Original Article

# Name Recognition to Identify Patients of South Asian Ethnicity within the Cancer Registry

Savitri Singh-Carlson[1], Frances Wong[2], Gurpreet Oshan[2], Harajit Lail[3]

[1]School of Nursing, California State University, Monterey Bay, CA, USA, [2]British Columbia Cancer Agency, Fraser Valley Centre, [3]Fraser Valley Centre, British Columbia, Canada

**Corresponding author:** Savitri Singh-Carlson
E-mail: savitri.singh-carlson@csulb.edu

Received: August 16, 2015, Accepted: October 21, 2015

## ABSTRACT

**Objective:** The goal of this project was to develop a list of forenames and surnames of South Asian (SA) women that could be used to identify SA breast cancer patients within the cancer registry. This list was compiled, evaluated, and validated to ensure comprehensiveness, accuracy, and applicability of SA names. **Methods:** This project was conducted by Canadian researchers who are immersed in conducting behavioral studies with SA women diagnosed with cancer in the province of British Columbia. Recruiting SA cancer patients for research can be a difficult task due to social and cultural factors. Methods used by other researchers to identify ethnicity related unique names were employed to filter surnames and forenames that were not common to this ethnic group. Co-author (Gurpreet Oshan) of SA ethnicity rigorously identified and deleted multiple lists and redundant entries along with common English forenames which resulted in a list of 16,888 SA forenames. All co-authors of Indian ethnicity (Gurpreet Oshan, Savitri Singh-Carlson, Harajit Lail) were involved in critiquing and manually reviewing the names list throughout this process. Comprehensive lists of SA surnames and women's forenames were reviewed to identify those that were unique to SA ethnicity. Accuracy was ensured by constantly filtering the redundancy by using an Excel program which helped to illustrate the number of times each name was spelled in different ways. **Results:** The final lists included 9112 surnames and 16,888 forenames of SA ethnicity. On the basis of the surname linkage only, the sensitivity of the list was 76.6%, specificity was 62.9%, and the positive predictive value was 58.5%. On the basis of both the surname and forename linkage, the specificity of the list was 88.6%. These lists include variations in spelling forenames and surnames as well. **Conclusions:** The list of surnames and forenames can be useful tools to identify SA ethnic groups from large population database in healthcare-related research. Ethnicity-specific population research is important in order to help identify how cancer care should be delivered for the SA population, as well as for planning and provision of other related health services. We are willing to share this list upon request to the authors.

**Key words:** Cancer registry, South Asian ethnicity, validation of names

Access this article online

| Quick Response Code: | |
|---|---|
| | **Website:** www.apjon.org |
| | **DOI:**<br>10.4103/2347-5625.170224 |

**Cite this article as:** Singh-Carlson S, Wong F, Oshan G, Lail H. Name Recognition to Identify Patients of South Asian Ethnicity within the Cancer Registry. Asia Pac J Oncol Nurs 2016;3:86-92.

## Introduction

Ethnicity (meaning "nation") originally derived from the Greek word "ethnos," (meaning the sharing of a common ancestry),[1] provides various contexts to which an individual could associate their belonging to such as: Language, religion, kinship, shared territory, geographical locatedness, nationality, and physical appearance.[2,3] When conducting cancer care related studies with migrant populations, it is vitally important to identify the study population's differing contextual factors such as acculturation, political and societal situatedness, geographical origins, age, and gender; however, it is imperative to capture as a large sample from the target population in order to effectively translate meaningful results into practice.[4,5] Beyond, this is the need to understand that almost every population group will have various subgroups that may be identified by language, religion, or common surnames and forenames, rituals, practices, beliefs, and values that are individual and/or shared with others within the same group.[3] Other researchers conducting behavioral studies with South Asian (SA) cancer patients have documented the importance of identifying a list of common surnames and forenames as a tool, which offers the possibility of inferential ethnic classification that can lend to identifying a certain population for healthcare-related research.[3-6] This form of classifying populations into ethnic groups through name recognition is an alternative to ethnicity self-identification when this is not available through cancer registries. The goal of this study was to develop a list of forenames and surnames of SA women that could be used to identify SA breast cancer patients within the British Columbia Cancer Agency (BCCA) cancer database for British Columbia (BC). This list was compiled, evaluated, and validated to maximize comprehensiveness, accuracy, and applicability of SA names.

This paper presents how we critiqued, categorized, and updated a directory of SA surnames and forenames when the need arose for us to identify SA breast cancer patients within the population utilizing the BCCA clinical information database. For this project, the SA population are those who identify themselves as being of Indian ethnicity and mainly of Indian origin, regardless of geographical locatedness. Authors recognize that the SA population will be subcategorized into those from Pakistan, Sri Lanka, or Bangladesh and into further subgroups within these categories.

Canadian Cancer Society statistics estimates that 25,000 women, representing 26% of all new cancer cases in women, will be diagnosed with breast cancer in 2015.[7] For women residing in BC, with breast cancer being the most frequently diagnosed type of cancer, an estimated 3400 women will be diagnosed with breast cancer in 2015.[8] Data by ethnicity is not identified in the Canadian Cancer Society statistics; however, research has shown that SA Canadians are among the under-screened ethnic groups that are significantly less likely to get tested in comparison to the general population for a variety of cultural, linguistic, and economic reasons.[9]

Statistics Canada indicates the influx of SA immigrant population in Canada, especially in BC within the last few decades.[10] This increase in the SA population is not only restricted to those from the Indian continent, but also from Fiji Islands, Caribbean, Sri Lanka, England, or the African continent. In BC, the SA subgroups are mainly from Pakistan, Sri Lanka, Bangladesh, and various parts of India, with the state of Punjab having the highest number.

Statistics Canada 2006 census data by "ethnic origin," reports two categories for population of Indian ethnicity: SA origin totaling 1,316,770 responses and East Indians totaling 962,670 responses in Canada.[11] Census data by provincial ethnic origins for BC report "SA" origin responses totaling 265,595 and East Indians as totaling 232,370.[12] Question targeted at "what were the ethnic origin or cultural origins of this person's ancestors" is descriptive where an individual could choose "as many origins as applicable," but up to 6 are retained by Stats Canada; whereas the question targeted at "this person is" provides choices with "SA (e.g., East Indian, Pakistani, Sri Lankan, etc.)" being one of the choices.[12] The "ethnic origin" census data is used to "paint a picture of Canada's multicultural communities" and refers to ethnic origin as "cultural origins of the respondent" or refers to "a person's roots" and should not be confused with the respondent's citizenship, whereas "this person is" data refers to the population group that the respondent belongs to, which is used to identify the particular population or visible minority group which is used to promote equal opportunity for everyone.[13] This format of census data collection critiqued here illustrates the complexity of collecting true ethnicity related data. Regardless of the method of identification, through self-identification or name-based extraction; the multifaceted nature and complexity of "ethnicity" always remains. The heterogeneity within ethnicity, due to subgroups among an apparently homogenous group of SAs who are of Indian descent is only adding richness to the research studies relating to how ethnic background may influence cancer patients in the presentation of the disease, understanding of the treatment and impact, and the preparation for survivorship.

# Methods

## *Study design*

This study was part of the multiphased project by the review ethics board at University of British Columbia. This study arose from a practical need to identify SA women who were of Indian ethnicity; resided in Canada; were between 18 and 85 years old; spoke Hindi, Punjabi, Urdu, or English and had a diagnosis of nonmetastatic breast cancer as per the TNM grading system.[14,15] SA women names were to be identified for the SA Breast Cancer Survivors Project[15-18] undertaken by the breast (Breast Research to Evaluate Access Research and Services) research team (BRT) from 19,896 records extracted from the BCCA clinical database.

## *Development of list of surnames*

Various resources were used to compile the BRT list of names. A preexisting list of SA surnames manually created in 2006 by the provincial cancer agency researchers for a cancer care related project[19] was used as a beginning point. These authors used the Screening Mammography Program of British Columbia (SMPBC), a provincial population-based breast cancer screening program that began in 1988, where the place of birth is self-reported.[15-17] These cancer agency researchers collated a list of SA surnames by manually identifying the SMPBC record list and the local telephone lists as resources.[7] To the cancer agency researcher's list of 2787 surnames, 623 surnames obtained from a local SA family physician's private practice were added. Another list of 198 unique SA surnames, obtained from the study Shah *et al*.,[20] 2606 surnames retrieved from Guhuray's SA name recognition project;[21] 126 Sri Lankan surnames acquired from an internet-based telephone directory; and various internet resources containing 5075 surnames were added to the original list. These internet resources and internet-based telephone directories were credible open sources with the longevity of these websites. At this time, a provincial-based SA surnames telephone directory (APNA) added 272 more unique surnames, and 809 surnames were added from the dictionary of Sikh names with common surnames listed in the book.[22]

Surnames collected from these various resources were merged into an Excel file. Using an Excel formula, the SA co-author (Gurpreet Oshan) counted the number of times a name appeared on the list and if the name was unique, the frequency count would be 1; twice, 2; and if three, then 3, and so on. Then the list was sorted by decreasing order with the name appearing the most number of times on the list at the top. Redundant names were then manually removed. This resulted in a compiled total of 9112 surnames after this rigorous name recognition procedure was completed.

## *Development of list of forenames*

A similar procedure was used to develop the list of forenames. An initial list of 617 forenames was identified for the SA Breast Cancer Survivors Project[13] using Hislop's SA surnames list as the first step. This was followed by various rigorous steps published in Singh-Carlson *et al*. study for which the names were identified in order to arrive at this list.[17] In this aforementioned study, authors created a list of common SA forenames and surnames using a variety of sources, such as the SA white-pages telephone book and the provincial screening mammography name list. The list was not exhaustive but was a fairly good indicator of common Indian surnames of persons residing in BC.

Of these names, the redundant names were identified using the Excel program and then were manually removed. By utilizing this process, 353 unique forenames were finally identified. At this time, various open resources websites listing SA baby girl names were identified as a resource listing for ethnic relevance and common SA female forenames. The female forenames were retrieved from these various websites and compiled into an Excel file. Co-author (Gurpreet Oshan) rigorously identified and deleted multiple lists and redundant entries along with common English forenames using the same procedure as mentioned previously for the development of the list of surnames. This process resulted in a list of 16,888 SA forenames. All co-authors of Indian ethnicity (Gurpreet Oshan, Savitri Singh-Carlson, Harajit Lail) were involved in critiquing and manually reviewing the names list throughout this process.

Development of a list of middle names was attempted with 86 middle names identified; however, since the SMPBC database does not collect middle names at this time, it was not possible to validate this list, therefore it was removed from subsequent analysis, although these unique middle names would be an added step in correctly identifying SA ethnic persons.

# Results

## *Validating the name list*

The final list of BRT SA surnames and forenames was validated in a two-step process by linking it to the SMPBC database as the SMPBC collects the self-ascribed ethnicity data. The demographic question asks, "To which group do you ancestors belong? Check all that apply." In the first step, the Data Surveillance and Outcomes Unit at the BCCA matched the two lists,

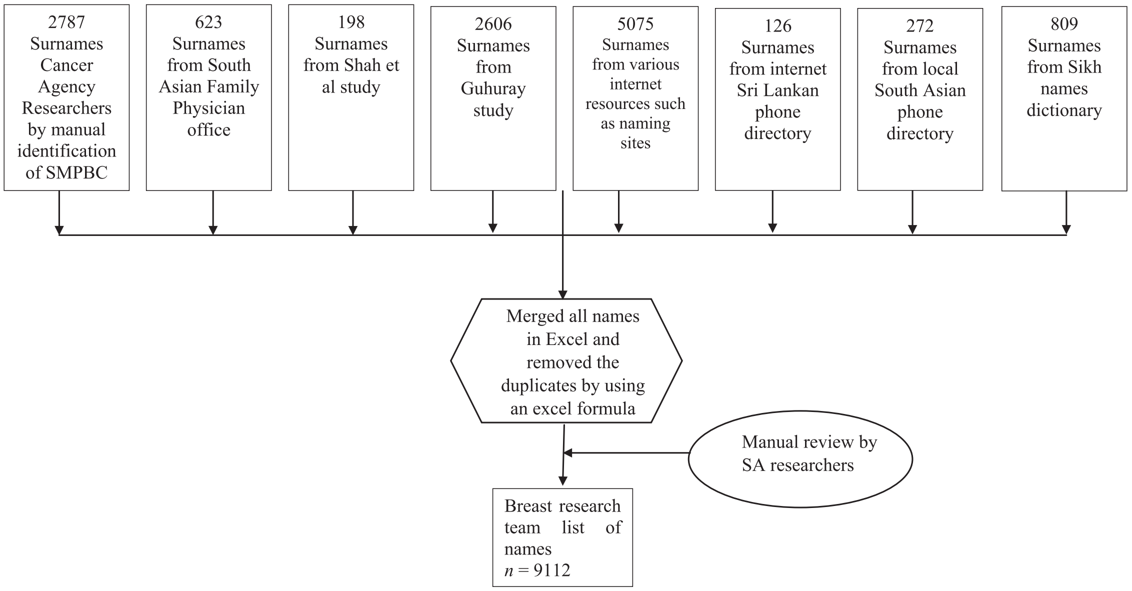
2787 Surnames Cancer Agency Researchers by manual identification of SMPBC
623 Surnames from South Asian Family Physician office
198 Surnames from Shah et al study
2606 Surnames from Guhuray study
5075 Surnames from various internet resources such as naming sites
126 Surnames from internet Sri Lankan phone directory
272 Surnames from local South Asian phone directory
809 Surnames from Sikh names dictionary
Merged all names in Excel and removed the duplicates by using an excel formula
Manual review by SA researchers
Breast research team list of names
n = 9112

Figure 1: Process for Identifying the Surnames

BRT names list and SMPBC dataset, on the basis of the surnames. By this process, 1822 surnames were identified as true positives, i.e., the surnames matched on the two lists and the self-ascribed ethnicity was SA. 1295 surnames were identified as false positives, i.e., the surnames matched but the self-ascribed ethnicity was not SA. The remaining names on the list ($n$ = 5995) were considered to be nonmatches on the basis of surnames. 2191 surnames were identified as true negatives, i.e., the surnames did not match and the self-ascribed ethnicity was not SA. 557 surnames on the SMPBC list were identified as false negatives, i.e., the surnames did not match to the surnames on the BRT list but the self-ascribed ethnicity was SA. However, upon manual review, it was identified that some of the surname variations due to different spellings were identified as false negatives. The surname variations and/or the surnames that were not on the BRT list originally were identified using the Excel program and these were considered to be the true, false negatives. Another reason for false negative was because people had provided their self-ascribed ethnicity as SA, despite the names being of South East Asian ethnicity.

The redundant surnames on the false negative list were removed. A manual inspection was completed, and three South East Asian surnames were identified which were also removed from the list of false negatives. In the end, the actual number of false negatives was 247. These names were later added to the BRT list of names.

In the second step, a further linkage of the list of the false positives ($n$ = 1295) was done with the BRT list of forenames. Initially, 269 names matched on the basis of forenames and were identified as false positives on the basis of surname and forename linkage, 700 were identified as nonmatches as the forenames did not match, and 326 were identified for manual review. Self-ascribed ethnicity data was unavailable for 329 of these records, so they were removed from further review. After the manual review process was completed, it resulted in a total of 283 false positives with matched forename and surname, and 683 records matched on the basis of surnames but failed to match on the basis of forenames. When the linkage was done using the forenames as an additional screen, the number of false positives was brought down from 1295 to 283. However, of these 283 records, 164 names were identified to be clearly SA. However, it is possible that people had listed a different ethnicity due to the basis of place of birth or place of living. Table 1 shows the sensitivity, specificity, and the positive predictive value of the list based on the surname linkage and the specificity of the list based on the surname and forename linkage.

Using the surnames only, 617 were extracted from the survivorship project.[13] Using the surname and forename combination, 16 more names were extracted, making it a total of 633 names. This indicates that by surname and forename function as an "either or" function, more names can be identified. When the study invitations were sent to these potential participants, only two breast cancer patients reported that they were not of SA ethnicity. Others identified that they were of SA ethnicity and continued to complete the survey.

## Discussion

Based on the diagnostic analysis and validation, it appears that the additional screening using forenames can be useful for excluding persons outside of the cultural group, therefore, rending the name list more specific. This may increase the number of false negatives, i.e., decreased the sensitivity of the tool. It is not possible to be certain of false positives, e.g., due to name change after marriage. For any population subjected to the name list tool, it is not possible to know the actual number of persons within the cultural group who are excluded (false negatives) due to variations in the mixture of subgroups. Increments of accuracy can be achieved by continuous approximation according to manual recognition.

As previously presented, the multidimensionality of ethnicity is complex due to sensitive nature of self-identification of characteristics one bestows on the self which is dependent on a number of social and cultural factors.[3] Researchers look for rigorous methods to identify and recruit a representative sample from a particular population group in order to conduct research that will help to illustrate health care needs; however, demographic data are not always collected at source, making it difficult to access population data. Building a reference list of surnames and forenames using name recognition which is a clue to the origin, kinship, language, shared religion, and territory have been reported as having the potential to improve the sensitivity and specificity of the said names.[3,4] No name list is entirely accurate and may include persons outside of the cultural group but with similar names, or those women who married into the cultural group. Hence, a manual review is needed for accuracy, and the process is time-consuming, but less costly than other alternatives.[18,19]

Breast cancer research for SA population is evidenced as being of importance in light of studies that report SA women being diagnosed at a later stage of cancer than other women; having lower rates of screening for breast and cervical cancer.[8,9,16-18] These results indicate that there is need to have as a large sample of SA women with cancer as possible and from as many age and subgroups as possible to translate knowledge into practice. Studies conducted with SA women have reported issues with recruiting participants from this ethnic group due to work, lack of transportation, family commitments, or social/cultural beliefs.[16-18] The name list created in this study is a purposeful name recognition that can be highly beneficial in identifying as a large sample as possible with the appropriate inclusion/exclusion criteria. Using this name recognition list for the SA population is of further benefit, especially in instances where cancer registries do not include ethnicity as patient demographic data as a practice. The generalizability of this tool is high since it can be utilized all over Canada in setting where there is a large SA residency.

Using multiple tools in order to increase the accuracy of the reference list will help to classify the target population is reported as increasing specificity and reducing redundancy, which will reduce costs when recruiting and identifying potential participants.[16-18] Using an expert from the community who can recognize names that fit the target population's origin along with experts for validating them are vitally important to improve accuracy; however, it is important to be constantly aware of subgroups within groups that will bring variations.[18-20] There is an advantage to using name recognition in identifying ethnicity because it does not mask important differences among the groups; however, further identification of subgroups and other demographic factors may have to be conducted to apply this name list to a particular population for healthcare-related research.[23-28] Studies report that name recognition is an important tool for identifying SA ethnicity for health-related research until such time that ethnic for monitoring becomes a part of routine demographic data collection for patients; however, it is important to remember the "mix" or the subgroups within the larger SA group within any given city.[23,24,29-34] Choi *et al*. also report the concerns of identifying subgroups within the Chinese ancestry names.[25,31]

Studies conducted in the United Kingdom where a large SA population resides confirm that intermarriage between non-SAs and SA along with those marriages within the subgroups of Pakistan, Bangladesh, Sikhs, and Hindus create difficulties when identifying SA subgroups who may be at health risk for a particular disease.[20,24,26-28,30,31] A further classification of names into SA subgroups would be important because of differences due to vegetarian or nonvegetarian diets, health-related exposures to social and cultural behaviors and disease risks.[28-29,32-37]

The immigration pattern changes with time. Hence, any name list needs to be updated to be as comprehensive as possible. Any such list may not be suitable for use in a different region or country with different immigration pattern. The other change that occurs is the acculturation and assimilation of a population, the personal meaning of their identity due to the socialization of self, age that may bring a different meaning to how the individual relates to her/his own ethnicity or kinship to the population.[20-22,28] Self-identification of ethnicity carries risks of preconceived ethnic group classification, so it is important to remember how this data is used in the context of reporting healthcare-related data.[4,18,38]

### Limitations

There are limitations to identifying SA women through surname identification that arises from marriage outside of ethnic groups and concomitant name change for women; however, having a list of forenames that can be used as an added filter is seen as strength when identifying SA women. Furthermore, breast cancer patients who may identify themselves as being of interracial status will confound research findings. Therefore, future research will have to make this identifier either an inclusion or exclusion criteria. Another limitation is the ability of this list to distinguish the variations due to spelling original surnames that were derived from indigenous SA due to immigration records matching birth records in the originating country.[26,32,33]

### Contributions to literature

A surname and forename list can be a useful tool to identify specific ethnic groups from large population database.[38-40] Useful insights can be gained into the understanding of the etiology of disease, as well as for planning and provision of health services. We are willing to share this list upon request to the authors.

### Financial support and sponsorship

Nil.

### Conflicts of interest

There are no conflicts of interest.

## References

1. Mateos P. A review of name-based ethnicity classification methods and their potential in population studies. Popul Space Place 2007;13:243-63.
2. Yavari P, Hislop TG, Abanto Z. Methodology to identify Iranian immigrants for epidemiological studies. Asian Pac J Cancer Prev 2005;6:455-7.
3. Singh-Carlson SW, Kaur H. Sikhism and nursing. In: Fowler MD, Reimer-Kirkham S, Sawatsky R, Johnson TE, editors. Religion, Religious Ethics and Nursing. New York: Springer Publishing; 2012.
4. Hislop GT, Bajdik CD, Regier MD, Barroetavena MC. Ethnic differences in survival for female cancers of the breast, cervix and colorectum in British Columbia, Canada. Asian Pac J Cancer Prev 2007;8:209-14.
5. Statistics Canada. Ethnic Diversity and Immigration; 2006. Available from: http://www.5.statcan.gc.ca/subject-sujet/subtheme-soustheme.action?pid=30000&id=-30000&lang=eng&more=0. [Last accessed on 2014 Mar 16].
6. Statistics Canada. Population by Selected Ethnic Origins, by Province and Territory (2006 Census); Canada, 2006. Available from: http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo26k-eng.htm. [Last accessed on 2014 Mar 16].
7. Available from: http://www.cancer.ca/en/cancer-information/cancer-type/breast/statistics/?region=bc. [Last Accessed on 2014 Mar 16]
8. Available from: http://www.cancer.ca/en/cancer-information/cancer-type/breast/statistics. [Last Accessed on 2014 Mar 16].
9. Available from: http://www.cancer.ca/en/about-us/for-media/media-releases/national/2013/south-asians-in-ontario-tremendously-under-screened-for-cancer/?region=bc. [Last Accessed on 2014 Mar 16].
10. Statistics Canada. Census Canada; 2011. Available from: http://www.12.statcan.gc.ca/census-recensement/2006/ref/rp-guides/ethnic-ethnique-eng.cfm. [Last accessed on 2014 Mar 16].
11. Aspinall PJ. The new 2001 Census question set on cultural characteristics: Is it useful for the monitoring of the health status of people from ethnic groups in Britain? Ethn Health 2000;5:33-40.
12. Bhopal R. Glossary of terms relating to ethnicity and race: For reflection and debate. J Epidemiol Community Health 2004;58:441-5.
13. Weber M. What is an ethnic group. In: Guibernau M, Rex J, editors. The Ethnicity Reader. Nationalism, Multiculturalism and Migration. Cambridge: Polity Press; 1997. p. 15-32.
14. Bulmer M. The ethnic group question in the 1991 census of population. In: Coleman D, Salt J, editors. Ethnicity in the 1991 Census. Demographic Characteristics of the Ethnic Minority Populations. Vol. 1. London: Office for National Statistics, HMSO; 1996.
15. Tucker DK. Surnames, forenames and correlations. In: Hanks P, editor. Dictionary of American Family Names. New York: Oxford University Press; 2003.
16. Singh-Carlson S, Neufeld A, Olson J. South Asian immigrant women's experiences of being respected within cancer treatment settings. Can Oncol Nurs J 2010;20:188-98.
17. Singh-Carlson S, Wong F, Martin L, Nguyen SK. Breast cancer survivorship and South Asian women: Understanding about the follow-up care plan and perspectives and preferences for information post treatment. Curr Oncol 2013;20:e63-79.
18. Singh-Carlson S, Nguyen SK, Wong F. Perceptions of survivorship care among South Asian female breast cancer survivors. Curr Oncol 2013;20:e80-9.
19. Screening Mammography Program of British Columbia. British Columbia Cancer Agency. Available from: http://www.bccancer.bc.ca/PPI/Screening/Breast/default.htm. [Last accessed on 2014 Mar 16].
20. Shah BR, Chiu M, Amin S, Ramani M, Sadry S, Tu JV. Surname lists to identify South Asian and Chinese ethnicity from secondary data in Ontario, Canada: A validation study. BMC Med Res Methodol 2010;10:42.
21. Guhuray S. Text Categorization of South Asian Last Names; 2003. Available from: http://www.dimax.rutgers.edu/~sguharay/. [Last Accessed on 2014 Mar 16].
22. Chilana RS, Chilana PK. Dictionary of Sikh Names. New Delhi: UBS Publisher's Distributors; 2001.
23. Coldman AJ, Braun T, Gallagher RP. The classification of ethnic status using name information. J Epidemiol Community Health 1988;42:390-5.
24. Lauderdale D, Kestenbaum B. Asian American ethnic identification by surname. Popul Res Policy Rev 2000;19:283-300.

25. Choi BC, Hanley AJ, Holowaty EJ, Dale D. Use of surnames to identify individuals of Chinese ancestry. Am J Epidemiol 1993;138:723-34.
26. Nitsch D, Kadalayil L, Mangtani P, Steenkamp R, Ansell D, Tomson C, *et al.* Validation and utility of a computerized South Asian names and group recognition algorithm in ascertaining South Asian ethnicity in the national renal registry. QJM 2009;102:865-72.
27. Mason D. Explaining Ethnicity Differences: Changing Patterns of Disadvantage in Britain. Bristol: Policy Press; 2003.
28. 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 Community Health 1998;52:336-7.
29. 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 Community Health 1986;40:364-8.
30. Cummins C, Winter H, Kar-Keung C, Maric R, Silcocks P, Varghese C. An assessment of the Nam Pehchan computer program for the identification of names of South Asian ethnic origin. Fac Public Health Med 1999;21:401-6.
31. 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.
32. Macfarline GJ, Lunt M, Palmer B, Afzal C, Silman AJ, Esmail A. Determining aspects of ethnicity amongst persons of South Asian origin: The use of surname-classification programme (Nam Pehchan). J R Inst Public Health 2007;121:231-6.
33. Ryan R, Vernon S, Lawrence G, Wilson S. Use of name recognition software, census data and multiple imputation to predict missing data on ethnicity: Application to cancer registry records. BMS Med Inform Decis Mak 2012;12:1-8.
34. Sheth T, Nargundkar M, Chagani K, Anand S, Nair C, Yusuf S. Classifying ethnicity utilizing the Canadian mortality data base. Ethn Health 1997;2:287-95.
35. Taylor VM, Nguyen TT, Do HH, Li L, Yasui Y. Lessons learned from the application of a Vietnamese surname list for survey research. J Immigr Minor Health 2011;13:345-51.
36. Harding S, Dews H, Simpson SL. The potential to identify South Asians using a computerised algorithm to classify names. Popul Trends 1999;97:46-9.
37. Ambekar A, Ward C, Mohammed J, Male S, Skiena S. Name-ethnicity classification from open sources. In proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining (Paris, France, June 28-July 1, 2009). KDD 2009. ACM, New York, NY, 49-58. DOI: http://doi.acm.org/10.1145/1557019.1557032
38. Chow TE, Lin Y, Chan WD. The development of a web-based demographic data extraction tool for population monitoring. Trans GIS 2011;15:479-94.
39. Bouwhuis CB, Moll HA. Determination of ethnicity in children in the Netherlands: Two methods compared. Eur J Epidemiol 2003;18:385-8.
40. Fiscella K, Fremont AM. Use of geocoding and surname analysis to estimate race and ethnicity. Health Serv Res 2006;41(4 Pt 1):1482-500.