DISCUSSION

Gordon Scott Bonham, National Center for Health Statistics

The papers in this session address a single subject: the combination of data from different sources to provide more accurate estimates than either source could provide alone. They report experience gained from the National Medical Care Expenditure Survey (NMCES). My discussion is based upon experience gained from the same survey. I will discuss the rationale and the basic issues underlying the combination of data from two different sources, and how the research presented in this session contributes to our understanding.

Objectives of the NMCES were to accurately measure health problems, medical visits, charges for health care, and sources of payment. Data were collected from a sample of households through personal and telephone interviews. In addition, data were collected from a subsample of providers of health care to interviewed households. Both the household and provider surveys measure the abstract constructs of visits, charges, and sources of payment with unknown degrees of accuracy. Comparability and accuracy in measurement is a substantial problem as demonstrated by the availability of at least three estimates of the average number of doctor visits per person during 1977: 4.8 visits per person from the National Health Interview Survey; 4.0 visits per person from the NMCES; and 2.7 visits per person from the National Ambulatory Care Survey. The numbers are not really comparable, however, since there are major differences in the surveys. It would be desirable to have a single accurate estimate. Information provided by physicians from their records should not be as subject to recall bias as information provided by households. Information provided by physicians should also be more accurate in terms of medical diagnosis than information provided by households. However, information provided by physicians will be limited to the individual visit, while information provided by the household can relate to the total person and all health care. Information provided by physicians is subject to errors such as incorrect initial recording, failure to record, looking at the wrong record, and incorrectly transcribing what was in the record. Information provided by households is subject to recall error, incorrect reporting, and lack of knowledge. Differences in estimates do not mean that one source of data is better or worse than the other, just that they are different and may have different errors.

A combination of household and physician provided information could capitalize on the strengths of both data types. However, there are many problems associated with combining the two types, and some of these problems are discussed in the papers of this session. Both sources of data probably underreport visits, and the visits they do report may not be the same. There are basic problems in combining data sources due to 1) coverage; 2) datum identification, definition and matching; and, 3) providing population estimates.

The same population coverage is required for combining data. This can be achieved by three methods: 1) collecting data on all visits reported through complete enumeration of all households and physicians; 2) collecting data on all visits reported by a sample of households and either a complete enumeration of all physicians or all physicians seen by sample households; and, 3) collecting data on all visits of a sample of households and of a subsample of physicians who provided care to these households (double sampling). The latter choice was made for the NMCES, and Folsom presents the theoretical equivalency of various ways to adjust for double sampling. Loss of precision occurs in a double sample, though, and Cox and Folsom find empirically that imputation is not equivalent to reweighting unless all analytic domains of interest can be used in defining weight classes. None of the papers in the session discuss why a double sample was chosen for the NMCES, nor the advantages of double sampling. Horvitz implies that there are some, but at this time there is no evidence that double sampling cost less than surveying all providers, that double sampling increased the quality of the data, or that double sampling was the only way the information could have been collected. The papers give guidance on what to do if double sampling is used to achieve coverage of the same population, but do not address the more basic issue of the best way to achieve coverage of the same population when combined data are desired from two different sources.

The second major issue with combining household and provider data is item identification and matching. Initially, it appears that a doctor visit is a doctor visit. The definition of visits, however, can cause problems. Consider the person who went to the oncology clinic and the radiology clinic of a hospital outpatient department on the same morning. Does the household consider this one or two visits? Does the hospital outpatient department report this as one or two visits? What are the chances that the household and the outpatient department report this the same way? Additionally, doctor visits must have a common identification if matching is to occur at the visit level. If the household reports a visit occurring on March 2 and the doctor reports a visit occurring on March 7, are they the same visit, or is one a followup to the other? The papers in this session assume a "truth set" of matched visits as they discuss matching algorithms, unreported visits, and imputation strategies. They discuss the outcomes of various processes, given that common definitions and identifications have been achieved. None of them warn, however, that there was probably more error involved in the process of determining "true" matches than in all the subsequent statistical processing. Cooley and Cox find that automated procedures can come close to replicating the match decisions of a half dozen or so government analysts, but that does not mean that either set of matches was close to "truth." Truth may be very elusive if only 30 percent of visits match by a "tight match" criteria: same person identification number, same medical provider

identification number, dates within one day of each other, and charges within $20 or 20 percent of each other (whichever was smaller). Figures in the Cox and Folsom paper show the implications of accepting the matching as true. An average of 2.0 visits per person reported by medical providers were not matched to visits reported by the household. Since households reported an average of 4.0 visits per person, households reported only 2/3 the visits they should have, if the matching is true. I suspect the problem may be as much or more in the matching as in underreporting. An illustration of an actual problem encountered in matching was how to handle two sets of three visits when the household reported dates of 3/5, 4/5, and 5/5, and the provider reported dates of 8/5, 9/5, and 10/5. Did these reported visits represent three or six true visits? The answer determines whether the household is judged to report all visits or to report only half of the visits. Cooley and Cox suggest the matching might be improved by taking clusters of visits into account rather than matching visits one at a time.

The third major issue is providing population estimates based upon data collected. Imputation is generally used as a way of handling a few missing items of information for a person or visit included in the survey, based upon other known information. The Williams and Folsom and the Cox and Folsom papers, however, evaluate imputation as an alternative strategy to reweighting for adjusting to a double sampling situation. The imputation was designed to answer the question, "what would have resulted if all rather than a sample of providers of medical care had been surveyed?" Imputation is found to perform satisfactorily in a 75 percent subsample where either the imputed variable is highly correlated to the available household data or where analysis variables are used for stratification. Unfortunately, most double sampling is at a rate less than 75 percent (32 percent in the NMCES Medical Provider Survey), few variables are highly correlated between households and providers, and only a half-dozen of the hundreds of analysis variables can be used for stratification. The papers conclude that imputation is really not appropriate in this type of double sampling situation, and that reweighting is the better strategy.

In conclusion, data from different sources can compliment one another and a combination of different data can greatly enrich the information available to address analytic topics. Different data sources have unique data and have different biases or limitations. The papers in this session were all generated by the attempt to combine household and physician provided data in the National Medical Care Expenditure Survey. The research focused on three topics: 1) ways of producing population estimates in a double sampling situation, 2) the accuracy of automated matching procedures in duplicating hand matching by experts, and 3) imputation of uncollected data. There are some more fundamental questions, however, that need to be addressed: 1) When does double sampling have advantages over complete enumeration? 2) How can true matches be made of records from different data sets? and 3) What is the best way to make population estimates with incomplete data?