Brief Reports

THE CASE OF THE MISSING EIGHTS AN OBJECT LESSON IN DATA QUALITY ASSURANCE

STEVEN D. STELLMAN

It seems self-evident that no epidemiologic study can be better than the data on which it is based. Nevertheless, quality assurance of data and control of the data entry process have received scant attention in the literature of epidemiology, and few epidemiologists receive systematic training in the mechanics of computerized data management.

Data editing procedures have occasionally been presented in outline (1), and even more rarely in extensive detail (2), particularly for large-scale studies (3). Few textbooks of biostatistics or epidemiology, however, have dealt systematically with this topic. In Cochran's classic *Sampling Techniques*, a chapter on nonrandom sampling variation no more than mentions "errors introduced in editing, coding, and tabulating results" (4, p. 359). Rothman, when discussing data editing in his textbook, *Modern Epidemiology*, enumerates important practices such as use of sophisticated data entry programs and editing checks, examination of univariate and joint frequency distributions of questionnaire items, and avoidance of repeated manual transcription, and gives some examples of simple data checks (5).

A researcher's control over data quality decreases with distance from the data entry source. The controls described by Rothman are most easily applied when data entry is done in house under the supervision of the researcher or other professionals used to

dealing with epidemiologic data. However, much data entry is done extramurally by commercial firms for whom survey research is not the principal business, and who cannot automatically be expected to appreciate the special needs of epidemiologic studies or to make the necessary checks as a matter of routine. Data checking specific to a study must be specified in detail when the job is negotiated.

At the greatest distance from the original data are the many investigators, including legions of graduate students in search of thesis topics, who first encounter their data in the form of a "data set," ordinarily on magnetic media (or, in bygone days, on punch cards), and undertake analyses without the least knowledge of—or control over—the manner in which the data came to reside on that medium. For such persons, thorough systematic data checking prior to analysis is an essential procedure for determining just how good the raw data actually are.

Even the most meticulous data scrutiny cannot routinely uncover all possible data entry problems. We describe here the circumstances surrounding the discovery and correction of an unusual data entry error in a large prospective mortality study, Cancer Prevention Study II (6). The baseline data consist of approximately 1.2 million questionnaires completed by study subjects during the fall of 1982. Each questionnaire resulted in a magnetic data record 605 bytes in length. The data were entered and key-verified under contract by a professional data entry firm to specifications developed during an extensive pretesting phase. The firm typically processed 20,000 forms per week. Upon delivery to us, each weekly batch was immediately read and subjected

Received for publication April 25, 1988, and in final form October 31, 1988.

From the American Cancer Society, New York, NY. Reprint requests to Dr. Steven D. Stellman, Division of Biostatistics and Epidemiologic Research, New York City Department of Health, 125 Worth Street, Box 22, New York, NY 10013.

to an exhaustive battery of range and logic checks, as well as content-sensitive tests (such as joint distributions of key variables). These quality assurance procedures occasionally detected drifts from acceptable error levels which were quickly corrected.

Diet was among the many topics covered. Dietary information consisted of a food frequency section (28 items) and a matrix of beverage consumption information. For each of the 28 food items, subjects were asked to enter the number of days per week each food item was consumed. Each response was punched as a single-digit number from 1 to 7, as well as 0 (if actually written); 8 was punched for less than weekly but more than monthly, and 9 was punched for seasonal consumption. If the item was left blank by the subject, the field was to be left blank. However, if a subject left the entire 28-character diet field blank, the data entry computer was to enter a special alphanumeric character in the first of the 28 food frequency fields as a flag. (The key operators had no control over this procedure and were, in fact, unaware of it; it was done by computer at a later stage.)

While data entry was still in progress, we began to prepare data work files for a factor analysis of diet and beverage consumption (7). Having determined that factor analysis interpretation was sensitive to the amount of information left missing by subjects, we sought to examine various patterns of missing data. Our strategy was first to see which foods tended to be omitted together, as a clue to interpreting factor analysis results. Because of the enormous size of the complete data base, testing was done on an SPSS-X (8) system file containing selected data fields for a random sample of 25,000 male subjects. SPSS-X has a built-in function, NMISS, which, given a list of variables, returns the number of variables that have missing values. It seemed natural to begin by checking the distribution of NMISS for the 28 food items.

Table 1 shows this distribution, which exhibits a peak at zero, indicating that 6,354 men (25 per cent) filled in the entire

TABLE 1
Frequency distribution of missing values for 28 food frequency items in 25,000 randomly selected male subjects, Cancer Prevention Study II, fall 1982

No. of missing items	Frequency
None	6,354
1	2,638
2	1,601
3	1,296
4	1,244
5	1,248
6	1,209
7	1,127
8	0
9	1,080
10	864
11	821
12	716
13	723
14	574
15	480
16	414
17	311
18	0
19	197
20	124
21	88
22	62
23	47
24	25
25	19
26	16
27	1,722*

* If all 28 items were left blank, a special alphanumeric character was to be recorded in the first of the 28 fields.

food frequency section. This was followed by a fairly regular tapering off. In this sample, some 2,638 men (11 per cent) filled in all food item questions except one, another 1,601 (6 per cent) filled in all questions but two, and progressively smaller numbers of respondents omitted progressively more answers. In addition, 1,722 (7 per cent) subjects had left the entire section blank, resulting in a flag code of the special character followed by 27 blanks.

There were, however, two unexplained exceptions to the otherwise smooth distribution. On no questionnaires were exactly eight questions left blank (whereas about 1,100 were expected, based on adjacent values in the table), nor were there any records

Values for 28 food
selected male
II, fall 1982

Frequency

,354
,638
,601
,296
,244
,248
,209
,127
0
,080
864
821
716
723
574
480
414
311
0
197
124
88
62
47
25
19
16
,722*

* special alpha-
1 in the first of

was followed
off. In this
percent) filled
cept one, an-
in all ques-
vely smaller
ted progres-
ion, 1,722 (7
ntire section
of the special
ts.

unexplained
nooth distri-
were exactly
ereas about
adjacent val-
any records

with exactly 18 responses missing (whereas about 250 were expected).

What does one do when confronted by such a bizarre observation? There was no epidemiologic reason to suspect such an extraordinary distribution. Nevertheless, experience dictates that distributions of real data do not contain giant holes. Suspicion turned to the data entry contractor. Despite the outstanding level of quality that we had heretofore enjoyed (due in good measure to rigorous performance standards and constant monitoring), we formed a small task force to find the solution.

As a first step, the contractor reviewed all data entry programs to see if some individual food item had been systematically miscoded through a program error. This was not the case. Next, in our own data file we recoded each food item to 0 if the answer was missing and 1 if the answer was present, and examined the Spearman correlation coefficients of all 28 items taken pairwise, to see whether absence of some variable might be systematically linked to absence of another. All coefficients were in the range of 0.3 to 0.7. No such linkage was found. Several dozen records were listed and missing values were counted manually to make certain that the SPSS-X missing-value function was operating correctly; it was. The distribution of NMISS was examined in the first 70,000 raw data records from the master data base (rather than an SPSS-X system file) to see whether the problem had originated in either the sampling procedure or in preparation of the system file; the distributions were, within sampling error, identical, and the larger sample had no observations at either eight or 18 missing items.

By this time, there were no further obvious tests that could be run internally on the computerized data. By interpolating within table 1, we theorized that about 6 per cent of the questionnaires that should have had exactly eight missing diet responses and about 1 per cent with exactly 18 missing diet responses had been systematically mispunched as something else.

Having no idea how they ended up, we had no way to identify them from the computer side. Consequently, examining original questionnaires was the only recourse. Clerks were instructed to pull questionnaires systematically from the files and record the number of missing diet responses on each. They were not told the purpose of the search. We could only hope that 1) the miscoded questionnaires were randomly distributed among the 1.2 million originals, and 2) once we identified and examined a few of the bona fide "missing eights" (of which we expected six in 100), the solution would be obvious.

The search paid off. The tenth questionnaire pulled was missing 18 responses, and questionnaires missing eight responses occurred in the files with their expected frequency. Upon examination, we found that the computerized records for every one of these questionnaires had been flagged as if the entire food frequency section were missing, that is, with the special character in the first column followed by 27 blanks.

The contractor's lead programmer was asked to inspect all code related to the flag in question, but could find no errors. This was an exceptionally capable individual, whose word could be accepted as final. Seeking a possible (but unlikely) flaw in our own data logging process, we examined originally delivered data tapes (from which our system files were derived), but these proved to be identical in content to the system files. The problem simply had to originate with the contractor. At the time we were reaching this conclusion, the programmer called back again with a sheepish tone to say she had discovered the problem in her program. After all data items had been entered, the number of missing items was subtracted from 28 and the result was tested against zero; if the numbers were equal, the first item in the series was output as the flag character and the remaining 27 were output as blanks. But the line of code with the test contained a misprint: A letter "O" had been typed instead of a zero (one of the hardest programming errors to de-

ted). In the machine level language of the contractor's computer, this mistyped instruction was still a legal one, but it gave a test result of "true" for any number of missing items that ended with the digit 8. The problem was now identified, and steps were taken to correct it.

CONCLUSIONS

This particular bug was dependent on the use of one particular data entry computer, and is unlikely to afflict other investigations or "data sets." Nevertheless, it exemplifies a number of important lessons in quality assurance that may be applicable to nearly any study involving computerized data.

1) A "data set" cannot automatically be assumed to be "clean," no matter how trustworthy its source. Newly acquired data must be checked carefully for errors.

2) Simple range and logic tests, while an essential ingredient of quality control, will not necessarily detect all possible data entry errors. Content-sensitive tests also provide useful quality checks.

3) The distribution of missing values for individual items and for related groups of items can provide important information on data quality. Distributions that appear unlikely on either epidemiologic or statistical grounds may be symptomatic of data entry or editing problems.

4) When all else fails in tracking down problems, read the source code carefully. Then read it again.

5) A single data field should not be made to double as a flag. This practice is a relic of the days of expensive punch cards. Magnetic media are now cheap enough that flags can be given their own fields, even for very large files. In our study, the savings of 1.2 megabytes of storage from double use of the first diet history column as both data

and a flag was ultimately outweighed by the cost of discovering and fixing a systematic error it generated.

6) Data should be scrutinized as soon after receipt as possible, and "final" versions should be compared with the raw data files from which they originated on an ongoing basis. A systematic data quality problem, once discovered, must be fixed. There was never any doubt, either on our part or on that of the contractor, that this problem had to be solved and its source determined and corrected. Uppermost in our minds was the question, "Is this a symptom of other deeper and more extensive data quality problems that could permeate the entire data base?" It was imperative to answer this question, and fortunate for subsequent analyses that we did.

REFERENCES

1. Royal College of General Practitioners. Oral contraceptives and health: an interim report. Appendix 1. London: Pitman Medical, 1976.
2. National Institute for Occupational Safety and Health. National Occupational Hazard Survey. Vol 2. Data editing and data base development. Cincinnati, OH: National Institute of Occupational Safety and Health, 1977:80-8. (US DHEW publication no. DHEW (NIOSH) 77-213).
3. Multiple Risk Factor Intervention Trial. The Multiple Risk Factor Intervention Trial: quality control of technical procedures and data acquisition. *Controlled Clin Trials* 1976;7(3).
4. Cochran WG. Sources of error in surveys. In: Sampling techniques. 3rd ed. New York: Wiley, 1977: 359-99.
5. Rothman KJ. Modern epidemiology. Boston: Little, Brown and Co, 1986:131-4.
6. Stellman SD, Garfinkel L. Smoking habits and tar levels in a new American Cancer Society prospective study of 1.2 million men and women. *JNCI* 1986;76:1057-63.
7. Stellman SD. Chairman's remarks on Session V: data analysis in cohort studies. In: Garfinkel L, Ochs O, Mushinski M, eds. Selection, follow-up, and analysis in prospective studies: a workshop. Bethesda, MD: US Public Health Service, 1985: 145-7. (National Cancer Institute monograph 67). (NIH publication no. 85-2713).
8. SPSS, Inc. SPSS-X user's guide. 2nd ed. New York: McGraw-Hill, 1986.