Discussion*

Thomas B. Jabine, Statistical Consultant

The papers in this session are by government statisticians (one has left government, but I would still put him in that category) who are working hard to find cost-effective ways to meet the data needs of policymakers and other users. Development of optimal data systems calls for consideration of relevant data from each of three major sources--censuses, surveys, and administrative records. This was billed as a session on administrative records, but what we should actually be doing when faced with a particular set of data requirements is to ask ourselves how we can best use relevant data from any or all of these three sources. In a very fundamental way, this is what government statisticians do--they integrate messy but potentially useful data from different sources to arrive at the best possible set of estimates for the purpose at hand. In some instances, as in the paper by Sailer and Weber, the goal is to create a multipurpose data base. In others, as illustrated by Scheuren and Winkler, the goal is to develop estimates for more narrowly defined purposes.

Obviously, one wants to optimize the quality of the estimates according to some set of criteria--not an easy conceptual problem when we are creating a multipurpose data system--but there are also two other important considerations: cost and acceptability to non-statisticians. Cost considerations may be nudging us in the direction of increased use of administrative records. On the other hand, the acceptability of different data sources and methods to policymakers, legislators, the courts, the media, and the general public may lead us in a different direction. Technically best solutions may involve record linkages, uses of income tax data, and complex model-based estimation procedures--techniques that are not always easy to explain convincingly to those who do not speak the language of statistics but do quite properly have a say in how government statistics are collected.

Unlocking the unrealized potential for greater statistical use of administrative records requires a willingness to admit that data systems based primarily on administrative records cannot produce exactly the same information as a traditional census or survey-based system. Let us suppose, however, that the resources used for the decennial census could be used instead to fund a system based primarily on income tax and Social Security records that would produce small-area data annually (as opposed to every ten years), with fewer data items and somewhat different definitions of concepts like family and income, but roughly equivalent. Unless we are willing to consider tradeoffs like these (as many Western European countries have), administrative records will continue to have a very minor role in our statistical system.

However, it need not be an all-or-nothing proposition. There appear to be two principal ways of using administrative records--a direct (or registry) approach in which the data for each unit of the target population come from one or more administrative records sources, and a symptomatic approach in which aggregate data based on administrative records along with census or survey data are used as variables in model-based estimates of the statistics of interest. There is also the possibility of combining these two approaches. I would argue that a combined approach may often be the one that makes the most sense. The plans for the 2000 Census call for primary reliance on traditional census methods, with administrative records perhaps being used peripherally to impute some missing data and to help identify missed persons or housing units. We might also conceive of a 2010 Census, which places primary reliance on administrative records, with traditional enumerative methods being used on a sample basis to adjust for undercoverage by the administrative records sources.

**This Discussion is an excerpt from a presentation that discussed four papers on statistical uses of administrative data at the 1996 Joint Statistical Meetings in Chicago. Only that portion of the presentation related to papers reprinted in this volume is included here.*

Having spent much of my allotted time on these general issues, I will now comment on the papers.

Scheuren and Winkler

This paper is concerned not with a specific application of administrative records but with a tool for linking data from two different files, of which one or both might be generated from an administrative records data system. The purpose of the linkage is to undertake analyses requiring variables from both files. The authors do not provide an explicit statement of their assumptions about the coverage of the two files, but it would appear, from the examples provided, that their intention is to work with census-type (rather than sample) files that cover the same population, although 100-percent overlap is not assumed. Their examples use business data, but, in principle, their methods are equally applicable to data for persons or households.

The approach is ingenious and appealing. It is a good illustration of my earlier statement that the task of statisticians is to make the best possible use of all the information at hand in developing estimates or conducting statistical analyses. To accomplish this, it helps to make creative use of tools and techniques that have been developed for other purposes, in this particular instance, record linkage, data editing, and imputation. The initial results look quite promising, and one hopes that some practical applications lie ahead. Perhaps the authors could try some applications with records for persons or households in ways that could aid in the development of the population data base proposed by the Census Bureau.

We must not forget that records linkage is also a tool that can be used by "attackers" to identify individuals in a data base that does not include specific identifiers. In saying this, I do not mean to discourage efforts to develop better records-linkage techniques for legitimate statistical and other purposes, but just to remind us that not all of the potential consequences of scientific and technical advances are necessarily benign.

Sailer and Weber

The project they describe is an extension of a well-established statistical system based entirely on administrative records: the IRS Statistics of Income Division's sample of individual income tax returns. The new feature provides sample data for "tax families," groups of closely related persons filing separate tax returns. Presumably, the resulting data will permit a better informed analysis of how our tax policies affect families, defined in a more traditional sense. This new family file illustrates what can be done, with suitable sampling and records-linkage methods, to overcome some of the differences between standard census and survey concepts and those used in administrative systems. The differences that will inevitably remain between the Census estimates of households or families and the IRS estimates of tax families remind us of the impossibility of exactly reproducing census concepts when working with administrative records. In spite of these differences, there are undoubtedly some data needs, especially the analysis of tax policies and how they affect families, for which the IRS Family File will be the best available source of information.

Unfortunately, it seems that some glitches in the matching and weighting procedures were discovered, and this delayed the comparisons of the new family file with the regular SOI file based on tax returns and with data on families from the Census Bureau. Thus, we are reminded of some of the difficulties of working with very large administrative data sets designed for nonstatistical purposes. Nevertheless, I remain convinced that there is much to be gained by better use of administrative records. But the payoff will not be obtained without both the understanding and cooperation of the custodians of administrative records systems and a commitment by statistical agencies to devote the resources needed for developmental work to solve the many technical problems that confront the authors of these papers. Their efforts deserve our support.