## DISCUSSION

M.P. Mi, University of Hawaii

It is indeed exciting to learn that so much progress has been made in the development of new methodology and applications of record linkage in Canada, as reported by the speaker of this session. I would like to take this opportunity to present our own experience with record linkage in Hawaii.

We are developing a unique population data base in Hawaii for demographic cancer research. Sources of data used are: the 1942-43 Population Registration of all residents in the territory of Hawaii, all vital statistics records including marriages, livebirths, fetal deaths, deaths and divorces during a period of about 4 decades from 1942 to date, and the Hawaii Tumor Registry. The latest additions are the State Voter Registration and Driver's License Registration. We now have a total of approximately three million recorded events of individuals in our population. The integration of various data sources involves two major operations: (1) the bringing together of recorded events belonging to the same individual and all individuals belonging to the same family based on identifying information; and (2) the management of a large data base for information storage and retrieval. Problems and techniques related only to the first operation are to be discussed here.

As the Hawaii population consists of many racial groups, we have a number of different problems in record linkage. A few examples are given as follows. You may or may not be aware that oriental names are all phonetically translated into English for recording purposes. On the one hand, it is not difficult to find that two or more names have the same English translation. On the other hand, it is equally common that the same name may have more than one translated form. Other difficulties are that the first name and middle initials are often varied. An event on the child may be reported by the parents with an oriental first name, but, in subsequent documents, an English first name may be used with or without the oriental name as middle initials. Orientals, mainly Chinese, traditionally accept age by lunar year and regard the newborn as one-year old by counting the gestation period.

In preparing data files for linkage it is necessary to subdivide the file into small blocks by selected criterion such as surname in original or coded form. Detailed pairwise comparisons of records are only made between two blocks sharing certain characteristics, one from the master and the other from the search file. It has been found that the frequency distribution of outcomes in certain linkage items or variables may vary greatly from block to block, creating a correlation between the grouping criteria and the linkage items. When it occurs, the statistical weights derived from frequency distributions of outcomes based on the total file or a sample, as in the conventional approach, may be less effective in discriminating between true and false linkages.

Recently, we have introduced a new approach using the block-based operation concept. This method treats the two blocks of records between which detailed comparison is to be made as a unique environment for linkage. For each linkage item the relative frequency distribution of outcomes is calculated, and the product of corresponding frequencies in the two blocks gives a measure of chance match for a specific outcome. When two records are compared, the result of agreement for the specific outcome may be represented by the corresponding probability of chance match. Common outcomes have a higher probability of match by chance than rare ones. When two or more items agree, a compound probability can then be formed by multiplying the two probabilities of chance match. When all pairwise comparisons are completed for a search record, the pair having the smallest value of chance probability is considered the best. This approach is justified on the ground that a true linkage can never be proved without further investigation. The chance probabilities of all outcomes in an item can also be used to derive the expected probability of chance match for the item. If several items are selected for linkage, the product of these expected values which characterizes the linkage environment between the two blocks can be used as an acceptance level. A linkage for acceptance is suggested when a linked pair for a search record has a smaller probability of chance match than the acceptance level. This development has minimized subjective decision-making in establishing the threshold value for acceptance in a record linkage operation.

We have tested the new procedure quite extensively using simulated data in which frequency distribution of outcomes in each linkage item can be pre-specified and true linkages are precisely known. In addition, many situations can be experimentally investigated. We have demonstrated that in successive iterative cycles of linkage between two blocks, the acceptance level becomes more relaxed due to the removal of linked pairs. In one of the experimental runs, seven linkage items were used, each of which had a different frequency distribution of outcomes. The number of possible outcomes varied from 2 to 100. We started with 500 master records and 50 search records. In the first cycle the acceptance level was $10^{-14}$; and 21 pairs of records were accepted. After these linked records were removed, a second cycle was initiated by re-calculating all frequencies and probabilities between the remaining 479 master and 29 search records. The acceptance level was $10^{-12}$; 7 records were accepted. In each of the subsequent cycles, additional linked records were removed. It is clearly shown that with successive removal of linked pairs there is a significant loss of information in the two blocks resulting in a much higher probability of chance match. After seven cycles of iteration, a single search record remained. It was finally accepted from a

comparison with 451 master records at a level of $10^{-5}$ in the 8th cycle.

We have also used our linkage procedure on real data. For example we have linked the total population registration, 1942-3, and all deaths registered during a 38-year period (1942-79), for a study designed to assess occupational cancer risks through mortality experience. The linkage criteria used were surname, first name and year of birth. On most death records, age at death instead of date of birth was recorded so that year of birth was estimated for comparison. We first divided the population cohort by sex. The linkage between 243,176 male subjects in 1942 and 89,885 male deaths identified 39,113 deaths in the original cohort. The total CPU time for this linkage operation on a HP3000 computer was 28.2 hours, giving an average time of 2.6 seconds per linked record at an estimated cost of 5.2 cents. The multiples are sets in which one search record linked with more than one master record or vice versa, each pair giving the same value of chance probability. These were later checked manually, using additional information, if available, on both records. A true linkage was successfully identified in 51% of these multiple sets. No decision could be made for the remaining sets.

Because most female subjects used their husband's name after marriage, the female cohort was further divided into two groups: those married and all others. The second group, representing 105,769 single, widowed and divorced women was linked with all marriages registered during 1942-1979. A total of 53,930 women were identified. More than 99% of the 1,051 multiples were women married more than once during the 38-year period. After the updating of married names, the linkage with 51,533 female deaths identified 22,197 deaths in the 1942 original cohort. The time estimate per link was similar.

We then considered the reliability of these accepted linkages. We selected six additional data items for comparison between the two records of each linked pair. These items, not used as linkage criteria, were middle initials, senior/junior designation, birth month, birth day, race and birthplace. The item-by-item comparison was made by computer, if and only if an additional item was available in both records. The results are grouped in five categories as follows:

1. All available items up to sex agreed between two records
2. More than one-half agreed
3. One-half agreed
4. Less than one-half agreed
5. None agreed

The first category strongly suggests that these are true linkages. This category represented 81% of the linked records for male subjects and 75% among females. We may combine the second category with the first producing 94% and 92%. The fifth category argues strongly against true linkage. These records represented 0.8 and 1.2 percent in the two operations. If we include the fourth category, the possible false linkage would increase to 1.9% in the male sample and 3.5% in females. The middle category could be used to argue either for or against true linkage. We have just completed a manual check of records sampled from each of these categories. Among male subjects there was no false linkage in the first and second category. The frequency of apparent false linkage was less than 5% in the middle category and 10-20% in the last two categories. The frequency of false linkage was found to be very similar in females for these five categories.

Another immediate question is how many deaths have we failed to link. We selected all individuals aged 60 and over from the 1942-43 cohort because all would have died by now. There was a total of 15,567 individuals. The linkage procedure identified 9,568 individuals or 61% as linked to death records by surname, first name and estimated year of birth. An extensive manual search was made for the remaining 5,999 individuals or 39% of the aged. Expressed as percentage of the total 15,567 individuals, 3% had their death records misplaced in other blocks because of surname and first name switched in recording. Eight percent had serious errors in spelling of surname or first name. To date death records have not been found for the remaining 29% of these individuals. Most of these people were Orientals born in their own countries, who could have returned and died in their homeland.