Biomedical Data Resources:

A Limiting Factor in Analytic Epidemiology

by R. A. Tjalma, D.V.M.* and H. L. Braun, Ph. D.*

Two of the more popular definitions of epidemiology are: "The science which is concerned with the natural history of disease as expressed in groups of individuals related by some common factor of age, sex, race, or occupation, as distinct from development of disease in individuals" (*1*) and "The study of the distribution and determinants of disease prevalence in man" (*2*). While most definitions of epidemiology specify disease as the study issue, other endpoints, including health, intellect, and socio-economic factors, are subject to epidemiologic investigation.

Epidemiology, by virtue of methodologic principal, does not provide proof in the context of cause and effect; rather, the science is limited to identification and quantitation of associations between cause and effect variables. In this respect, it is of consequence to note that the complex pathobiology which underlies many current health problems serves to preclude laboratory demonstration of cause and effect. For example, prior to recent reports which confirm the development of lung cancer in animals regularly exposed to tobacco smoke, the generally accepted causal association between smoking and lung cancer was predicated on the strength of epidemiologically determined associations; likewise, the causal association between smoking and cardiovascular disease is based on epidemiologic evidence (*3*, *4*).

*Epidemiology Program, National Institute of Environmental Health Sciences, P. O. Box 12233, Research Triangle Park, North Carolina 27709.

Epidemiology serves to identify and evaluate problems which occur under natural conditions and permits specification of particular aspects of the problem with high potential for productive laboratory investigation. It is not possible, for practical reasons of limited time and resources, to subject to laboratory study all aspects of every problem observed; however, following epidemiologic assessment of problem parameters, such as environmental-occupational exposure, peculiarities of disease distribution, incubation, and course, appropriate laboratory effort can frequently be brought to bear. Conversely and equally important, laboratory observations can often be extended and complemented by epidemiologic studies. With reference to the progressive complexity of current biomedical considerations, epidemiology is the method by which the existence of a problem is decided, the magnitude of the problem defined, and the focus of further research effort determined.

Basic Requirements

Epidemiologic research requires the availability and interaction of two components: data and personnel. Individuals must ask relevant questions—the answers to which can be secured by analysis of available data. Many epidemiologic observations, including the teratogenic effects of rubella virus (*5*) and thalidomide (*6*), the causal association between exposure to radium and subsequent risk of bone cancer (*7*), and the

carcinogenic effect of occupational exposure to mustard gas (*8*), were accomplished by "non-epidemiologists" who asked correct questions and had access to appropriate data. It is obvious that training cannot compensate for inadequate data resources and, more important, training is of limited potential in the absence of appropriate data resources. While it is not difficult to pose important and intellectually rational epidemiological questions, it is progressively more difficult to generate meaningful questions for which data are available to provide an answer. Unfortunately, detailed knowledge of the limitations intrinsic to available resources and ability to modify questions as required by the configuration of such resources, rather than objective responses to problems in terms of national needs, constitute the sine qua non of current epidemiologic function.

Historical Development

It is instructive to review briefly the historical development of epidemiology as a function of data availability. Prior to the early part of the 19th century, epidemiology was, by virtue of the absence of organized medical data collection and reporting systems, restricted to crude description of isolated episodes. Irrespective of the observer's inclination or ability, the absence of data precluded any precise measurements and resulted in subjective evaluation of anecdotal information. The first half of the 19th century witnessed development of comparative epidemiology with emphasis focused on comparison of contemporary epidemics with earlier or concurrent episodes in the same or different areas. Such methods were more meaningful as regards the identification of epidemic phenomena and provided an objective basis for predicting the distribution, duration, and effect of disease. It is important to realize this advance in epidemiologic methodology was possible only as a result of the improved availability of more reliable data on disease occurrence. The last half of the 19th and early years of the 20th centuries were dynamic in that vital statistics evolved to the level of a discipline. The improved precision of available birth, morbidity, and mortality data encouraged the development of more sensitive biometric procedures and, as a direct consequence, epidemiology rapidly evolved to the status of an exact analytic science with the capacity to test hypotheses.

Progress in the prevention and control of the epidemic infectious diseases resulted in a dramatic shift of research emphasis to the chronic degenerative diseases during the fourth and fifth decades of this century. It is important to recall, at this point, the basic differences between the acute infectious and chronic degenerative diseases: acute infectious diseases are usually characterized by abrupt onset, defined (brief) incubation, short course, discernible progression among population groups, high frequency, and obvious termination in recovery or death; conversely, chronic degenerative diseases are usually characterized by insidious onset, undefined (extended) incubation, protracted course, absence of discernible progression among population groups, low frequency, and questionable criteria of termination.

Early epidemiologic efforts in the chronic disease area, limited in large part to exploitation of resources developed in response to acute disease research needs, quickly demonstrated the profound implications of the differences between the two classes of illness. In effect, it was realized that available data, while appropriate to support epidemiologic investigation of the acute infectious diseases, were frequently inadequate to support analytic epidemiologic study of the chronic diseases. Further, it became apparent that both the quantity and quality of data would need to be improved to compensate for the inherent, confounding characteristics of the chronic diseases. Consequently, a major commitment was made to develop new data resources and improve the quality of existent data resources. Efforts involved national and international standardization of disease nomenclature and classification systems, expansion of census coverage

to include items of socio-medical consequence, requirements for improved hospital record systems, evaluation of reliability of death certificate data, improved design of birth and death certificates, establishment of chronic disease registries at the institutional and state levels, development of animal disease registries, and automation of data resources to improve accessibility.

While it is generally recognized that the magnitude and sophistication of chronic disease research efforts during the last 30 years are unparalleled in the history of medicine, the great progress in epidemiology achieved during this period is somewhat less appreciated. Epidemiologic exploitation of improved data resources resulted in a number of important observations, the implications of which may not be fully appreciated for some time. Some of the more important observations on neoplastic disease include the following: absence of classical time-space cluster effect for leukemia (*9*, *10*); leukemia concordance among identical twins (*11*); excess frequency of leukemia and other cancers among individuals with certain congenital conditions (*12*, *13*); absence of association between the occurrence of cancer in man and pet animals (*14*); significant and consistent variation in cancer rates in different geographic areas and among different ethnic groups (*15*); apparent relationship between age at first pregnancy and subsequent risk of mammary carcinoma (*16*); the association between body stature and risk of osteogenic sarcoma (*17*, *18*). These and other epidemiologic observations have served to provide important leads for laboratory research efforts.

Limitations of Available Resources

The studies which yielded the aforementioned and other epidemiologic observations usually required large quantities of highly refined data and would not have been possible in the absence of earlier efforts, as previously indicated, to improve the quality and quantity of available data. It seems clear that the epidemiologic usefulness of available data resources has, in large part, been exhausted. Progressive awareness of the complex nature of the chronic diseases prompts progressively more complex questions which, in turn, tend to exceed the quality of available data. As a direct consequence, inordinate time and material resources are invested in an effort to compensate for the inadequate nature of available data. It appears obvious that future and continued progress in chronic disease epidemiology will depend on a real and practical improvement in the quality of national biomedical data resources.

The national commitment to protect man from environmental hazards, ranging from air pollution to food additives to biocides, raises a serious question regarding the ability of the biomedical research community to respond to the epidemiologic challenge. Progressive emphasis on environmental health problems has resulted in a profusion of reports regarding the real or imagined danger of many or most environmental agents. Of primary concern is our collective ability to identify significant problems. As previously indicated, problem definition constitutes one of the basic functions of epidemiology.

While the need for epidemiologic definition and evaluation of environmental health problems is obvious and urgent, the problems in question are of great complexity and evaluation will require more data of better quality than has generally been required for epidemiologic investigation of the categorical chronic diseases. In addition to the intricate characteristics of the chronic diseases, as previously noted, investigation of most environmental health problems is further complicated by factors which include low levels of exposure over long periods of time, interaction and synergistic effects, delayed manifestations of effect over many years or generations, variability of pathologic response, and population mobility.

Many qualified individuals and expert groups have called attention to current and anticipated problems attributable to the inadequacy of existent biomedical data resources (*19*, *20*, *21*). Contrary to the all too

general reaction of the inexperienced or misinformed, medical research cannot be uncomplicated or reduced in sophistication to a level dictated by the nature of currently available data. By the same token, progressively greater investment of time and fiscal resources, in an effort to compensate for the marginal utility of existent data, is unrealistic and of limited merit. Equally important, no knowledgeable person proposes creation of data resources as ends in themselves, to a level which exceeds potential return, to an extent which violates or is incompatible with individual privacy, or as a locked system to stand for all time as a source of pertinent information to answer all questions. Indeed, data is a fluid commodity and it is essential that data mechanisms and resources be maintained subject to change as warranted by rational and progressive variation in the nature of research demand.

Proposed Improvements

As indicated in an excellent but largely unappreciated report of the Subcommittee on Use of Vital and Health Statistics in Epidemiologic Research, U.S. National Committee on Vital and Health Statistics, National Center for Health Statistics, Series 4, No. 7, March 1968, the appropriate course of action at this time involves mainly the integration and improvement of existent data resources rather than creation of new resources per se. There is no deficiency of data generation and collection activities of all varieties, at all levels of human endeavor, and by all manner of commercial, municipal, occupational, medical, and Federal agencies. The basic problem involves the failure to integrate the many existent data resources to achieve continuity of subject identification and reciprocal validation of information from multiple sources. What has been consistently and correctly proposed is implementation of record linkage mechanisms. Such mechanisms, possibly involving the assignment at birth and subsequent routine use of Social Security or other lifetime personal identification numbers, would effectively and efficiently improve the quantity and quality of data from a variety of existent resources including birth, death, military, Social Security, insurance, and hospital records.

The proposed National Death Index constitutes an example of a new and urgently needed resource (*19, 22, 23, 24*). At present, there is primary need for long-term studies of occupational, genetic, medical, and other population groups suspected to be at risk of excess mortality. Such studies are essential to test hypotheses of possible causal associations between some factor of initial exposure or disease episode and subsequent mortality experience. The difficulty has to do with the fact that 10, 20, or more years may be required for expression of excess mortality in the study group. During such extended study periods, a high proportion of the subjects can be expected to retire, move, change jobs, marry, or divorce. At termination of the study period, it is exceedingly difficult, and often impossible, to ascertain the living or dead status of individuals who migrated or were lost to follow-up for any reason. At the present time, as many as 30 steps may be required to ascertain the fact and cause of death of a single individual. Such efforts involve searches of union and company records, school rosters, Social Security and tax files, vital statistics records in a number of states, credit files, etc. The investigator has little choice but to avoid such studies, which is often the case, or, as indicated, resort to the only resources available, which are inefficient and costly. Further, utilization of such inappropriate resources might well be considered to constitute misuse of confidential information and possible invasion of personal privacy. Creation of a National Death Index, which would consist of a registry of all deaths (2 million per year), would facilitate needed research by reducing costs, improving precision, simplifying procedures, and precluding the need of acquisition of data from questionable sources. Indeed, a National Death Index would permit ascertainment of the fact of individual death in one step. The recorded cause of death

could then be ascertained through the usual mechanism of direct inquiry to the appropriate death registration area officials. The cost of the Index would be returned many-fold each year through reduction of research costs. Other possible improvements include modification of death certificates to include available information on all disease residua present at death, uniform encoding of congenital malformations on birth certificates, inclusion of parents' Social Security or personal identification numbers on birth certificates of progeny, uniform encoding of occupational category on medical records and death certificates, development of a national twin registry and organization of Medicare, commercial insurance, and union medical records to facilitate data retrieval.

Conclusions

Society, quite properly, has come to expect protection from environmental health hazards and continued progress in the prevention and cure of the categorical chronic diseases such as cancer and cardio-vascular disease. Experience of the last several decades has prompted the biomedical community to a position of enhanced appreciation of the actual and potential health hazards associated with a broad array of environmental pollutants. The complex nature of many such emergent problems precludes clear documentation of laboratory proof of cause and effect; consequently, the need for epidemiologic research is unprecedented and will become more acute. Unfortunately, the data resources required to support the sophisticated epidemiologic efforts in question are largely unavailable. While available data resources are of a quality which serve frequently to hamper chronic disease epidemiology, future environmental health epidemiology will be critically restricted in the absence of an early and significant improvement of national biomedical data resources. Legitimate concern has been expressed regarding creation of national data banks and attendent danger of invasion of individual privacy; however, it seems clear that a distinction can and must be made between those data resources with unacceptable potential for erosion of individual rights and those required to support medical research responsive to national needs.

REFERENCES

1. Top, F. H. (ed.). Communicable Diseases. 3rd Edition. The C.V. Mosby Co., St. Louis. 53.
2. MacMahon, B. and Pugh, T. F. 1970. Epidemiology Principles and Methods. 1st Edition. Little, Brown and Co., Boston.
3. U.S. Public Health Service. The Health Consequences of Smoking. A Report of the Surgeon General: 1971. Washington, U.S. Department of Health, Education, and Welfare, DHEW Publication No. (HSM) 71-7513.
4. U.S. Public Health Service. 1972. The Health Consequences of Smoking. A Report of the Surgeon General: 1972. Washington, U.S. Department of Health, Education, and Welfare, DHEW Publication No. (HSM) 72-7516.
5. Gregg, N. M. 1941. Congenital cataract following German measles in the mother. Trans. Opthal. Soc. Aust. 3:35.
6. Lenz, W. and Knapp, K. 1962. Thalidomide embryopathy. Arch. Environ. Health 5: 100.
7. Martland, H. S., Conlon, P. and Knef, J. D. 1925. Some unrecognized dangers in the use of and the handling of radioactive substances. J. Amer. Med. Assoc. 85: 1769.
8. Miller, R. W. 1965. Environmental agents in cancer. Yale J. Biol. Med. 37: 487.
9. Ederer, F., Myers, M. H. and Mantel, N. 1964. A statistical problem in space and time: do leukemia cases come in clusters? Biometrics. 20: 626.
10. Miller, R. W. and Fraumeni, J. F., Jr. 1967. "Leukemia houses." Ann. Intern. Med. 67: 675.
11. MacMahon, B. and Levy, M. A. 1964. Prenatal origin of childhood leukemia-evidence from twins. New Eng. J. Med. 270: 1082.
12. Miller, R. W. 1966. Relation between cancer and congenital defects in man. New Eng. J. Med. 275: 87.
13. Fraumeni, J. F., Jr. and Miller, R. W. 1967. Epidemiology of human leukemia: recent observations. J. Nat. Cancer Inst. 38: 593.
14. Tjalma, R. A. 1968. Implications of animal cancers to human neoplasia epidemiologic considerations. Int. J. Cancer. 3: 1.
15. Doll, R. 1972. Cancer in five continents. Proc. Roy. Soc. Med. 65: 49.
16. MacMahon, B. et al. 1970. Age at first birth and breast cancer risk. Bull. Wld. Hlth. Org. 43: 209.
17. Fraumeni, J. F., Jr. 1967. Stature and malignant tumors of bone in childhood and adolescence. Cancer. 20: 967.

18. Tjalma, R. A. 1966. Canine bone sarcoma: estimation of relative risk as a function of body size. J. Nat. Cancer Inst. 36: 1137.
19. Chase, H. C. 1972. Report on a national death index—pros and cons. Amer. J. Public Health 62: 719.
20. National Institute of Environmental Health Sciences. 1970. Man's Health and the Environment—Some Research Needs. U.S. Department of Health, Education, and Welfare.
21. Report of the Secretary's Commission on Pesticides and Their Relationship to Environmental Health, Parts I and II. 1969. U.S. Department of Health, Education, and Welfare. 661.
22. Recommendation. 1968. National Advisory Cancer Council. National Cancer Institute.
23. Recommendation. Task Force on Research Planning in Environmental Health Science. National Institute of Environmental Health Sciences. 1970.
24. Federal Statistics: Report of the President's Commission. 1971. Vol. I. 73.