates that sales of thermometers on that day were within 0.5 standard deviations of expected, and the color *blue* (the darker shade) indicates sales were between 0.5 and 2.0 standard deviations.

granularity of sales by UPC number, by day, and by store. Due to varying capabilities and preferences, different retailers are at different stages of evolution toward the ideal goal of UPC level of granularity, and this heterogeneity is reflected in the user interface shown in Figure 4. The initial two retailers currently transmit daily counts of OTC sales in five product categories: pediatric electrolytes, cough-and-cold products, thermometers, stomach remedies, and antifever medications. These retailers have agreed to create UPC-level feeds. The third and fourth retailers are transmitting a finer-grained set of categories that contain pediatric-adult distinctions. These retailers have agreed also to create UPC-level feeds. The 18 categories can be aggregated for analytic purposes into the original five categories and are so aggregated in the 'Day Old' categories shown in Figure 4. The fifth, sixth, and seventh retailers—who have agreed to participate but have not yet built data interfaces—will build UPC-level feeds.

Every day, three retailers transfer the previous day's sales data by secure file transfer protocol over the Internet by 3 PM EST. The fourth retailer transmits data approximately every two hours around the clock. The combined data volume currently approximates 14 megabytes (for four retail chains) per day in comma-separated format, although this number will increase when the feeds are converted to UPC-level granularity.

Data Storage and Security

The data are stored in a secure, firewall-protected facility that has been described previously.[4] The facility meets high availability and capacity requirements. The design assumes that, in the event of a bioterrorism-related outbreak anywhere in the country, thousands of users at health departments will query the site repeatedly and continuously. The technical approach therefore involves (1) fault tolerant network and server configurations; (2) utilization of hardware that supports future mirrored operation at a second site; (3) server clustering that facilitates fault tolerance and load balancing; and (4) early creation of a second mirrored site to ensure against loss through fire or other building event.

Mapping Universal Product Codes into Product Categories

Many of the distinctions represented by UPC codes have little importance in public health surveillance. It is unlikely that the sale of a 4-ounce bottle of grape-flavored cough syrup is a better indicator of cough than the 8-ounce bottle of cherry-flavored syrup. Moreover, sales of individual products have high variability due to marketing factors such as coupons, discounts, and shelf placement in the store. Therefore, the authors developed a method to aggregate products into analytic classes (*product categories*) for detection. Currently, two different product categorizations are used because there currently is a transition from reporting the original five categories (mentioned earlier) to a new set of 18 more finely grained categories (Table 1). The authors (as clinicians) constructed the 18 categories by reviewing all OTC health care products available on the market across the United States (as defined by the ACNielsen current listing of products). The first step was to eliminate products of no apparent value for outbreak detection such as *sports cream analgesics*. There then remained 7,554 unique OTC health care products (each represented by a unique UPC). Next, each remaining product was assigned to exactly one of the 18 categories by using information that was either explicitly available as a coded characteristic of the product or derivable by simple lexical processing methods from the information provided by ACNielsen. If the latter automated assignment did not work, manual assignment was done based on the same information sources. The assignment process guaranteed that 18 categories were exhaustive and mutually exclusive (i.e., all 7,554 health care products of interest are included, but each only maps to one category). As a result, combinations of the 18 categories can be merged into broader categories (e.g., counts of pediatric and adult cough liquids can be accurately merged into counts of *cough liquids* by simple addition).

Table 1 ■ OTC Health Care Product Analytic Categories

Category	UPC Count
Cold relief, adult, liquid	709
Cold relief, adult, tablet	2,467
Cold relief, pediatric, liquid	323
Cold relief, pediatric, tablet	74
Cough syrup, adult, liquid	592
Cough syrup, adult, tablet	32
Cough syrup, pediatric, liquid	24
Nasal product internal	371
Throat lozenges	364
Antipyretic, pediatric	274
Antipyretic, adult	1,340
Bronchial remedies	43
Chest rubs	78
Diarrhea remedies	165
Electrolytes, pediatric	75
Hydrocortisones	185
Thermometer, pediatric	125
Thermometer, adult	313
TOTAL	7,554

Maintaining UPC-to-category Mappings

Product changes ("new improved formula") result in the assignment of new UPC codes to products, creating a need to update UPC-to-category assignments. Store brands are a particularly difficult instance of this problem, because the retailer, as the manufacturer of the product, has no real need to share an "internal" UPC code outside of the corporation. Because of these considerations, updating UPC-to-category mappings is an important requirement. As a general rule, the majority of UPC code changes in the OTC drug industry occur in August and September when the industry launches new initiatives and releases new products each year. Similar to the technology industry hosting the annual COMDEX trade fair, the drug industry hosts annual food and drug buying "shows" in which the majority of the upcoming merchandising year's buying decisions are made and contracts executed. Fortunately, ACNielsen and IRI monitor UPC-code changes as part of their normal business, and the authors plan to incorporate a similar process for the purpose of ongoing maintenance. The current approach for maintaining UPC data integrity over time compares the existing UPC map with a new universal map from an industry partner on at least a monthly basis to identify new codes. For the future, several experts who represent top manufacturers, retailers, syndicated data analysts, and industry standards governance organizations—including the Uniform Code Council, which manages the distribution of UPCs and all electronic commerce standards in the industry—have agreed to make their expertise available to design a process model that will work throughout the industry.

Integrating Real-time and Day-old Data

The existence of two different time latencies in the data feeds (three retailers send data at 3 PM about the previous day's sales, and one retailer sends data every two hours) presented a design challenge. Either the real-time data could be held in abeyance until the following day (a lowest common denominator approach), or the real-time data could be handled as a separate data source, facilitating much earlier analysis, with the potential benefit of earlier detection. Because detecting cohort exposures as early as possible is important, the authors chose to treat the real-time data as a separate data stream. Analyzing the data in this manner asks the question, *Can the monitor detect anything unusual about* ***today****, using only data from the real-time stores?* The authors have not yet implemented detection algorithms on the real-time data stream, although the approach will be identical to how the day-old data are analyzed. Currently, detection algorithms operate on the day-old data stream, which includes data from all four retailers. This analysis asks the question, *Can the monitor detect anything unusual about* ***yesterday****, using data from all the stores?*

User Interfaces and Routing Data to Health Departments

Health departments can access data collected by the National Retail Data Monitor in two ways: (1) through secure Web interfaces that provide temporal and geographic plotting capabilities and (2) through secure raw data feeds that end-user sites can analyze using their own surveillance software or analytic packages. Most health departments use the Web interfaces. Any health department may obtain user accounts for its staff by contacting <nrdmaccounts@cbmi.pitt.edu>. A staff member will be asked to execute a simple data use agreement that limits the use of the data to public health surveillance purposes. Five entities receive raw data on a daily basis at 5 PM EST. They are New York State, New York City, New Jersey, the ESSENCE/JHAP project in the National Capital Area (comprising Washington, DC, Virginia, and Maryland), and the CDC.

Currently, users are instructed to log into the Web interface once a day at 5 PM local time when the data and analyses from the previous day's sales of OTC health care products become available.

Maps

Users can visualize sales of OTC health care products on maps to detect spatial patterns such as clusters of zip codes with increased sales or linear clusters of zip codes with increased sales. Figure 3 is an example of a map generated by the National Retail Data Monitor. There are currently five such maps (one per product category) per geographic region per day for review. When the monitor is converted to include 18 categories and augmented to map analyses of real-time data, there potentially will be 36 maps to review per day. This number is large. Although some of these product categories may not be of interest to public health and can be eliminated, a more general solution that the authors plan to implement in the near future will be to screen the maps automatically with spatial scan statistics to identify those with anomalies suggesting a need for human review.[29–31]

The maps represent a novel approach to presenting surveillance data. They plot for each zip code—using the colors green, blue, yellow, orange, and red to indicate increasing levels of concern—how "unusual" sales were for the day in question relative to historical patterns of sales for that zip code. In particular, the colors represent the number of standard deviations by which the observed sales of a product category in a zip code deviate from the expected counts. In presenting the data in this fashion, the map serves as a device to focus the user's attention on the degree(s) of anomaly. A user can quickly spot whether the map is predominantly green with a scattering of blue zip codes as would be expected, or whether there are confluent or linear patterns of blue, yellow, orange, or red indicating "unusual" sales activities. The map monitor computes the number of standard deviations relative to a residual signal that has zero mean and constant variation after removal of weekly and longer trends in the data by wavelet transformation. This procedure is intended to produce a "normalized" map that is very sensitive to sudden increases in product counts as would be the case in a medium- to large-scale contamination of the air, food, or water. Alternative transformations of the data are possible using different signal processing approaches focused on detecting more gradual increases. The authors discuss below the reasons that they do not attempt to plot population- and sampling-adjusted mapping.

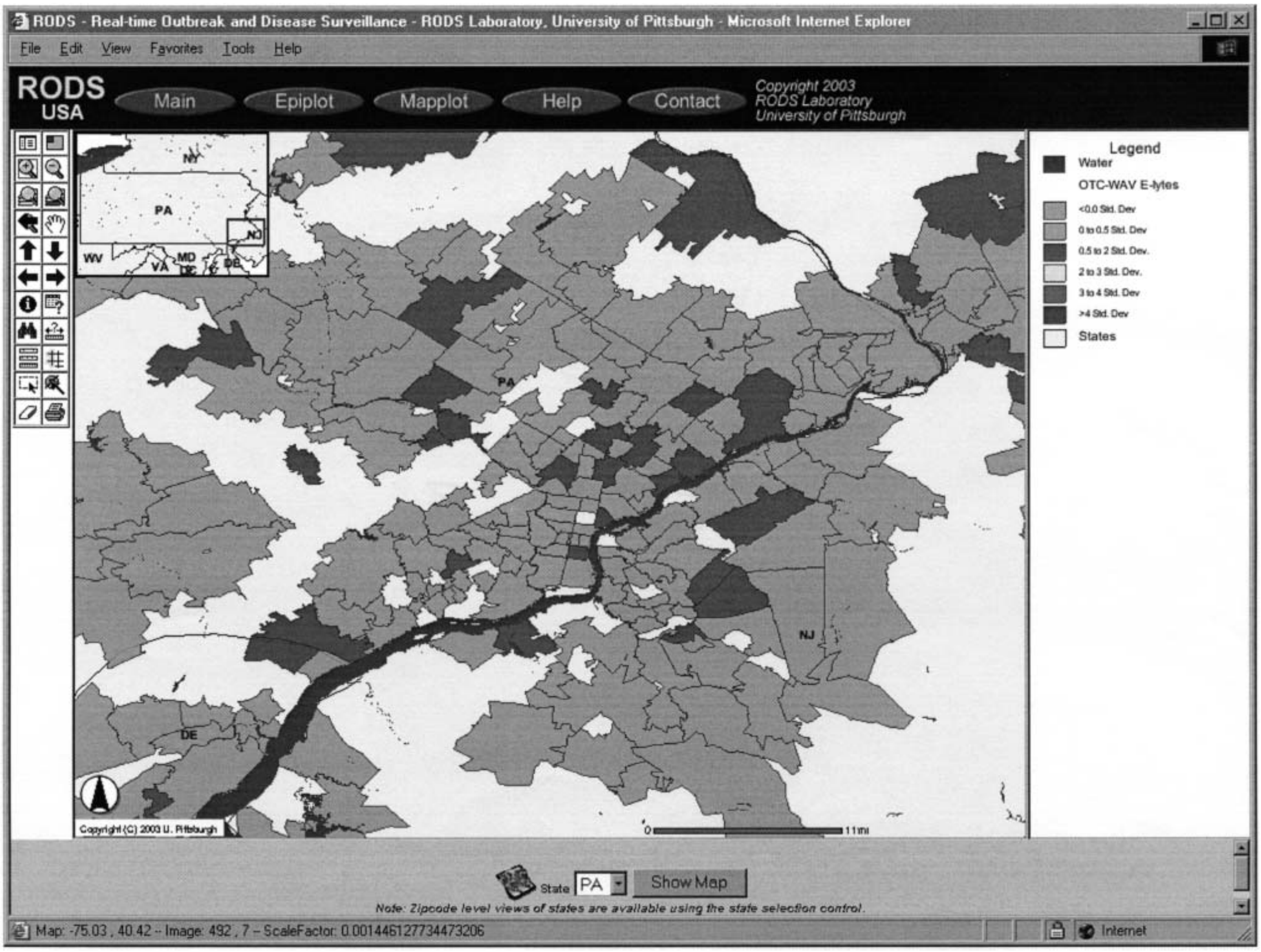
Normalization Issues

Because populations and market share coverage for sales of OTC health care products differ between zip codes, plotting raw sales counts is uninformative. The raw sales data for a zip code in midtown Manhattan in the absence of any epidemic might vastly exceed sales in a rural zip code in New Jersey even if every individual in the New Jersey town were to be ill and make purchases.

Epidemiologists are accustomed to adjusting for sampling rate and population differences by plotting the incidence of disease per 100,000 population. Transformations of sales data that approximate this metric would have the desired property of familiarity. If one assumes that the purchase of an OTC health care product is a reasonable indicator of disease (on a one unit to one case basis), one could in theory estimate disease incidence by the following equation that normalizes the raw counts by both market share and population:

$$\begin{aligned} incidence = (&raw\ daily\ sales\ in\ zip\ code/ \\ &market\ share)/population\ in\ zip\ code \end{aligned}$$

Thus, if the market share being monitored were only 50% in one zip code, after normalizing by market share, the adjusted counts would be double the observed counts from that zip code, making the counts comparable to those from a zip code with 100% market share. After further adjustment by population, the resulting "incidence" would have the desirable property of being comparable across zip codes with different market share and vastly different populations.



Note: Zipcode level views of states are available using the state selection control.

Map: -75.03 , 40.42 – Image: 492 , 7 – ScaleFactor: 0.001446127734473206

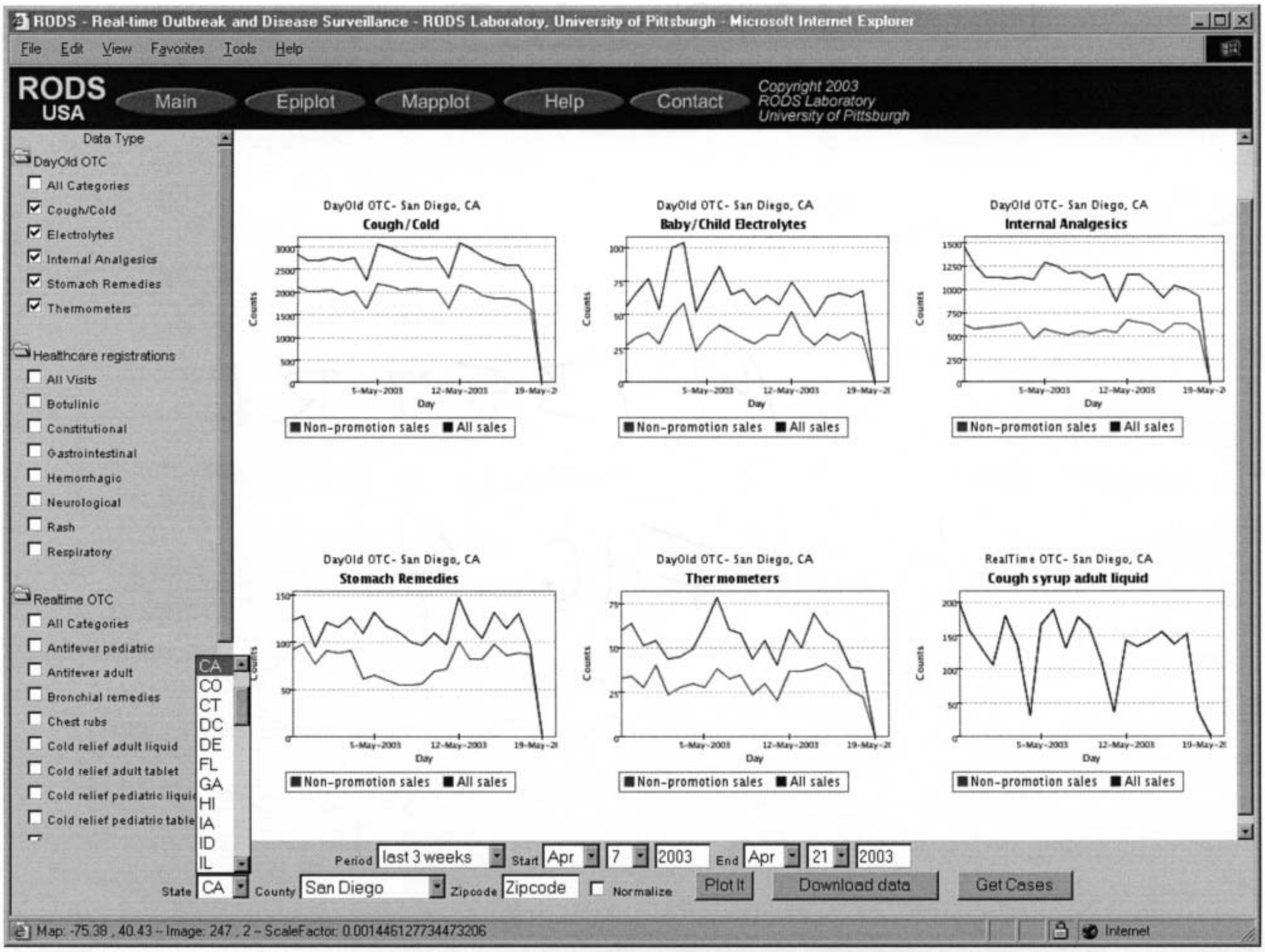
Figure 3. National Retail Data Monitor: Sales of pediatric electrolytes in Philadelphia on May 17, 2003. The color coding indicates the level of sales of pediatric electrolytes for each zip code, relative to historical trends for that zip code. Most of the zip codes are colored *green* (*the lighter shade*) indicating that sales are within 0.5 standard deviations of expected. A few zip codes are colored *blue* (*slightly darker shade*), indicating sales are 0.5 to 2.0 standard deviations higher than expected. There are no clusterings of blue zip codes (or yellow, orange, or red areas) that would be indicative of more anomalous sales activity perhaps warranting investigation.

However, there are practical problems that make normalization a nontrivial task. The authors cannot yet obtain good estimates of *market share* at the zip code or even county level. Highly reliable estimates of market share are available at the regional (multicounty) level, but that level of analysis sacrifices spatial granularity. There is a more fundamental problem—people living in a particular zip code may make most of their purchases at stores in a different zip code. For example, commuters to New York City may buy OTC health care products near their place of work, or patrons of a store located on the border of a zip code area may predominantly reside in an adjacent zip code region. For these reasons, the authors took the more general approach to spatial analysis described in the previous section.

Time-series Data (Epi Curves)

Figure 4 shows the epidemic curve plotting capability of the National Retail Data Monitor. The user can review sales of OTC health care products for any product category, region, or time interval. The user can plot raw counts or counts that are normalized (divided) by the total number of OTC health care products sold on that date in the region in question. Normalization in theory is desirable because sales are influenced by nondisease factors such as store hours and bad weather (e.g., blizzards). Such factors potentially could be adjusted for through measures of overall store traffic (as indicated by unique cash register checkouts). The monitor does not yet obtain such measures from all retailers so that the monitor can only normalize by measures of overall sales activities (available as total sales of OTC health care products). A problem with this approach is that sales of OTC health care products are dominated by the cough-and-cold category, so normalization of the cough-and-cold signal itself by total sales will tend to remove any real spikes in cough-and-cold sales. For this reason, the authors recommend that users look at raw data in the current interface.

Store and newspaper product promotions are another potential confounder, although retailers know that it is very difficult to promote increased consumption in the absence of disease for many products. Promotions may affect which specific products consumers purchase within a category but


RODS - Real-time Outbreak and Disease Surveillance - RODS Laboratory, University of Pittsburgh - Microsoft Internet Explorer
File Edit View Favorites Tools Help
RODS USA
Main
Epiplot
Mapplot
Help
Contact
Copyright 2003
RODS Laboratory
University of Pittsburgh
DayOld OTC- San Diego, CA
Cough/Cold
DayOld OTC- San Diego, CA
Baby/Child Electrolytes
DayOld OTC- San Diego, CA
Internal Analgesics
DayOld OTC- San Diego, CA
Stomach Remedies
DayOld OTC- San Diego, CA
Thermometers
RealTime OTC- San Diego, CA
Cough syrup adult liquid
Counts
Day
5-May-2003
12-May-2003
19-May-2
■ Non-promotion sales ■ All sales
Plot It
Download data
Get Cases
Map: -75.38 , 40.43 -- Image: 247 , 2 -- ScaleFactor: 0.001446127734473206
Internet

Data Type

DayOld OTC
- □ All Categories
- ☑ Cough/Cold
- ☑ Electrolytes
- ☑ Internal Analgesics
- ☑ Stomach Remedies
- ☑ Thermometers

Healthcare registrations
- □ All Visits
- □ Botulinic
- □ Constitutional
- □ Gastrointestinal
- □ Hemorrhagic
- □ Neurological
- □ Rash
- □ Respiratory

Realtime OTC
- □ All Categories
- □ Antifever pediatric
- □ Antifever adult
- □ Bronchial remedies
- □ Chest rubs
- □ Cold relief adult liquid
- □ Cold relief adult tablet
- □ Cold relief pediatric liqui
- □ Cold relief pediatric table

CA CO CT DC DE FL GA HI IA ID IL

Period last 3 weeks Start Apr 7 2003 End Apr 21 2003

State CA County San Diego Zipcode Zipcode □ Normalize

Figure 4. National Retail Data Monitor: Sales of six product categories in San Diego for a three-week period. This screen shows daily sales of five *DayOld OTC* product categories and the *Cough syrup adult liquid Realtime OTC* category (*bottom right of screen*). The upper line in each graph represents total sales and the lower, nonpromotional. A dip in Sunday sales of cough and cold products is visible. Users can view an arbitrary number of such graphs by checking the desired graphs on the list on the left of the screen and then clicking "Plot It." At the bottom of the screen are controls for selecting the geographic region, time interval, and normalization (by total OTC sales). Controls also allow download of 30 days of data to a file for off-line analysis. The "Get Cases" function is not available for OTC sales data.

do not affect category-level sales. Nevertheless, the data feeds from retailers distinguish between sales of promoted and nonpromoted items (i.e., there are two records for every category—that day's sales of promoted products and that day's sales of nonpromoted product). To expose the potential effects of promotion on the time series, the user interfaces plot both nonpromoted sales and total sales (the lower and upper lines, respectively, in the graphs in Figure 4).

Status Report

The National Retail Data Monitor has been in continuous operation since December 2002. The authors consider the project to be in its developmental "build" phase due to ongoing work with the retail industry to achieve 70% data coverage, reduce time latency, and move to UPC-level data feeds. As of May 15, 2003, the data coverage of the system is 23% of total national sales of OTC health care products. The time latency is one day or less. The project has created 119 user accounts for health department employees in 17 states, Washington DC, and at the CDC. Five entities receive raw data feeds from the system including New York City, JHAP/Essence in the National Capital Area, New York State, New Jersey, and the CDC.

The goal for system usage is that the product category geographic maps for each jurisdiction be reviewed every day when they become available at 5 PM local time. This map-based analysis is a relatively recent feature and there has not been sufficient time for education about the schedule of availability and its proper use. Recent usage statistics indicate that on weekdays, at least 15 unique users log in per day, dropping to three on weekends. The authors have not analyzed this usage data sufficiently to understand how many jurisdictions are monitoring the data routinely. Ideally, differential rates for weekend and weekday usage should not exist in an operational surveillance system. The data suggest that, possibly related to temporal staffing patterns, some users may perceive the system as something to consult in the event of an epidemic or a heightened level of alert but not as an early warning system (its intended use). The long-term status of the system is under discussion with a coalition of states that are considering creating shared surveillance resources. During the ongoing "build" phase, the National Retail Data Monitor will continue to be supported by grant funding from the Commonwealth of Pennsylvania. The authors have actively sought additional support from foundations, the Department of Homeland Security, Department of Health and Human Services, and CDC.

Discussion

When completed, the National Retail Data Monitor will make available to health departments, in near to real time, UPC-level product sales data in both raw and analyzed form representing at least 70% of sales in their jurisdictions. These data may be useful for public health surveillance for bioterrorism, infectious diseases, and chronic diseases. The data will also be useful for prospective validation studies of new detection methods. The National Retail Data Monitor represents a model for developing a "data utility" that can serve public health surveillance. As such, the authors' experiences might facilitate future development of public health surveillance tools using data from nurse call centers, health maintenance organizations, national laboratory companies, and poison call centers.

The authors found that a key element for success included the deep understanding of the industry provided to them by an industry expert. This knowledge was invaluable in crafting all aspects of the project, from the "80–20 sufficing" approach, to obtaining the data, to designating product categories, and to maintenance issues. The authors also found that presenting the scientific case for the value of the data was important. Equally key was a personal invitation, sent to the CEO of relevant corporations, for participation (sharing of otherwise proprietary data), authored by a highly respected government or public health official. Another success factor involved the development of an interdisciplinary team with expertise in medical informatics, computer science, law, and engineering.

Many of the problems encountered in analysis and presentation of surveillance data are not unique to retail data. For example, the problem of merging similar data arriving from multiple sources with different time latencies will be common in new surveillance approaches (e.g., hospital registration data may be available in real time from some sources and in batch mode with a one-day delay from other sources). The issues of normalization for spatial and temporal analysis are generic, and we have already encountered them with analysis of hospital chief complaint data in our own work and anticipate that they will be found in analysis of call data to nurse call lines. The method of spatial analysis involving plotting standard deviations for each spatial cell, based on that cell's historical expectation, may be widely applicable in other domains.

Other lessons learned include that data sharing agreements should allow redistribution of data to any public health authority and permit data to be used in research. Data sources that are amenable to a "national" approach should be formed into data utilities—services independent of any particular user interface—and should be industry based. Health departments differ in their needs. Some agencies prefer to receive raw data because they already have surveillance "front ends" or are more comfortable using "off-line" analytic packages. Although there is very significant value in creating analytic applications for users, delivering the data through only a monolithic, vertical application would not meet the needs of a significant subset of end-users (health departments).

The project's immediate future plans are to achieve the target of 70% data coverage and to reduce time latencies toward "real time." The project intends to add a rapid spatial scan algorithm to automate more fully the review of maps. The project will develop an evaluation strategy to document prospectively the system's ability to detect naturally occurring outbreaks, should they occur.

Longer-term project plans include the expansion of monitoring to the level of selected prescription medications. As in the case of retail products, a standard coding system (NDC) that is used in industry data systems provides the basis of feasibility. The national retailers' store prescription data are stored separately from retail sales data. Prescription data are subject to HIPAA controls, but aggregation of the data by zip code, similar to how sales of OTC health care products are aggregated, should satisfy HIPAA regulations.

Future plans also include extension of the project to international scope. Just as the Uniform Code Council administers and manages UPCs in the United States and Canada, EAN International provides the same service outside of North America by using European Article Numbering (EAN) codes. EAN is a UPC-compatible system, which began in the 1970s and eventually merged into the current EAN·UCC System, used as a standard by 45 countries including Japan, and most of the European Union for coding pharmaceutical products.

Conclusions

The National Retail Data Monitor is a general model for a class of surveillance approaches that leverage existing commercial data collections. The promise of such systems derives from the inherent early availability of their data (reflecting early illness behaviors) and the extreme efficiency with which the data can be obtained, relative to more traditional surveillance data methods. Only with real-world experience in detecting various types of outbreaks will the true utility of the National Retail Data Monitor become known.

References ■

1. Wagner M, Tsui F-C, Espino J, et al. The emerging science of very early detection of disease outbreaks. J Public Health Manag Pract. 2001;6(6):50–8.
2. Kaufmann A, Meltzer M, Schmid G. The economic impact of a bioterrorist attack: are prevention and postattack intervention programs justifiable? Emerg Infect Dis. 1997;3(2):83–94.
3. Lober WB, Karras BT, Wagner MM, et al. Roundtable on bioterrorism detection: information system-based surveillance. J Am Med Inform Assoc. 2002;9:105–15.
4. Tsui F-C, Espino JU, Dato VM, Gesteland PH, Hutman J, Wagner MM. Technical description of RODS: a real-time public health surveillance system. J Am Med Inform Assoc. 2003;10: 399–408.
5. Talan DA, Moran GJ, Mower WR, et al. EMERGEncy ID NET: an emergency department-based emerging infections sentinel network. The EMERGEncy ID NET Study Group. Ann Emerg Med. 1998;32:703–11.

6. Effler P, Ching-Lee M, Bogard A, et al. Statewide system of electronic notifiable disease reporting from clinical laboratories: comparing automated reporting with conventional methods. JAMA. 1999;282:1845–50.
7. Lewis M, Pavlin J, Mansfield J, et al. Disease outbreak detection system using syndromic data in the greater Washington DC area. Am J Prev Med. 2002;23:180.
8. Lazarus R, Kleinman K, Dashavsky I, et al. Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. Emerg Infect Dis. 2002;8:753–60.
9. Lazarus R, Kleinman KP, Dashavsky I, et al. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. BMC Public Health. 2001;1(1):9.
10. Zelicoff A, Brillman J, Forslund DW, et al. The Rapid Syndrome Validation Project (RSVP). Proc AMIA Symp. 2001:771–5.
11. Gesteland PH, Wagner MM, Chapman WW, et al. Rapid deployment of an electronic disease surveillance system in the State of Utah for the 2002 Olympic Winter games. Proc AMIA Symp. 2002:285–9.
12. National Electronic Disease Surveillance System (NEDSS): a standards-based approach to connect public health clinical medicine. J Public Health Manag Pract. 2001;7(6):43–50.
13. Labrie J. Self-care in the new millennium: American attitudes towards maintaining personal health. Consumer Healthcare Products Association, 2001, p 76. <http://www.chpa-info.org/pdfs/CHPA%20Final%20Report%20revised%20(03-20)_.pdf>. Accessed July 12, 2003.
14. McIsaac WJ, Levine N, Goel V. Visits by adults to family physicians for the common cold. J Fam Pract. 1998;47:366–9.
15. Hogan WR, Tsui F-C, Ivanov O, et al. Detection of pediatric respiratory and diarrheal outbreaks from sales of over-the-counter electrolyte products. J Am Med Inform Assoc. 2003;10 (in press).
16. Magruder S, Florio E. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of public health. Johns Hopkins University Applied Physics Laboratory Technical Digest. 2003;24(4) (in press).
17. Welliver RC, Cherry JD, Boyer KM, et al. Sales of nonprescription cold remedies: a unique method of influenza surveillance. Pediatr Res. 1979;13:1015–7.
18. Goldenberg A, Shmueli G, Caruana RA, Fienberg SE. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proc Natl Acad Sci U S A. 2002;99: 5237–40.
19. Reis BY, Mandl KD. Time series modeling for syndromic surveillance. BMC Med Inform Decis Mak. 2003;3(1):2.
20. Broome C, Pinner R, Sosin D, Treadwell T. On the threshold. Am J Prev Med. 2002;23(1):229.
21. Halperin W, Baker ELJ (eds). Public Health Surveillance. New York: Van Nostrand Reinhold, 1992.
22. Teutsch S, Churchill R. Principles and Practice of Public Health Surveillance (ed 2). Oxford: Oxford University Press, 2000.
23. Yasnoff WA, Overhage JM, Humphreys BL, et al. A national agenda for public health informatics. J Public Health Manag Pract. 2001;7(6):1–21.
24. Goldstein A. Strategic tracking of sniffles: scientists on alert for terrorism monitor area health factors. Washington Post. March 28, 2003: A10. <http://www.washingtonpost.com/wp-dyn/articles/A39629-2003Mar27.html>. Accessed July 12, 2003.
25. Perez-Peña R. An early warning system for diseases in New York. New York Times. April 4, 2003. <http://www.nytimes.com/2003/04/04/nyregion/04WARN.html?ex=1053489600&en=a50eec53d50d3da6&ei=5070>. Accessed May 18, 2003.
26. Wein LM, Craft DL, Kaplan EH. Emergency response to an anthrax attack. Proc Natl Acad Sci U S A. 2003;100:4346–51.
27. Panackal AA, M'ikanatha NM, Tsui F-C, et al. Automatic electronic laboratory-based reporting of notifiable infectious diseases. Emerg Infect Dis. 2001;8:685–91.
28. Ewert DP, Westman S, Frederick PD, Waterman SH. Measles reporting completeness during a community-wide epidemic in inner-city Los Angeles. Public Health Rep. 1995;110: 161–5.
29. Kulldorff M, Athas WF, Feurer EJ, et al. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos, New Mexico. Am J Public Health. 1998;88:1377–80.
30. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. Stat Med. 1995;14:799–810.
31. Glaz J, Balakrishnan N. Scan statistics and applications. Statistics for industry and technology. Boston: Birkhèauser, 1999, pp xxi, 324.