DATA STORAGE, RETRIEVAL, AND REUSE IN EPIDEMIOLOGIC STUDIES

A. Wouter Voors, M.D., Dr.P.H.

EACH original epidemiologic study can serve to test some hypothesis and to provide a body of primary data for use in any applicable future test.

In a sense, every epidemiologic study is a test of some hypothesis, however inexplicit. Aside from serendipity, the selection of characteristics to be recorded is motivated by more than whim, and the numerical description of the study material, if unambiguous, either supports or weakens the implicit hypothesis. Thus, each study may produce new theories which, in turn, are testable. The observations for the new tests can often be obtained by rearranging old data. This is especially true when studying diseases of low incidence, where combining data from several studies can be both advantageous and feasible.

Such reuse, however, usually requires that the readers have access to the full data. For example, a publication might include two tables: disease incidence by age and by race and disease incidence by age and sex. The reader might wish to examine the disease by race and sex. This cross tabulation could not be obtained from the two published tables alone; only the full data for these characteristics would suffice.

In most instances multiple use of collected data requires communication with the author of the original article and depends on his willingness and ability to keep the full data and duplicate it in answer to requests for an indefinite time. These requirements strain both provider and demander. To avoid this strain the most economical way to store the full data and the most accessible services for storing, retrieving, and duplicating this data should be found. Whether or not the amount of data collected in an average epidemiologic study is small enough to remain within the practical limitations of such services should also be determined.

Data Storage

Storage of data is usually accomplished by punchcards, magnetic tapes, or simple lists. Table 1 is an example of a data list. For purposes of duplication and dispatch, the list is probably the least expensive of the alternative storage media, although reuse of the data usually requires punching a new deck of cards. A simple method to minimize errors in the copying process is therefore desirable. In table 1 this is accomplished by adding marginal subtotals to both rows and columns. Thus any punching error shows up as a difference in the subtotals of both its row and its column when a printout (which can include these subtotals as obtained by an appropriate computer program) of the newly punched cards is compared with the original list of data.

Facilities for Data Storage and Retrieval

The American Documentation Institute, a service of the Library of Congress' Auxiliary Publications Project, which for a nominal fee provides photoduplicates of data supplementary to papers in certain scientific journals, is a suitable facility for data storage and retrieval. The editor of a journal makes arrangements for this service for suitable papers in his journal with the American Documentation Institute,

Dr. Voors is research associate, department of epidemiology, University of North Carolina School of Public Health, Chapel Hill. He was formerly senior resident, epidemiology residency program, New York State Department of Health, Albany. This study was supported in part by Public Health Service training grant GM–558.

Library of Congress, Washington, D.C. 20540. An author can channel requests for data to this facility by including the address and the amount of money due per request in his paper. Naturally, there are certain practical restrictions regarding the amount of data duplicated for each request. However, the restrictions are not defined officially, and there seem to be no practical obstacles against handling up to 20 typewritten or printed pages of the usual format as a single request.

Amount of Data

To obtain some idea of the amount of data in an average epidemiologic paper, all reports of original epidemiologic research were chosen from the papers published in the *Journal of Chronic Diseases*, January–June 1964. The number of persons or cases (indicating the number of rows in a list of the full data) and the number of characteristics measured (indicating the number of columns in such a list) are given for each paper in table 2.

In table 1 each characteristic coded has only two classes, so that one digit suffices to indicate the class in which each person belongs. If the characteristic "age" included as classes all years from 10 to 99, two digits would be needed to express the corresponding numerical code.

Table 1. Coded data list with marginal subtotals added

Person or case identification No.	Age [1]	Race [2]	Sex [3]	Disease [4]	Row subtotal
01	0	0	1	1	2
02	0	0	1	0	1
03	0	0	1	0	1
04	0	0	1	0	1
05	1	0	1	1	3
06	0	1	1	1	3
07	0	0	1	0	1
08	1	1	1	0	3
09	1	0	1	1	3
10	1	0	1	0	2
11	0	1	0	1	2
12	0	1	0	0	1
13	0	0	0	1	1
14	1	0	0	0	1
15	1	1	0	1	3
16	1	1	0	1	3
17	1	1	0	0	2
18	1	1	0	0	2
19	1	0	0	1	2
20	1	0	0	1	2
Column subtotal	11	8	10	10	39

[1] 0–29 years =0, 30 years and over =1.
[2] White =0, nonwhite =1.
[3] Male =0, female =1.
[4] Disease absent =0, disease present =1.

Table 2. Number of persons, characteristics, and digits needed for coding all classes within each characteristic for each original article published in the *Journal of Chronic Diseases*, January–June 1964

Principal author	Initial page number	Number persons or cases	Number characteristics	Total digits needed for class codes
Chazan	9	460	6	13
Brindley	19	131	12	22
Okun	31	19	11	40
Skinner	57	857	13	15
Wardwell	73	609	7	7
Hoehn-Saric	91	72	8	21
Larkin	109	35	10	10
Grizzle	127	227	6	19
Wallace	153	2, 141	9	9
Tyroler	167	14, 710	6	6
Wiener	191	128	11	20
Veterans Administration	207	716	9	9
Petrinovich	225	40	28	28
Ostfeld	265	1, 885	30	75
Syme	277	609	12	12
Kasl	325	331	22	38
Caffey	347	348	30	58
Wylie	359	411	12	15
Mabry	371	181	9	11
Julius	391	100	12	36
Julius	397	208	6	18
Harburg	405	74	27	59
Stazio	415	242	20	23
Solomon	439	129	6	7
Maddox	449	182	9	13
Schroeder	483	637	6	16
Kinch	503	1, 849	10	20
Rodstein	515	168	7	8
Wallace	527	52	12	17
Rogoff	539	46	9	9

The last column in table 2 shows that no paper in this sample needed more than 75 digits for this purpose. A maximum of 80 digits, including 5 for patient identification, can be accommodated on a standard IBM punchcard, and, similarly, 80 characters can be printed by a regular typewriter on one line of paper. The total number of persons or cases in the 30 sampled papers is less than 1,000 in all but 4 instances. Hence, for approximately 9 of 10 studies, the complete list of data can be typed on less than 20 sheets of 50 lines each. If the sample is representative of current literature, relatively few epidemiologic studies present mechanical obstacles to the general accessibility of the full data.

In summary, reuse of epidemiologic data can be an economical and useful step in the scientific process. Reuse can be promoted by the adoption of an inexpensive method of data storage, which includes some redundancy to permit detection of copying errors, and by the use of an existing specialized library service to provide accessibility. In a sample of 30 epidemiologic papers, the data of all but 4 had a bulk which remained within the practical limitations inherent in such library service.

Food Chemicals Codex Published

The National Academy of Sciences—National Research Council has announced publication of the "Food Chemicals Codex," which defines standards of identity and purity for more than 500 food additives now in common use.

The Codex will provide chemical manufacturers and food processors with uniform release, procurement, and acceptance specifications comparable to those that have been available for drugs through the U.S. Pharmacopeia and the National Formulary.

Publication of the first edition of the Codex culminates a 5-year effort, initiated by the Food Protection Committee of the NAS-NRC Food and Nutrition Board, in which scientists from Government, industry, universities, and private research organizations have cooperated. Prior to this time, sections of the Codex were issued in loose-leaf form.

Funds for the project were provided by the U.S. Public Health Service, supplemented by grants from more than 100 organizations concerned with the manufacture of food chemicals and their application in food processing.

Dr. James L. Goddard, Commissioner of the Food and Drug Administration, has stated that FDA will regard specifications in the Food Chemicals Codex as defining chemicals of an appropriate food grade within the meaning of relevant sections of the Food Additive Amendment to the Food, Drug, and Cosmetic Act of 1938.

Flavoring agents, antioxidants, preservatives, sequestrants, nutrient supplements, and emulsifying, stabilizing, and thickening agents are among the many classes of food additives covered in the Codex. These include such common ingredients as salt, baking powder, citric acid, and monosodium glutamate, but exclude sugar and starch which are regarded as basic nutrients.

The Government has, by regulation and informal statements, established quality requirements for about 575 food chemicals generally recognized as safe and has set use tolerances for others. However, these requirements are not always sufficiently detailed to serve as procurement and release specifications for industry. Food processors, when ordering chemicals from primary manufacturers, need one source of standards accepted by manufacturers, Government regulatory agencies, and purchasers. The Codex is expected to fill this need.

The 832-page Codex consists of a series of monographs. Each monograph provides the name, description, endorsed purity standards, and test methods for determining the purity of the subject chemical. The specifications, in most instances, are more rigid and more definitive than those published in the U.S. Pharmacopeia and the National Formulary compendia. As a result, new methods were developed for determinations at Codex purity levels.

The methods, described in a special technical section, include tests for melting range, loss on drying, distillation range, specific gravity, and procedures for determining such impurities as arsenic, heavy metals, and lead which are present in many natural foods, and are safe in trace amounts.