Discussion

J. William Lloyd - National Cancer Institute

The four papers presented this morning have given us a comprehensive view of record linkage procedures and some of the problems encountered in their use.

Rather than review the papers individually, I would prefer to offer some general comments on record linkage procedures with reference to what our speakers have presented.

My first job in the field of statistics was with a local health department where the compilation of vital statistics had for many years consisted of single entry of events in a set of ledgers. Each of these ledgers being maintained by geographic area and finely subdivided by the demographic characteristics of interest. Before my arrival on the job, it had been decided that the volume of events to be recorded was too much for the ledger system and should be replaced by a mechanized system of counting--namely, a key punch machine and a 250-card-a-minute sorter.

As you may have anticipated, what should have been a much more rapid and efficient system of enumeration was at least for a short while an uncontrollable Frankenstein. Coding schemes had to be developed, the clerk who had made the Journal entries had to be trained to key punch--a task he never did master. He was convinced that the new-fangled system wouldn't work. And the sorter, in addition to throwing cards in the wrong pocket, would without warning chew up cards by the handful.

Now we are in the day of the "black box"--the computer that can pair 200,000 records of one system with 300,000 records of another, and quite rapidly select the 38,000 pairs that meet some specification. Our old ledger clerk would be amazed! While the problems generated by our new equipment, and our new technique of record linkage are not quite analogous to these encountered with our earlier mechanized procedure, we have responded in quite the same manner as we did to the 250-word-a-minute sorter. In our eagerness to make use of the high speed computer, we sometimes forget that we are still doing exactly the same thing as the ledger clerk. Each of the decisions made by the ledger clerk when he visually reviewed a pair of records is a decision that we must make. The problem that we encounter in working with the high-speed computer is that we have to anticipate the kinds of situation which might have arisen in a manual review, and specify decisions for each situation.

Before cataloging decisions of this sort, however, we should first ask ourselves if it is reasonable in terms of expected productivity to even consider matching of the two systems. While no detailed set of rules can be devised to cover every set of records which could be matched, there are some general criteria which should be considered before any record linkage is attempted.

The first of these, which may seem obvious, is that each record in one group should have a chance to appear in the other group. For example, in the HIP study we might expect some attrition in the death file over time among persons retiring to Florida and therefore not reported as New York residents at the time of death. There will be times, of course, when this criteria is not strictly met but can be corrected for by estimating the occurrence of an event among the nonmatched cases. In the HIP study such estimates might be based on out of state death claims filed with the insurance company. The same sort of a procedure was suggested in the paper on psychiatric admissions. The bias resulting from failure to meet this criteria can be considerable. In one of the studies being conducted by our agency we have noted quite different patterns of mortality between those who have remained in a local community and those who have left.

As other criteria for successful record linkage, I would suggest those pointed out by the authors of the first paper. They bear repeating. The common identifying information of the two systems should have

(1) high discriminating power
(2) low probability of change during an individuals lifetime and
(3) low likelihood of being recorded erroneously.

The unexpected low rate of matching in the psychiatric admissions study may be due in part to the dependence on information which does not fit these criteria, i.e. address at admission. Dr. Pollack, it should be noted, also requested the patient's address at the time of the census. Some of the nonmatches may therefore be due to faulty recall. The probation records, in contrast, showed a much higher rate of matching for an eighteen month period around the census with no apparent tendency to decrease over time.

One other factor which might account for the different match rates of these two studies is that there may be considerable differences in levels of enumeration and/or identification by enumeration district for large metropolitan areas as compared to complete states (which may have a large rural component). The authors of the last paper have pointed out that problems exist in the classification of rural addresses by enumeration district. Finally, I would raise the question as to whether matching of survey records to those of the decennial census, using name and address as the primary information, appears to be sufficiently fruitful for estimating vital statistics rates.