American College of Epidemiology

Policy Statement on Sharing Data from Epidemiologic Studies

May 2002

I. Preamble

The American College of Epidemiology (ACE) believes that science conducted in the public interest is best served when data produced in epidemiologic studies are shared with other investigators. The principles that guide appropriate data sharing are set forth in this document.

The value of data sharing for epidemiologists is common to other branches of science [1], but special considerations in population sciences require additional thought and attention. The issue of sharing epidemiologic data was addressed by Hogue [2] in 1991, on behalf of the Society for Epidemiologic Research through its ad hoc committee on data access and sharing of epidemiologic research. They concluded with a quote that "much analysis and debate will be needed to arrive at a comprehensive taxonomy of data sharing situations, methods for constraining sharing, alternative ways to satisfy needs for data, and agreement on when refusal to share is appropriate [3]." They noted also that this effort would be well worth the cost.

It is the purpose of this ACE Policy Statement to 1) present some of the pros and cons of data sharing from the epidemiologist's perspective and 2) set forth principles on which policies and practices can be developed.

Increased attention is being directed to the issue of data sharing because of several factors. These include the ease of electronic data transfer, concerns about privacy and the confidentiality of individually identifiable data, requests for epidemiologic data from special interest groups [4], and new legislation that expands access to research data through the Freedom of Information Act [5]. In addition, a new statement on sharing of research data from the National Institutes of Health (NIH) will require a plan for data sharing as part of a grant application [6]. This statement expands existing policy on data sharing that was produced to assist in preparing data for sharing and archiving by the NIH [7].

Data sharing has many advantages for epidemiology. It supports the principle of the openness of scientific inquiry, provides opportunities for confirmation of results, allows for new research perhaps not anticipated by the original investigators, encourages the re-examination of data from diverse perspectives by researchers from different scientific disciplines, permits the creation of new data resources by linkage with other data sets, and facilitates training and education. Data

sharing also ensures that the considerable investment in constructing large human populations studies is maximized.

Despite these benefits, there are real challenges to data sharing for population research conducted by epidemiologists that must be considered. The privacy of participants in research must be protected [8] and informed consent must include provisions for future data sharing. However, there are concerns that detailed descriptions of possible future uses could alarm prospective research participants, decrease response rates, and thereby undermine the validity of the research. The effort and costs that go into constructing large population data sets, which frequently take many years, represent significant intellectual investments. These investments and the intellectual rights of the primary investigators must be respected in the development of data sharing practices. The costs involved in the process of data sharing may be substantial and, while data sharing may conserve research costs in the long run, in the short-term the increase in budgets may make less money available for primary research. Because of the complexity of epidemiologic data sets, data sharing ideally involves a process of becoming familiar with the data that goes beyond documentation procedures; the non-financial burden of this interaction must also be taken into account. Finally, data sharing needs to protect proprietary information and agreements, and procedures need to address the concerns of third party agreements.

The principles that follow attempt to guide epidemiologists toward incorporating data sharing into their thinking and actions. For some issues, there are no clear answers and time and experience must be our teachers.

II. Definitions

- Data - Final research data comprise the recorded factual material commonly accepted in the scientific community as necessary to validate research findings. Final research data do not include: laboratory notebooks, partial data sets, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as gels or laboratory specimens.
- De-identified data - Data from which individual identifiers have been removed, but linkage to original identifiers is maintained by the original collector of those data.
- Anonymized data - Data from which individual identifiers have been removed and personal identification is not possible.

III. Principles of Data Sharing

1. Benefits of Data Sharing

Science and the utility of the epidemiologic approach to generating new knowledge benefit when data are shared with other investigators.

The purpose of data sharing includes both the verification of original results and the use of the data to pursue questions or topics not the subject of the original study. Some data are unique resources and the value of data sharing is increased. In general, new analyses of existing data are a cost-effective use of limited research resources.

2. Collaboration and Data Sharing

Whenever possible, secondary users of existing data sets should work with the original investigators. It is desirable that those who collected or produced the original data be co-authors or be given the right of refusal.

Because of the complexity of many epidemiologic data sets, it is highly preferable for secondary users to collaborate with the originators of the data (sometimes on site) to learn about the nuances involved in data collection, the reliability of variables, and the particularities of the data set. Mandatory co-authorship imposed by those who produced the original data may introduce a coercive element to the data sharing that is contrary to the intent to encourage a free and independent re-analysis of the data. Rather, the primary investigators should be given the right of declining co-authorship.

3. Privacy of Human Subjects

The rights and privacy of people who participate in epidemiologic research must be protected and the confidentiality of research data should be safeguarded at all times.

Data intended for expanded uses should be de-identified and free of information on individuals that would permit identification of the research participants. Reasonable efforts to protect the identity of individuals must be balanced with the value of individual level data in answering research questions. Anonymization of data is superior to de-identifying data for protecting confidentiality, but may make the data set much less valuable as a resource.

4. Informed Consent

Informed consent for future use of data beyond the original study should be obtained at the time of the original study.

Any future use should involve general approval appropriate to the relevant Institutional Review Board. Attempts to provide detailed descriptions of possible future uses are not feasible and should not be required.

5. Timing of Data Sharing

Data should be made available for sharing as soon as possible after publication of the main results of a study.

There is a legitimate concern on the part of original investigators who may wish to benefit appropriately from their investment of time and effort. The principle is to avoid prolonged, exclusive use. Data sharing may be appropriate with unpublished data in cases

where, for whatever reason, publication is delayed for a prolonged period.

6. Archiving of Data Sets

Archiving of data sets, especially those that are unique, is to be encouraged when possible and appropriate.

Archives provide stable, reliable, and cost-effective means to maintain and distribute data. Archiving requires anonymization, or at least de-identification, of the data and documentation to make data sets publicly available. Once data are archived, additional requirements on the producer of the original data are minimized. Data protection, documentation, technical assistance, and funding requirements should be provided for in the original application.

7. Cost Considerations

The costs of data sharing may be borne by the entities funding the original research or those supporting the secondary use of the data.

When data are collected with the intention of making them available later through sharing or archiving, the anticipated costs should be borne by the funding agencies. Costs for data documentation and archiving should be covered in the original grant or competing renewal. Costs of making data available to secondary users, when not planned by the original investigators and funding agencies, should be borne by the secondary user. A new funding mechanism may be needed to cover administrative and other costs associated with unanticipated long-term data storage and management.

8. Time Commitment

Data sharing arrangements should take the time requirements, beyond pure costs, into account.

The time required to document data sets and negotiate secondary use, even if it is for archiving of the data or making them available on the web, may be extensive. Such requirements will likely compete with other research priorities of the original investigator. As a partial offset to this issue, the value of the service provided to science needs to be recognized by the profession in considering the advancement of knowledge and academic achievement.

9. Data Sharing is not Always Appropriate

Data sharing may not be appropriate in some circumstances.

In some cases, data may be proprietary or protected by a legal requirement between the original sponsor and the investigator. Such requirements should be reviewed by the investigator and the Institutional Review Board in advance of data collection to assure the ethical conduct of the original research. In other cases, it is difficult to anonymize or de-identify data sets because individuals can be identified on the basis of the epidemiologic

variables collected, or the uniqueness of the variables might identify an individual. Data sharing may also be inappropriate if a large proportion of the participants have specifically declined consent for the future use of their data. A judgment about the appropriateness of data sharing should be made by the appropriate IRB.

IV. References

1. Rockwell RC, Abeles RP. Sharing and archiving data is fundamental to scientific progress. Guest editorial. *J Gerontol* Series B: Psychological Sciences and Social Sciences. 1998;53B:S5-S8.
2. Hogue CJR. Ethical issues in sharing epidemiologic data. *J Clin Epidemiol* 1991;44:103S-107S.
3. Sieber JE. Data Sharing: Defining problems and seeking solutions. *Law Hum Behav* 1988;12:199-206.
4. Deyo RA, Psaty BM, Simon G, Wagner EH, Omenn GS. The messenger under attack -- intimidation of researchers by special-interest groups. *N Engl J Med* 1997;336:1176-1180.
5. http://grants.nih.gov/grants/policy/a110/a110_guidance_dec1999.htm
6. http://grants2.nih.gov/grants/policy/data_sharing/index.htm
7. http://grants.nih.gov/grants/policy/nihgps_2001/part_iia_6.htm
8. http://www.acepidemiology.org/policystmts/DataAccess.htm

5/12/2002, bs:cc:vs