DISCUSSION

Gilbert W. Beebe, National Cancer Institute

In my discussion I want to stress applications in Canada and in the U.S., the need for them, their limitations, and the very great advantage enjoyed by Canada vis-à-vis the U.S. Although I've depended heavily on record linkage in my own research, especially in my studies linking U.S. military service records with those of the Veterans Administration, I've had the advantage of unique service serial numbers or, in recent years, of Social Security numbers which are almost unique. And in Japan I've had the advantage of the family registration system that is tied to a fixed home address. I've not had direct experience with probabilistic matching algorithms that seem so necessary in the absence of unique identifying numbers, but have been able to rely on exact matches. Hence, I'll not comment on the methodologic aspects of the matching procedures in the paper but will talk about the epidemiologic significance of the program as a whole.

The need for empirical risk estimates based on human experience is, as Smith and Silins point out, an urgent, growing one. In regard to cancer risks alone, the need appears to be insatiable. We are asking more and more questions about the risks arising from our environment, our life styles, even our genes. And we are sensitive to increasingly small risks, which, when multiplied by our large populations, yield non-negligible numbers of possibly affected individuals, even of premature deaths. And, in the face of this growing need for information, in the U.S. we impose almost indiscriminate restraints upon access to individually identifiable records in government files because of abuses that arise, not in the course of medical research, but in other uses of these records. For example, medical records today play a large role in social, financial, and legal affairs directly impinging on the individual, and illegal or questionable practices may be employed to gain access to them.

In the U.S. we are developing an analogue to the Canadian Mortality Data Base, the National Death Index, now in a final testing phase and shortly to be in operation at the National Center for Health Statistics. John Patterson had a luncheon roundtable on this today. But the NDI begins, not with 1950 deaths, but with 1979 deaths, and whether it can be made retroactive even for five years is very much in doubt, especially in these times of federal economizing. The recent budget-slashing by the Administration and the Congress has severely restricted new NCHS programs and, if the NDI is to go forward in the short run, it will be on the basis of transfers of funds from other agencies with a stake in its success.

On a more technical note, the NDI will begin with an algorithm that will require an exact match on a name and a number, e.g., the Social Security Number and either the surname or the given name, with the name represented in either alpha or soundex format, or, lacking the SSN, the month and year of birth + both surname and given name. When the SSN is not available, this initial algorithm may identify an insufficient number of true matches in the user's file, if I interpret Smith and Silins' data correctly. You may recall that, in one study, they report that among definite links there was full agreement on first name in only 83 percent, and on year-of-birth in only 81 percent. The NDI may have to modify its matching procedures in the direction of those we have heard about today if it is to deal effectively with users' files that lack SSN. It might be instructive at this stage for the NCHS to ask Statistics Canada to test the intitial NDI algorithm against the GIRLS routines on one of their large files.

One consequence of the NDI exact match routine is that the consumer, presented with multiple "hits" per name, will have no convenient score with which to evaluate the list of candidates and will, in all likelihood, resort to subjective criteria in evaluating them.

Another difference between the NDI and the Canadian Mortality Data Base is that the NDI will not itself provide the consumer with cause of death; this he must obtain from the state where the death was registered. In a large file this can be an expensive operation. I think of Dorn's file of WW I veterans interrogated 25 years ago about their smoking history as well as their employment history, now being followed up by the NIH and the Medical Follow-up Agency of the National Academy of Sciences. That's a file of almost 300,000 men, most of whom are dead. To manually obtain and code the cause of death, when the information was at one time available in machinable form, is a task that gives one pause.

Smith and Silins speak of a new era in epidemiology opening up in Canada, and the list of studies they have underway is most impressive. We have, of course, in the U.S. Social Security files, a considerable capacity for ascertaining mortality over a period of 40 years. Recent work with the file suggests that it may be 90 percent or more complete in recent years, but it is probably much less complete for deaths that occurred, say, 25 years ago. And, since SSA has not been requiring that a death certificate be filed as proof of death, its files do not contain cause of death. Again, one must go back to the states and request cause of death once SSA files have indicated the fact of death. To end this thought on an even more discouraging note, the budget reconciliation bill just passed by the Congress has gutted the mortality reporting system of the SSA by virtually wiping out the routine burial allowance on which the system has depended. This makes us completely dependent on the NDI in future years; should that fail, we would again have to fall back on the 50-odd registration areas for death clearance of national rosters.

We have in the U.S., I would suppose, as much statistical information per capita as in Canada. Why, then, is it so hard to envision a U.S. record-linkage capacity on the order of what has just been described as fulfilling a mandate to Statistics Canada? I think the root cause is the fragmentation, some would say decentralization, of our federal statistical responsibilities and programs without an effective integrating

mechanism with legislative authority and funds to produce data serving needs that lie outside the missions of the individual statistical agencies. This is not to denigrate the work of the Office of Federal Statistical Policy and Standards and its predecessors that, since the Federal Statistics Board was established 50 years ago, have sought to prevent the collection of unnecessary information, to maximize the use of collected data, to standardize and upgrade statistical procedures, and the like. But none of them has ever had the authority, the funds, and the personnel to do what would seem to be needed, to do what Statistics Canada is doing today. Recently, the National Center for Health Statistics was given a legislative mandate to "develop a plan for the collection and coordination of statistical and epidemiological data on the effects of the environment on health." To the extent that the activation of such a plan will require large-scale linkage of the existing administrative and statistical files of other agencies, the Center will have its troubles.

Over the years there have been many surveys and commissions concerned with the federal statistical program. In the early years there were proposals for a central statistical bureau. In the mid-1960's, efforts to establish a National Data Service were rejected, for fear that it would lead to dossiers on individual corporations. Data banks had begun to have a bad name. But in the main the recommendations have been for a decentralized statistical system, with data-collection and analysis closely articulated to the mission of each agency. I thought we might have an acceptable compromise two years ago when I received for comment a draft bill entitled, "Confidentiality of Federal Statistical Records" under which the major statistics-producing agencies would have formed an "enclave" within which their files could have been linked for purely statistical purposes. Unfortunately, this bill never got beyond the Office of Management and Budget.

Bullish as I am about record linkage, I've had enough experience to know that it is merely one tool, and not a sufficient one for most purposes. Smith and Silins mention, but do not stress, that the statistical associations developed through the linkage of large data files are useful primarily as, in their words, "pointers to possible trouble spots, or to investigate selected suspected high-risk groups." In my experience, what goes on a large tape file of administrative and statistical data is likely to be pretty thin from the standpoint of the epidemiologist, especially in regard to quantitative measures of exposure and to variables one may wish to control in making comparisons. We badly need better input data on exposure, not just the names of occupations or even of chemicals used by occupational groups, but quantitative measures of exposure at the level of the individual worker, as we have, e.g., for ionizing radiation.

I would like to take exception to one inference I perceive in the paper, namely, that only through cohort studies can one reliably investigate late effects of low frequency and long latent period. I think most epidemiologists would agree that sometimes the case-control approach is more efficient, sometimes the cohort approach, and that both can have their problems. I doubt that the recently completed NCI study of bladder cancer and artificial sweeteners could have been done nearly as well with a cohort approach as it was with the case-control approach.

In their manuscript Smith and Silins remark that the combination of an escalating need for information on health risks and a scarcity of government funds made it necessary to exploit fully all available resources within Statistics Canada, minimizing cost and respondent burden. In the U.S. we are beginning to experience a scarcity of government funds and we certainly need far more information on health risks. Is it too much to hope that an economy-minded administration will perceive the economic desirability of large-scale linkage of administrative and statistical records? Or, am I grasping at straws? Will we have to ask Statistics Canada to do some of our studies for us?