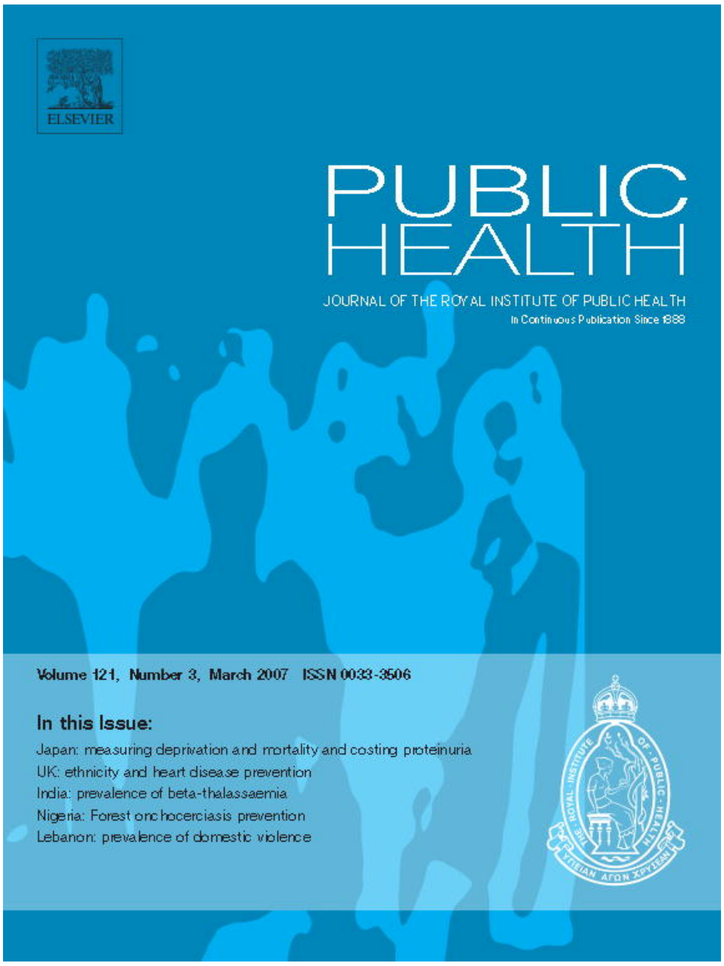


ELSEVIER

PUBLIC HEALTH

JOURNAL OF THE ROYAL INSTITUTE OF PUBLIC HEALTH

In Continuous Publication Since 1888

Volume 121, Number 3, March 2007 ISSN 0033-3506

In this Issue:

Japan: measuring deprivation and mortality and costing proteinuria
UK: ethnicity and heart disease prevention
India: prevalence of beta-thalassaemia
Nigeria: Forest onchocerciasis prevention
Lebanon: prevalence of domestic violence

Public Health (2007) **121**, 231–236

PUBLIC HEALTH
JOURNAL OF THE ROYAL INSTITUTE OF PUBLIC HEALTH
www.elsevierhealth.com/journals/pubh

Determining aspects of ethnicity amongst persons of South Asian origin: The use of a surname-classification programme (Nam Pehchan)

Gary J. Macfarlane[a,*], Mark Lunt[b], Benedict Palmer[b], Cara Afzal[b], Alan J. Silman[b], Aneez Esmail[c]

[a]*Aberdeen Pain Research Collaboration, Epidemiology Group, Department of Public Health, University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen, AB25 2ZD, UK*
[b]*The Arthritis Research Campaign Epidemiology Unit, Division of Epidemiology and Health Sciences, University of Manchester, UK*
[c]*Division of Primary Care, University of Manchester, UK*

Received 29 March 2006; accepted 7 July 2006

KEYWORDS
Ethnicity;
Language;
Religion;
Nam Pehchan;
Validity;
South

Summary *Objective:* Name-based classification systems are potentially useful in identifying study samples based on probable ethnic minority group. The aim of the current study was to assess the validity of the Nam Pehchan name classification programme of religion and language against subject self-report.

Study design and setting: A population-based cross-sectional survey conducted in areas of the North-West and West Midland regions of England with a relatively high density of South Asian ethnic minority groups. The sampling frame was age–sex registers of selected general practices and subjects were classified according to language and religion using the Nam Pehchan programme. These were compared with responses by subjects on a self-complete postal questionnaire.

Results: One thousand nine hundred and forty-nine subjects who participated, classified themselves as South Asian. Sensitivity in identifying religion was high amongst Muslims (92%) and Sikhs (86%), and somewhat lower in Hindus (62%). Specificity exceeded 95% for all ethnic groups. The vast majority of subjects assigned Punjabi or Gujarati as their main South Asian language indicated that they did in fact speak these languages (97% and 94%, respectively). Subjects assigned Urdu or Bengali, however, were less likely to do so (61% and 35%, respectively).

Conclusions: The name-based classification system Nam Pehchan has demonstrated high levels of accuracy in some sub-groups of the South Asian population in

[*]Corresponding author. Tel.: +44 1224 552495; fax: +44 1224 550925.
E-mail address: g.j.macfarlane@abdn.ac.uk (G.J. Macfarlane).

doi:10.1016/j.puhe.2006.07.001

determining subjects likely language spoken and religion—and is likely to be a useful additional tool when information on ethnicity is not already available.

Introduction

There is a significant literature showing that ethnic minorities in the UK have poorer health when compared to Caucasians.[1–4] If we are to research the reasons for the poor health outcomes and the factors associated with these outcomes, then we have to have better methods of identifying and classifying ethnic minority groups.

In epidemiological studies, classification of ethnic group can be done in a variety of ways. The first is on the basis of respondents choosing an ethnic minority group from a list, and whilst this form of classification is useful it also has its limitations[5]. The second method is to group persons according to demographic information, e.g. on their (and their parents and ancestors) place of birth. This has the potential disadvantage that it does not take account of individual opinions on their ethnicity and the beliefs and attitudes that may therefore be associated with this. Both these methods require collecting information at the level of the individual. In terms of record linkage studies or in identifying suitable sampling frames of ethnic minority groups either for research studies or health service provision, it is useful to be able to classify subjects into probable ethnic sub-groups. Current techniques for identifying sub sets of ethnic minority populations include programs such as SANGRA[6] and Nam Pehchan (NP)[7]. NP is a computerized classification system used to identify all South Asian names and to assign likely religious and language groups to which they belong.

The accuracy of the NP method of classification in distinguishing South Asians from non-South Asians has been reported[7] comparing programme classification versus programme classification plus visual inspection of names. What is unknown however is, within South Asian ethnic groups the ability of the programme to determine aspects of ethnicity. The aim of the current study was therefore to assess the validity (against self-report) of a computerized name-based classification system in determining religion and language used within communities of persons identifying themselves as South Asian.

Methods

Population survey

A population-based cross-sectional survey of individuals aged 18–75 years living in selected areas of England was conducted. The population sampling frame was derived from the age–sex registers of 13 general practices in areas with high densities of persons of South Asian origin. Ten general practices in the North-West towns of Bolton, Oldham and Ashton/Tameside, and three from the West Midlands city of Birmingham agreed to participate. All registered patients from the 13 study practices were mailed an English language questionnaire, with a note in each of the appropriate South Asian languages (Urdu, Punjabi, Bengali and Gujarati) offering a version of the same questionnaire in any of those languages, if requested by mail or telephone. Non-responders received a postcard reminder 2 weeks after the initial mailing, and a further questionnaire 2 weeks later. If there was still no response, a contact visit by a linkworker was made to assist in completing the questionnaire.

The questionnaire asked subjects 'What is your ethnic group?' and used the classifications of the United Kingdom 2001 census (http://www.statistics.gov.uk/cci/nugget.asp?id=455) and the present analysis is restricted to those describing themselves as within the sub-categories of Asian or British Asian ('Indian', 'Pakistani' or 'Bangladeshi'). Ethnicity was explored by enquiring about the factors which form an individual's cultural identity. Thus information about an individual's religious affiliation and command of South Asian languages was collected. Data regarding religion was collected by requesting 'Which of the following best describes your religious affiliation?': the options for reply were 'Christian', 'Muslim', 'Hindu', 'Sikh', 'Buddhist', 'not religious' or 'other'. The language section of the questionnaire asked subjects 'Here is a list of languages which you may speak. For each please indicate with a tick if you are able to understand, speak, read or write the language' The languages listed were 'English', 'Urdu', 'Punjabi', 'Bengali', 'Sylheti' and 'Gujarati'. The present analysis concentrated on the ability to speak such languages.

NP classification system

The NP computerized classification system (version 1) identifies names likely to be of South Asian origin, and based on the forename and surname assigns a likely religion and language. It categorizes the religious origin of a South Asian name as Muslim, Hindu or Sikh. Language may be classified as Urdu, Punjabi, Bengali, Gujarati, Hindi or other. If no specific religious or language origin can be assigned, the program may offer two options, or alternatively one of the following categories may be applied: 'common', indicating an Asian name of no specific origin, 'ambiguous', where only the name stem (rather than the whole name) is matched, or 'clash' where the origin of a forename and surname are incongruous. For each individual, the classification of religious and language origin was then compared against the gold standard of the subjects' questionnaire responses. The questionnaire was designed to allow subjects to list multiple languages but only one religious affiliation.

Sensitivity and specificity analyses were carried out to compare the NP results with the self-reported information on the questionnaire. To determine whether age (or gender) affected the ability of NP to identify religious groups, logistic regression models were used, one for each religion. The outcome variable was whether the individual described themselves as having that religious affiliation, the predictor variables were age or gender.

Whereas subjects could identify multiple languages on the questionnaire, the NP programe assigns only one language to a name, and hence the research question here differed slightly from the sensitivity and specificity analysis used to address religious affiliation. The approach in this case was to assess whether a subject did in fact speak the language assigned to their name by the computer programe, irrespective of any other languages spoken by that subject. To determine if age or gender affected the ability of NP to identify the language used by an individual, multinomial logistic regression was used. The outcome variable was language status (spoke assigned language with/without other South Asian language, spoke only English, or spoke a different South Asian language), and the predictors were the language assigned by NP, age (or gender) and their interaction.

Results

Population survey

The study questionnaire was mailed to 7668 individuals. Following a repeat mailing and link-worker visit to non-responder addresses, 1774 addressees could not be traced and were assumed not to be living at the registered address, leaving 5894 successfully delivered questionnaires. Of these, 2998 were returned, giving a response rate of 51%. One thousand nine hundred and forty-nine subjects classified themselves as Asian, and were included in the present analysis. The remaining 1049 were excluded: 933 subjects who described themselves as Caucasian, 73 subjects who were self-classified as belonging to other ethnic groups (other Asian, Black or Black British, Chinese, mixed race or other), and a further 43 subjects did not report their ethnic group.

Demographic information on the respondents is provided in Table 1. Indians were oldest, most likely to have been born in the UK and most likely to have been educated for longer. Bangladeshis were the youngest, least likely to have been born in the UK and had the lowest level of education.

Classification of religion and language

Self-reported religious affiliation was provided by 1905 (98%) South Asian participants. The vast majority of South Asian subjects (96%) identified themselves as being affiliated to one of the religions recognized by the NP classification (Muslim, Hindu or Sikh). Using self-reports of ethnicity

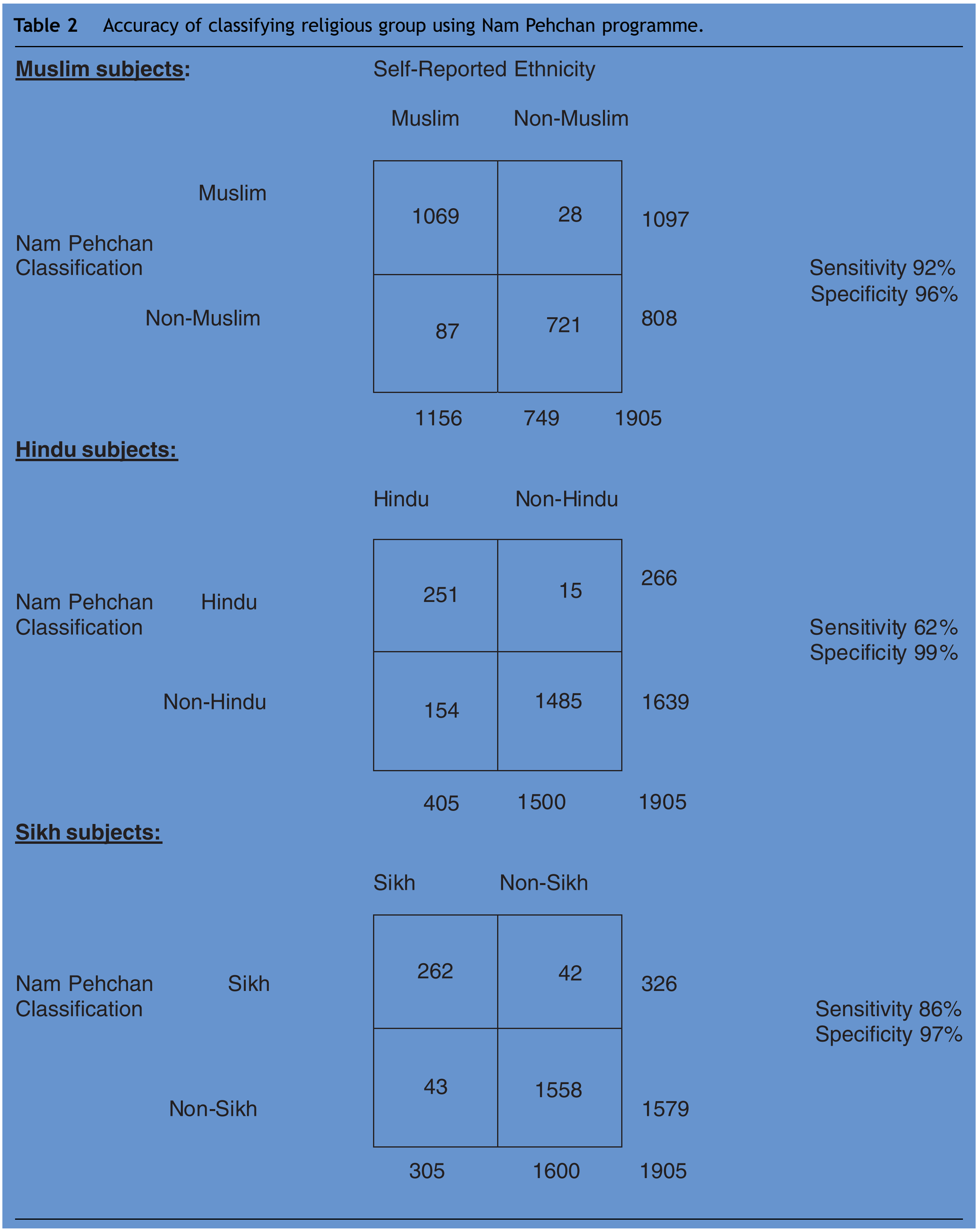
Table 1 Demographic characteristics of the study population.

			Indian	Pakistani	Bangladeshi
Females		*n* (%)	612 (52)	221 (53)	132 (38)
Born in UK		*n* (%)	364 (31)	114 (27)	51 (14)
Years of education	None	*n* (%)	177 (15)	113 (28)	106 (31)
	1–7 years	*n* (%)	250 (22)	62 (16)	86 (25)
	8–12 years	*n* (%)	280 (24)	90 (22)	63 (18)
	>12 years	*n* (%)	445 (39)	133 (33)	90 (26)
Years spent in UK		Median (IQR)	29 (22, 35)	24 (19, 31)	18 (12, 24)
Age		Median (IQR)	41 (30, 51)	36 (26, 48)	29 (24, 40)
Percentage of life spent in UK		Median (IQR)	68 (52,100)	67 (50,100)	58 (36, 100)

as the gold standard, the sensitivity and specificity of the classification system was determined (Table 2).

Sensitivity was high amongst Muslims (92%) and Sikhs (86%), and somewhat lower in Hindus (62%). Much of this lack of sensitivity in the Hindu group was due to ambiguous classification, with 98 (24%) classified as having a 'common' Asian name of no specific religious origin. Specificity exceeded 95% for all ethnic groups. Age or gender of the subject did not influence performance in identifying religious group.

Table 2 Accuracy of classifying religious group using Nam Pehchan programme.



Muslim subjects:

		Self-Reported Ethnicity			
		Muslim	Non-Muslim		
Nam Pehchan Classification	Muslim	1069	28	1097	Sensitivity 92% Specificity 96%
	Non-Muslim	87	721	808	
		1156	749	1905	

Hindu subjects:

		Hindu	Non-Hindu		
Nam Pehchan Classification	Hindu	251	15	266	Sensitivity 62% Specificity 99%
	Non-Hindu	154	1485	1639	
		405	1500	1905	

Sikh subjects:

		Sikh	Non-Sikh		
Nam Pehchan Classification	Sikh	262	42	326	Sensitivity 86% Specificity 97%
	Non-Sikh	43	1558	1579	
		305	1600	1905	

Table 3 The performance of Nam Pehchan in predicting language.

Subject report/NP classification	Urdu ($n = 458$)	Punjabi ($n = 282$)	Bengali ($n = 468$)	Gujarati ($n = 575$)
Subject reported speaking assigned language only *n* (%)	71 (16)	224 (79)	46 (10)	320 (56)
Subject reported speaking assigned language and another South Asian language *n* (%)	212 (46)	52 (18)	118 (25)	217 (38)
Subject didn't report speaking assigned language *n* (%)[a]	159 (35)	3 (1)	288 (62)	24 (4)
Subject reported speaking only English *n* (%)	16 (3)	3 (1)	16 (3)	14 (2)

[a] Of subjects who reported speaking at least one of South Asian language.

The proportion of individuals who self-reported speaking the language assigned to them by NP is given in Table 3. The vast majority of subjects assigned Punjabi or Gujarati as their main South Asian language indicated that they did in fact speak these languages (97% and 94% respectively). Subjects assigned Urdu or Bengali, however, were less likely to do so (61% and 35%, respectively). The low proportion associated with the Bengali language may in part be due to NP not having a marker for the Sylheti language, and thus misclassifying Sylheti speakers as Bengali speakers. Combining these two languages gave the result that 50% of subjects classified as Bengali speakers self-reported speaking Bengali or Sylheti. Further amongst subjects classified as speaking the Bengali language by NP, there was an interaction between performance of the classification and age (above or below the median age of 37 years). Older subjects were more likely to speak a different South Asian language and not Bengali (72% ($n = 136$)) than younger subjects (54% ($n = 152$)) and correspondingly less likely to speak Bengali (25% ($n = 48$); compared to 42% in the young ($n = 116$)). There was no effect of gender.

Discussion

This study has confirmed that in populations of South Asians, a computerized name-based classification programme (NP) offers a useful additional tool in classifying some sub-groups of populations on the basis of language and religion.

There are a number of methodological issues to consider. Firstly, whether the proportion of non-participants has influenced the validity of the results. Despite intensive efforts (repeat mailing, questionnaires available in a variety of languages, home visits by linkworkers) the participation rate was modest (but similar to or better than many comparable studies). The current population survey included Caucasians, although their data did not contribute to the current analysis, and amongst this group the participation rate achieved was actually slightly lower, suggesting that low response was not a particular feature of the South Asian groups but possibly common to areas which generally experienced above-average levels of deprivation. Non-participants will only have influenced the external validity of the study if the ability of the computerized name–classification system to correctly identify their religion or language spoken was different. Names of non-participants were available and in terms of language and religion assigned by NP to these non-participants, the distribution did not differ from participants (data not shown).

This data adds to information confirming that name classification systems are potentially useful in providing information on likely ethnicity—although the performance of systems is specific to particular geographical areas. A study in the Netherlands found that a combined name (first and surnames) classification system performed well in identifying Turkish and Arabic (Moroccon) children in a routinely registered database of Dutch children—but not for Surinamese children[8]. In the United States the use of surnames in addition to available information on race greatly enhanced researchers' ability to identify Hispanics and whites[9].

How can the results of this study be used? Firstly it has previously been shown that in population-based epidemiological studies, use of the NP classification system can (together with visual inspection) provide a mechanism for identifying populations of people of South Asian origin. This current study has identified that overall it has reasonably good accuracy within such population groups in identifying aspects of ethnicity (language and religion). These results may be useful for studies conducted at practice level—for example access to and use of services, health experiences and health status. We recognize that the sensitivity and specificity will be affected by the population mix in any area: for example, the presence of people who are Muslim but not from South Asia (such as Somalia) but information available at practice level may allow such groups to be identified more easily. It may also be useful for record linkage studies whereby only demographic information is available on subjects and this would allow an additional classification in terms of aspects of ethnicity.

In summary, the name-based classification system NP has demonstrated reasonably good accuracy in determining subjects likely language spoken and religious affiliation, and hence ethnic sub-group, amongst some persons of South Asian origin in an area of the United Kingdom with relatively high density of such ethnic minority groups. This therefore may be a useful additional instrument, where information on ethnicity is not already available, for conducting research into the health of South Asian populations.

References

1. Bhopal R. What is the risk of coronary heart disease in South Asians? A review of UK research. *J Public Health Med* 2000;**22**:375–85.
2. Littlewood R, Lipsedge M. Psychiatric illness among British Afro-Caribbeans. *BMJ* 1998;**296**:950–1.
3. Allison TR, Symmons DP, Brammah T, et al. Musculoskeletal pain is more generalised among people from ethnic minorities than among white people in Greater Manchester. *Ann Rheum Dis* 2002;**61**:151–6.
4. Gill PS, Kai J, Bhopal RS, Wild S. Health care needs assessment: black and minority ethnic groups (http://hcna.radcliffe-oxford.com/bemgframe.htm).
5. McKenzie K, Crowcroft NS. Describing race, ethnicity, and culture in medical research. *BMJ* 1996;**312**:1054.
6. Nanchahal K, Mangtani P, Alston M, dos Santos Silva I. Development and validation of a computerized South Asian Names and Group Recognition Algorithm (SANGRA) for use in British health-related studies. *J Public Health Med* 2001;**23**:278–85.
7. Cummins C, Winter H, Cheng KK, et al. An assessment of the Nam Pehchan computer program for the identification of names of south Asian ethnic origin. *J Pub Health Med* 1999;**21**:401–6.
8. Bouwhuis CB, Moll HA. Determination of ethnicity in The Netherlands: two methods compared. *Eur J Epidemiol* 2003;**18**:385–8.
9. Morgan Ro, Wei H, Virnig BA. Improving identification of Hispanic males in Medicare: use of surname matching. *Med Care* 2004;**42**:810–6.