Journal of Public Health Medicine

Vol. 23, No. 4, pp. 278–285
Printed in Great Britain

Development and validation of a computerized South Asian Names and Group Recognition Algorithm (SANGRA) for use in British health-related studies

Kiran Nanchahal, Punam Mangtani, Mark Alston and Isabel dos Santos Silva

Abstract

Background Studies on ethnic variations in health have played an important role in aetiological and health services research. Most routine datasets, however, do not include information on ethnicity. South Asians, one of the largest minority ethnic groups in Britain, have distinctive names that also allow differentiation of the main sub-groups with their important differences in health-related exposures and disease risks.

Methods A computerized name recognition algorithm (SANGRA) was developed incorporating directories of South Asian first names and surnames together with their religious and linguistic origin. SANGRA was validated using health-related data with self-ascribed information on ethnicity.

Results SANGRA was successful in recognizing South Asian origin in reference datasets, with sensitivity of 89–96 per cent, specificity of 94–98 per cent, positive predictive value (PPV) of 80–89 per cent and negative predictive value (NPV) of 98–99 per cent. Religious origin was correctly assigned in the majority of cases: sensitivity, specificity and PPV were 94 per cent, 91 per cent and 90 per cent for Hindus; 90 per cent, 99 per cent and 98 per cent for Muslims; and 76 per cent, 99 per cent and 94 per cent for Sikhs. SANGRA correctly identified 76 per cent Gujerati and 70 per cent Punjabi names, although only 62 per cent of Gujerati names were sufficiently distinct to be allocated to the Gujerati-only category and only 53 per cent Punjabi names were allocated to the Punjabi-only category. However, specificity and PPV were high for both languages (respectively 97 per cent and 93 per cent for Gujerati, and 99 per cent and 97 per cent for Punjabi).

Conclusions SANGRA provides a practical and valid method of ascertaining South Asian origin by name and, to a lesser degree of accuracy, of differentiating between the main religious and linguistic subgroups living in Britain. This algorithm will be useful in health-related studies where information on self-ascribed ethnicity is not available or is of a limited nature.

Keywords: South Asian, ethnicity, computer algorithm

Introduction

Studies on ethnic variations in health have played an important role in aetiological and health services research. Changes in disease risk in populations subsequent to their migration to areas with markedly different environmental exposures have been widely used to infer the relative importance of environmental factors and inherited predisposition in disease aetiology, and to identify the age at which disease risk is set.[1–9] Furthermore, the ability to monitor ethnic differentials in health and health service use allows effective planning and evaluation of health care services.[9–17]

South Asians (i.e. persons whose families originated from India, Pakistan, Bangladesh or Sri Lanka, irrespective of the individual's place of birth) are one of the largest minority ethnic groups in Britain, representing 2.7 per cent (almost 1.5 million) of the total population.[18] Aetiological and health services related studies in South Asians have, however, been hampered by the fact that most routine, nationally representative data do not include information on ethnicity. Some recent developments include the collection of data on ethnicity in the 1991 Census and, since April 1995, for all NHS hospital admissions. However, ethnicity is not recorded in birth registrations, death certificates, general practitioner patient lists, and most disease registers. Information on country of birth is collected by some routine data collection systems and has had to be used in many studies as a marker for ethnic origin.[9,19] It is well known,

Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT.

Kiran Nanchahal, Lecturer in Medical Statistics

Punam Mangtani, Clinical Lecturer in Epidemiology

Mark Alston, Computer Officer

Isabel dos Santos Silva, Clinical Senior Lecturer in Epidemiology

Address correspondence to Kiran Nanchahal.
E-mail: kiran.nanchahal@lshtm.ac.uk

however, that country of birth does not accurately reflect ethnicity. For example, only 44 per cent of the South Asians living in Britain in 1991 were born in the Indian subcontinent. The proportion of South Asians born in Britain is similar, although higher at young ages (e.g. 68 per cent at ages under 30 years).[20] Conversely, 15 per cent of those born in South Asia were white, 88 per cent of whom were over the age of 50.[18] Use of country of birth as a marker of ethnicity will thus lead to substantial misclassification and to dilution of true differences in disease risk between South Asians and Europeans.[19]

South Asian names are distinctive and their use has long been recognized as an alternative approach to ascertaining ethnicity.[21–23] Moreover, inspection of names allows identification of the main South Asian population groups living in Britain because the religious and regional origin of these communities are usually reflected in the choice of first and second names, and in the surnames.[24] The largest South Asian communities living in Britain are Sikhs from Punjab, Hindus from Gujerat and Muslims from Pakistan or Bangladesh.[25] Identification of these communities in health-related datasets is important because of differences between them in socio-economic status, health-related exposures or behaviours, and disease risks.[3,26] For instance, first-generation Gujerati Hindus are usually vegetarian whereas Sikhs do not usually eat beef and Muslims do not usually consume pork or alcohol.

Visual inspection of names by people familiar with South Asian names is, however, subjective and time-consuming, particularly when dealing with large datasets. In this paper, we describe the development and validation of a computerized algorithm (SANGRA) that identifies names of South Asian ethnic origin and also classifies them according to their religious and linguistic origin.

Methods

Compilation of directories of South Asian first names and surnames

A directory of surnames was compiled from a number of different sources, including lists of names provided by South Asian voluntary organizations,[27] the Office for National Statistics (Population Statistics Division, unpublished data), and researchers who had conducted studies among South Asian communities. These lists were complemented by the use of telephone directories, which were particularly useful as sources of variation in spelling caused by transliteration of names into English. A directory of first names was compiled from similar sources, with additional first names extracted from books of South Asian baby names.[28–31]

The two directories of names were reviewed by a panel comprising members of voluntary organizations representing the various South Asian communities living in Britain, and members of the research team who were of South Asian origin. The panel was asked to confirm whether the names were South Asian and, if so, to classify them according to their religious and linguistic origin. The revised directories of names including the religious and language categories were used to develop a computer algorithm using Microsoft Visual Basic. The South Asian Names and Group Recognition Algorithm (SANGRA) identifies South Asian subjects in a dataset by matching their names to the names in the directories and creating four new variables. The first variable indicates whether or not the subject was recognized by the program as being South Asian. The second and third variables give the religious and language categories, and the fourth indicates whether the subject was identified as being South Asian on the basis of both first name and surname, or on the basis of first name only, surname only, or middle name only.

SANGRA combines the information derived from all available names to provide the final religion and language classification for each subject. If the first name ends in -bhai, -bai or -ben (and is not, for instance, Rueben or Ben) SANGRA assigns final religion to 'Hindu' and language to 'Gujerati', regardless of the religious and language categories assigned to the surname. Similarly, if the middle name is Singh or Kaur SANGRA automatically classifies religion as 'Sikh' and language as 'Punjabi', whereas if the middle name is Bibi, religion and language are assigned as 'Muslim' and 'Bengali, Gujerati, Punjabi or Urdu', respectively. If none of the above applies and there is a conflict between the categories assigned to first name and surname, SANGRA selects as final the categories assigned to first name. The choice of first name is more likely to reflect a family's identification with a particular linguistic and/or religious origin than the surname. For instance, Ismaili Muslims tend to have Muslim first names and Hindu surnames, with the religious origin of their first names reflecting more accurately their lifestyle and health-related behaviours.

Validation of SANGRA

Validation of SANGRA was carried out using data that included information on self-ascribed ethnicity from interviews or from hospital records.

To validate South Asian ethnicity, London and Midlands hospital in-patient admissions data from mid- to late 1990s were used. Self-ascribed ethnicity in these data had been recorded and categorized according to the 1991 Census classification. In this classification South Asians are identified according to their country of origin as 'Indian', 'Pakistani' and 'Bangladeshi'. For the purposes of this validation, these categories were combined into a single 'South Asian' group. The admissions were from catchment areas that covered West and North West London, where many Gujeratis and Punjabis live, East London, where a significant proportion of Pakistanis and Bangladeshis reside,[18] and the Midlands, where a greater proportion are from other, often poorer, regions of Pakistan.[32] Multiple entries and records with missing ethnicity or missing first names for newborns were not included. The proportion of South Asians was much higher in the London (20 per cent) than in the Midlands (9.7 per cent) datasets. The composition of the South Asians was also different, with London having a higher proportion of Bangladeshis (49 per cent versus 7 per cent) and a smaller proportion of Pakistanis (9 per cent versus 28 per cent). In addition, data were available from an obstetric research survey conducted in the early 1990s in London. A total of 293 women were interviewed for this study, 27 per cent of whom were South Asian (mainly Bangladeshi and Pakistani).

To validate the ability of SANGRA to ascribe religious and linguistic origin, data from a population-based dietary study conducted among 731 South Asian women resident in Greater London (76 per cent), West Midlands (19 per cent) and Glasgow (5 per cent) were used. Detailed interview-based information was collected on the participant's religious and linguistic origin.

The areas covered by these datasets include varying proportions of non-South Asian ethnic groups whose names are also of Muslim origin (e.g. people of Northern African, Arab, Iranian, Turkish and Eastern European origin) and who could, therefore, be potentially classified as South Asians based on an analysis of names.

Results

Directories of South Asian first names and surnames

The directories of South Asian first names and surnames included 9917 and 9422 names, respectively. Additional information included different aspects of ethnicity: ethnic origin (South Asian); religion (Buddhist/Christian/Hindu/Muslim/Sikh); and language (Bengali/Gujerati/Hindi/Punjabi/Sindhi/Sinhalese/Tamil/Urdu). Some names are associated with more than one religion and/or language. Names of Muslim origin cannot generally be differentiated by region of origin. Thus, for South Asian Muslims, their language roots can be Urdu, Punjabi (spoken in Pakistan and India), Gujerati (spoken in the western region of India) or Bengali (spoken in West Bengal and Bangladesh). Muslim names were therefore assigned by SANGRA to a mixed language group defined as 'Bengali, Gujerati, Punjabi or Urdu'. The minor languages were included within the appropriate major linguistic category; for example, Sylheti was included in the Bengali and Kutchi in the Gujerati group.

Validation of SANGRA

The results of the validation of South Asian origin assigned by SANGRA are given in Table 1. Both sensitivity and specificity were very high for the various reference datasets, ranging from 89 per cent to 96 per cent and from 94 per cent to 98 per cent, respectively.

Ninety-one per cent of the South Asian in-patients in London were recognized as such by SANGRA (Table 1). Specificity was high, with only 1570 (5.7 per cent) of the 27 384 non-South Asian names in the hospital data being classified as South Asians by SANGRA. Half of these latter names had been allocated in the hospital dataset to the 'Other' category and 14 per cent to the 'Black-African' category. Thus, the true number of South Asians in this dataset was slightly overestimated, corresponding to a relative increase of 13 per cent [= (7879/6959) × 100]. The positive and negative predictive values for South Asian ethnicity were 80 per cent and 98 per cent, respectively. Eighty-nine per cent of South Asian patients in the Midlands' admissions data were identified as such by SANGRA. Only 1969 (2.3 per cent) of the names recorded as non-South Asians in this dataset were classified as South Asians by the program. Therefore, the true number of South Asians was overestimated by 10 per cent. The positive and negative predictive values were 81 per cent and 99 per cent, respectively. Ninety-five per cent of the South Asian women in the obstetric dataset were recognized as such by SANGRA, with only 10 (4.8 per cent) of the non-South Asians being classified by the program as South Asians. The positive and negative predictive values were 89 per cent and 99 per cent, respectively. Ninety-six per cent of the South Asian women in the dietary study were identified by SANGRA as such.

Visual inspection of the names recorded as non-South Asian in the reference datasets but classified as South Asian by SANGRA revealed that many were typical South Asian names. The majority of them probably represent East African Asians or Indo-Caribbeans, European women married to South Asian men and people of mixed heritage, whereas some may be Muslims who were probably not of South Asian origin. A number of names recorded as South Asians in the hospital datasets but not recognized as such by SANGRA are in the dictionary files but with a different spelling (e.g. Harbujan, Sathya), some are

Table 1 Validation of SANGRA against self-ascribed ethnicity as given in the reference datasets

	Number of subjects							
		South Asian					Predictive values (%)	
Reference datasets	All	Reference	SANGRA	Reference & SANGRA	Sensitivity (%)	Specificity (%)	Positive	Negative
London in-patients	34343	6959	7879	6309	90.7	94.3	80.1	97.5
Midlands in-patients	95596	9262	10239	8270	89.3	97.7	80.8	98.8
London obstetric survey	293	83	90	80	96.4	95.2	88.9	98.5
Dietary study	761	761	731	731	96.1	–	–	–

common to both South Asians and the white population (e.g. Anita, Gill), whereas others had both first name and surname of European origin.

The ability of SANGRA to ascertain the religious origin of the South Asian women taking part in the dietary study is given in Table 2. Ninety-four per cent of Hindus were recognized as such by SANGRA; this percentage increasing to 98 per cent if allocation to mixed categories containing Hindu (i.e. 'Hindu or Muslim', 'Hindu or Sikh' and 'Hindu or Christian') was also regarded as correct. Specificity and positive predictive value for the Hindu-only category were 91 per cent and 90 per cent, respectively. Sensitivity for the Muslim-only category was slightly lower at 90 per cent (94 per cent if allocation to mixed categories was included), but with higher specificity (99 per cent) and positive predictive value (98 per cent). SANGRA correctly classified 76 per cent of the Sikh subjects, with an additional 8.5 per cent being classified as 'Hindu or Sikh'. Specificity (99 per cent) and positive predictive value (94 per cent) were, however, both very high for the Sikh-only category.

SANGRA correctly identified 76 per cent Gujeratis in the dietary study, although only 62 per cent could be allocated to the Gujerati-only category, with the rest being common to other northern Indian linguistic groups (Table 3). Specificity (97 per cent) and positive predictive value (93 per cent) for the Gujerati-only category were high. SANGRA correctly identified 70 per cent Punjabis, although only 53 per cent could be allocated to the Punjabi-only category. The proportion recognizable by the program was 63 per cent if names assigned to the 'Hindi or Punjabi' category were also included. The specificity (99 per cent) and the positive predictive value (97 per cent) for the Punjabi-only category were again high. There were relatively small numbers of Hindi-only and Bengali-only speakers in this dataset. Muslims, mainly Urdu speakers, have *a priori* been assigned in SANGRA to a mixed category of 'Bengali, Gujerati, Punjabi or Urdu'.

Discussion

The results presented here indicate that SANGRA provides a quick, user-friendly valid method to recognize people of South Asian ethnic origin in health-related data and of categorizing them by religious and linguistic origin. The sensitivity for ascertaining South Asian ethnicity ranged from 89 per cent to 96 per cent. The specificity of SANGRA was also high – over 94 per cent for all reference datasets. This was true even in London where the number of non-South Asian Muslims is known to be greatest. A high specificity is important, as a large proportion of people in routinely collected datasets in Britain will be of non-South Asian origin. Thus, small reductions in specificity would lead to substantial numbers of false positives and overestimation of the true numbers of South Asians. In the present study, the true proportion of South Asians in the various reference datasets ranged from 10 per cent to 27 per cent, and these

Table 2 Validation of SANGRA against self-assigned religion in the dietary study data

		Religious origin assigned by SANGRA						All groups including specified religion		
Self-assigned religious origin (reference)	Total number (%)	Hindu	Muslim	Sikh	Hindu or Christian	Hindu or Muslim	Hindu or Sikh	Sensitivity	Specificity	Positive predictive value
Total number	731	349	218	124	3	19	18			
Sensitivity										
Hindu*	336 (100)	94			0.60	2.4	1.2	98	85	85
Muslim†	238 (100)		90		0.42	4.2	0.42	94	97	94
Sikh	153 (100)			76		0.65	8.5	84	98	91
Other‡	4 (100)									
Specificity		91	99	99						
Positive predictive value		90	98	94						

Figures are percentages unless stated otherwise.
*Includes 11 Jain.
†Includes five Ismaili Muslims.
‡One Buddhist, two Christian, one 'none'.

Table 3 Validation of SANGRA against self-assigned linguistic origin in the dietary study

Self-assigned linguistic origin (reference)	**Linguistic origin assigned by SANGRA**		
Gujerati (n = 264)	Gujerati-only (n = 176)	Gujerati-only/Gujerati, Hindi or Punjabi (n = 245)	
Sensitivity	62	76	
Specificity	97	91	
Positive predictive value	93	82	
Negative predictive value	82	87	
Punjabi (n = 262)	Punjabi-only (n = 145)	Punjabi-only/Hindi or Punjabi (n = 183)	Punjabi-only/Hindi or Punjabi/ Gujerati, Hindi or Punjabi (n = 252)
Sensitivity	53	63	70
Specificity	99	96	85
Positive predictive value	97	90	72
Negative predictive value	79	82	83
Urdu (n = 124)	Bengali, Gujerati, Punjabi, Urdu (n = 214)		
Sensitivity	94		
Specificity	84		
Positive predictive value	54		
Negative predictive value	98		

Figures are percentages unless otherwise stated.

were slightly overestimated by SANGRA (relative increases of around 10 per cent).

The sensitivity, specificity, and predictive values reported here depend, however, on the quality of the ethnicity information recorded in the reference datasets. Differences in the accuracy of this information may explain the slightly lower sensitivity achieved by SANGRA for hospital-in-patient compared with research-based data. The two research-based datasets were likely to be accurate as the interviewers were trained to collect detailed information on ethnicity using a standardized approach. The quality of the hospital in-patient datasets is probably lower, given the lack of completeness and lower accuracy of most routine hospital data collection systems. The performance of SANGRA would also have been affected if subjects who regarded themselves as being 'East African Asians or Indo-Caribbean' or, simply, from the 'Indian subcontinent' were allocated to the 'Other' category in the hospital datasets. Unlike the 1991 Census classification, which allowed separation of this group into the category 'Other Group – Asian', this is not the case in the hospital system.[33] Visual inspection of the names allocated to the 'Other' category in the hospital datasets who were identified as South Asians by SANGRA showed that the majority of them were names common among South Asians. The values for specificity reported here for hospital data are, therefore, likely to be underestimates.

Most South Asian names are distinctively associated with a specific religious group although some names are common to several South Asian communities; for example, Kamal, Gulab or Malik are common among both Muslims and Hindus. Despite this, SANGRA was able to correctly identify the majority of people in each of the three major South Asian religious groups in Britain: Hindus, Muslims and Sikhs. The ability of SANGRA to correctly ascertain linguistic origin, as a marker of region of origin, was, however, more problematic, as most names are common to a number of linguistic groups. Only 62 per cent of Gujerati and 53 per cent of Punjabi names were sufficiently distinct to be allocated to a Gujerati-only and Punjabi-only category, respectively, but the specificity and positive predictive values in both cases were high. Sensitivity could be improved by including people who could only be allocated to the probable categories (for example, for Gujeratis the mixed category 'Gujerati, Hindi, or Punjabi'). In addition, categorization by both religious and linguistic origins will, for example, be able to differentiate Sikh Punjabis from Hindu Punjabis. Thus, although SANGRA is unable to identify all the Gujeratis and Punjabis in a given population, it provides a useful tool to identify unbiased samples (for instance, for cross-sectional and case–control studies) in order to calculate population-based rates from these South Asian communities in Britain. The reference datasets used to validate South Asian ethnicity included both male and female names but the reference dataset used to validate the religious and linguistic origin included only female names, as we did not have access to similar data on men. Thus, further studies are needed to establish whether these latter results can be generalized to male names.

Visual inspection of names as a way of identifying subjects of South Asian ethnicity has been used before[21] but the validation of this method was based on a small study conducted in Bradford. In this study, the sensitivity and specificity of this approach were near 100 per cent relative to self-ascribed ethnicity, but the population examined was well known to the panel inspecting the names and there were few non-South Asian Muslims.[21] Another computerized method (Nam Pehchan) has been developed, incorporating a smaller directory of names than SANGRA and relying on stem as well as full matching of names.[22,34] Validation of this program against self-ascribed ethnicity showed that it performed well (with sensitivity nearly 100 per cent) for data from Bradford, where the directory of names originated, but less so (sensitivity of 61 per cent) when validated on a national dataset.[22] Nam Pehchan has also been validated using the opinion of a panel of experts as the reference;[23] this indicated a good performance in populations from West Midlands and Yorkshire but not in populations in South East England, Leicester and Nottingham.

Strengths and limitations

The usefulness of SANGRA depends on its ability to produce accurate and consistent results across the various South Asian communities living in Britain. For this reason, we selected reference datasets from areas in Britain with the highest proportion of people with origins in the Indian subcontinent and including South Asian communities with different religious and linguistic origins.[18] South Asians living in Greater London and West Midlands account for 58 per cent of Indians, 37 per cent of Pakistanis, and 64 per cent of Bangladeshis living in Britain.[32] The results presented here are encouraging and indicate that SANGRA will be able to produce valid results across Britain, although further studies are needed to confirm this.

There are limitations to the use of name analysis as a sensitive and specific method of identifying South Asians. First, transliteration of South Asian names into English often results in alternative spellings, for example, Bhavana or Bhavna, Rafiq or Rafique, and Choudhury, which has a number of different spellings. We tried to overcome this problem by including different spellings of each name in the directories. Attempts were made to incorporate Soundex into SANGRA, and to modify the program so that it would identify South Asian names on the basis of both full and stem matches. These were discarded because they led to the identification of an unacceptable number of false-positives with a consequent decline in the positive predictive value. Further names and spellings identified in the present validation will be added to the directories, which should lead to achievement of higher sensitivity. Second, some names are common to both South Asians and Europeans, for example, Rita, Ayesha, Sonia, Sheila, and the surname Gill. These names were deleted from the directories to maintain high specificity. Third, South Asians with Anglicized or Christian names (common in southern Indian states such as Kerala) and those originating from Goa (who have names of Portuguese origin) will not be recognized by SANGRA. The number of people from these groups residing in Britain is very small and, hence, the resulting misclassification is likely to be negligible. Fourth, South Asian Muslim names do not allow distinction by country of origin, as Muslims may originate in India, Pakistan or Bangladesh. In addition, many South Asian Muslim names are common to other population groups, for example, North Africans, Arabs, Iranians, Turkish and Eastern Europeans. Thus, people from these groups may be misclassified by SANGRA as South Asians. This is a problem only in areas where there are significant numbers of people from these non-South Asian communities. As an indication of the extent of this potential misclassification we have incorporated unpublished data from the 1991 Census showing the proportion of non-South Asian Muslim people by local health authority in look-up tables in SANGRA. This proportion is higher in Greater London, but even here, only 7 per cent of the total number of people who may potentially be categorized as South Asian by SANGRA are likely to be Muslims with origins in non-South Asian countries. Finally, change of a woman's surname on marriage may result in misclassification if the woman's partner is from a different ethnic group. Data from the 1991 Census showed, however, that this is unlikely to be a major problem as over 95 per cent of Indian, 95 per cent of Pakistanis and 99 per cent of Bangladeshi women were married to men belonging to the same ethnic group, with the overall percentage decreasing slightly between first- and second-generation migrants.[35]

An important issue to bear in mind is that, despite the high levels of sensitivity and specificity achieved by SANGRA, considerable misclassification may still occur in populations with a low proportion of South Asians. In these circumstances, specificity needs to be close to 100 per cent to avoid large numbers of non-South Asians being misclassified by the program as South Asians, with a consequent dilution of any association between ethnicity and health outcomes. If, for instance, SANGRA is used in a population where the true proportion of South Asians is 2.5 per cent, the positive predictive value will be 71 per cent if sensitivity is 95 per cent and specificity 99 per cent, but only 33 per cent if specificity drops to 95 per cent. An increase in sensitivity from 90 per cent to 99 per cent would make little difference to the magnitude of the positive predictive values. Visual inspection of the names identified by the program as South Asians could be used in populations where the true prevalence of South Asians is relatively low to reduce the number of misclassifications. This approach would still be much more cost-effective than having to visually inspect all the names, particularly so for large datasets.

For studies that rely on SANGRA to ascertain cases and on the Census to obtain population figures, there is a possibility that bias may be introduced as different methods are used to define numerators and denominators. The results for the hospital in-patient datasets, which used the Census classification to categorize ethnicity, seem to indicate that this numerator–denominator bias is likely to be small. The degree of numerator–denominator mismatch will vary from study to study, however, depending not only on the sensitivity and specificity of SANGRA, but also on the actual ethnic structure of the population being studied. More critical for epidemiological studies is the lack of Census denominators for the various religion and language categories identified by SANGRA, as the 1991 Census allows differentiation of South Asians into only very broad categories defined according to country of origin (i.e. as 'Indian', 'Pakistani' or 'Bangladeshi'). Part of this problem will be overcome using the questions on religion included in the 2001 Census.

Conclusion

Direct collection of information on ethnic group should remain the aim in all health-related studies where this is possible. Some non-South Asian migrant groups, whose disease experience may also help in elucidating disease aetiology and issues that directly help the communities concerned, are not distinguishable by examination of names. However, the identification of people of South Asian origin on the basis of their names through SANGRA represents a valid and cost-effective alternative, especially so for historical health-related datasets that do not contain information on ethnicity. Its use will be dependent on obtaining appropriate ethical approval, given the implications of the new Data Protection Act. The use of SANGRA combined with information on country of birth will also allow examination of changes in risk across generations. SANGRA also provides a valid tool to identify groups originating in South Asia who have different religious, regional and linguistic backgrounds, different experiences and opportunities, and hence, different health-related exposures and disease risks. The development of SANGRA is a continuing process, as new South Asian names and spellings, including those identified in the present validation study, are being incorporated into its directories.

Acknowledgements

We would like to thank the following for their help with this study: Mrs N. Khurana, Ms L. Sevak, Mrs R. Kabbawala, Mrs K. Manji, Ms J. Wilkinson (LSHTM); Mrs T. Rana, Mrs A. Sarna, Mrs S. Walia; Mrs S. Aggarwal (UK Asian Women's Conference); Mr M. Bharadia (Gujerat Welfare Association); Mr K. Chopra (Hindhu Cultural Trust Centre); Mrs A. Dhillon (Indian Workers Association); Mr Janarthan (Tamil Welfare Association); Mr P. Jariwala (Gujerati Literary Academy); Ms R. Lobo (Goan Welfare Association); Mr D. Loney (Hyndburn and Rossendale Racial Equality Council); Ms J. Mian (Pakistan Women's Welfare Association); Mr J. Patel (Brent Indian Association); Mr J. Rajput (Indian Association of Harrow); Ms Y. Sadiq (Roshni Asian Women's Resource Centre); Mr D. Sharma (Acton Asian Association); Dr N. Srinivasan (Federation of Ambedkarites and Buddhist Organisations). Funding for this work was provided by the NHS National Cancer Research & Development Programme. Copyright of the program is held by I.S., K.N. and P.M. (LSHTM). Further details on SANGRA can be obtained from I.S.

References

1 Buell P, Dunn JE. Cancer mortality among Japanese Issei and Nisei of California. *Cancer* 1965; **18:** 656–664.

2 McKeigue PM, Ferrie JE, Pierpoint T, Marmot MG. Association of early-onset coronary heart disease in South Asian men with glucose intolerance and hyperinsulinemia. *Circulation* 1993; **87:** 152–160.

3 Balarajan R, Adelstein AM, Bulusu L, Shukla V. Patterns of mortality among migrants to England and Wales from the Indian subcontinent. *Br Med J* 1984; **289:** 1185–1187.

4 Pedoe HTT, Clayton D, Morris JN, Brigden W, McDonald L. Coronary heart attacks in East London. *Lancet* 1975: 833–838.

5 Donaldson LJ, Taylor JB. Patterns of Asian and non-Asian morbidity in hospitals. *Br Med J* 1983; **286:** 949–951.

6 McKeigue PM, Pierpoint T, Ferrie JE, Marmot MG. Relationship of glucose intolerance and hyperinsulinaemia to body fat pattern in South Asians and Europeans. *Diabetologia* 1992; **35:** 785–791.

7 Marmot MG, Adelstein AM, Bulusu L. *Immigrant mortality in England and Wales 1970–78. Causes of death by country of birth. Studies on medical and population subjects no. 47*. London: HMSO, 1984.

8 Matheson LM, Dunnigan MG, Hole D, Gillis CR. Incidence of colo-rectal, breast and lung cancer in a Scottish Asian population. *Health Bull* 1990; **43**(5): 245–249.

9 Balarajan R, Soni Raleigh V, Botting B. Sudden infant death syndrome and postneonatal mortality in immigrants in England and Wales. *Br Med J* 1989; **298:** 716–720.

10 Acheson D. *Independent inquiry into inequalities in health*. London: The Stationery Office, 1998.

11 Soni Raleigh V, Balarajan R. Suicide and self-burning among Indians and West Indians in England and Wales [see comments]. *Br J Psychiat* 1992; **161:** 365–368.

12 Gillam SJ, Jarman B, Law R. Ethnic differences in consultation rates in urban general practice. *Br Med J* 1989; **299:** 953–957.

13 Balarajan R, Raleigh VS, Yuen P. Hospital care among ethnic minorities in Britain. *Health Trends* 1991; **23:** 90–93.

14 Lear JT, Lawrence IG, Pohl JE, Burden AC. Myocardial infarction and thrombolysis: a comparison of the Indian and European populations on a coronary care unit. *J R Coll Phys London* 1994; **28:** 143–147.

15 Bhopal RS, Samim AK. Immunization uptake of Glasgow Asian children: paradoxical benefit of communication barriers. *Community Med* 1988; **10:** 215–220.

16 Hoare T. Breast screening and ethnic minorities. *Br J Cancer* 1996; **74**(Suppl XXIX): S38–S41.

17 Luke K. Cervical cancer screening: meeting the needs of minority ethnic women. *Br J Cance*r 1996; **74**(Suppl XXIX): S47–S50.

18 Karn V. *Ethnicity in 1991 Census Volume 3. Social geography and ethnicity in Britain: geographical spread, spatial concentration and internal migration*. London: HMSO, 1996.

19 Balarajan R, Bulusu L. Mortality among immigrants in England and Wales 1979–83. In: Britton M, ed. *Mortality and geography. A review in the mid 1980s. England and Wales. OPCS DS no. 9*. London: HMSO, 1990: 104.

20 Office of Population Censuses and Surveys and General Register Office for Scotland. *1991 Census. Ethnic group and country of birth. Great Britain*. Vol. 1 of 2. London: HMSO, 1993.

21 Nicoll A, Bassett K, Ulijaszek SJ. What's in a name? Accuracy of using surnames and forenames in ascribing Asian ethnic identity in English populations. *J Epidemiol Commun Hlth* 1986; **40:** 364–368.

22 Harding S, Dews H, Simpson SL. The potential to identify South Asians using a computerised algorithm to classify names. *Population Trends* 1999; **97:** 46–49.

23 Cummins C, Winter H, Cheng K, *et al*. An assessment of the Nam Pechan computer program for the identification of names of South Asian ethnic origin. *J Publ Hlth Med* 1999; **21:** 401–406.

24 Martineau A, White M. What's not in a name. The accuracy of using names to ascribe religious and geographical origin in a British population. *J Epidemiol Commun Hlth* 1998; **52:** 336–337.

25 Brown C. *Black and white Britain*. Policy Studies Institute. London: Heinemann, 1984.

26 Modood T. Culture and identity. In: Modood T, Berthoud R, eds. *Ethnic minorities in Britain. Diversity and disadvantage*. London: Policy Studies Institute, 1997: 290–338.

27 Confederation of Indian Organisations (UK). *Directory of Asian voluntary organisations*. London: Confederation of Indian Organisations (UK), 1994.

28 Koul RK. *Indic names (a documentation list)*. New Delhi: Seemant Prakashan, 1980.

29 Mehrotra RR. *The book of Indian names*. New Delhi: Rupa, 1994.

30 Kaushik A. *Hindu baby names*. New Delhi: Star Publications, 1994.

31 Dimpy MK. *Sikh baby names*. New Delhi: Star Publications, 1996.

32 Peach C. *Ethnicity in the 1991 Census. Volume 2. The ethnic minority populations of Great Britain*. London: HMSO, 1996.

33 Aspinall P. Department of Health's requirements for mandatory collection of data on ethnic group inpatients. *Br Med J* 1995; **311:** 1006–1009.

34 Winter H, Cheng KK, Cummins C, *et al*. Cancer incidence in the south Asian population of England (1990–92). *Br J Cancer* 1999; **79:** 645–654.

35 Coleman D, Salt J. *Ethnicity in the 1991 Census. Volume 1. Demographic characteristics of the ethnic minority populations*. London: HMSO, 1996.

Accepted on 21 May 2001