A Public Health Information System for Conducting Community Health Needs Assessment

Elio F Spinello, MPH,* Ronald Fischbach, PhD
Department of Health Sciences, California State University, Northridge, CA

Abstract

Students in the health education program at California State University, Northridge (CSUN), like students in similar programs at other universities, are often trained through the use of practical projects as part of the curriculum. Health education projects typically require that students select a geographically defined community, conduct a community health education needs assessment, and then produce a program plan that addresses the health needs of that community. The purpose of this project was to develop a community health database and mapping system that geographically integrated demographic, housing, morbidity, and mortality data, using specially written software to facilitate analysis and mapping of the data. The creation of this public health information system (PHIS) represents a step toward using geographic information system technology to improve the process of developing public health needs assessments. The PHIS is also a valuable resource for training public health educators, making hard-to-find needs assessment data readily available to them as well as giving them the ability to analyze the data spatially. After pilot testing in fall 1997, the PHIS was implemented for use in community health education classes at California State University. The CSUN Department of Health Sciences intends to find additional applications for the PHIS database. Currently, the PHIS is being tested by a national medical center as part of a program to determine its effectiveness as a decision-support tool. The system also has been made available to the Los Angeles County (California) Department of Health Services for testing in the Border Health program as a needs assessment tool.

Keywords: assessment, community health, health education, morbidity, mortality

Background

A critical step in the development of a health education program is the needs assessment. At the community level, conducting a needs assessment requires collecting a substantial amount of information about the demographic, socioeconomic, and health characteristics of the population. For most communities, this process is a time-consuming, largely manual task. A number of methods are typically used to collect community-level data. These often include focus groups; in-depth interviews; and surveys of residents, key informants, health care providers and community leaders. Indirect methods are often used to gather statistical data about a community. The sources for statistical data can include the US Census Bureau, state and local departments of public health, law enforcement agencies, and health regulatory agencies.

* Elio F Spinello, Dept. of Health Sciences, California State University, Northridge, 18111 Nordhoff St., Northridge, CA 91330-0001 USA; (p) 818-831-7607; (f) 818-831-9078; E-mail: elio.spinello@csun.edu

One of the limitations of statistical data gathered from multiple sources is that the data must be standardized and integrated geographically. This means, for example, that demographic data by census tract would be matched up with housing data by census tract to produce a combined database containing both housing and demographic data. The linkage for combining the information would be the unit of geography—in this case, census tracts. A problem arises when one agency produces data at the census tract level while another produces data at an overall, county, community, zip code, or other level of geography. Various agencies are also often inconsistent with respect to coding variables such as age and ethnicity (1,2). Data collection protocols often differ between agencies; this can cause different agencies to collect data for different time periods, or can mean that some agencies do not collect some needed portion of the data, such as information from a particular geographic region. Many of the difficulties involved in collecting statistical data for communities were noted by Paulu, Ozonoff, Coogan, and Wartenberg (3), when they observed that clinical medicine and public health have, in some ways, become more similar over time. While physicians have become more prevention-oriented, public health professionals are more frequently being called upon to engage in what might be called community diagnosis and treatment. While physicians have patient histories available to them when they conduct individual medical needs assessments, though, there is often no readily available community health history for use by the public health official. Sources of information about a community do not reside in the memory or records of an individual person, but rather in institutional arrangements made by a number of agencies for many different purposes. Given the lack of an integrated public health information resource for Los Angeles County (California), a need was identified for the development and creation of a computerized community-level database that contains demographic, morbidity, and mortality data.

Purpose of the Project

The purpose of this project was to design and implement a public health information system (PHIS) for Los Angeles County, organized by zip code and containing health, housing, and demographic variables. This PHIS was to be used by health education students in the preparation of community health needs assessments. Zip codes were selected as the basic unit of geography because all of the required datasets were believed to be available at the zip code level. Software was designed and developed that would facilitate access to the database by enabling users to rank zip codes, query the data, and produce reports and maps for selected zip codes and census block groups.

Database and System Design

Major foundation courses at California State University, Northridge (CSUN), that prepare professional health educators emphasize the concepts of program planning, implementation, and evaluation as they apply to health education in the community. The primary teaching strategy in these courses involves assigning students a project in which they must select a community, assess the health needs of that community, and then design a health education program to address those needs. Traditionally, students select a community and identify its boundaries, then begin researching the health, demographic, and socioeconomic data available for that area. This process involves gathering statistical data from local government and private sources, and identifying and interviewing key informants. It also involves other primary and secondary research. Once the research is completed, the student must design a program plan with measurable health education objectives, an action plan for a health intervention, an implementation scheme for conducting the program, and a selection of appropriate evaluative measures. The premise behind the PHIS project was that all of these activities, which are quite time-consuming, could be accomplished much more quickly with the help of a computerized database system.

The concept for the database came from the need for students to be able to assess the health needs of a community. Based on the requirements for undergraduate and graduate students' health education projects, the needs assessment database had to include the following information (4):

- Community backdrop: including demographics, geographic features, transportation systems, and political structure.
- Health care system: hospitals, clinics, health department sites, emergency services, voluntary health agencies, etc.
- Community health status: including morbidity and mortality measures of populations in the community.
- Social assistance system: programs available in the community.

The geographical element common to all of the data sources was a 5-digit zip code. Because the zip code was provided by all of the data sources and was a concept assumed to be easily understood by users, it was adopted as the definition of a community. The demographic data sources were actually found to provide census data for much smaller areas, such as census tracts and block groups. Based on its availability, demographic information in the database was included for block groups, because they represent much smaller geographic areas than zip codes and therefore enable users to conduct deeper analyses. Once the data sources were assembled, a relational database approach was adopted—a design that would allow multiple tables that have a common variable (zip code) to be integrated as if they were a single combined table.

Specific tables used in the database included:

- Hospital discharges for Los Angeles County by major diagnostic categories (MDCs) and *International Classification of Diseases, Ninth Revision* (ICD-9) (5) codes, by zip code.
- Los Angeles County death certificates summarized by MDCs and ICD-9 codes by zip code.
- Hospital point locations for Los Angeles County.
- Locations of cases of chlamydia and gonorrhea treated at Los Angeles County clinics.
- Current-year and projected demographic variables for Los Angeles County zip codes.
- 1990 demographic and housing data for Los Angeles County block groups.

Upon receipt of the various tables mentioned above, the documentation for all of the files was examined to determine the best way to standardize the tables and link them together. Because the coding conventions for the different files varied, coding standardization issues were also identified at this stage. In the death certificate table, for example, ethnicity was assigned to each case using one of 16 codes, whereas in the hospital discharge file there were only 8 possible codes. The final database design consisted of two levels of summary. The first level applied to the discharge dataset (6). Using an algorithm developed for this project, records were coded as to whether they represented individual or repeat discharges for the same diagnosis. Once coded, the dataset was reduced to one record per individual rather than one record per discharge (4,7,8,9). The second level involved summarizing each discharge and death certificate record down to one record per discharge or death per 5-digit zip code. This reduced the size of the database substantially while still maintaining a reasonable degree of detail for reporting and analysis (see Figures 1 and 2). After the discharge data and death certificate data had been reduced to the zip code level, the processed discharge data, hospital information, death certificates, and demographic tables were assembled into a relational database.

Figure 1 Example of raw data.

Case Number	Zip Code	ICD-9 Diagnosis	Disposition
1	91324	101.1	Normal Discharge
2	91324	101.1	Normal Discharge
3	91324	101.1	Deceased
4	91324	302.0	Normal Discharge
5	91324	302.0	Deceased

An example of raw data. One record is present for each discharge, indicating the zip code, diagnosis, and disposition.

Figure 2 Example of aggregated records.

Zip Code	ICD-9 Diagnosis	Normal Discharges	Deaths
91324	101.1	2	1
91324	302.0	1	1

An example of aggregated data. Raw data are reduced to one record per diagnosis per zip code.

The programming language used to create the system was Microsoft Visual Foxpro 3.0. Visual Foxpro was chosen because it can query large databases quickly and produce reports based on those data. The PHIS was also designed to produce geographic maps of an area depicting the morbidity and mortality data graphically. The mapping system that was integrated into the PHIS software was created with MapObjects LT (ESRI, Redlands, CA).

System Features

The data access software was designed to enable students to select zip codes and then produce tabular reports and maps to describe the communities. With the inclusion of a table of population estimates by zip code, morbidity and mortality rates could be calculated by age cohort and ethnicity category. To facilitate its use, the software was designed with three major analytical components: a zip code ranking module, zip code query and reporting modules, and an integrated GIS module (10).

The zip code ranking module allowed students to select a particular health problem as defined by a single MDC or ICD-9 code. Additionally, a custom definition could be created by combining ICD-9 codes (e.g., all ICD-9 codes relating to sexually transmitted diseases). Once a pre-defined or custom disease was selected, one of eight ranking measures could then be selected (e.g., the crude death rate, prevalence rate, or total number of cases). The ranking report produced by this module could be used to identify the best-to-worst or worst-to-best zip codes in Los Angeles County based on the selected disease and measure. The results could then be displayed on a theme map.

The query module of the PHIS software was designed to allow users to select one or more zip codes from a list. Users could then press a button that would select different types of available data, and choose tabular reports to be produced for each of the selected zip codes. The zip code reports available included:

- The demographic trends report, which indicated basic demographic characteristics such as average household size and per capita income from 1994 to a 1999 projection.
- The demographic profile report, which contained a breakdown of the population by age and ethnicity as well as households by income category.
- The MDC profile, which summarized the number of cases and other key measures and rates for each of the 27 MDC categories.
- The ICD-9 profile, which summarized the number of cases and other key measures and rates for each of the 909 3-digit ICD-9 categories.
- The hospital profile, which detailed each of the hospitals contained in the selected zip code(s) together with key information such as name, address, number of beds, and relative size of emergency room.
- The morbidity and mortality report, which provided a detailed analysis of the morbidity and mortality of a selected disease in the selected zip code(s), including morbidity and mortality rates by age category and ethnicity.

In addition to producing reports for specific zip codes, students could also produce reports summarizing data for all of Los Angeles County. The county-level reports could then be used as a benchmark when comparing rates for various diseases. Additionally, students could select all of the block groups in a census tract, zip code, or custom-defined area and produce a report that would summarize all of the demographic data for the combined area.

The integrated GIS module was designed to give users the ability to produce a color-coded thematic map of their chosen community for purposes of determining the locations of zip codes, census tracts, block groups, hospitals, and highways. The mapping system also provided the ability to display concentrations of gonorrhea, chlamydia, and tuberculosis cases treated at Los Angeles County clinics. Users could select which types of geographic feature to display on a map and change the map view by zooming in or out and by panning across areas. They could also select individual geographic features such as zip codes, hospitals, or highways. Once they selected a geographic feature, users could view information about it by clicking on the feature with a mouse. The geographic features contained in the mapping system included:

- California county regions
- California 5-digit zip code regions
- California interstate highways
- Los Angeles County census tract regions
- Los Angeles County block group regions
- Los Angeles County chlamydia cases shown as point locations
- Los Angeles County gonorrhea cases shown as point locations
- California hospital point locations

The program and database were documented in a user manual that accompanied the software. The manual provided procedures and a step-by-step process for use of the program. The user manual also contained a number of illustrations of the software's interface; this helped users follow along with the manual's examples. A case study was included as a tutorial exercise. Once written, the manual was converted into a Microsoft Windows help file.

System Integrity and Usability Testing

To identify programming bugs and data errors, both the software and database were subjected to an initial test. This process uncovered a number of problems, including installation problems, errors in calculations, and errors in reports.

Upon completion of the initial test, the software and database were pilot-tested by two undergraduate community health education classes, as well as a graduate community health education class. Students from the three classes attended a 30-minute review session of the database and the software's features conducted by the project investigator. Pilot test feedback was collected using three methods. Users were directly observed by the project investigator, who filled out an observational checklist as they used the software. All users were also provided with test reports that they were asked to complete and return. Finally, focus groups were held with all users.

In general, feedback from all three collection methods was relatively consistent. Overall, users found the system to be a useful tool in completing their projects. The project investigator was already aware of a number of the issues that surfaced, such as a missing manual and bugs affecting two of the reports. A number of other issues that were identified had not been previously considered. These included the need for additional training, the need to train lab assistants, the need to include more specific directions in the manual and online help system, and the existence of bugs affecting the mapping system.

What We Learned

Based on a review of the pilot test results, a number of conclusions were reached with respect to the use of the PHIS at CSUN:

- The need to fully explain the constraints and limitations of the database cannot be dismissed. Although the database methodology has some clear limitations, the ability to quickly and easily develop health status profiles made it tempting to take the results at face value without further research. In reality, the database has significant biases for some diseases. Records of drug and alcohol abuse, for example, only reflect cases in which individuals died (with drugs or alcohol identified as the cause of death) or cases in which individuals either overdosed or admitted themselves for treatment. The vast majority of untreated cases are not recognized in the database. Many acute and chronic diseases that require hospitalization or tend to have high case fatality rates, however, are better represented by the database, because a larger percentage of those cases will be represented in the discharge or death certificate file.

- Due to the number of tasks that can be accomplished with the PHIS, a number of students felt they needed some initial direction to help them best take advantage of it. The approach most commonly suggested was the use of a tutorial or case study, something that would give students an example to work through.
- The next update to the PHIS database will need to incorporate a more accurate method of estimating repeat discharges by the same individual (7,11,12). Although the method used in the initial database was believed to be at least somewhat effective at reducing the dataset from one record per discharge to one record per individual, a more sophisticated statistical approach will likely produce much more accurate results.
- A more effective method of disseminating the data and analytical system will likely improve the effectiveness of the application. Approaches including Internet access to the database and functionality are being explored (3).

References

1. Andres N, Davis DE. 1995. Linking public health data using geographic information system techniques: Alaskan community characteristics and infant mortality. *Statistics in Medicine* 14(5–7):481–90.
2. Avery C, Zabel D. 1995. *Gathering client data: What works?* Proceedings from the International Conference on TQM and Academic Libraries, Washington, DC, April 20–22, 1994.
3. Paulu C, Ozonoff DM, Coogan P, Wartenberg D. 1995. Making environmental data accessible for public health aims: The Massachusetts Environmental Database Project. *Public Health Reports* 110(6):776–83.
4. Murray SA, Graham, LJC. 1996. Practice based health needs assessment: Use of four methods in a small neighborhood. *British Medical Journal* 310(6992):1443–8.
5. World Health Organization (WHO). 1977–78. *Manual of the international statistical classification of diseases, injuries, and causes of death: Based on the recommendations of the Ninth Revision Conference, 1975, and adopted by the Twenty-Ninth World Health Assembly.* Geneva: WHO.
6. California Office of Statewide Health Planning and Development. 1996. *California healthcare data programs.* http://www.oshpd.cahwnet.gov.
7. Jaro MA. 1995. Probabilistic linkage of large public health data files. *Statistics in Medicine* 14(5–7):491–8.
8. Kirby RS. 1996. Toward congruence between theory and practice in small area analysis and local public health data. *Statistics in Medicine* 15(17–18):1859–66.
9. O'Hara D, Hart W, Robinson M, McDonald I. 1996. Mortality soon after discharge from a major teaching hospital: Linking mortality and morbidity. *Journal of Quality Clinical Practice* 16(1):39–48.

10. Roos LL, Walld R, Bond R, Hartford K. 1996. Record linkage strategies, outpatient procedures, and administrative data. *Medical Care* 34(6):570–82.
11. Huff L, Bogdan G, Burke K, Hayes E, Perry W, Graham L, Lentzner H. 1996. Using hospital discharge data for disease surveillance. *Public Health Reports.* January.
12. Scholten HJ, de Lepper MJ. 1991. The benefits of the application of geographical information systems in public and environmental health. *World Health Statistics Quarterly* 44(3):160–70.