(CN) Chapter ~~X~~ 11

(CT) Computerizing Public Health Surveillance Systems

(CA)

Andrew G. Dean

Robert F. Fagan

Barbara Panter-Connah

(Epi) Everything should be made as simple as possible. But to do that you have to master complexity. ~~"We only conquer what we wholly assimilate."~~

(Epi-S) Butler Lampson ~~André Gide~~

In this chapter on informatics or computerization of surveillance systems, we will first explore what is technically possible in computerization of surveillance, ~~finding~~ measuring ~~an enormous~~ the gap between this and the best of today's actual systems. The barriers to optimal use of computers in surveillance---mostly social, organization, and legal---are explored. The remainder of the chapter explores some of the problems that must be

confronted in thinking about microcomputer-based surveillance, leaning heavily on examples from the notifiable disease system in the United States.

~~OVERVIEW OF A SURVEILLANCE SYSTEM IN THE FUTURE~~

(1) An Ideal Future Surveillance System

Ideally the epidemiologist of the future will have, a computer and communications system capable of providing management and surveillance information ~~on all these phases and~~ that is also capable of being connected to individual households and medical facilities to obtain additional information.

¶ Suppose that the epidemiologist of the future has a computer with automatic input from all inpatient and outpatient medical facilities, with standard records for each office or clinic visit and each hospital admission. S/he chooses to compare today or this week with a desired period, perhaps the past 5 years, and the computer produces ~~displays or prints~~ a series of maps for all conditions with unusual patterns. One of the maps seems interesting, and the epidemiologist may point to a particular area and request more information. A more detailed map of the area appears, showing the data sources that might provide the desired information, with estimates of the cost of obtaining the items desired. A few clicks of the mouse button select the sources, types of data, and format for a display, and the computer spends a few minutes interacting with computers in the medical facilities involved—extracting information and paying the necessary charges from the epidemiology division's budget. Soon the more detailed information is displayed on the epidemiologist's computer screen.

The pattern of hospitalizations and outpatient visits for asthma stands out, and the epidemiologist requests a random sample of specified size, of persons who have ever had asthma in the same area, matched by age and gender, to serve as controls for a case-control study. The video-cable addresses of these "controls" and of the ~~"case"~~ patients are quickly produced through queries to appropriate local medical-information sources. The epidemiologist formulates several questions about recent experiences, types of air conditioning, visits to various public facilities, and the like, adapts these to a previously tested video questionnaire format, and requests that video interviews be performed for case-patients and controls. Each household is contacted or left a FAX-like request to tune to a particular channel and answer a 5-minute query from the

state health department on a matter of importance to public health. Eighty-five percent of the subjects respond to the first query, and the computer automatically follows up with the rest, bringing the response to 92%, with half of the remainder reported to be absent from their homes for at least 2 days.

The odds ratio for persons with recent hospitalizations for asthma who work in or visit in a particular neighborhood is considerably higher than 1.0, and the epidemiologist connects by local-area network to the state occupational surveillance system and requests a display of all factories in the relevant area. Selecting those that deal with possibly allergenic materials, s/he issues a request for a more detailed investigation of activities at the plants in a selected time interval. The epidemiologist also requests information from the weather bureau on wind direction and velocity, temperature, and rainfall.

Within a few hours, a plant is identified that is in the process of moving a large pile of by-products with a bulldozer. A request is issued that the by-product be sprayed with water to prevent its particles from becoming airborne, and the plant manager readily agrees, when shown the maps that depict hospitalization rates for asthma downwind from the plant. To monitor progress and widen the investigation, the epidemiologist asks the computer to do similar studies for conjunctivitis and for coryza or hay fever over the previous and next 2 weeks. Selecting several maps and tables to include in the report, s/he asks the computer to write a description of the studies performed and the findings, and then dictates a brief summary of the problem and several follow-up notes to the voice port of the computer. ¶ At the end of 2 weeks, the number of cases of asthma has fallen to normal for the area. ~~and~~ The computer calculates on the basis of the number of medical visits during the outbreak that $55,000 has been saved at a total cost of a few hours of the epidemiologist's effort, a site visit to the plant, and charges of $9,500 for the data and the communication facilities used to perform the interviews.

(1) ~~(2)~~ Barriers to the Ideal Surveillance System

Obviously, we are a long way from implementing the system described above. It may be helpful in thinking about the future to explore what barriers must be surmounted before this scenario can be enacted. Strangely enough, few of them are technical; all of the necessary systems could be built today with fairly conventional equipment and

software, with the exception of the two-way interactive video connection with each household. This hook-up with the individual household is more likely to be available within the next 10 years than is the connection between the physician's record files and the health department. In fact, the two-way interactive video link between the household and the outside world is simply awaiting the government's or the marketplace's decision on what format will be used and on the realization of the benefits of such a connection on the part of the entrepreneurs and the public.

However, there are some difficult problems to be solved before the "ideal system" can be implemented. ~~They include the following:~~

~~a) *The rapid availability of standardized, computerized medical records*. Several issues need to be addressed before such a system is possible.~~ In the United States, for example, there is a profusion of computerized medical-record systems for inpatient and outpatient records as well as insurance and other purposes. ~~have been developed~~ The~~se~~ existing systems contain a plethora of different variables and use many different formats. Until a simple core public health record of age, gender, geographic location, diagnosis, and a few other items is created for each outpatient visit and each hospitalization—and is available in a standard format without delay—the responsive interactive system described above will remain~~s~~ an unrealistic pipe dream. ~~An additional problem is that~~ Most medical records are still not more than partially computerized.

full measure

The barriers to establishing standardized public health output from computerized medical records are primarily political and administrative; most large retail organizations create records ~~of similar size~~ for each item sold, and the average item~~s~~ carr~~y~~ies ~~on average~~ a much lower price than the cost of a visit for medical care. Once there is the will to establish a national computerized medical record system, the technical hurdles will be readily overcome. The needs include standard but suitably flexible record formats, solutions to problems associated with confidentiality, incentives to create the records (including the assurance of appropriate and cost-effective use of the records), and voice ~~output.~~ input.

Another problem is the lack of recognition that information about patients, except for legally designated "reportable diseases," is useful in public health and should be available to public health agencies. The level of awareness could be heightened if technical solutions to problems of confidentiality were publicized and understood by the public and their legislative representatives. Such solutions as one-way encoding algorithms could provide ~~partial~~ solutions to matching and follow-up problems, if properly used without turning public health agencies into ~~carbon copies of dreaded~~ "big brother."

full measure

~~Third~~ Common problem is the A pervasive feeling among those in charge of data that their data base must be "clean" before anyone else can use it. Months or even years are consumed while corrections and updates are made to make the data as accurate as possible. Although from one perspective this quality control is necessary and important, the concept of "surveillance" includes rapid turnaround, a realization on the part of everyone concerned (even the media and the public) that the data are preliminary, and the understanding that in order to look at today's data today, one must be willing to accept today's imperfections. ~~This~~ Several kinds of mental shifts, as well as corresponding technical developments, will be necessary before a computerized system can be used to examine automatically a "time slice" of disease and injury records that originate in clinics and hospitals. Imperfections will be everywhere, and methods must be found to cope with reality—even if it includes warts—on an immediate basis.

(1) The Technology of the Future

As stated above, today's technology, given enough social and organizational development, is adequate to allow the creation of miracles in public health information and communication. Nevertheless, it seems likely that development in technology will continue to ~~reflect~~ be more of a driving force in public health computing than progress in political and social organization.

Technologic developments over the next decade will probably include the areas ~~shown~~ discussed below.

High capacity storage devices

CD ROMs (compact disk with read-only memory) similar to those used for music make it possible to have access to large bibliographic data bases anywhere there is electricity. The MEDLARS data base of the U.S. National Library of Medicine can be searched from a clinic in Africa; once there are lower prices for books on CD ROM and they include needed illustrations, it will be possible to take a medical library anywhere in a briefcase. Past data bases from the United States and elsewhere will become available on CD ROM, although the process of cleaning them up for this purpose often reveals gaps and inconsistencies that reflect changing definitions and diminish their value as consistent anchors for comparison.

Networks

A local area network (LAN) is a system linking microcomputers, terminals, and workstations with each other and~~/or~~ a mainframe computer to facilitate sharing of equipment (e.g., printers), programs, data, or other information. LANs are transforming the way many agencies do business. The most noticeable effect is the transmission of written memoranda that could or would not have been typed, packaged, and sent through a paper system. The cost of installing and supporting a LAN is not small, particularly in terms of support personnel. Uses for surveillance include entering data at multiple computers connected by a LAN. This requires special software to protect against errors. Special precautions to protect confidentiality are necessary in a network, if several people enter data in the same file at the same time.

New user interfaces

The parts of programs that interact with users have become easier to understand, and more attractive, with pull-down menus, windows, and pointing devices such as the "mouse." This elegance has its cost in terms of requirements for faster computers, for more memory, and particularly for greater skill to produce such programs. Some new programs cause unexpected problems when run with older programs or on older computers. All in all, the trend is toward a standard set of screen "controls," like those in modern cars, but the path in that direction is replete with experiment and minor failures.

New programming tools

It is widely recognized that software production is the narrow point in the implementation of new ideas in computing. Useful software still requires hundreds of thousands of lines of hand-written and highly personal "coding." Many new trends such as fourth-generation data bases, computer-assisted software design (CASE) tools, and object-oriented design have made programming more productive, but this area of new tools is one in which major advances would create revolutionary changes.

Higher-capacity processors and more memory

The almost miraculous advances in computer speed and memory capacity in the last decade have removed many of the limits that required use of mainframe computers or minicomputers rather than microcomputers. Now almost any project can be done on a microcomputer or several microcomputers connected by a LAN if there is sufficient motivation.

Video and computer integration

Photographs and fully functional video will soon be appearing on our computer screens. Although this may have the greatest impact in pathology, ~~and~~ radiology, and education, it also increases ~~alters on~~ opportunities to use color and three-dimensional dynamic displays for epidemiologic data. The possibilities for computer interaction via ordinary television sets are exciting, because every epidemiologist (and market researcher) can savor the possibility of interviewing citizens via cable television with the results captured immediately in computerized form. The medium offers new challenges in identifying responses that result from the various stages of humor, exasperation, or intoxication that citizens may undergo in the privacy of their homes.

Au: OK?

Voice and pen input

Systems are available now that identify thousands of spoken words (for tens of thousands of dollars) and allow for a crude interaction between voice and computer. Computers that recognize handwritten text of reasonably structured type are being sold currently. Presumably the rather elementary state of computerization of medical records will undergo a quantum leap once such systems allow medical staff to dictate to the computer without typing and preferably without being near a computer. When medical handwriting is replaced by voice dictation into a lapel microphone, real progress may occur in the use of computers in both clinical medicine and public health

settings. As stated above, however, realizing real public-health benefit from such technology will require dramatic social and legal changes.

(1) ~~BACK TO THE PRESENT:~~ COMPUTERIZED PUBLIC HEALTH SURVEILLANCE TODAY ~~IN 1992~~

Beginning in ~~Since~~ 1985, Centers for Disease Control (CDC) staff have installed and maintained customized disease-surveillance software in 40 ~~36~~ state health departments and a number of county, district, and territorial departments. The software has been based on *Epi Info*, a public-domain word-processing, database, and statistics package for IBM-compatible microcomputers that is a joint product of CDC and the Global Programme on AIDS, World Health Organization (*1,2*). These systems have made possible the participation of all 50 states in the National Electronic Telecommunications Surveillance System (*3,4*). Benefits cited in a recent evaluation include improved access to data and improvement in both quality of data and access associated with decentralized entry of data (*5*).

Although reportable-disease systems are a specific kind of surveillance system and *Epi Info* is only one type of database ~~/~~ and statistics program around which a system can be built, many of the principles of computerization apply to other systems. ~~To avoid empty generalization,~~ Much of the rest of this chapter is based on CDC's experience with reportable-disease surveillance using *Epi Info*. The information is directed to those considering computerization of a disease-surveillance or similar system of records, whether they ~~wish~~ plan to do their own system design or will be working with a professional computer-systems designer.

Computerizing a surveillance system for disease is not easy. Since the success of computerization depends as much on the administrative and epidemiologic environment as on the software, it is vital that public health practitioners understand the details of a new system and participate in its design. The most important step in developing a computerized surveillance system is identifying the public health objective for the system. In some cases, the objective~~(s)~~ will have been clear for decades in a manual system ("Identify and treat or isolate cases of X and evaluate results," or "Assess results of immunization programs and identify new cases for special control efforts"). Computerization can then be directed toward accomplishing the same task more efficiently or in greater volume or detail.

¶ The most successful computer systems, however, are those that change methods by which an agency operates rather than those that merely automate a manual task (*6*). In establishing a new surveillance system or reexamining an existing system, it may be useful to address the following question: "What key pieces of information do I want to see on my desk (or computer screen) every day, week, month, or year that will make my work easier or more effective?" The same question can be asked at several levels of management--from epidemiologic technician to epidemiologist to director of a public health agency.

~~Given~~ If a surveillance system that is already in place has a public health goal and to some extent achieves the goal, why computerize? Sometimes the answer is obvious-- ~~because~~ the annual report takes a herd of clerks 2 years to process," or ~~"we~~ employees like the graphs health department A turns out so easily with their computer." Potential benefits relate to quality of data or of reports, quantity of data that can be processed, and speed of processing. ~~Dissemination~~ Transmitting (copying) of surveillance records to another site is one reason disease reports in all 50 U.S. states are computerized.

We were unable to find systematic studies on the benefits of computerizing public health surveillance systems, although numerous articles describe individual systems that have been computerized (*7*--*10*), and Gaynes *et al.* (*11*) describe methods for evaluating a computerized surveillance system. In literature about the commercial world, benefits of computerization have been examined from the viewpoint of financial savings. Savings by automating a manual information process may amount to 20% or so, but the real benefits are achieved if computerization transforms the entire process concerned, giving a competitive advantage in the commercial world--which would correspond to a new order of service in the public health world (*6*). So far, most public health applications have automated manual systems, although some--such as the spreadsheet calculation of the impact of smoking on populations--verge on establishing new and previously unknown styles of doing business (*12*).

One problem ~~cited~~ associated with developing a computerized surveillance system also occurs in other "vertical markets" (industries with specialized practitioners) such as the construction, meat-packing, and real estate industries. With only 7,000 epidemiologists in the United States, relatively few commercial developers feel that it is financially worthwhile to develop software for this market

alone, since applications such as spreadsheets, languages, and word processors may sell millions of copies to the general public (*13*). ✓

2 Basic Needs

The first requisite for computerization is a paper system or operational design that works reasonably well or would do so if the process were speedier and more accurate. Chaos computerized is not necessarily an improvement over what is already in place, although the process of computerization offers a chance to rethink some of the features of a system and to make improvements. If the surveillance system is a new one, it may be desirable to evolve the computer facilities in small stages with minimal investment until the system proves to be useful and well-conceived. This requires a careful plan (including provision for changing the plan if necessary), but will minimize the expense of adaptation as the epidemiologic design of the system undergoes the inevitable adaptation to external reality. After the "bare-bones" system has proven its worth and the probability of expensive changes is lower, the "bells and whistles" can be added later.

¶ Personnel to do the collection of data, data entry, analysis, and system maintenance are important contributors to the system. Many of the tasks can be learned by current employees, particularly if they find this challenge welcome. If possible, those chosen should be long-term employees to assure stability of the system, although they may be aided by students and other temporary employees. The epidemiologist who will use the results should participate in the planning of the system and should understand how it is constructed. A staff member with some programming skills ~~and/~~or aptitude for microcomputing should be involved in designing and setting up the system, even if an outside consultant does the actual programming.

If several computers are to interact and share data, a set of standards is necessary (e.g., just as humans carrying on a conversation need a common language). In the United States, the states and CDC chose a standard record format so that computers of different types could reformat data to a set of standard records and send these to the central agency. This standard, first devised in 1984 and revised in 1991, has served the purpose well, without placing unnecessary restrictions on the type of hardware or the format of records kept within each state. One state maintains 20 times more information for local use than do other states, but all export the same standard

record formats to the national level. The new standard record format allows for standard demographic and diagnostic information, attachment of variable-length detailed reports for selected diseases, mixture of summary with individual records, and automatic comparison of state and national data bases with each transmission.

Most government settings have an organization in charge of computer programming, approval of new systems, and purchasing of computers and software. It is important to maintain liaison with this organization and to arrange its assistance ahead of time with difficult areas such as purchasing computers. In some organizations, purchases are limited to particular types of computers-occasionally with unique characteristics-or to centrally administered systems. We recently encountered a network of "diskless" workstations that presented numerous problems in trying to load or run software or back-up files from a particular station without a removable storage device. If such problems are present, it is prudent to discover and, if possible, to surmount them at an early stage through patient negotiation and collaboration, or other methods if necessary. The technical difficulties that arise in setting up a computer system are usually the easy problems; the difficulties that lead to months and years of delay and unhappiness usually reflect misunderstanding and miscommunication among individuals or organizational entities.

2 ~~Some Key Concepts:~~ Files, Records, and Fields

Computerized records are stored in files. A file is a collection of records, usually one record per case, that has a name (e.g., GEPI.REC, for General EPIdemiology) and can be manipulated as a unit. Files, like books, can be opened, closed, read, written to, or discarded. They are stored on nonvolatile media such as hard or floppy disks or magnetic tape.

¶ Records correspond to one copy of a completed questionnaire or form, such as a disease-report card. Usually, one disease report or questionnaire is stored in a file as a single record. Records can be displayed on the screen, searched for by name or some other characteristic, saved (written) to a disk, or marked as deleted. Many records can be stored in each file.

A field is one item of information within a record. NAME, AGE, and DATEONSET might be fields within a disease-report record. Records in a particular file all have the same

fields. Each field has a name, a type (text, upper-case text, numeric, date, etc.), and a length, such as 22 characters for NAME or 3 for AGE. During analysis, fields may be called variables, and commands such as "TABLES DISEASE COUNTY" are used to instruct the system to process a particular file and construct the desired table by tabulating the fields or variables called DISEASE and COUNTY. In this case, the result in *Epi Info* would be a table that lists DISEASE down the left side and COUNTY across the top, with numbers of reports by county indicated in the cells of the table.

Hardware: What Size Computer is Appropriate?

With microcomputers costing a few thousand dollars, ~~being available for much less than $5000~~, it is possible to process more than 100,000 records in reasonable time periods. Processing time tends to reflect the record length as well as the number of records, however, and the size of each record should be kept short if large numbers will be processed. Since the total number of disease reports for the United States is several hundred thousand per year, states and counties should find it possible to build most systems on a microcomputer if desired.

Minicomputers and mainframes can serve as the basis for surveillance systems if available at reasonable cost and if programming and support staff are available to work creatively with staff of the surveillance system. The greater technical skill required to run and program such computers often resides in an organization other than the one running the surveillance system, and close coordination becomes much more important than in the do-it-yourself situation with a microcomputer.

Systems that seem to require processing of millions of records, such as hospital discharge or Medicare records for a state, can be reduced by sampling to a manageable size for the microcomputer. The mainframe can be used to select a sample of records (e.g., particular age groups or diseases, or every tenth record~~, or persons born in decade years~~). Files are then exported for processing on a microcomputer that is more responsive to the epidemiologist's wishes. Epidemiologists are usually acutely conscious of sample size when performing interviews but sometimes fail to recognize how unnecessary it is to process 6 million records to estimate a simple proportion.

Software

¶ The type of software used to perform the computerization is often less crucial than the skills of those who will program and run it. Usually, there are several types of data base or statistical packages that will do a given task well if properly programmed. Beware of the "indispensable programmer" syndrome, in which a single expert programmer writes a system in his or her favorite language and then departs for greener pastures, leaving the users without resources for further ~~maintenance~~ support and modification.

Data base packages such as ~~dBase~~ dBASE, Paradox, Foxbase, and Clipper are designed to allow data input, storage, retrieval, and editing. Most will count records but do not easily do such statistics as odds ratios. They require a skilled programmer to produce a customized system.

Statistics packages, such as Statistical Analysis System (SAS) and Statistical Package for the Social Sciences (SPSS), focus on producing statistical reports, usually from single files of data. They are less convenient for data entry. Both SAS and SPSS now have mainframe and microcomputer versions. They contain many routines rarely used by epidemiologists and occupy large amounts of disk space (tens of megabytes for SAS).

Epi Info fits on a single 1.4 megabyte diskette and provides a combination of data-base and statistical functions, allowing relational linking of several files during data entry or analysis. Questionnaires or forms may be up to 500 lines, with hundreds of numeric or text fields, and the number of records is limited only by disk storage space. Frequencies, cross tabulations, customized reports, and graphs can be produced through commands contained in a program file or interactively from the keyboard. Commonly used epidemiologic statistics are part of the statistical output. Although it takes little experience to use *Epi Info* for investigating outbreaks, producing a complete surveillance system from the beginning takes both skill and time. It ~~may,~~ is however, ~~be much~~ simpler to modify the surveillance software supplied with the program.

It is important to realize the limitations of software packages before they are used. Both statistical and data base packages typically cost at least several hundred dollars and therefore are not likely to be feasible for classes of students or large numbers of ~~[illegible]~~ computers.

no paragraph

Some database packages limit the number of fields in a record or the number of records in a file, and few will do statistics without advanced programming or purchase of a supplementary package. Statistics packages, on the other hand, may have limitations in handling textual ("alpha") data, and most allow processing of only one file at a time. A complete surveillance system may require the functions of both database and statistical programs.

(5.01b) That's a l.c. "B" not a six

The current version of *Epi Info* has limitations on the number of records that can be sorted or relationally linked at one time (tens of thousands), a limitation that will be removed in version 6. ~~however, and~~ Since text fields are limited to 80 characters, *Epi Info* would not be a good choice if large amounts of text are to be stored, as in a complete clinical system containing dictated notes.

(2) Designing Entry Forms

In a surveillance system, data items are usually entered in a standard format (e.g., a questionnaire or report form). The information is stored in files containing one record per individual. In *Epi Info*, the format of the database file is specified by typing a questionnaire or form in the word processor. The result resembles a paper form, with entry blanks indicated by special symbols (e.g., underlined characters for text fields and number signs for numeric fields). The computer reads the form and constructs a file in the proper format.

¶ In designing a form, it is useful to include a unique case identifier as a number ~~of~~ or combination of letters and digits. This may include meaningful information, such as the year, but should not include any item that may need to be changed, such as a disease code. It must be designed so that a new and unique number will always be available for each record.

The amount of data entry and computer storage required may be minimized by computerizing only information that will actually be used. If follow-up information such as name, address, and telephone number can be used from the paper form, there may be no need to enter it into the computer. If contact tracing is recorded, the computer record may summarize the number of contacts named and the number found or treated, with the details on each and progress of the follow-up efforts relegated to the paper forms used by field investigators. When including an item on the input form, it is helpful to ask, "How will this be analyzed?" and "How would the result

look after processing? Computers around the world are full of data items that someone entered "just in case we need it." Most are never needed.

Textual material can be printed from a computer file, but it is usually difficult or impossible to process such entries as "Pen, Strep, and Ampicillin," to produce meaningful tabulations. For serious analysis a more usable format would be

Dis

Penicillin	<Y>
Streptomycin	<Y>
Ampicillin	<Y>

in which "<Y>" represents a blank for a "Y" or "N" response.

¶ A common problem in designing entry forms is that several data items may be similar. Suppose you want to record name and treatment (RX) status for up to 12 contacts of each [redacted] patient. One possible approach is to create fields called NAME1 through NAME12 and RX1 through RX12. This approach allows the data to be entered, although it creates a very large data-entry record (say 12 x 22 characters for NAMEs and 12 x 1 characters for RX=276 characters, even if no information about contacts is entered). However, analyzing the information becomes a programming nightmare, as determining the number of contacts or their treatment status requires examining at least 12 different fields in each record to see whether they have been filled in and keeping a running tally of the results. In computer database jargon, the record is not "normalized." These repeating groups of fields should be placed in separate records—one for each contact—linked to the main file as described below in the section on linking special-purpose records. Then a [redacted] patient with one contact has one record in the case file and one record in the contact file rather than the equivalent of these plus 11 empty records in a single file.

This problem is resolved by rethinking what is really the best unit around which to build an individual record. The simple answer is that if you intend to tabulate cases, build a case record; if you will tabulate contacts or follow-up visits, then you need a contact or follow-up record. If both are necessary and the system is large or permanent, records should be placed in separate files and linked using relational database features as described below.

2 Data Entry

¶ The details of data entry should be determined and documented, including who will prepare the paper records (if needed) for entry, who will enter them, and at what intervals. The status of the report as "suspected" or "confirmed" may determine whether it is entered, and this must be determined at the outset. Most disease reports are entered in batches—once a week, for example—and in many states not more than an hour or two is needed to enter the data for a week, although the ~~quantity~~ number of records varies ~~[illegible]~~ greatly ~~in size in~~ ~~different~~ among states, and ~~correspondingly in~~ time required to enter data.

Records linked to more extensive specialized forms can be sent as partial submissions and revised later to avoid delays in reporting caused by the slower progress of data collection for the more detailed forms. This issue needs to be considered and resolved in advance.

2

Cleaning and Editing the Data

Errors or duplications inevitably occur during data entry, and additional information may arrive that requires changes or additions. The data can be "cleaned" during data entry or with the help of analytic programs that display "outliers," and data can be checked visually by browsing through records in the ENTER program or by scanning a list printed by the ENTER or ANALYSIS programs. Records can be viewed and corrected in a spreadsheet format in ANALYSIS. Finally, a program called VALIDATE can be used to compare files entered in duplicate by different operators. Records showing differences in the two files ~~different entries~~ are printed out for reconciliation.

¶ *Epi Info* **allows extensive programming** of error checks on data entry. Each field can be set to **accept only specified** codes, and, if necessary, multiple fields can be checked for **inconsistencies** such as gynecologic conditions recorded for males. Unfortunately, many errors cannot be caught by such systems, and one can still enter the wrong code for a less gender-specific disease. Another method involves running a special checking program after all records have been entered.

Regardless of the method used, errors should be caught and corrected near the time of data entry if possible, since they can create much larger problems if left for the end of the year. The choice depends largely on the orientation and number of personnel available, and perhaps on their preferences after trying different methods.

Analysis of Data

The type of output desired should be planned in advance, since the inputs and outputs usually specify fairly precisely what kind of processing is needed to achieve the result. Dummy tables and graphs should be sketched on paper. *Epi Info* and many other data~~-~~base programs can be programmed to print a table or mixture of text and tables in almost any format, using a feature called the "report generator."

It is not necessary to design reports to cover all possible needs, since ad hoc queries are an important part of any system, and additional reports can be added later if they are deemed useful. In *Epi Info*, an epidemiologist can learn to do simple queries (READ GEPI; TABLES RACE COUNTY) in a short time and to limit these to particular time periods (SELECT REPORTWK = 34) almost as easily.

Sometimes a simple report such as a list~~ing~~ of this week's ~~reports~~ entries, sorted by disease, may be as useful as a ~~number of~~ large table~~s~~ with very small numbers in each cell. The number of records ~~available~~ should be considered in designing reports and in determining how often they will be produced.

Distributed Data Base

So far, we have described a surveillance system in a single microcomputer. As more community health departments obtain computers, however, the trend is toward networks of computers within a state, connected by modem in ways analogous to those used in the National Electronic Telecommunications Surveillance System (NETSS), ~~with~~ which have more than ~~its~~ 50+ state and territorial participants. Each participating site enters data and sends them periodically to a computer at the next level up.

This process would be simple if all data were entered at the local level and sent to the state level, and if no changes were made later. However, in practice, not only are changes made, but in some states records are entered at both state and local levels. Some method must be in place to see that both levels of staff eventually have the same records.

Ideally, only one copy of the records would be considered the "master" copy, and each user would know its location and provide updates only at the designated time. The

best way to accomplish this objective is still being worked out, and experiments of several types are likely. Designating only one of the sources as the "owner" and rightful editor of the data is one possibility. At present, we favor indicating on each record the site at which it was created and allowing only that site to make changes that are transmitted weekly to the other sites to update their copies of the records.

State health departments use the latest software to transmit year-to-date summary information on the state data base to the national level each week. These ~~data~~ summaries are compared automatically with the contents of the national data base, and any discrepancies are reported.

(2) Transmitting Data

In NETSS, most states transmit reports each week through a commercial telecommunications network. The 50 plus reports stay in the network computer until they are picked up on Tuesday morning by CDC staff, stripped of comments and address material, and joined together in a single file for processing on the CDC mainframe. Error checking is done to test for invalid codes and other problems, and error notices are sent back to the states.

¶ Another method that eliminates errors caused by telephone noise involves transmission directly from computer to computer by means of modems and software that retransmits if errors are caused by noise. Several states are using this method to connect with CDC microcomputers that, in turn, send the files to the CDC mainframe.

A third less elegant but often practical solution is physical transfer of floppy diskettes by mail or messenger at intervals. This allows large files to be transferred with minimal inconvenience, and may be appropriate if the additional trouble of setting up modems and software is not yet warranted or in ~~developing~~ countries where telephones are unreliable or unavailable.

In any case, the result is that a copy of a file of records from the peripheral site arrives at the central site. The records must then be merged into the main data base. If all are new records, this task is straightforward. If the incoming records contain updates for records previously transmitted, the process is more complex.

(2) Correcting and Updating Records from Another Site

In NETSS, only state participants are allowed to update records; CDC staff do not do so, although they may enter temporary telephone reports. Updates are sent as records with the same identification number as that for the original record. If a new record has the same identification number as a record in the data base, the existing record is updated so that all non-blank fields of the new record prevail. To change an age, for example, a state would send a record containing the case identification number and the new age. To delete a record, the state, year, and identification numbers are sent in a special "Delete" record. When errors are found at CDC, the information is transmitted to the state staff, who then correct~~s~~ the errors and transmit update records the following week. This system is quite dependent on adequate staffing and attention to detail.

(2) Individual and Summary Records

Many systems function with a record for each individual case report. In some, however, there is a need for summary records, each of which represents a number of case reports. This is helpful if large numbers of similar records (e.g., cases of gonorrhea in a big city) are processed, or if only summary numbers are available. It also allows records from entire years to be summarized in condensed format, so that a 5-year trend can be calculated without reading and processing each record for the previous 5 years.

¶ A summary record is similar to a case record, but it contains an additional field called "COUNT," which contains a number. The number indicates how many records with the same information are represented by the summary record. *Epi Info* contains commands called SUMTABLES and SUMFREQ to process summary records. These commands sum the contents of the count field rather than counting individual records. Since a record with COUNT equal to 1 is an individual case record, files that are mixtures of summary and individual records can be processed as a single unit.

(2) Linking Special-Purpose Records to the Main Data Base

As mentioned above, sometimes it is necessary to link related records in different files together in order to allow easy processing of (for example, [redacted] patients and contacts who are related to [redacted] patients). This requires that a common [redacted] identification number be included in each record. *Epi Info* and other database

programs, such as dBASE, allow automatic linking of records through such a common identifier. On data entry, answering Y to the question Contacts (Y/N)? might cause another form, representing the contact file, to appear on the screen. The operator can then enter one or many contact forms for this case, pressing a function key (F10) to return to the main form. A separate record is created for each contact.

In *Epi Info*'s ANALYSIS program, the CONTACT file is READ, and the CASE file is linked (related) to it. Each contact record then contains information about the patient as well as about the contact, and questions such as how many contacts of female case-patients were treated? can be answered easily. The CASE file can also be processed alone to answer questions such as how many cases of syphilis were there?

We also link disease-specific forms to the main data base of reports. Hepatitis, for example, requires a full page of extra information used to define further the epidemiology of a report. By linking a hepatitis file to the main case file, records are created only if the disease is hepatitis, thus saving a great deal of storage space over the single-file method, in which all the questions on hepatitis would be left blank in a non-hepatitis record. Current systems, including the one distributed as an example on the *Epi Info* disks, contain related files for hepatitis, meningitis, and ~~enteric~~ Lyme disease, each of which only appears if a relevant disease code is entered.

2 **Dissemination of Data**

Dissemination of results is an important element of the surveillance cycle. Computerization can assist by making new methods of analysis or presentation practical. Use of tabular or graphics software in conjunction with desk-top publishing technology can make the preparation of results not only faster but more accurate and meaningful. A graphic method for comparison of current results with those for the past 5 years has been introduced to the *Morbidity and Mortality Weekly Report* in the United States (Figure 5-12) (*14*). This method would have been too cumbersome for manual processing.

Computer software greatly simplifies and improves the production of maps and graphs. *Epi Map*, a public domain companion to *Epi Info*, to be released in 1993, will make mapping available to anyone with an IBM-compatible microcomputer.

Tables, maps, graphs, text, and data files may be made available either on-line via modem connections or by distributing floppy or CD-ROM disks. The latter are particularly useful in remote areas or for large volumes of data that ~~can be~~ cannot easily be sent over low-speed modems.

Data Disasters

Destruction or damage ~~or~~ of data on hard disks should be expected and planned for. During the first 4 years of NETSS (and during the 3-year tenure of its predecessor, the Epidemiologic Surveillance Project), a number of hard disks have "crashed." In most cases, back-up files on floppy diskettes had been properly prepared and stored, and they were used to restore the data once the disk had been replaced.

Recently, some state programs began to reuse case-identification numbers from several years ago, not realizing that the new records would overwrite the old records in the national data base. It is important to be clear about the time period for which updates will be accepted. Recycling of identification numbers should be avoided if at all possible.

Upgrading either hardware or software is a frequent cause of problems. stet ~~when the~~ The new items may have unexpected features, occupy more memory space, or require that protocols for functions, such as communications, be changed.

Computer viruses are an increasing cause of problems. They can cause a variety of difficulties ranging from erratic behavior of software to complete loss of files. They may be introduced from networks, by accessing other computer bulletin boards, or by loading copied software from unknown sources. Programs to detect and eradicate computer viruses are available commercially. It is essential to install one of these and to be sure that any disk from an external source is scanned for viruses before it is copied or used as a source of new programs.

Backup Methods

Methods for disaster prevention center around regular backup of data files onto floppy diskettes (or tape if available, but beware of tape backups with only one compatible tape drive in the same institution). The back up copies should be rotated so that several circulate in turn and so that the one overwritten has at least two more recent

relatives. To protect against fire, water damage, and damage by panic-stricken personnel, it is wise to keep at least one backup in a site remote from the computer. Setting the write-protection feature on the diskettes after making the backup is an additional protection.

¶ Upgrading hardware or software should be done at a time when use of the system is least critical, and care should be taken to allow for replacing the old system exactly as it was if problems occur with the new one. Thus, before installing a new version of software, the old one should be thoroughly backed up or, preferably, left in place in another directory so that it can be used if necessary.

2 Training of Staff and Transition Techniques

We have found that the most effective staff training occurs by having potential operators participate in the design of the system and receive short demonstrations and hands-on lessons at the time the system is installed. Usually installation of a system takes ~~two~~ 2 or ~~three~~ 3 days for planning and decision making, ~~two~~ 2 or ~~three~~ 3 days for programming, and a similar period for staff training, trial runs, and revisions.

¶ National meetings and training sessions for operators of state surveillance systems have been helpful in providing extra training and motivation and in ~~surfacing~~ revealing ~~[illegible]~~ problems that need to be addressed and new ideas for software improvements.

During the transition from a paper to a computerized system, both systems are run in parallel for a period until the results are satisfactory and staff feel comfortable with the new system.

1 CONCLUSIONS ~~DISCUSSION~~ (SC)

Au: OK?

The old image of the computer expert in an expensive suit handing the client the keys to the new "turn-key" system perfectly adapted to his ~~or her~~ needs was probably always a fantasy, but with modest budgets, small data bases, and a desire for "hands-on" access to data, it certainly has little relevance to public health needs. Although in some ways centralized computers and instant interactivity for updating records would present fewer problems than the distributed systems we have described, public health workers usually do not require and cannot financially afford the instant updates

needed for law enforcement, banking, or airline reservations. Microcomputers and local data bases can maintain the data and analytic results closer to the professionals primarily responsible for prevention and control.

We are convinced that participation of all 50 state health departments in the national computerized system would have been impossible without (a) software for states that allow~~ed~~ customization for use of local forms and procedures, (b) participation of each state epidemiologist's staff in designing a system unique to the state, and (c) a standardized record format. Each state has a different input form, although the records sent to CDC are restructured and variable values are recoded by *Epi Info* programs so that they are in the uniform national format.

As systems become more complex, however, it is important to standardize as many features as possible from state to state so that a thoroughly debugged core system can be used by all. We are gradually achieving this with a new *Epi Info*-based system that has a series of standard modules, accompanied by other modules that are highly customizable.

As pointed out in this chapter, there is an enormous gap between what is technologically possible with the use of computers in public health and what is actually going on at the grass-roots level of public health practice. Until the keeping of medical records in clinical practice is computerized to a much greater extent, ~~it would be difficult to imagine that~~ our scenario of the future cannot become ~~will actually move closer to~~ reality.

Other key issues remaining to be resolved include (a) the balance between confidentiality and free access to clinical records for public health purposes, (b) the cost of data entry and of programming and processing, ~~and~~ (c) the ability of both professionals and the public to deal with incomplete ~~"dirty"~~ and preliminary data. (d) methods for unifying data with diverse formats.

OK

Many of these issues have both technical and social solutions. A great deal of work in both realms remains to be done before computerized public health surveillance can be said to have achieved its full potential.

REFERENCES

260 1. Dean AD, Dean JA, Burton AH, Dicker RC. Epi Info, version 5: a word processing, database, and statistics program for epidemiology on microcomputers. Atlanta~~, GA~~: Centers for Disease Control, ~~Atlanta~~, 1990.

260 2. Dean AD, Dean JA, Burton AH, Dicker RC. Epi Info: a general-purpose microcomputer program for public health information systems. *Am J Prev Med* 1991;7:178-82.

260 3. Graitcer PL, Burton AH. The epidemiologic surveillance project: a computer-based system for disease surveillance. *Am J Prev Med* 1987;3:123-7.

260 4. Centers for Disease Control. National Electronic Telecommunications System for Surveillance--United States, 1990-1991. *MMWR* 1991;40(29):502-3.

260 5. Odell-Butler ME, Ellis B, Hersey JC. Final report for task 8, an evaluation of the National Electronic Telecommunications System for Surveillance (NETSS). Arlington, Va.: Battelle, June 1991:49-50.

261 6. The big pay-off (benefits of computerizing a business) (node supplement). *IBM System User* March 1990:S20.

261 7. Mary M, Garnerin P, Roure C, et al. Six years of public health surveillance of measles in France. *Int J Epidemiol* 1992;21:163-8.

Stet

~~Au: List all authors~~

261 8. Centers for Disease Control. Surveillance of influenza-like diseases through a national computer network--France, 1984-1989. *MMWR* 1989;38(49):855-7.

261 9. Watkins M, Lapham S, Hoy W. Use of a medical center's computerized health care database for notifiable disease surveillance. *Am J Public Health* 1991;81(5):637-9.

26 10. Bernard KW, Graitcer PL, van der Vlugt T, Moran JS, Pulley KM. Epidemiological surveillance in Peace Corps volunteers: a model for monitoring health in temporary residents of developing countries. *Int J Epidemiol* 1989;18(1):220-6.

261 11. Gaynes R, Friedman C Copeland TA, Thiele GH. Methodology to evaluate a computer-based system for surveillance of hospital-acquired infections. *Am J Infec Control* 1990;18:40-6.

261 12. Shultz JM, Novotny TE, Rice DP. Quantifying the disease impact of cigarette smoking with SAMMEC II software. *Public Health Rep* 1991;106(3):326-33.

262 13. Call B. The ones that got away: why some industries have not yet computerized. *PC Week* June 24, 1986;3:125.

272 14. Centers for Disease Control. Proposed changes in format for presentation of notifiable disease report data. *MMWR* 1988;38(47):805-9.