ENCYCLOPEDIA OF MICROCOMPUTERS

EXECUTIVE EDITORS

Allen Kent *James G. Williams*

UNIVERSITY OF PITTSBURGH
PITTSBURGH, PENNSYLVANIA

ADMINISTRATIVE EDITORS

Rosalind Kent *Carolyn M. Hall*

PITTSBURGH, PENNSYLVANIA

VOLUME 6

Electronic Dictionaries in Machine Translation to Evaluation of Software: Microsoft Word Version 4.0

MARCEL DEKKER, INC. • NEW YORK and BASEL

EPIDEMIOLOGY AND MICROCOMPUTERS

INTRODUCTION

Epidemiology, the study of "the distribution and determinants of disease in populations," became a science rather suddenly around 1850, when Dr. John Snow [*1*], a London physician, studied cholera cases and their water sources and concluded that specific water supplies were associated with disease. His removal of the handle on the Broad Street pump, an implicated water source, has been a model of effective disease control efforts for 140 years.

In contrast to other health sciences, epidemiology focuses on populations rather than individuals and uses rates and other statistics as a principal tool. Careful counting of cases and of "exposed" persons is the basis of epidemiologic methods.

"Cases" are defined as persons in a given population having particular diagnoses, symptoms, laboratory findings, or other characteristics describing their state of health over a defined time period. "Exposed" persons are those with defined personal, temporal, and geographic characteristics considered to have possible influences on disease rates. Exposure variables may include food intake, use of drugs or devices, sexual practices, presence in a particular place, work history, genetic type, or other factors that may be associated with an increased risk of disease. Other variables that define populations with different rates of both disease and exposure are called "confounding variables," and these are also, of necessity, the subject of epidemiologic study, since they can cause incorrect results if not controlled for in the analysis.

Most epidemiologic studies are based on comparison of exposure in ill and non-ill people (a case-control study), or on comparison of illness rates in exposed and unexposed persons (a cohort study). Studies that compare two groups in this way are called analytic studies. Descriptive studies, which describe characteristics of only a single group of ill or exposed persons, are also frequently performed, sometimes to generate hypotheses before an analytic study is undertaken.

In the physical sciences, units of measurement tend to be continuous variables like mass, distance, and time. Epidemiology, like the social sciences, deals more frequently with categorical data such as sex (M/F), baked ham consumption (Yes/No), income category (1–5), age group (by 5 or 10 years), or illness (Yes/No).

Sometimes the categories are ordered (ordinal) as in age groups, sometimes not, as in race. Statistical methods to deal with categorical data are different from those for continuous variables, and epidemiologists have their own favorite procedures for statistical analysis that differ in emphasis, name, and application from those in other sciences. Computer statistics packages for general use often contain abundant resources for analysis of continuous data, but lack such epidemiologic staples as the Mantel–Haenszel

Submitted: March 1988.

analysis for stratified tables, the Fisher exact test for 2-by-2 tables, and logistic regression. They usually do contain chi-square tests, which are much used in epidemiology. In a later section on investigative epidemiology, we will mention the common procedures for categorical data analysis and several microcomputer packages that are designed specifically for epidemiologic work.

Epidemiology can be divided into at least two broad areas, which we will call "inductive" and "deductive" epidemiology. Public health investigation and public health research usually begin with data gathered about a population and proceed to form conceptual models which are compared with the actual situation to test hypotheses. This can be considered inductive (or investigative) epidemiology.

"Deductive" epidemiology, or mathematical modeling, constructs a hypothetical population from assumptions about its members, and then describes the consequences of a disease or exposure having defined characteristics. The model is sometimes tested for validity by comparing the results with data from real populations.

Both approaches use mathematical models to describe disease and risk factors in a population. The investigative method begins with data on a particular population and extracts rates or fits curves. The deductive method produces a hypothetical population and a mathematical model, and does not necessarily involve data from an actual situation, although this is often done for validation purposes.

INDUCTIVE EPIDEMIOLOGY: EXTRACTING MODELS FROM NATURAL POPULATIONS

Although within epidemiology "epidemic modeling" is often taken to mean the study of theoretical populations as described in the section of deductive models, it also applies in a broad sense to the construction of models, however simple, from populations as found in nature, using classification, counting, arithmetic, and sometimes more sophisticated statistics. Most epidemiologic investigations start with an affected population, attempt to count persons with and without exposure and disease, and then use statistics to examine possible associations. All of these processes can be assisted and made more accurate by computers. Many textbooks of epidemiology [*2—4*] and of epidemiologic statistics [*5—9*] are available.

In doing an epidemiologic study, the first step is to verify that the problem being studied actually exists and that the numbers of cases or exposures are truly in excess of what is expected for the population. The epidemiologist then performs the all-important political/administrative step of establishing contact with appropriate authorities and enlisting cooperation in performing a study. He or she then decides *what information to collect* by defining hypotheses, *from whom to collect it* by developing a study design, and *how many persons to study* by deciding on a sample size.

Often a questionnaire is developed to collect and organize the appropriate information. Collection of data usually is through telephone, mail, or face-to-face interviews or abstraction of information from medical records.

The data records are "cleaned up" by verifying, selecting, and recoding. Analysis is performed manually or on a calculator, microcomputer, large computer, or by a combination of means. In some cases, more than one study is performed simultaneously, or a descriptive study is performed

to generate hypotheses followed by an analytic study to test these hypotheses. When conclusions have been reached about the associations observed, control measures are instituted and a written report is prepared.

All of the processes can be assisted by computer programs that provide access to information or provide methods for manipulating data or textual material. Most available programs focus on analytic procedures and provide a variety of statistical procedures used in epidemiology, but computers can be used for many other steps in an investigation—either through specialized programs like the two "expert systems" described later or through creative use of spreadsheets and word processing programs.

Hypothesis Generation and Literature Searching

The early stages of epidemiologic investigation involve a quick review of the problem, perhaps through a descriptive study as described later, and then the generation of hypotheses through knowledge of the subject matter and/or searching the scientific literature. In their simplest form, these steps may be performed by public health workers, practicing physicians, citizens, or parents in communities, hospitals, industries, or schools. One or more cases of illness with known cause are questioned about contacts or recent exposures, a conclusion is reached, and appropriate action is taken. Any statistics can usually be done on the fingers of one hand, and the need for computer assistance lies in the realm of record keeping and access to medical information.

Descriptions of common infectious diseases and public health responses are found in *Control of Communicable Diseases in Man* [*10*], a printed handbook of information for public health use. For more comprehensive assistance in hypothesis generation, the National Library of Medicine MEDLARS system provides an on-line database of over six million reference citations to to journal articles, books, and audiovisual materials, that can be accessed through major telecommunciations networks, using a terminal or a microcomputer.

Grateful Med [*11*] is a user-friendly front-end program for MEDLARS which runs on IBM PC-compatible microcomputers and makes searches easy for persons without library training. Grateful Med is available from the National Technical Information Service, Washington, D.C.; the user must apply for an account with MEDLARS to use the service. A number of other commercial bibliographic services such as Paper Chase, Dialog, and BRS Colleague, offer access to MEDLARS, usually at higher cost.

Gathering and Quality Control of Data

Data for epidemiologic studies must meet several requirements not necessary for individual patient records. For a sample to be useful, it must bear a known relationship to the population being sampled, as either a random or a systematic sample. For an analytic study of ill people, the population of non-ill people must also be represented. Studies of exposed and unexposed persons must include both populations. Distortions of results caused by uncertain relationship of the sample to the denominator population are called "selection bias." Other types of bias are caused by inconsistent or inaccurate data collection, errors in data entry, laboratory inaccuracies, misdiagnosis, and other difficulties that can influence the accuracy of results. Much of the epidemiologist's work is in recognizing and minimizing various types of bias.

A major limitation in epidemiologic studies is the difficulty in obtaining data from defined populations. Information about ill persons and their exposures or risk factors (the two terms are used interchangeably) is still obtained largely from face-to-face interviews or from visual review of medical records. Large computerized databases such as death certificate files, insurance claim records, and computerized hospital records are useful for obtaining overall frequencies or trends, but are often disappointing inaccurate or incomplete for purposes other than those originally intended.

The diversity and lack of standardization of both medical services and medical records in most settings requires abstraction of data from handwritten records or oral interviews into a form useful for study. In-patient records in the United States—even laboratory data, consisting primarily of numbers—are recorded in hundreds of formats in different institutions. Patient identification numbers usually apply to a person only within a single institution and sometimes only for a single hospital stay. Billing records may describe families without distinguishing individuals. Computerized medical databases useful for epidemiology are found only in a few isolated institutions, and often do not relate to a population defined more precisely than as members of the institution's clientele.

In an epidemic, many ill persons may not visit a physician or they may visit many different physicians and so must be interviewed or examined or have their records abstracted uniformly to permit epidemiologic study. A portable microcomputer can be used to enter questionnaire information directly or to transfer the information from a paper record to the computer. Data entry may be done in a statistics or database program, or in a dedicated data entry and file creation program.

Analytic packages with specific epidemiologic features and data entry capability include (in alphabetical order) CRISP, Epi Info, EpiLog, Epistat, and True Epistat. Other more general packages are SPSS, SAS, and dBADE III and III Plus. Not all handle nonnumeric data equally well, and one should evaluate this as well as program size, price, and compatibility with other programs before deciding which one to use.

It is important in field investigations to use software that the investigator can operate comfortably and has employed for a similar purpose in the past. Unfamiliar computer programs have a way of consuming time unexpectedly, and learning time is at a premium in many epidemic investigations, particularly if one epidemiologist is doing the investigation and also trying to operate the computer. The ease of setting up or changing a questionnaire in the system is important, since investigations of acute problems often have to be done in a hurry, and hypotheses may change after even a few hours in the field.

Skilled programmers or those planning long-term studies may want to write dedicated data entry programs in BASIC, Pascal, or other languages, but this is seldom necessary, since several of the statistics packages allow data entry to be programmed with considerable specificity, and most packages provide means to transfer files to other formats.

More than 40 epidemiologists from the Centers for Disease Control assigned to work in states have been provided with portable IBM PC-compatible microcomputers to assist in epidemic investigations. They and a number of other epidemiologists now routinely enter data into microcomputers while in the field. Some use portable computers in hospital record rooms to abstract records without using an intervening paper form.

Data entry programs that incorporate range checking and other types of error prevention can play a useful role in providing accurate data for large studies where time permits setting up the necessary error checking. Software packages that permit rapid analysis of data can be used to screen for errors by producing tables showing extreme values and illogical conditions (e.g., male pregnancies), while they can still be corrected from the original data source.

Large research studies such as those assessing the effect of risk factors on cardiovascular disease, spend millions of dollars on carefully controlled interviews and examinations of defined populations so that accurate data can be stored for comparison with future disease rates. One study of risk factors for atherosclerosis funded by the National Heart, Lung, and Blood Institute uses microcomputers for data collection in centers in four different cities [*12*]. Each examination center has a number of microcomputers for entry of data as a study participant is interviewed and examined. The subject carries his own floppy disk from station to station, and paper forms are not used. Customized data entry programs provide error checking and formatting of the data so that ambiguities can be resolved while the subject is still present. At intervals the records are sent to a national center for processing on a mainframe computer. Coordination between centers is assisted by a microcomputer bulletin board accessed daily by each of the four centers.

Study Design

Most investigations involve careful collection of data from which a mathematical model is constructed of an ill or exposed population and possibly of a non-ill or nonexposed population. If a single group of ill or exposed persons is involved, the study is "descriptive." Adding a comparison group produces either a case-control study (ill versus non-ill) or a cohort study (exposed versus nonexposed). The four major types of epidemiologic study are shown in Table 1.

Another type of study, called the cross-sectional study, begins with an entire population, such as persons who attended a particular banquet, or were included in a survey. Conceptually, however, the information is analyzed as either a case-control or a cohort study, depending on how the problem is viewed.

The most common model employed for analysis of case-control or cohort studies is the 2-by-2 table, often set up as follows:

		Illness	
		+	−
Exposure	+	A	B
	−	C	D

Each letter represents the number of persons having the characteristics in the margin. Thus, C is the number of ill persons who were not exposed to the factor under study.

For case-control studies, the odds ratio AD/BC is used to reflect the degree of association between illness and the specified exposure. Counts

TABLE 1 Types of Epidemiologic Studies

Starting point	Descriptive study	Analytic study
Ill persons	Describe illness and exposure in the ill population (Case series)	Case-control study: Com Compare Ill and well persons with respect to exposure
Exposed persons	Describe exposure and illness in the exposed population	Cohort study: Compare exposed and unexposed persons with respect to illness

in the A and D cells support the hypothesis of association: those in B and C weaken it. If the sick are exposed and the well are not the association is supported. The chi-square test and an associated p (probability) value are often used to test for association, sometimes with a correction for small numbers (Yates' correction). Methods for calculating confidence limits provide a range of values of the odds ratio which are consistent with the data. The confidence interval may also be used as a test of significance. If the confidence limit includes 1.0, the level reflecting no association, the association is considered "not significant."

The measure of association most used in exposure-based (cohort) studies is relative risk, which is the proportion of illness in exposed persons, divided by the proportion of illness in the unexposed

$$\text{Relative Risk} = \frac{\frac{A}{A + B}}{\frac{C}{C + D}}$$

Chi-square tests, p-values, and confidence intervals can also be calculated for relative risks.

This simple model would be sufficient if the illness and a single type of exposure were the only differences between cases and controls. Except in carefully matched studies, this is rarely true. Suppose that 2-by-2 tables show associations between the rate of illness and the eating of three foods—ham, mustard, and raisin sauce, and also with age (adult versus child). In the study of any one of these factors the others are potential *confounding factors*, since they are associated both with disease and with exposure. It is easy to postulate that adults ate ham in preference to hot dogs, and that ham eaters were more likely to use mustard and raisin sauce; in fact, it may be hard to find either ill or well ham eaters who did *not* use mustard or raisin sauce.

One way of evaluating the significance of cases' consumption of ham alone is to choose matched controls who have the same age/mustard/raisin sauce values. This would require a separate study for each factor being considered, and is generally impractical because of the limited number of attendees at the event who can be used as controls. Matching for common

confounding variables, such as age and sex, with 1, 2, 3, or 4 comparison subjects per case or exposed person is often used in research studies, however.

If cases and controls are not matched for confounding variable status, the fair way to study the association of disease with ham is to choose groups with the same mustard-sauce-age group and study ham within each group. This splitting into subgroups is called "stratification." Mantel and Haenszel [*13*] showed how to combine the results from each stratum into an overall estimate of the odds ratio, and this procedure is still the recommended way to begin an analysis of several categorical variables in epidemiology.

If the case rate in ham eaters differs with age, sauce, or mustard, one or more of the factors would be said to *interact* with ham eating.

Interaction is usually handled in epidemiology by statistical modeling, which will be described in the next section. There is not universal agreement on the definition of interaction, since there are two kinds (additive and multiplicative) [*14*], and factors may interact in one model and not in the other. Asbestos exposure and smoking are known to interact strongly in producing lung cancer in an additive model, but the interaction is minimal or absent in a multiplicative model.

Multivariate Regression Analysis

In regression analysis, an equation containing variables and constants is "fitted" to the data describing disease and exposure in a population. The process involves finding the values of constants or "coefficients" that give the best fit, a task that is usually practical only when done on a computer.

Experts say that this is one area where a little knowledge is a dangerous thing, and that regression analysis, like French cookery, requires more than a recipe for its successful performance and interpretation. Determining which data sets should be used for regression, and exactly how to proceed with analysis and interpretation, is an area requiring statistical judgment. Although it seems to be an ideal area for development of an "expert system," no one appears to have reported such a development in the medical literature. General strategies and a flow chart for doing the analysis have been published [*15*].

Some of the problems involved in using regression models have been discussed by Vanderbroucke [*16*], who raises the question, for the sake of argument, of abandoning regression models altogether. He concludes that results of regression modeling should at least be checked against those of stratificiation with the Mantel-Haenszel analysis discussed above. Rothman in his textbook on epidemiology, expresses the same view by saying, "The epidemiologist is better off to rely primarily on the more straightforward procedures of stratified analysis whenever possible, since these procedures engender greater familiarity with the data for both the investigator and the reader" [*17*].

A form of regression often used in epidemiology is logistic regression, which fits an equation of the following form to the data.

$$\text{Probability of disease} = 1/\{1 + \exp[-(a + \beta_1 \cdot x_1 + \ldots + \beta_j \cdot x_j)]\}$$

Disease must have values of 0 and 1 (binary outcome). The x values represent individual risk factors, x_1 being the exposure under study, and the

other x's either risk factors or confounding variables, since both are treated as confounding variables in this analysis. An x term may be replaced by the product of two risk factors thought to interact, such as $\beta_6 x_1 x_2$. One of the goals of the analysis is to choose the minimal number of such terms that will produce a good "fit" to the data.

In performing logistic regression, information on disease outcome and individual risk factors for a number of individuals is fed to the computer. Two methods of calculation are available—conditional and unconditional logistic regression. Conditional logistic regression is more accurate for studies with small numbers of subjects, but consumes more computer resources. The results of interest are the beta coefficients and the odds ratios and confidence limits that may be derived from them. An odds ratio of 2 for the risk factor *smoking* in a study of myocardial infarction would mean that smokers were twice as likely as nonsmokers to have myocardial infarctions, with all other risk factors and confounding variables (e.g., cholesterol, blood pressure, exercise, age, sex) held constant. If interaction terms are included, their values must be specified in the results, since the odds ratio changes with different values of the interacting terms.

Multivariate Poisson regression is used for grouped data in diseases of low frequency, and survival analysis is used to study deaths or other events in exposed and unexposed groups over a defined period of follow-up. Computer programs for doing many of these procedures are available on microcomputers, and all are parts of mainframe packages such as SAS, BMDP, and SPSS.

LOGRESS [*18*] is a public domain program for logistic regression on IBM PC-compatible microcomputers, written in FORTRAN and assembly language. Data values to be input from a file or the keyboard include the names of variables, codes for missing values, and frequencies for each category of each variable. LOGRESS then produces odds ratios and 95% confidence limits associated with each risk factor. An odds ratio of 1.4 for "smoking (yes/no)" suggests that smokers have a 40% greater frequency of disease than nonsmokers who are otherwise similar in all respects entered into the model. Other statistics, such as beta coefficients, standard errors, and "z" scores are also given.

Loglinear modeling is another technique used in epidemiology for analysis of several variables. In logistic regression, the outcome variable, such as disease (ill/well), must have only two values, but loglinear methods allow more than two categories of outcome. Koepsel [*19*] describes a BASIC program called LOLA which does loglinear modeling on microcomputers.

A program called PC Fit for fitting a variety of functions to observed data points has been described by Smith [*20*]. Like LOGRESS, it uses the method of maximum likelihood estimation. Wallenstein and Bodian [*21*] have recently provided methods for using weighted least-squares analysis rather than maximum likelihood estimation for multivariate analysis. The calculations can be performed with current microcomputer versions of SAS and SPSS.

Cyrus Mehta and colleagues [*22*] have developed network algorithms for calculation of exact confidence limits on the odds ratio, a procedure which was formerly practical only on mainframe computers. Extension of similar methods to logistic regression analysis is under way in Dr. Mehta's group.

Other Programs for Epidemiologic Statistics

A description of 18 "free or inexpensive" programs for performing epidemiologic work has been published [23]. Most of the programs perform statistical operations. One should assume that they are not all of equal quality in terms of the user interface and documentation.

A public domain program for performing nutritional surveys has been produced by the Division of Nutrition at the Centers for Disease Control, Atlanta, Georgia. Known as CASP, it compares height, weight, and age measurements with appropriate standards and also tabulates results for a population.

"Expert Systems" for Epidemiologic Work

Epi Info and EPIAID form an experimental microcomputer system for epidemiologic field investigations developed at the Centers for Disease Control [24]. It uses interactive word processing to assist the investigator in drafting a report of an investigation. The programs contribute sentences or phrases to the report and then prompt the user to add words or sentences describing the current investigation. The result is a series of text summaries which form the introduction, methods, discussion of hypotheses, results, tables, references, and sample-size calculations for the written report of the investigation. The data handling and statistics package in Epi Info is being used in many epidemiologic settings.

Another application of "expert system" methods to epidemiology in Italy has been described by DeRosis, Pizzutilo, and Greco [25]. Their MICRO-IDEA system provides assistance in selecting correct methods for analysis of data. It uses decision trees to arrive at appropriate statistical methods and then performs the necessary calculations. MICRO-IDEA was written in compiled BASIC on an IBM PC-compatible microcomputer.

Disease Surveillance Systems

In the United States, a number of local health departments and the majority of state health departments now maintain records of "reportable" diseases on microcomputers. Almost any file-handling or database system, including a spreadsheet program, can be used at the local level. In January 1990, 31 state health departments were using the program called Epi Info, developed at the Centers for Disease Control, for maintaining reports of disease. Others use dBASE III or other database systems on computers of all sizes.

PUBLIC HEALTH COMMUNICATION

A public health network administered by the Public Health Foundation, Washington, D.C., on a commercial telecommunication network has (January 1990) 783 users, including most state health departments. All 50 states send weekly disease reports to the Centers for Disease Control in Atlanta in the form of electronic mail messages containing structured data records.

Microcomputers are used in state health departments for maintaining records on sexually transmitted disease, acquired immunodeficiency syndrome

(AIDS), immunization, and random telephone surveys of chronic disease risk factors. Each of these programs is coordinated nationally, and records are sent from microcomputers at the state level to the Centers for Disease Control, where the files are merged and then processed on a mainframe computer.

Mapping of Disease

Mapping of disease cases and disease rates has been performed on mainframe computers for many years. Commercial microcomputer programs such as Mapmaster and Atlas AMP produce maps of the United States down to zip code level, and of other countries in less detail. Producing maps other than those supplied as options with the programs is not yet convenient, and files containing detailed coordinates are often expensive.

DEDUCTIVE OR A PRIORI MODEL BUILDING: CONSTRUCTING A WORLD AND THEN A MODEL

Mathematical models of communicable disease in hypothetical populations were pioneered by Sir Ronald Ross for malaria in 1909 and are the topic of hundreds of articles and at least ten books. Reviews of the subject have been published by Fine [*26*], Bailey [*27*], and others [*28, 29*].

Models of infectious disease begin with assumptions about the source, route, and probability of infection. Most models assume infection from human cases, as in influenza, a popular subject for modeling. "Mass-action" models assume that risk of infection is proportional to the product of the number of active cases and the number of susceptible persions, much as in models of chemical reactions. The Reed–Frost model gives similar results if the population is large and the cases are widely scattered, but it also adjusts for multiple exposures (only one of which can be effective) in populations of smaller size. Programs incorporating the Reed–Frost model for calculators [*30*] and for IBM PC-compatible microcomputers [*31*] have been described.

Models assuming society to be a series of cases, susceptibles, and immunes bouncing off each other randomly, like molecules of a gas, have clear limitations. A series of Reed–Frost models with increasing flexibility for compartmentalization of subpopulations and inclusion of multiple risk factors has been developed by Ackerman et al. [*32*]. The "Resource for Simulation of Stochastic Micropopulation Models" at the University of Minnesota offers collaborative research facilities for those wishing to explore particular models on a minicomputer within this framework.

The term "stochastic" refers to the use of random events generated by the comptuer to simulate the real world. If the probability of an event is 0.70, for example, a random number is generated between 0 and 1.0. If it falls in the range 0 to 0.70, the event "occurs"; otherwise, it does not. This type of simulation requires the generation of many epidemics—often 100—to describe the range of statistical possibilities that can occur. More computer time is likely to be needed for these models than for "deterministic" models in which a series of equations is solved, although this depends on the methods used for their solution.

Time is handled differently in different models. "Discrete-time" models use slices of time, calculating all the nececessary results for a finite

period such as an hour or a day. Calculations for the next interval are based on these results, each cycle being the basis for giving the model another "crank" for the next time period. Continuous-time models use differential calculus to work with changes in rates at each point on a curve. Theoretically, the results of discrete-time models should approach those of continuous-time models as the intervals used become smaller.

We have briefly described models of society as random individuals arranged in family, school, or work groups. A variant on compartmentalized groups occurs when the groups are different cities and the determinant of contact is the amount of air traffic between the infected and uninfected cities. A microcomputer model of this type has been described by Longini, Fine, and Thacker [*33*].

Airborne and close-contact infection are the easiest situations to model. Epidemics of foodborne disease do not appear to have been the subject of deductive modeling, since they arise through a series of rare events (e.g., slicing the cooked turkey on the same cutting board used for raw turkey, followed by storing it several hours at the wrong temperature), and the microscopic components of those events usually can only be imagined. It is rare that uniform mixing of the food vehicle occurs to allow necessary assumptions about dose.

Malaria has been extensively modeled since 1909. Bailey's book on the subject [*34*] concludes with pleas for field validation of mathematical models and comments that "It may well be that genuinely new and useful applications of malaria biomathematics will only be effectively fostered and promoted by carrying them directly into the areas where their contribution to the alleviation of human suffering will be most keenly appreciated." In other words, improved health is the real test, one that has not yet produced a good score for mathematical modeling in the malaria field, despite the large amount of thought which has been stimulated by the process.

AIDS is a problem of immense concern to world health that appears to offer several favorable features for modeling. A number of modeling groups have begun to work on this important problem. The difficulties, as pointed out by May and Anderson [*35*], lie in the lack of data to describe the behavior of the causative virus in even one population, since studies have not continued long enough to discover the 10-year outcome of infection, for example. The results of infection may vary by route of infection, genetic background, or other factors. Data collection is impaired by the fact that major risk factors, such as homosexuality and intravenous drug use, are difficult to monitor accurately.

Modeling of Chronic Disease

Many chronic illnesses such as heart disease and cancer can be measured through their effect on mortality, which is comprehensively (although crudely) reflected in the death certificates kept by every state health department. Surveys of the prevalence of risk factors such as smoking, hypertension, obesity, and lack of exercise are also commonly available [*36*]. Rice and Hodgson et al. [*37*] have combined these and other types of information into an estimate of "attributable risk," or the fraction of disease attributable to smoking. Shultz [*38*] has incorporated these models into spreadsheets for IBM PC-compatible computers, and they have been used in a number of states to estimate the fraction of illness attributable to smoking for use in public policy decisions.

Models for Teaching

Simulation offers obvious advantages for teaching epidemiology, particularly for illustrating the results of random sampling. Bland [*39*] has used microcomputers to simulate a cancer treatment trial in classes of medical students. An interactive microcomputer program for teaching simple epidemiologic concepts has been developed by Pastides et al. [*40*].

The potential exists for realistic simulation of epidemics using computers to present illnesses, interviews, and other factors that could be sampled randomly by each member of a class. Video disc technology has been used in clinical teaching of medicine [*41*], but apparently the first epidemic on a computer-controlled video disk has not yet appeared.

Evaluation of Epidemic Models

Most mathematical models, if they have been evaluated against the real world at all, have been compared with data sets from past epidemics, often from isolated populations such as those of Iceland or Tristan da Cunha. Even if such epidemics fit the model within, say, 95% confidence limits, the value of a single epidemic for such validation is limited, as is any other sample of one in the statistical world. Assuming that epidemics are driven by random processes (as most models assume), one would expect natural epidemics to match the central tendencies of models very seldom; in fact, it would be very surprising to see an exact match. There is no way of knowing whether the epidemic observed came from the right tail, the left tail, or the center of nature's normal curve, and the same is true of individual elements that were measured to form the basis for the model. Only when numbers are very large does randomness begin to approach some kind of predictability. With large populations, frequent, detailed observations of sick and well people are not affordable—throwing a kind of uncertainty principle in the way of deductive epidemiological modeling.

Mathematical modeling played a substantial role in planning the World Health Organization's unsuccessful worldwide malaria eradication compaign of the 1950s and 1960s, a major example of unsatisfactory forecasting. Empirical discoveries on vaccination strategy, not modeling, were the basis for the successful worldwide smallpox eradication which began in 1966 [*42*]. The "swine flu affair" [*43*] of 1976–1977 in the United States arose from predictions of a worldwide epidemic of influenza, in a field that had been the subject of numerous mathematical models. The shortcoming is usually not in the model, but in the information which is available to feed into the model, for example, the infectivity of a new strain of virus in human populations, or political and economic factors which affect health programs. Other predictions based on modeling have been more accurate, such as those of Hethcote on measles and rubella trends in the United States [*44*].

It appears that deductive modeling has raised interesting questions, contributed to teaching and understanding of epidemic processes, and given rise to hundreds of articles and books. It has not yet become a regular part of disease control programs. Fine, an expert in the field, says, "An enterprise cannot survive on potential alone, and one may well question to what extent the large literature on mathematical models has lived up to any potential usefulness. It is probably correct to say that it has not . . . " [*45*]. Nevertheless, contributions of mathematical modeling to teaching and statistical analysis of disease patterns remain, and probably

will increase if microcomputers make modeling available to a wider audience at lower cost.

FUTURE POSSIBILITIES FOR MICROCOMPUTERS IN EPIDEMIOLOGY

Neural Network Models

Computer hardware and software patterned after networks of neurons in the central nervous system has lately become a subject of intense interest in artificial intelligence circles [*46*]. A neural network is built of "neurons," each of which has many inputs and usually a single output. The output is a function, such as summation, of the inputs, and each of the inputs can have a weight or more complex function attached to it. Unlike a sequential processor, the "program" is embodied in the connections between elements and the weights assigned to each connection. Neural networks are particularly adapted to optimization and equilibrium problems and to "learning" from repeated exposure to similar patterns. They function by parallel rather than sequential processing, and work largely with analog rather than digital variables. "Neurons" can be constructed in hardware form or simulated in software on a conventional computer. Neural network programs are available for both the Macintosh [*47*] and IBM PC-compatible microcomputer [*48*].

In epidemiology, illness (or lack of illness) might be viewed as the sum of a complex series of risk factors. It might be possible to represent society as a collection of "neurons" representing persons, and infection as the propagation of impulses from one neuron to another. If suitable time delays could be introduced, such a system might offer an alternative format for modeling human disease in a population, although connecting the model with the real world would present the same challenge as it does with present models.

The ability of neural networks to learn by incorporating new data to change the strength of connections between elements might be used for disease surveillance. A disease surveillance system in a public health department someday might consist of a neural network model of society, updated daily by births, deaths, and illness information obtained from the real world. The model would remain as imperfect as our data and our knowledge of the causes of disease, but constructing it would at least pose some interesting questions.

Expanded Role of Microcomputers

Microcomputers are already being used to collect and process data from outbreak investigations and to maintain surveillance records for selected diseases. It is practical to handle files of 100,000 or 200,000 modest-sized records on an IBM-compatible computer. Larger files are usually handled on mini- or mainframe computers. The level at which this crossover occurs will undoubtedly rise as processors become faster and disc capacity grows larger and cheaper.

At present speeds, microcomputers can handle files containing *samples* of larger files adequate for most analytic purposes. Given a file of many million Medicare records, for example, a sample of the file can be selected and downloaded to a microcomputer for analysis. Considering the inaccuracies in the original data due to sampling, data input, and provider

bias, there seems to be little reason to process the entire file. Since samples of a few thousand records often represent a population quite well, the actual processing now falls within the capabilities of microcomputers, making such data handling technically possible for persons outside the mainframe world.

Clearly, this expanded access to processing of larger data sets invites the inexperienced user to make mistakes, but it also expands the epidemiologic community's pool of experienced "data crunchers." In developing countries, where mainframe computers are in short supply, microcomputers potentially can make even larger contributions to epidemiology.

Improved Data Collection

Computerized epidemiologic data files are almost all created by manual typing of information from other sources. This has advantages in condensing and cleaning up the data, but essentially limits epidemiologic studies to situations in which a commitment has been made to do a special study, or to use large data files collected for other purposes. Medical records are not currently accessible in machine-readable form from most general hospitals and clinics.

A medical record is often several inches thick, the equivalent of a megabyte of information (300 pages × 3,000 characters per page = 900k). On some of today's microcomputers and workstations, it would be possible to manipulate such a document, extract information, and produce a standardized record—perhaps on optical storage media—which could be used for clinical, archival, and epidemiologic purposes. Although research on methods of extracting useful information from the free-flowing form of most medical records has begun [*49*], the task of producing and using records in machine-readable form is a major obstacle. Voice-input dictation and improved text and cursive-handwriting scanners will probably be needed to carry this effort to the practical stage.

The next decade will produce the necessary technical tools for making medical records machine readable and available for epidemiologic analysis. Radical changes in medical practice and new approaches to maintaining both confidentiality and useful access to records will be needed before epidemiologic work can benefit from these advances. Many legal and political problems will have to be solved to provide for adequate systematic sampling of computerized medical records for public health purposes.

SUMMARY

During the past decade, microcomputers have come into routine, but not universal, use for data collection and processing in epidemiology. Two experimental "expert systems" have been developed for epidemiologic work. A variety of microcomputer programs for data entry and statistical processing are available, several of which are designed specifically for the categorical data analysis frequently used in epidemiology.

Advances in computer technology over the past few years have allowed tasks formerly confined to mainframe computers to be done on microcomputers, and many epidemiologists now make extensive use of microcomputers. A major advance in public health surveillance and epidemiologic study can

be expected if patient medical records become routinely available in machine-readable form, and if epidemiologic analysis of national or other population-based samples of such records becomes possible. Microcomputers alone or connected with larger computers will play an increasing role in medical record keeping and in collection and analysis of epidemiologic data.

REFERENCES

1. J. Snow, *Snow on Cholera*, The Commonwealth Fund, New York, 1936.
2. J. P. Fox, C. E. Hall, and L. R. Elveback, *Epidemiology: Man & Disease*, Macmillan, New York, 1970.
3. G. D. Friedman, *Primer of Epidemiology*, 2nd ed., McGraw Hill, New York, 1979.
4. A. M. Lilienfeld and D. E. Lilienfeld, *Foundations of Epidemiology*, 2nd ed., Oxford University Press, New York, 1980.
5. J. L. Fleiss, *Statistical Methods for Rates and Proportions*, 2nd ed., John Wiley, New York, 1981.
6. J. J. Schlesselman, *Case-Control Studies*, Oxford University Press, New York, 1982.
7. N. E. Breslow and N. E. Day, *Statistical Methods in Cancer Research*, Volume 1, *The Analysis of Case-Control Studies*, International Agency for Research on Cancer, Lyon, France, 1980.
8. D. G. Kleinbaum, L. L. Kupper, and H. Morgenstern, *Epidemiologic Research*, Lifetime Learning Publications, Belmont, CA, 1982.
9. B. Rosner, *Fundamentals of Biostatistics*, Duxbury Press, Boston, 1982.
10. A. S. Benenson, *Control of Communicable Diseases in Man*, American Public Health Association, Washington, DC, 1985.
11. H. M. Schoolman, "The Physician and the Medical Literature. From Index Medicus to MEDLARS to GRATEFUL MED and Beyond," *Arch. Dermatol.*, *122*(8), 875–876 (1986).
12. A. L. Dannenberg, Personal communication, 1987.
13. N. Mantel and W. Haenszel, "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease," *J. Nat. Cancer Inst.*, *22*, 719–748 (1959).
14. D. G. Kleinbaum, L. L. Kupper, and H. Morgenstern, *Epidemiologic Research*, Lifetime Learning Publications, Belmont, CA, 1982, pp. 403–418.
15. D. G. Kleinbaum, L. L. Kupper, and H. Morgenstern, *Epidemiologic Research*, Lifetime Learning Publications, Belmont, CA, 1982, pp. 447–456.
16. J. P. Vandenbroucke, "Should We Abandon Statistical Modeling Altogether?," *Am. J. Epidemiol.*, *126*, 10–13 (1987).
17. K. J. Rothman, *Modern Epidemiology*, Little, Brown, Boston, 1986, p. 284.
18. D. L. McGee, "A Program for Logistic Regression on the IBM PC," *Am. J. Epidemiol.*, *124*, 702–705 (1986).
19. T. D. Koepsell, "Loglinear Modeling with Inexpensive Computing Equipment," *Am. J. Epidemiol.*, *120*, 777–787 (1984).
20. W. P. Smith, "PC FIT: A Distribution Fitting Program," *Am. Statist.*, *39*, 141 (1985).

21. S. Wallenstein and C. Bodian, "Inferences on Odds Ratios, Relative Risks, and Risk Differences Based on Standard Regression Programs," *Am. J. Epidemiol., 126,* 346–355 (1987).
22. C. R. Mehta, N. R. Patel, and R. Gray, "Computing an Exact Confidence Interval for the Common Odds Ratio in Several 2 × 2 Contingency Tables," *J. Am. Statis. Assoc., 80,* 969–973 (1985).
23. "Inventory of Free or Inexpensive IBM-PC Compatible Software for Epidemiologists," *Epidemiol. Mon., 8,* 1–8 (1987).
24. A. G. Dean, "EPIAID," *Byte*, October, 225–231 (1985).
25. F. DeRosis, S. Pizzutilo, and D. Greco, "MICRO-IDEA: Improving Decisions in Epidemiological Analysis by a Microcomputer," *Med. Inform., 11,* 225–236 (1986).
26. P. Fine, "Applications of Mathematical Models to the Epidemiology of Influenza: A Critique," in *Influenza Models: Prospects for Development and Use* (P. Selby, Ed.), MTP Press, Hingham, MA, 1982, pp. 15–85.
27. N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases and Its Applications,* Hafner, New York, 1975.
28. K. Dietz and D. Schenzle, "Mathematical Models for Infectious Disease Statistics," in *A Celebration of Statistics: The ISI Centenary Volume* (A. C. Atkinson and S. E. Fienberg, Eds.), Springer-Verlag, New York, 1985, pp. 167–203.
29. N. Becker, "The Uses of Epidemic Models," *Biometrics, 35,* 295–305 (1979).
30. E. L. Franco and A. R. Simons, "Simple Algorithms for the Representation of Deterministic and Stochastic Versions of the Reed-Frost Epidemic Model Using a Programmable Calculator," *Am. J. Epidemiol., 123,* 905–915 (1986).
31. "Inventory of Free or Inexpensive IBM-PC Compatible Software for Epidemiologists," *Epidemiol. Mon., 8,* 1–8 (1987).
32. E. Ackerman, L. R. Elveback, and J. P. Fox, *Simulation of Infectious Disease Epidemics,* Charles C. Thomas, Springfield, IL, 1984.
33. I. M. Longini, P. E. M. Fine, and S. B. Thacker, "Predicting the Global Spread of New Infectious Agents," *Am. J. Epidemiol., 123,* 383–391 (1986).
34. N. T. J. Bailey, *The Biomathematics of Malaria,* Charles Griffin, London, 1982.
35. R. M. May and R. M. Anderson, "Transmission Dynamics of HIV Infection," *Nature, 326,* 137–142 (1987).
36. E. M. Gentry, W. D. Kalsbeek, G. C. Hogelin, et al., "The Behavioral Risk Factor Surveys: II. Design, Methods, and Estimates from Combined State Data," *Am. J. Prevent. Med., 1,* 9–14 (1985).
37. D. P. Rice, T. A. Hodgson, P. Sinsheimer, et al., "The Economic Costs of the Health Effects of Smoking, 1984," *Milbank Memorial Fund Q., 64,* 489–547 (1986).
38. J. M. Shultz, SAMMEC: Smoking-Attributable Mortality, Morbidity, and Economic Costs (Computer Software and Documentation), Minnesota Department of Health, Minneapolis, MN, 1986.
39. J. M. Bland, "Computer Simulation of a Clinical Trial as an Aid to Teaching the Concept of Statistical Significance," *Statis. Med., 5,* 193–197 (1986).

40. H. Pastides, K. A. Mundt, and C. B. MacKnight, "Microcomputer-Based Instruction: Applications in Epidemiology," Program and Abstracts, American Public Health Association 115th Annual Meeting, New Orleans, October 18–22, 1987, session #2118, APHA, Washington, DC, 1987.
41. W. G. Harless, M. A. Zier, and R. C. Duncan, "A Voice-Activated, Interactive Videodisc Case Study for Use in the Medical School Classroom," *J. Med. Educ.*, *61*, 913–915 (1986).
42. The Global Eradication of Smallpox: Final Report of the Global Commission of the Certification of Smallpox Eradication, Geneva, December 1979, World Health Organization, Geneva, 1980.
43. R. E. Neustadt and H. Y. Fineberg, *The Swine Flu Affair*, Washington, U.S.D.H.E.W., Government Printing Office, Washington, DC, 1978.
44. H. W. Hethcote, "Measles and Rubella in the United States," *Am. J. Epidemiol.*, *117*, 2–13 (1983).
45. P. Fine, "Applications of Mathematical Models to the Epidemiology of Influenza: A Critique," in *Influenza Models: Prospects for Development and Use* (P. Selby, Ed.), MTP Press, Hingham, MA, 1982, p. 31.
46. M. Gaudill, "Neural Networks Primer, Part 1," *AI Expert*, *2*, 46–52, (1987).
47. *McBrain: Neural Network Simulation System, User's Manual*, Neuronics, Cambridge, MA, 1987.
48. "Hecht-Nielsen's ANZA," *IEEE Expert*, Winter, p. 92 (1987).
49. S. M. Downs, M. G. Walker, and R. L. Blum, "Automated Summarization of On-Line Medical Records," in *MEDINFO 86, Proceedings of the 5th Conference on Medical Informatics*, Washington, October, 26–30, 1986 (R. Salamon, B. Blum, and M. Jorgensen, Eds.), Elsevier Science Publishers, Amsterdam, 1986, pp. 800–804.

ANDREW G. DEAN