Journal of Epidemiology and Community Health, 1988, **42**, 390–395

The classification of ethnic status using name information

ANDREW J COLDMAN, TERRY BRAUN, AND RICHARD P GALLAGHER

From the Cancer Control Agency of British Columbia, Vancouver, Canada

SUMMARY Methodology is developed to classify ethnic status by name using a simple probabilistic model. This method involves the consideration of four rules which may be used to classify individuals using three name components (first, middle and last names). In order to do this, conditional probabilities of ethnic status are estimated from a sample in which the ethnic status is known. Using a split sample technique the sensitivity and specificity of this methodology were examined in a data set of death registrations. Each of the classification rules performed well on the data from which they were constructed but were not as efficient when applied to another population. Nevertheless a model (linear), in which the sum of the conditional probabilities of each home component is used, achieved a sensitivity and specificity of 97% and 100% respectively in males and 89% and 100% in females.

Information on the occurrence of specific diseases in different ethnic and racial populations can provide valuable clues to their aetiology. Such information may be obtained in a number of ways, eg, from published data of incidence and mortality rates[1] in different countries. Even then, many geographically defined populations contain several ethnic groups and it is of some interest to determine their separate disease rates. Unfortunately in many cases the identification of ethnic groups within populations is hampered by a lack of information. For example, population or disease registries may not contain information on ethnic origin. In many situations it would be useful to be able to identify ethnic status using information which is routinely collected, such as the patient's name.

Previous investigators in England,[2] have used experienced observers to classify individuals into Asians/non-Asians using name information alone. This method was very successful but may be time consuming in situations where large numbers of individuals are involved. In large data sets computerised approaches are clearly of interest.

We wished to examine death rates from various diseases over the period 1950–1984, in individuals of Chinese ethnicity resident in British Columbia. Ethnic status was routinely reported on death certificates in British Columbia but this practice was discontinued in 1978. We will report here the methodology developed to classify individuals and the results of a test of this methodology on a group with known ethnic status.

Methods

We shall consider an individual to be in one of two ethnic groups, designated C and $\bar{C}$ which we will refer to as Chinese and non-Chinese respectively, although they can represent any two mutually exclusive groups. We shall assume that an individual's name, N, consists of three components, a first (N_1), a middle (N_2) and a last (N_3) name. We have available two sets of data containing the names of individuals; in one we know the ethnic status of each person (the referents) and in the other we do not and wish to classify them (the targets). If we view names as if they were random variables, then from elementary decision theory,[3] classification of individuals is best done using the value of the ratio of conditional probabilities:

$$\frac{P(N|C)}{P(N|\bar{C})} \gtrless k \qquad (1)$$

where k is a constant determined by the "cost" of misclassifications. If we assume that the distribution of names is the same (in each ethnic group) in both populations it is a straightforward matter to use the empiric density function of names in the referent population to estimate the required probabilities for the target population. However, two practical difficulties arise in attempting to do this: (1) many names are unique so that no matter what the size of the referent population a large number of names in the target population will not exist in the referent

population; and (2) because there are so many unique names, the data from the referent population are sparse, so that the resulting estimates of probabilities are unlikely to be accurate.

One technique for reducing both these problems is to consider individual name components separately since the number of unique components is much smaller than the number of possible combinations. Thus we consider the quantities

$$\begin{aligned} F_1(i) &= P_1(N_i|C), \\ G_1(i) &= P_1(N_i|\bar{C}), \end{aligned} \qquad i = 1,2,3, \tag{2}$$

where the subscript 1 refers to the distribution of the reference population. We wish to use the relationships in equation (2) to construct the ratio contained in equation (1). Unfortunately this may not be simply done and we instead suggest consideration of the ratio

$$\frac{F_1^*(N)}{G_1^*(N)} \tag{3}$$

where $F_1^*(N)$ and $G_1^*(N)$ are functions of $F_1(i)$ and $G_1(i)$ respectively. We considered four possible definitions of $F_1^*(N)$ as follows:

$$\begin{aligned} &\text{I} \quad F_1^*(N) = \prod_{i=1}^{3} F_1(i) && \text{— multiplicative} \\ &\text{II} \quad F_1^*(N) = \tfrac{1}{3} \sum_{i=1}^{3} F_1(i) && \text{— linear} \\ &\text{III} \quad F_1^*(N) = \max_{1 \leqq i \leqq 3} \{F_1(i)\} && \text{— maximum} \\ &\text{IV} \quad F_1^*(N) = F_1(3) && \text{— last name} \end{aligned} \tag{4}$$

where $G_1^*(N)$ are defined analogously with $G_1(i)$ replacing $F_1(i)$ in equation (4). The rules of equation (4) do not all have a formal probabilistic interpretation and are chosen arbitrarily. Many other choices can be imagined and these four rules are presented as examples.

Estimating the probabilities of equation (2) using the empiric distribution and replacing the ratio in equation (1) by that of equation (3) it only remains to select k. The choice of k will depend upon the particular application, since it affects the likelihood of false negatives and false positives. If the proportion of the ethnic group in the target population is known, then k may be chosen so that the proportion classified agrees with the true proportion. If there is no information on the true proportion in the ethnic group it is helpful to have some data dependent method for estimating it. This may be done as follows. For any name N we have

$$P_2(N) = P_2(N|C)P_2(C) + P_2(N|\bar{C})(1 - P_2(C))$$

where the subscript 2 refers to the target population. If, as before, we assume that the distribution of names within each group of the reference population is the same as that of the target population, then we have

$$P_2(C) = \frac{P_2(N) - P_1(N|\bar{C})}{P_1(N|C) - P_1(N|\bar{C})} \tag{5}$$

We may replace the single name, N, in equation (5) by a set of names, A say, and the relationship still holds. As before we would suggest replacing the probabilities in (5) by the rules in (4), so that

$$\widehat{P_2(C)} = \frac{H_2^*(A) - G_1^*(A)}{F_1^*(A) - G_1^*(A)} \tag{6}$$

where $F_1^*(A) = \sum_{N \varepsilon A} F_1^*(N)$ etc, and $H_2^*(A)$ is alculated by replacing $F_1(i)$ by $P_2(N_i)$ in equation (4).

A particularly simple choice for A is the set of names for which $G_1^*(N) = 0$. Use of rule (IV) maintains the probabilistic interpretation of (5) in (6). It should be emphasised that estimation of $P_2(C)$ using equation (6) is likely to be very approximate and should not be used unless it is necessary.

We have discussed to date the decision of allocating individuals to Chinese/non-Chinese groups. This may be generalised to include a group of "uncertain cases" if the intention was to identify a group of Chinese and a group of non-Chinese for comparison rather than classify everybody as Chinese or non-Chinese. If individual allocation was not important, but it was desired to calculate disease rates for example, then we could calculate rates in the Chinese population by directly using the estimated probabilities:

$$\begin{aligned} \text{rate} &= \frac{\sum_{\{N\}} 1_D(N) P_2(C|N)}{\sum_{\{N\}} P_2(C|N)} \\ &= \frac{\sum_{\{N\}} 1_D(N) P_1(N|C)/P_2(N)}{\sum_{\{N\}} P_1(N|C)/P_2(N)} \end{aligned} \tag{7}$$

where $1_D(N)$ is a variable which takes the value 1 if the individual with name N has disease and 0 otherwise, and summation is carried out over the target population. Once again it seems reasonable to replace the probabilities in equation (7) with the rules in equation (4) so that

$$\widehat{\text{rate}} = \frac{\sum_{\{N\}} 1_D(N) F_1^*(N)/H_2^*(N)}{\sum_{\{N\}} F_1^*(N)/H_2^*(N)}$$

where $H_2^*(N)$ is as given in equation (6). In cases

where equation (7) is appropriate, it is much easier than having to classify individuals since it does not require knowledge of the proportion Chinese or the specification of k and only requires the distribution of names of the Chinese from the referent population.

A rate for non-Chinese may be calculated by replacing C by $\bar{C}$ in equation (7). As when attempting to classify individuals, it is possible to define more than two groups so that individuals only contribute to the Chinese rate if $P_2(C|N) > c$ where c is some positive constant. By this method rates can be calculated in groups with high probability of being Chinese, etc.

Results of evaluation

A computerised file was available of all death records in British Columbia for the period 1950–1973. Each record included the full name, sex and ethnic status of the subject. The data set was split into two sets by year of recorded death: 1950–1964 and 1965–1973. These two groups were chosen for convenience as the record format was different in these two periods. The data for 1950–1964 were used as the referent population and those for 1965–73 as the target population. We shall present results for distinguishing Chinese from non-Chinese.

Table 1 *Composition of Chinese and non-Chinese populations, 1950–1973*

	Male	*Female*
Data for 1950–1964		
Non-Chinese		
Total	120 935	76 989
Number of unique last names	32 767	22 288
unique first names	6404	4355
unique middle names	8762	5421
Chinese		
Total	5016	424
Number of unique last names	415	129
unique first names	767	225
unique middle names	596	136
Data for 1965–1973		
Non-Chinese		
Total	90 430	61 994
Number of unique last names	28 952	20 800
unique first names	5533	4268
unique middle names	7495	5111
Chinese		
Total	2667	538
Number of unique last names	296	133
unique first names	625	284
unique middle names	496	217

Descriptive information on the composition of the two populations is contained in table 1. For the total period, 3·5% of male deaths and 0·7% of female deaths registered were Chinese. In both periods there were more male than female deaths and substantially more Chinese male deaths than Chinese female deaths. In the following analysis we shall present the results separately for males and females.

The empiric frequencies of equation (2) were calculated using the referent population. If any name component was absent, that name was assigned the null character and treated as a legitimate name. As a method of determining the maximum accuracy which could be expected from the four rules of equation (4), they were calculated for the names in the referent population. The value of k, in equation (1), was then selected so that the four rules classified (approximately) the correct proportion of individuals as Chinese. The results of this classification are given in table 2. Examination of this table shows that the sensitivity of each index was very high, exceeding 90% in all cases except where the surname alone was used in women. In all cases the specificity was approximately 100%, which results from the requirement that the percentage classified as Chinese should equal the true percentage of Chinese and the small proportion of Chinese in the data. It is likely that intermarriage between the two groups causes the sensitivity of each index to be lower for females than for males. This is particularly marked for rule IV when only the last name is used. Unfortunately there were insufficient numbers of single Chinese women in the reference and target populations for separate analysis to see if their classification rates were more similar to those of the males.

In attempting to use the techniques previously described on the target population one is immediately faced with two problems: (1) the proportion of Chinese is (presumed) unknown; and (2) the target population contains many names not present in the referent population. Table 3 indicates the diversity of names present in the target population, with 19% of surnames not present in the referent population. Using the method discussed in the development of equation (6) the proportion of Chinese in the target population was estimated. To do this, attention was restricted to those with a surname present in the referent population and all considerations were based using surname alone; ie, the fourth rule in equation (4) was used in equation (6).

In examining the efficiency of the various rules in classifying individuals as Chinese/non-Chinese it was necessary to restrict attention to those whose names were available in the referent population; such names will be referred to as recognised names. It was therefore necessary to estimate the proportion of Chinese in those with recognised names, which was done by assuming this to be the same as the proportion estimated for the total population. Restricting attention to those whose last name was recognised (75 717 males and 49 682 females) resulted in estimates

Table 2 *Results of classification using the four rules on referent population*

*Rule**	*Classification by rule*	*True classification* C	$\bar{C}$	*Sensitivity (%)*	*Specificity (%)*
Males					
I	C	4985	32	99	100
	$\bar{C}$	31	120 903		
II	C	4958	63	99	100
	$\bar{C}$	58	120 872		
III	C	4931	86	98	100
	$\bar{C}$	85	120 849		
IV	C	4742	293	95	100
	$\bar{C}$	274	120 642		
Females					
I	C	405	18	96	100
	$\bar{C}$	19	76 971		
II	C	388	31	92	100
	$\bar{C}$	36	76 958		
III	C	387	30	91	100
	$\bar{C}$	37	76 959		
IV	C	333	39	79	100
	$\bar{C}$	91	76 950		

* For definition see equation (4) in text

Table 3 *Number of individuals from target population whose name components were present in the referent population*

	Number of individuals in target population							
	Chinese				*Non-Chinese*			
	M	*(%)*	*F*	*(%)*	*M*	*(%)*	*F*	*(%)*
All names found in referent	2295	(86)	235	(44)	67 973	(75)	45 112	(73)
Last name but not all names found in referent	293	(11)	228	(42)	5156	(6)	4107	(7)
Last name not found in referent	79	(3)	75	(14)	17 307	(19)	12 775	(21)
Total	2667	(100)	538	(100)	90 430	(100)	61 994	(100)

of the number of Chinese as 2103 males and 419 females. Using these estimated numbers the population was classified, using the techniques previously described, so that the number classified equalled (or was as close as possible to) the estimated numbers. In cases where a name component was not recognised, this component was ignored in the calculation of each rule. The results of the classification are given in table 4, which shows that the first three indices were almost equally good at distinguishing between Chinese and non-Chinese.

Evaluation of the relative qualities of each rule is obscured by the error in the estimated numbers of Chinese, which tends to reduce the efficiency of each classification. The analogous results of classification using the true proportion of Chinese are given in table 5. Use of the correct proportion of Chinese is seen to bring about a considerable increase in the sensitivities of each index, with a smaller decrease in their predictive value positives. Examination of the results of table 5 shows that correct classification rates for women are considerably lower than those for men. This was also found in table 2, indicating that women are more difficult to classify using the methods described. Judging from the results in table 5 the additive index (II) would appear to be the best all around index for these data.

Discussion

The example shows that the method is quite successful at classifying names into Chinese and non-Chinese. The overall accuracy of the method is strongly governed by estimates of the proportion of Chinese which exist in the population. We presented one method for estimating this quantity and no doubt other methods, which are possibly more accurate, may be devised. With an accurate estimate of the proportion Chinese, comparison of tables 2 and 5 shows that the additive index performs nearly as well on new data as on the data from which it was derived.

Table 4 *Classification for four rules in target population using estimated number of Chinese*

Sex	*Rule**	*Classification by rule*	*True classification* C	$\bar{C}$	*Sensitivity (%)*	*Specificity (%)*	*Predicted value pos. (%)*
Males	I	C	2333	24	90	100	99
		$\bar{C}$	255	73 105			
	II	C	2091	13	81	100	99
		$\bar{C}$	497	73 116			
	III	C	2076	50	80	100	98
		$\bar{C}$	512	73 079			
	IV	C	2075	139	80	100	94
		$\bar{C}$	513	72 990			
Total			2588	73 129			
Females	I	C	341	25	74	100	93
		$\bar{C}$	122	49 194			
	II	C	389	22	84	100	95
		$\bar{C}$	74	49 197			
	III	C	376	53	81	100	88
		$\bar{C}$	87	49 166			
	IV	C	303	50	65	100	86
		$\bar{C}$	160	49 169			
Total			463	49 219			

* For definition see equation (4) in text

Table 5 *Classification for four rules in target population using true number of Chinese*

Sex	*Rule**	*Classification by rule*	*True classification* C	$\bar{C}$	*Sensitivity (%)*	*Specificity (%)*	*Predicted value pos. (%)*
Males	I	C	2467	131	95	100	95
		$\bar{C}$	121	72 998			
	II	C	2507	96	97	100	96
		$\bar{C}$	81	73 033			
	III	C	2403	179	93	100	93
		$\bar{C}$	185	72 950			
	IV	C	2364	213	91	100	92
		$\bar{C}$	224	72 916			
Total			2588	73 129			
Females	I	C	341	25	74	100	93
		$\bar{C}$	122	49 194			
	II	C	414	62	89	100	87
		$\bar{C}$	49	49 157			
	III	C	378	52	82	100	88
		$\bar{C}$	85	49 167			
	IV	C	369	122	80	100	75
		$\bar{C}$	94	49 097			
Total			463	49 219			

* For definition of rules see equation (4) in text

In the analysis of the referent population the multiplicative model proved the best; however it was not so good in the target population. This results partially from the fact that if any name component was not held by a Chinese in the referent population, then $P_i^*(N) = O$ and the individual will be classified as non-Chinese, even if the other name components indicate that person is likely to be Chinese. It is well known that sparse data do result in these types of problems. One possible solution is to use empirical Bayes estimates[4] for the marginal distributions of the name components, which would have the effect of centring estimates. Clearly many other schemes for combining the name components into a single rule can be considered and the ones we have presented are only examples. Nevertheless with names like Lee and King which commonly appear as both Chinese and non-Chinese last names, it would seem that substantial improvement over the additive rule is unlikely.

The greatest difficulty with this technique is the handling of names, particularly last names, for which there are no data in the referent file. We could not attempt to classify 19·4% of cases whose surname was not recognised. If it is critical that all individuals be classified, then it will be necessary to examine these cases manually. In these circumstances the technique we have presented will be labour saving but will only provide part of the required solution. If, on the other hand, it is only necessary to classify a Chinese group in the data and not everybody then this technique may prove sufficient. In other situations, such as a cohort study, one could use equations like (7) to identify the disease experience among Chinese and not worry about classifying individuals. Such a calculation is particularly simple since equation (7) requires only knowledge of the marginal distribution of the name components among the Chinese in the referent population and does not require knowledge of the proportion which are Chinese in the target population. In this case individuals with unrecognised names would be ignored although care must be taken to avoid possible biases.

It is not possible to predict with confidence whether the methodology we have described will distinguish between groups in other regions. The Chinese represent an interesting example since their names are quite different from those of Europeans and their incidence of disease is of considerable interest. Unfortunately the major difference in their names makes detailed analysis unnecessary since the untrained observer can usually classify most individuals correctly. British Columbia does not provide an ideal situation for using these techniques because of the highly heterogeneous nature of the non-Chinese population. The rapid growth of population, in recent years by migration from other countries, has led to a very heterogeneous name base. it would be of interest to test these techniques in other countries where the name base has been more stable.

The techniques we have described rely on having a considerable amount of background information (proportion Chinese in target population, distribution of names in referent population, etc.) which may not be available in many practical situations. In cases where it is desired only to estimate rates, the distribution of disease in a Chinese referent population is all that is required (making the same assumptions), but even this may not be available. In situations where there is little or no information and expert inspection of each name is not practical it may be feasible to compile registers of ethnic names and use these to sort the large data sets into more homogeneous subgroups which may then be selectively examined further. Such registers can be compiled by experts or by computer analysis of data sets with individuals of known ethnic status, or by a combination of these methods. They can be tested in a way similar to that used to test the techniques described in this paper.

In conclusion, it does seem to be possible to classify all individuals as Chinese or non-Chinese using the techniques described. These techniques are likely to be useful for the analysis of medium to large size target data bases. However, the large size of the reference data bases required limits their general application. Furthermore the likelihood that substantial numbers of individuals will be unclassifiable using these techniques alone suggests that they will be most useful as part of a comprehensive classification system which will involve some element of manual resolution.[2]

AJC was supported by a grant from the British Columbia Health Foundation.

Address for correspondence and reprints: Dr A J Coldman, Cancer Control Agency of British Columbia, 600 W 10th Avenue, Vancouver BC V5Z 4E6, Canada.

References:

[1] Segi M, Kurihara M and Matsuyama T. *Cancer mortality for selected sites in 24 countries*. Sendai, Japan: Dept. of Public Health, Tohuku School of Medicine, 1969.
[2] Nicoll A, Basett K, Vlijaszek SJ. What's in a name? Accuracy of using surnames and forenames in ascribing Asia ethnic identity in English populations. *J Epidemiol Community Health* 1986; **40:** 364–8.
[3] Ferguson T. *Mathematical statistics: A decision theoretic approach*. New York: Academic Press, 1967: 201.
[4] Barnett V. *Comparative statistical inference*. 2nd ed. New York: John Wiley and Sons, 1982: 217.

Accepted for publication July 1988