



Syndromic Surveillance Using Minimum Transfer of Identifiable Data: the Example of the National Bioterrorism Syndromic Surveillance Demonstration Program

Richard Platt, Carmella Bocchino, Blake Caldwell,
Robert Harmon, Ken Kleinman, Ross Lazarus,
Andrew F. Nelson, James D. Nordin,
and Debra P. Ritzwoller

ABSTRACT *Several health plans and other organizations are collaborating with the Centers for Disease Control and Prevention to develop a syndromic surveillance system with national coverage that includes more than 20 million people. A principal design feature of this system is reliance on daily reporting of counts of individuals with syndromes of interest in specified geographic regions rather than reporting of individual encounter-level information. On request from public health agencies, health plans and telephone triage services provide additional information regarding individuals who are part of apparent clusters of illness. This reporting framework has several advantages, including less sharing of protected health information, less risk that confidential information will be distributed inappropriately, the prospect of better public acceptance, greater acceptance by health plans, and less effort and cost for both health plans and public health agencies. If successful, this system will allow any organization with appropriate data to contribute vital information to public health syndromic surveillance systems while preserving individuals' privacy to the greatest extent possible.*

KEYWORDS *Acute disease epidemiology, Ambulatory care, Bioterrorism, Cluster analysis, Disease outbreaks, Human, Medical informatics applications, Medical records systems, Population surveillance methods, Statistics and numerical data, Surveillance.*

INTRODUCTION

One potentially valuable source of information for syndromic surveillance is diagnoses assigned during routine ambulatory care, including office visits and nurse telephone triage lines. Several such systems have been described,¹⁻⁵ and a national

Drs. Platt and Kleinman are with the Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, Boston, Massachusetts; Drs. Platt and Lazarus are with Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; Ms. Bocchino is with the American Association of Health Plans, Washington, DC; Dr. Harmon is with Optum, Golden Valley, Minnesota; Dr. Caldwell is from Savannah, Georgia; Dr. Lazarus is from the University of Sydney School of Public Health, Australia; Mr. Nelson and Dr. Nordin are with HealthPartners Research Foundation, Minneapolis, Minnesota; and Dr. Ritzwoller is with Kaiser Permanente Colorado, Denver.

Correspondence: Richard Platt, 133 Brookline Avenue, Sixth Floor, Boston, MA 02215. (E-mail: richard.platt@channing.harvard.edu)

demonstration project involving multiple health plans and approximately 20 million covered lives is currently being developed. This surveillance system will rely principally on reporting by health plans to public health agencies of aggregated (count) data rather than on reporting of encounter-level data. This article presents a brief description of the program and discusses the reasons for adopting this method of data sharing.

THE NATIONAL DEMONSTRATION PROGRAM

The Centers for Disease Control and Prevention, the American Association of Health Plans, Harvard Medical School, five health plans or physician groups (Harvard Pilgrim Health Care/Harvard Vanguard Medical Associates in Massachusetts, HealthPartners in Minnesota, Kaiser Permanente Colorado, Scott and White Healthcare System in Texas, and the Austin Regional Clinic in Texas), and Optum, a nationwide consumer health information company, are collaborating to develop a syndromic surveillance system that will cover more than 20 million individuals with prepaid health care in all 50 states. The system will use encounter-level data from routine and urgent office visits to the first five health plans and from the nurse telephone triage and health information system of the rest.

Although this system is under development, it will be based on one created by Harvard Pilgrim Health Care/Harvard Vanguard Medical Associates in collaboration with the Massachusetts Department of Public Health.⁵ That system uses a number of features of the US Department of Defense ESSENCE (Electronic Surveillance System for the Early Notification of Community-Based Epidemics) system,¹ plus other features developed by other organizations, including HealthPartners in Minnesota and Kaiser Permanente Colorado in Denver. The new system will have several appealing features from a public health perspective. First, the fact that the source population will be known will allow greater flexibility for signal detection than is possible when only the affected individuals are known. Second, it will use electronic information that is already collected by the practices, health plans, and call centers as part of routine operations. Therefore, no clinicians or other health plan personnel are required to perform any nonroutine assignment of diagnoses, perform any classification, or initiate daily reporting of syndromic data. The fact that no additional labor will be required is important for a surveillance system that is intended to operate permanently. Third, because the information is available electronically, the incremental cost will be small to extract data of interest, manipulate it, and make it available to public health agencies. A substantial fraction of the US population and a larger fraction of population centers are covered by one or more health plans that have some electronic information that could be useful for this type of surveillance.

The principal features of this surveillance system based on health plans will include different activities performed by the health plans, a data center, and public health agencies (Figure). The activities based on the health plans will be (1) annual or semiannual enumeration of health plan members; (2) assignment of each member to a geographic area, for instance, census tract or ZIP code; (3) daily or more frequent creation of a data extract containing new records of encounters with diagnoses of interest; (4) identification of only new episodes of illness by excluding people who have had recent encounters with a diagnosis in the same syndrome; (5) assignment of new episodes to the individual's ZIP code or census tract; and (6) transmission to the data center of the counts of new episodes in each area.

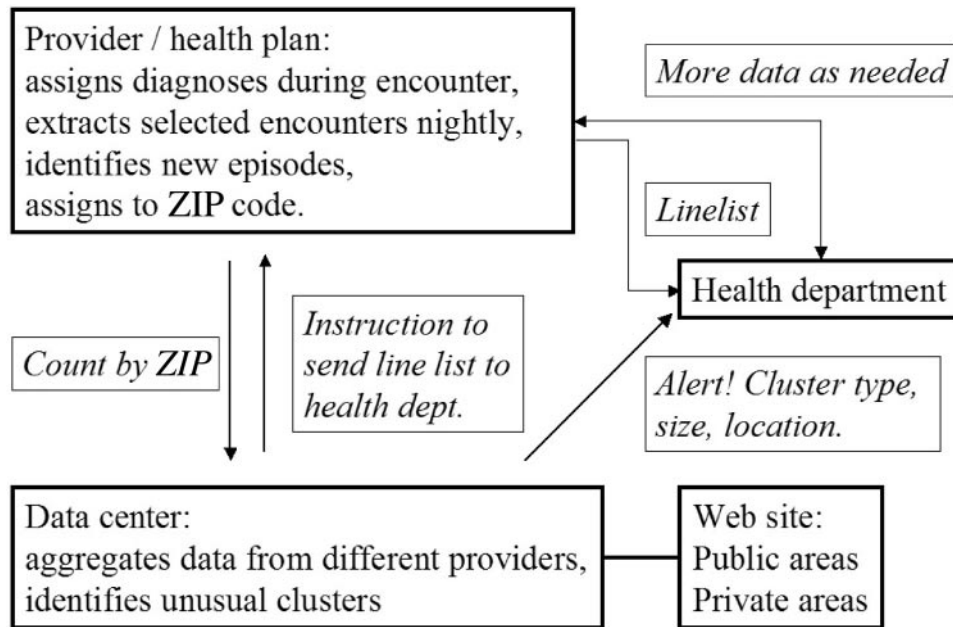


FIGURE. Data flow for the National Bioterrorism Syndromic Surveillance Demonstration Program of the Centers for the Centers for Disease Control and Prevention.

The identification of new episodes and assignment to geographic areas will be performed using programs provided by the data center. Secure transmission of count data to the data center will use messaging software provided by the Centers for Disease Control and Prevention. The health plans will retain the clinical data extracts on a computer (server) inside their firewall. These extracts contain confidential identifying information that may be required if the event is suspected to be part of a cluster of illness needing further investigation.

When a cluster is detected, these extracts will be used to generate line lists that can be sent automatically to the health department in the affected area. These line lists will contain additional demographic and clinical information, including the patients' area of residence, age, sex, recorded temperature, diagnoses assigned, and diagnostic tests performed. The line lists will not contain the patients' names, addresses, or other information about the patients' health care.

The health plans will also designate an individual who will be able to respond promptly to requests by public health agencies for more information about specific patients than is contained in the line lists. Note that health plans only provide to public health authorities information that can be traced to individuals in the unusual event that the individuals are believed to be part of an epidemiologically important cluster. This system thus will be consistent with other mandated public health reporting activities.

The data center will maintain information about each health plan's membership in each geographic region. It also will maintain historical information about the daily counts of new episodes of each syndrome in each ZIP code or census tract. Each day, it will combine reports from different health plans (with service areas that may overlap) to provide a combined total number of episodes in each syndrome in

each area under surveillance. In each area, it will compare the observed number of new episodes of illness to the expected number, using the historical data and a variety of modeling techniques, and it will identify areas with an unusual numbers of events (*signals*), and report these signals to public health agencies. At the same time, the data center will instruct the health plan's data server to send the line list described above to the designated recipient in the health department who has responsibility for the geographic region involved.

This method of aggregating reports from multiple health plans with overlapping catchment areas should allow detection of signals that are too weak to be observed in any single health plan's data. It also will obviate the need for each health plan to develop the capacity to report to multiple public health agencies. This is particularly important for health plans with national or multistate coverage.

Public health agencies will work with the data center to develop an acceptable reporting format, and they will indicate their preferred thresholds for reporting of signals that require immediate attention. They will query the designated responders of the health plans to obtain additional information about the individuals who contributed to a signal. This reporting sequence will begin differently from usual reports by clinicians since the first notice to the public health agency will include only a count of individuals rather than a report from a clinician who is concerned about one or more specific individuals because of their unusual clinical or historical features.

With the new system, the epidemiological feature of interest will be the total number of affected individuals within limited geographic regions; most of these individuals will have symptoms consistent with common illnesses (e.g., cough or headache). Many or all of the individuals who contribute to a signal will have benign explanations for the diagnosis, and clinicians would be unlikely to identify any of them with reportable illnesses. It is also very likely that moderate clusters of illness will be detected by this method of aggregation even though most clinicians will see no more than a single extra symptomatic person.

ADDITIONAL CAPABILITIES

This data structure has flexibility that can serve several important purposes. It is relatively simple to modify the syndrome definitions or to create new syndromes because the health plans retain diagnosis-level data that can be manipulated by new programs supplied by the data center. Examples of modifications that may be of interest are reports involving specific segments of the population, such as children. It will also be possible to perform ad hoc queries through programs distributed by the data center; such queries will be subject to the agreement of the health plans. This ad hoc query capability can be automated so that it can operate essentially in real time.

DATA-SHARING CONSIDERATIONS

These data-sharing provisions have several advantages compared with a case-based reporting system. We believe the system based on routine reporting of counts of ambulatory encounters is simpler, quicker, and less expensive to develop and maintain than an encounter-based system since much of the work of data reduction is performed by health plans that already possess the encounter-level data. Thus, there

is no need for public health agencies to develop and maintain a separate capability for receiving many millions of encounter records per year to identify signals of interest.

More important, this system conforms to the general expectation that confidential personal health information will be used as sparingly as possible to accomplish the mission of the public health agencies. Even if providing encounter-level identified information is permitted by public health reporting law, this will be essentially new terrain both for the public and for public health agencies since it will affect nearly every person and because the majority of reported episodes will not involve a condition that has traditionally been considered to be of public health interest. In addition, public health agencies will need to develop robust capabilities to keep confidential the large amount of data they will receive. It is difficult to predict with confidence the public's reaction to such broad reporting, but many individuals may perceive this intensity of reporting of protected health information to be inappropriate. It is worth noting that these concerns may be less important for syndromic surveillance based in emergency rooms because the number of visits is so much smaller, a much smaller proportion of the population is directly affected, individual visits are typically not linked to any other health care the individual receives, and the proportion of events that may be of public health interest may be higher.

It is possible for health plans to provide individual encounter-level data with encrypted identifiers. However, developing and maintaining this capacity will entail considerable investment of human and financial resources by the health plans. Even without explicit identifiers, like name, date of birth, or address, it is often possible to identify individuals through patterns of care they receive, particularly in geographic areas with relatively few members of a specific health plan. Techniques exist to minimize the risk of reidentification, but they are not perfect, and they require additional investment. It is likely, therefore, that health plans will be slow to join an encounter-based syndromic reporting system because of their concerns about the extra work involved and their interest in appearing to their members to be good stewards of protected health information.

The trade-off for not providing individual-level data routinely is that health plans must respond promptly to requests for additional information about the individuals represented in the clusters of concern. This responsibility is typically assigned to an individual within the health plan who is authorized to access clinical information. Because participating health plans are large, such an individual is usually on duty at most times; when not on duty, it is important for such an individual to be available. Access to the required information is ordinarily straightforward because it exists in electronic form.

LIMITATIONS

There is no proof that syndromic surveillance will be useful, either for early detection of bioterrorism events or to support other public health activities, although there is some evidence suggesting that our system provides a valid representation of seasonal events.⁴ Even under optimal conditions, syndromic surveillance may yield insufficient information, such as a diagnosis of cough with no additional clinical data, to allow adequate sensitivity to find events of interest or to find them in a timely manner. In addition, the specificity may be too low to make this a useful routine screening activity; even a moderate volume of false alarms may make this approach unusable on a sustained basis.

It is unclear how a syndromic surveillance system based in an ambulatory setting will perform compared to a system based in emergency rooms. In some circumstances, one would expect the first signal to be observed in the ambulatory setting. This will be the case if a substantial number of people with relatively mild symptoms seek care. However, the first signal might be detectable in emergency rooms, for instance, if a smaller number of people become seriously ill more quickly and use emergency rooms as their first point of contact with the health care system.

Because the demonstration project described here uses medical record information rather than billing data, we expect the accuracy of the information to be as good as clinicians create in actual practice. Nonetheless, there is likely to be a substantial amount of misclassification. Inaccuracies in the data are likely to be greater in systems that use administrative data.

Many logistical issues will need to be addressed for this system to work as intended. The electronic data created as part of routine care will need to be available promptly and without interruption. Communications links will need to function efficiently between the health plans and the data center and between the data center and the health departments. Clinician responders in health plans will need to respond quickly to requests for additional data from health departments.

More generally, it will be necessary to demonstrate that the system described here can provide information that is as useful for public health purposes as one that provides encounter information directly to health departments.

Because of the many uncertainties, even a large ambulatory care-based syndromic surveillance system should be considered to provide information that is complementary to other surveillance systems.

CONCLUSION

Experience in individual health plans indicates that data from medical offices and telephone triage lines can be a useful component of syndromic surveillance systems. These have prompted the development of a national demonstration program. A principal design feature of this system is reporting of counts of individuals with syndromes of interest in small geographic regions, with retention by the health plans of individual-level data except in the relatively unusual event of signals that require follow-up by public health agencies.

ACKNOWLEDGEMENT

This work was supported by the Centers for Disease Control and Prevention grants (UR8/CCU115079 to the Eastern Massachusetts Prevention Epicenter and U90/CCU116997 to the Massachusetts Department of Public Health). Several participating health plans also received support from their state and local health departments.

This report is in the public domain. Additional information about this research is available at www.btsurveillance.org.

REFERENCES

1. US Department of Defense. *Annual Report, Fiscal Year 1999*. Silver Spring, MD: Walter Reed Army Institute of Research; 1999.
2. Martinez B. Questions of security: HealthPartners use reach, speedy data to hold watch for bioterrorism attacks. *Wall Street Journal*. November 1, 2001;sect A:10.

3. Rodman J, Frost F, Jakuboski W. Using nurse hot lines for disease surveillance. *Emerg Infect Dis.* 1998;4:329–332.
4. Lazarus R, Kleinman K, Dashevsky I, DeMaria A, Platt R. Using automated medical records for rapid identification of illness syndromes: the example of lower respiratory infections. *BioMed Central Public Health.* 2001;1:1–9.
5. Lazarus R, Kleinman K, Dashevsky I, et al. Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. *Emerg Infect Dis.* 2002;8:753–760.