DISCUSSION

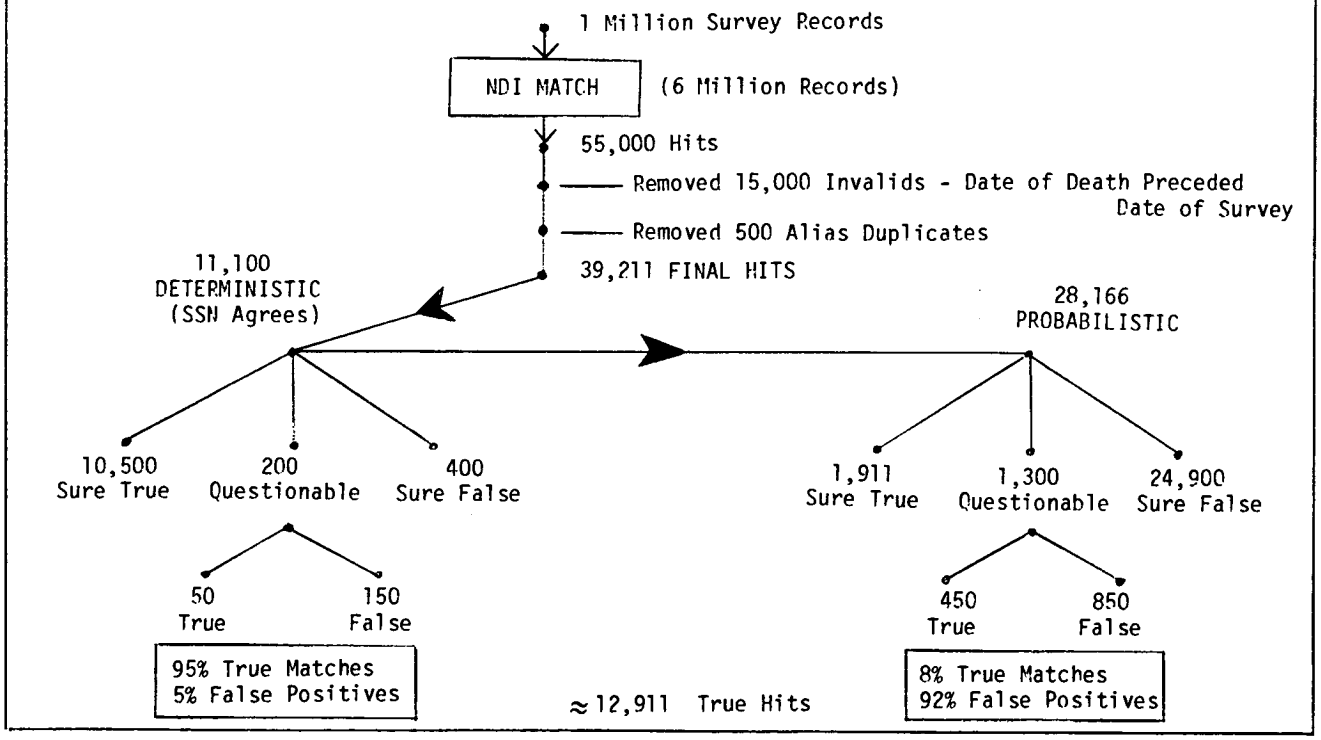
Norman J. Johnson, U.S. Bureau of the Census

I would like to present my discussion of these three papers in terms of points which we have encountered in an application of matching from our project. I have been working on developing the data base for The National Longitudinal Mortality Study (NLMS). This study is being conducted jointly by the National Heart, Lung, and Blood Institute, the National Center for Health Statistics and the U.S. Census Bureau. The primary objectives of the NLMS are to analyze socioeconomic, demographic and occupational differentials in mortality within the United States. A major interest of our analysis will be to compare survival rates of different subsets of the cohorts.

The study population consists of eight cohorts of selected Census samples. Deaths in this population are identified through periodic matching to the National Death Index (NDI), the index discussed in the first paper by Mr. Patterson. As pointed out in that presentation, in terms of number of records submitted for matching, our project is a major user of the National Death Index. The National Longitudinal Mortality Study currently consists of approximately 1 million records from eight cohorts. One match has been made to the NDI, which at the time consisted of approximately 6 million records. We intend to conduct follow-up matches approximately every two years.

The process we used to obtain the final matched records was completed in two steps. First, our files were matched to the NDI using the NCHS criteria. Then, an extensive screening was made of the resulting match using some of the methodologies discussed in presentations given earlier in these sessions to determine the final true match status. This second step involved both computer and manual matching. Our approach in the computer matching phase was similar to that used in the CAMLIS project of Mr. Arellano, the presenter of the second paper of this section. A link was made deterministically for all matches in which there was an exact agreement on social security number. Records not matched deterministically were then matched probabilistically using a modified Newcombe model. Weights for this model were estimated from a subsample of records from the NCHS match which had been reviewed manually to establish correct match status. Three categories of records from the probabilistic match resulted: true, false and questionable matches. Questionable matches were decided on the basis of a manual review. This process and the final results have been schematically diagrammed in Figure 1. From the initial one million records, approximately 12,900 links occurred. The information in the figure also indicates the substantial difference in the true match rate between the deterministic and the probabilistic steps.

Figure 1. -- National Longitudinal Mortality Study



PATTERSON AND BILGRAD

As I mentioned in my introduction, our project is a major user of the National Death Index. Deaths in our cohorts are determined by linking our records to records in this Index. The NDI matching algorithm is, in a sense, deterministic. It uses combinations of five major variables in seven criteria to determine a link. These criteria are soon to be expanded to twelve. A link is made if any one of the seven criteria is satisfied. As other studies continue to match using this index, the NDI may wish to incorporate some probabilistic components into their matching procedure based on the experience of their users. Results from our project may be helpful in this regard.

Five major categories of users were summarized in the presentation. The major users identified are in health-related fields. In many health studies, analysis is done by comparing survival of cohorts, as is the case in our study. Rare events are often of interest and small counts may be greatly affected by match rates. For this reason, in our study, we feel that matching algorithms should put emphasis on detecting true matches, with willingness to manually review more questionable matches, in order to rule out false positives. The additional criteria made available in the new NCHS matching algorithm are a step in the right direction. The expanded criteria will generate more true links as well as more false positives.

The paper presents results of studies to measure the improvements in the match rate to the NDI due to the replacement of the Soundex Code for matching of names by the NYSIIS code. If the NCHS studies of the effects of this change are true, that is, 18 percent fewer true matches and 31 percent fewer false matches could be expected, then, in view of the comments which I made earlier, the Soundex Code would be preferable to us.

ARELLANO

I will focus my discussion on the three points mentioned in the conclusion section of the paper. The paper deals with the use of the Fellegi-Sunter approach in the CAMLIS project to link user files to death certificates from the state of California. The first point discussed concerns the potential for making estimates of error terms in the Fellegi-Sunter model. The estimation of error terms is a major difficulty encountered in application of the theory. In some applications, making simplifying assumptions is the only way to obtain estimates of errors. The similarity of the CAMLIS study and the National Longitudinal Mortality Study may enable us to exchange estimated parameter values once they are obtained.

The conclusion on the robustness of error probability estimates is important and potentially very useful. This quality of the estimates would allow the use of approximate values without great risk of poor matching results and permit a more frequent borrowing of parameter values from other studies. A nice collection of results in the literature demonstrating this robustness would be very useful.

The third point covered in the conclusion deals with the effects of bias. We have observed a positive bias in our scoring algorithm. It would be helpful for us to know if the CAMLIS project has identified any consistent bias in their procedure. If so, what explanation do they have for it?

COHEN

The findings of this particular study are based on the results of a match of two files performed by a Government agency. The match was based on an apparently deterministic match procedure using a certain identification number. The provider of such match results should advise clients of error rates and nonmatch results of similar studies. Error rates of such matches should be required as part of publications and presentations in order to give the reader a chance to determine if any biases have resulted due to the matching procedure. This is similar to documenting which computer and software were used when publishing papers based on computer simulation. In this paper, matching determines the study and data base. What is the error rate in the identification number in both files? Errors in deterministic match variables are more important than in probabilistic match variables. The paper does compare the finding of this study with those of other sources to demonstrate that the match was effective.

The question of what impact effective matching algorithms have on the confidentiality of person records was mentioned in the paper. The law provides specific statements on this subject. Some confidentiality problems were discussed in an earlier session. By linking data from several sources, individual records can be identified more easily. In the case of data collection at the Census Bureau, there is an additional concern. The Bureau is a passive collector of data. Cooperation of the respondent is of crucial importance in obtaining reliable information. As the public becomes aware of our ability to link records from several Governmental agencies, response rates to our questionnaires may decrease, become biased, and possibly inaccurate due to the fear of person-record identification. This is in spite of the potential to provide more beneficial information than would exist without the linked records.