# Classification of Canadian immigrants into visible minority groups using country of birth and mother tongue

**Mohammad R. Rezai, Laura C. Maclagan, Linda R. Donovan, Jack V. Tu**


## ABSTRACT

**Background:** The Permanent Resident Database of Citizenship and Immigration Canada (CIC) contains socio-demographic information on immigrants but lacks ethnic group classifications. To enhance its usability for ethnicity-related research, we categorized immigrants in the CIC database into one of Canada's official visible minority groups or a white category using their country of birth and mother tongue.

**Methods:** Using public data sources, we classified each of 267 country names and 245 mother tongues in the CIC data into 1 of 10 visible minority groups (South Asian, Chinese, black, Latin American, Filipino, West Asian, Arab, Southeast Asian, Korean, and Japanese) or a white group. We then used country of birth alone (method A) or country of birth plus mother tongue (method B) to classify 2.5 million people in the CIC database who immigrated to Ontario between 1985 and 2010 and who had a valid encrypted health card number. We validated the ethnic categorizations using linked self-reported ethnicity data for 6499 people who responded to the Canadian Community Health Survey (CCHS).

**Results:** Among immigrants listed in the CIC database, the 4 most frequent visible minority groups as classified by method B were South Asian (n = 582 812), Chinese (n = 400 771), black (n = 254 189), and Latin American (n = 179 118). Methods A and B agreed in 94% of the categorizations (kappa coefficient 0.94, 95% confidence interval [CI] 0.93–0.94). Both methods A and B agreed with self-reported CCHS ethnicity in 86% of all categorizations (for both comparisons, kappa coefficient 0.83, 95% CI 0.82–0.84). Both methods A and B had high sensitivity and specificity for most visible minority groups when validated using self-reported ethnicity from the CCHS (e.g., with method B, sensitivity and specificity were, respectively, 0.85 and 0.97 for South Asians, 0.93 and 0.99 for Chinese, and 0.90 and 0.97 for blacks).

**Interpretation:** The use of country of birth and mother tongue is a validated and practical method for classifying immigrants to Canada into ethnic categories.


**Mohammad R. Rezai,** MD, PhD, is a Postdoctoral Fellow in the Cardiovascular Research Program, Institute for Clinical Evaluative Sciences, Toronto, Ontario. **Laura C. Maclagan,** MSc, is an Epidemiologist in the Cardiovascular Research Program, Institute for Clinical Evaluative Sciences, Toronto, Ontario. **Linda R. Donovan,** BScN, MBA, is a Project Manager with Sunnybrook Research Institute, Toronto, Ontario. **Jack V. Tu,** MD, PhD, FRCPC, is a Tier 1 Canada Research Chair in Health Services Research and a Senior Scientist with the Institute for Clinical Evaluative Sciences and the Clinical Epidemiology Program, Sunnybrook Research Institute; a Staff Physician, Division of Cardiology, Schulich Heart Centre, Sunnybrook Health Sciences Centre; and a Professor in the Faculty of Medicine and the Institute of Health Policy, Management and Evaluation at the University of Toronto, Toronto, Ontario.

**Competing interests:** None declared.

**Funding:** This study was supported by operating grants from the Public Health Agency of Canada (PHAC), the Heart and Stroke Foundation of Ontario, and a Team Grant (TCA 118349) to the Cardiovascular Health in Ambulatory Care Research Team (CANHEART) from the Institute of Circulatory and Respiratory Health, Canadian Institutes of Health Research. Additional support was provided by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The immigration data used in the study were provided to ICES by Citizenship and Immigration Canada. These data sets were held securely in a linked, de-identified form and were analyzed at ICES. Dr. Tu is supported by a Canada Research Chair in Health Services Research and a Career Investigator award from the Heart and Stroke Foundation of Ontario. The opinions, results, and conclusions reported in this paper are those of the authors and are independent from, and should not be attributed to, the funding sources. No endorsement by PHAC, ICES, the Ontario MOHLTC, or Citizenship and Immigration Canada is intended or should be inferred.

**Correspondence:** Dr. Jack V. Tu, Institute for Clinical Evaluative Sciences, G-106, 2075 Bayview Ave., Toronto ON M4N 3M5; tu@ices.on.ca

➢ As one of the most ethnically diverse countries,[1] Canada is home to individuals of over 200 ethnic origins.[2] Canada's growing diversity is due primarily to high levels of immigration. Since the 1990s, about 250 000 immigrants have arrived annually.[2] The major sources of Canada's immigrants are Asia, Europe, the Caribbean, South and Central America, Africa, and the United States.[1] In Ontario, Canada's largest province, the 2006 Census

identified that 23% of the population belonged to an ethnic minority group, with the largest groups being South Asian, Chinese, and black.[1] From 2007 to 2011, 42% of all Canadian immigrants landed in Ontario.[3]

The increasingly multi-ethnic nature of society in Canada and other countries is fuelling a need for ethnicity data to permit better understanding of these diverse populations. For example, in health research, ethnicity classifications can be used to better understand the etiology of disease, the respective roles of environment and genetics in health and disease, and the health status of disadvantaged groups, as well as to improve health care delivery and target specific public health interventions toward high-risk populations.[4,5] However, such classifications may also have associated weaknesses, such as contributing to racialized identities, a social concept denoting power inequality between ethnic or racial groups, which has been suggested to have negative health implications.[6]

The concept of ethnicity is complex and its definition challenging.[5,7,8] The concepts of ethnicity and race are sometimes used synonymously, although they do not overlap completely[7]:

- Ethnicity has been defined as "the social group a person belongs to, and either identifies with or is identified with by others, as a result of a mix of cultural and other factors including language, diet, religion, ancestry, and physical features traditionally associated with race."[7]
- Race has been defined "by historical and common usage," as "the group (sub-species in traditional scientific use) a person belongs to as a result of a mix of physical features such as skin colour and hair texture, which reflect ancestry and geographical origins, as identified by others or, increasingly, as self-identified."[7]

Although ethnicity has long been recognized as an important covariate in health research, individual-level ethnicity data are rarely collected in Canadian health care data sets. Similarly, although some ethnicity data are captured in Canada's census, these data are restricted to Statistics Canada and therefore cannot be linked to many other administrative data sets available in Canada's provinces and territories. In an effort to address this gap, alternative ethnicity classifications have been used.

The various methods used to define and assign ethnicity include surname-based approaches, geocoding of residential address, and classification based on country of birth, language, or a combination of these.[9–13] Country of birth in particular has been widely collected in many administrative and government data sets and represents an objective and potentially valuable source of ethnicity information.

Given Canada's high immigration rate, the Permanent Resident Database of Citizenship and Immigration Canada (CIC) may be a useful source of ethnicity data for health research. In the past decade, this database has been used for socio-economic and health studies of immigrants in various Canadian provinces.[14–20] The CIC data provide detailed prelanding demographic and socio-economic information, including country of birth, for all Canadian immigrants. However, this data set lacks self-reported ethnicity, and the large number of options for country of birth and mother tongue within this data set (over 200 options for each variable) can also make it challenging to use for such purposes. In an attempt to improve the practical use of this database for ethnicity-related research projects, we describe here a method for classifying Ontario CIC data records into Canada's 10 official visible minority ethnic groups (plus a white group) using either country of birth or country of birth plus mother tongue variables. We also report validation of this method using information from Statistics Canada's Canadian Community Health Survey (CCHS), a large population-based telephone survey of the Canadian population which includes self-reported ethnicity information, the current "gold standard" for ethnicity classification.

## Methods

**CIC Permanent Resident Database.** The CIC Permanent Resident Database provides detailed socio-demographic information for all legal immigrants to Canada, including country of birth, citizenship, country of last permanent residence, and mother tongue. For this analysis, we used the CIC data set held at the Institute for Clinical Evaluative Sciences, which pertains to Ontario immigrants who arrived between 1985 and 2010. This data set includes 267 options for country of birth and 245 options for mother tongue. The Ontario CIC database has been used as a source of ethnicity data for previous health research studies.[14–16]

**Study population.** The CIC data set used for this study contains records for 2.9 million immigrants who landed in Ontario over the period from 1985 to 2010. We excluded about 400 000 records because of an inability to identify a valid health card number in the

Ontario Registered Persons Database; the health card number was required for record linkage to the self-reported ethnicity data that we used for validation purposes. The reasons for absence of a valid health card number are multifactorial and include immigrants' departure from the province shortly after arrival (i.e., before registering for a health card number), as well as typographic inconsistencies in the CIC database or the Registered Persons Database (or both). Landed immigrants become eligible for health care benefits after a 3-month waiting period. Records for the remaining 2.5 million Ontario immigrants could be linked to other administrative databases available at the Institute for Clinical Evaluative Sciences. All data were de-identified and health card numbers were encrypted to protect privacy.

**Classification of ethnic groups**. The CIC database lacks ethnic or visible minority group classifications. To facilitate use of the CIC data for health research, we tested 2 methods for classifying the immigrants in this data set into 11 ethnic categories, specifically the 10 official visible minority groups used by Statistics Canada (South Asian, Chinese, black, Latin American, Filipino, West Asian, Arab, Southeast Asian, Korean, and Japanese) and a white category. According to the Employment Equity Act, visible minorities are defined as "persons, other than aboriginal peoples, who are non-Caucasian in race or non-white in colour."[21] For this study, we first tested country of birth alone (method A) and then country of birth plus mother tongue (method B) to classify the Ontario CIC data set.

**Method A (country of birth).** We mapped each of the 267 country-of-birth names in the Ontario CIC data set (including previous county-of-birth names changed for political reasons) to 1 of 12 categories: the 10 visible minority groups specified by Statistics Canada, a white category, and an "excluded" category. We used a combination of publicly available resources for this purpose, including Statistics Canada's ethnic origin categories for the 2006 Census of Population (our preferred source),[22] the United Nations *Standard Country or Area Codes for Statistical Use* (also known as the M49 list),[23] the World Bank list of economies (as of July 2012),[24] and *The World Factbook* of the US Central Intelligence Agency.[25] These resources consider the ethnic mix of countries and provide additional information needed to appropriately assign each country to its predominant ethnic group.

The 10 visible minority categories used by Statistics Canada are heterogeneous. Whereas some categories are associated with a single country, and classification is straightforward (e.g., the country of Japan was assigned to the Japanese category), other categories, such as South Asian and Latin American, relate to geographic regions and include multiple countries. For example, we assigned the countries in South America and most of those in Central America to the Latin American category. In contrast, categories such as black and Arab may be considered primarily ethnocultural classifications associated with overlapping geopolitical boundaries (see methodological details in online Appendix A).

The white category was used for European countries and those with populations of predominantly European origin (e.g., Australia). The "excluded" category was created for immigrants whose countries of birth were not accounted for by the 10 major visible minority groups defined by Statistics Canada or the white category as defined above and those whose CIC data were irregular (e.g., "country not stated" or "British Overseas Citizen"). Further details are provided in online Appendix A.

**Method B (country of birth plus mother tongue).** In an effort to further refine the classification based on country of birth, we then completed a second classification based on country of birth plus mother tongue. The ethnic makeup of many countries is heterogeneous, and there may be individuals whose ethnic background differs from the predominant ethnocultural group (or groups) of their country of birth. For instance, a person may be born to South Asian parents in a country with a predominantly white population (e.g., the United Kingdom). In such cases, a person's mother tongue may be more representative of his or her ethnic background than his or her country of birth.

We mapped each of the 245 mother tongues in the Ontario CIC data set to 1 of 15 categories: the 10 Statistics Canada visible minority groups, a white category, and 4 additional language categories ("world language," "other," "excluded," and "unknown"). Publicly available data sources, such as *Ethnologue: Languages of the World*[26] and *The World Factbook*,[25] were used to gather language information and assign each mother tongue to an ethnic group. For instance, Cantonese and Mandarin were categorized as Chinese, and Persian and Kurdish were categorized as West Asian.

The "world language" category, created to account for languages spoken in multiple categories and by various visible minority groups, comprised English, French,

Spanish, Portuguese, and Russian. Less specific language options were assigned to the closest category (e.g., "other European languages" to white) or to the "other" category (e.g., Hebrew) (see methodological details in online Appendix A). The "excluded" language category was created for 3 languages (Busan, Uzbek, Samoan) associated with the countries in the "excluded" category defined in method A (as described in the previous subsection). The "unknown" category was created for languages for which a region of origin or single ethnic group could not be identified. The number of immigrants speaking languages categorized as "excluded" or "unknown" was relatively small (< 0.1% of total sample).

We developed an algorithm to determine a final category for each individual immigrant record using both methods: country of birth (method A) and country of birth plus mother tongue (method B) (see the flow chart in online Appendix B).

**Validation of classification accuracy.** For validation, we compared the ethnic group assigned by each of our 2 methods with self-reported ethnic group data in Statistics Canada's CCHS, a population-based cross-sectional health telephone survey of Canadians aged 12 years and older. More specifically, we used respondents' answers to the CCHS question, "People living in Canada come from many different cultural and racial backgrounds. Are you [white, South Asian, etc.]?" We used encrypted health card numbers to link data from 4 cycles of the CCHS (2000/2001 to 2007/2008) with the CIC data set for Ontario.

We calculated percent agreement and simple kappa statistics to compare classification by methods A and B with the CCHS self-reported ethnic classification (the reference standard). Overall percent agreement was defined as the number of similar ratings by the 2 methods divided by the total number of ratings. Sensitivity, specificity, and positive and negative predictive values were calculated for each visible minority category for comparisons of methods A and B with the CCHS classification. We used SAS version 9.2 (SAS Institute Inc., Cary, North Carolina) for all statistical analyses.

**Geographic visualization.** We used ArcGIS Desktop software version 10 (ESRI, Redlands, California) to create a map showing the global distribution of major ethnic groups associated with the countries of birth of Ontario immigrants, as recorded in the CIC database. A data set for world country boundaries was obtained from the website thematicmapping.org.[27]

**Ethics approval.** This project received ethics approval through the Research Ethics Board of Sunnybrook Health Sciences Centre.

## Results

The study sample from the Ontario CIC data set consisted of 2 500 514 immigrants with mean age ± standard deviation (SD) of 30 ± 17 years at the time of landing, of whom 51% were female. The top 3 countries of origin were India, China, and the Philippines, and the top 3 mother tongues were English, Mandarin, and Cantonese (Table 1).

Figure 1 displays the world distribution of the ethnic groups assigned to countries of birth in our sample. A list of all countries and mother tongues in the Ontario

Table 1

**Top 20 countries of birth and mother tongues of immigrants recorded in the Citizenship and Immigration Canada (CIC) Permanent Resident Database who landed in Ontario from 1985 to 2010**

| | Top 20 countries of birth | | Top 20 mother tongues | |
|---|---|---|---|---|
| Rank | Country of birth* | No. of immigrants | Mother tongue* | No. of immigrants |
| 1 | India | 296 805 | English | 365 194 |
| 2 | China, People's Republic of | 263 450 | Mandarin | 170 317 |
| 3 | Philippines | 163 223 | Cantonese | 166 533 |
| 4 | Pakistan | 134 967 | Tagalog | 143 603 |
| 5 | Sri Lanka | 96 110 | Arabic | 135 219 |
| 6 | Hong Kong | 94 038 | Punjabi | 134 238 |
| 7 | Poland | 78 368 | Urdu | 129 566 |
| 8 | Iran | 74 957 | Spanish | 123 156 |
| 9 | Jamaica | 72 782 | Tamil | 96 376 |
| 10 | United States of America | 60 155 | Russian | 81 746 |
| 11 | United Kingdom and Colonies | 58 180 | Polish | 78 601 |
| 12 | Guyana | 50 643 | Gujarati | 58 070 |
| 13 | Korea, Republic of | 41 005 | Chinese | 48 702 |
| 14 | Vietnam, Socialist Republic of | 39 666 | Portuguese | 48 365 |
| 15 | Romania | 38 223 | Hindi | 45 879 |
| 16 | Trinidad and Tobago, Republic of | 34 819 | Farsi | 43 637 |
| 17 | Yugoslavia | 34 788 | Korean | 41 489 |
| 18 | Russia | 34 523 | Romanian | 37 099 |
| 19 | Iraq | 33 648 | Bengali | 37 024 |
| 20 | Bangladesh | 32 331 | Persian | 33 692 |

*The country and language labels are as per the CIC data formats and therefore may refer to old names in some cases.

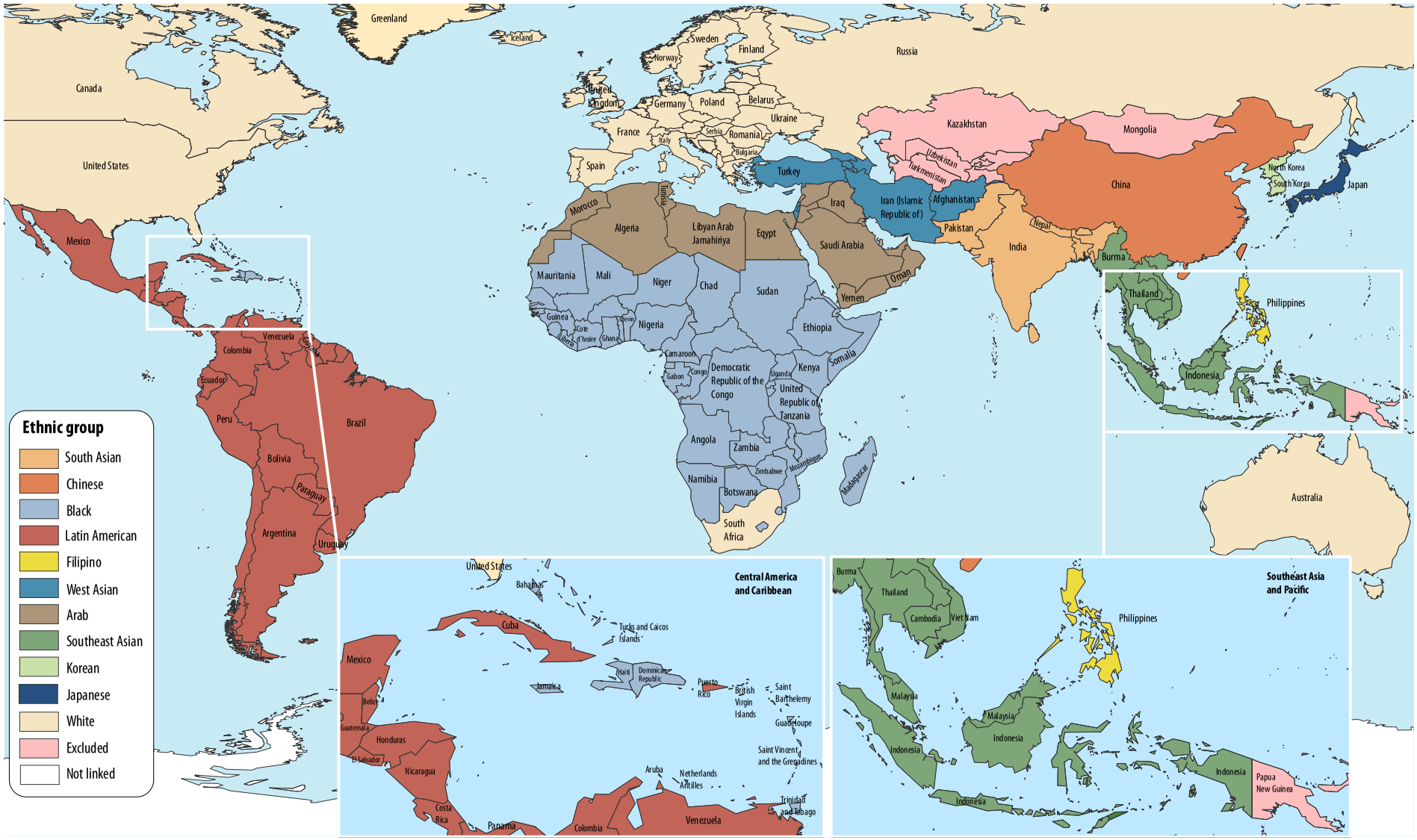


Figure 1

**World distribution of major ethnic groups associated with countries of birth of the immigrants in Citizenship and Immigration Canada's Permanent Resident Database for Ontario (1985–2010).** The national boundaries shown on the map are not an expression of the authors' views on the legal status of territories or the definition of the boundaries.

CIC data with the assigned ethnic categories is available by contacting the corresponding author.

The 2 methods used to classify immigrants (i.e., on the basis of country of birth alone or on the basis of country of birth plus mother tongue) resulted in some differences in categorization (Table 2). For instance, among 523 855 immigrants classified as white by country of birth, 8271 and 2502 individuals were classified as South Asian and Chinese, respectively, by country of birth plus mother tongue. Methods A and B showed agreement for 94% of the ratings (kappa coefficient 0.94, 95% confidence interval [CI] 0.93–0.94).

From the Ontario CCHS data set (n = 134 567), we linked 6585 records to the CIC data set. Of these, 86 individuals with multiple ethnicities were excluded, leaving 6499 for the validation analysis. For these 6499 CCHS respondents, the mean age ± SD was 29 ± 15 years at the time of landing, and 52% were female. For the vast majority of the respondents, self-reported ethnicity in the CCHS data matched the ethnic group assigned by our method B (Table 3) or method A (see online Appendix C).

Ethnic categorization by either method A (country of birth alone) or method B (country of birth plus mother tongue) agreed with the self-reported CCHS ethnic group for 86% of respondents (kappa coefficient 0.83, 95% CI 0.82–0.84, for both comparisons).

When the classification accuracy of method B was compared with self-reported ethnicity from the CCHS, consistently high specificity and negative predictive values were found for all groups (Table 4). Sensitivity for the Southeast Asian category and positive predictive values for the Latin American, Southeast Asian, and West Asian categories were relatively lower. For the majority of classification indices, method B (country of birth plus mother tongue) showed a slight improvement in categorization over method A (country of birth alone) or no change (see online Appendix D for validation results for method A).

Table 2

**Frequency of immigrants to Ontario (1985–2010) in each ethnic category, classified by country of birth alone (method A) and country of birth plus mother tongue (method B), based on data in the Citizenship and Immigration Canada Permanent Resident Database*†**

| | Method B (using country of birth plus mother tongue) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method A (using country of birth) | Excluded | White | South Asian | Chinese | Black | Latin American | Filipino | West Asian | Arab | Southeast Asian | Korean | Japanese | Total |
| **Excluded** | **11 406** | <70 | <70 | <70 | <70 | <70 | <70 | <70 | <70 | <70 | <70 | <70 | 11 406 |
| **White** | 1 312 | **501 235** | 8 271 | 2 502 | 1 785 | 2 007 | 600 | 2 357 | 3 063 | 310 | 331 | 82 | 523 855 |
| **South Asian** | 3 449 | 164 | **547 439** | 5 768 | 541 | <70 | 94 | 4 538 | 286 | 1 068 | <70 | <70 | 563 361 |
| **Chinese** | 1 592 | <70 | 320 | **372 298** | 174 | <70 | 212 | <70 | <70 | 736 | 175 | <70 | 375 670 |
| **Black** | 2 737 | 498 | 9 687 | 445 | **249 547** | <70 | 124 | <70 | 8 160 | 233 | <70 | <70 | 271 499 |
| **Latin American** | 2 053 | 3 251 | 74 | 569 | 123 | **177 111** | <70 | <70 | 156 | <70 | 86 | <70 | 183 491 |
| **Filipino** | 405 | <70 | 548 | 1 699 | <70 | <70 | **148 013** | <70 | <70 | 12 115 | <70 | <70 | 162 968 |
| **West Asian** | 4 413 | 1 168 | 1 564 | <70 | 143 | <70 | <70 | **119 945** | 3 202 | 1 544 | <70 | <70 | 132 067 |
| **Arab** | 8 559 | 329 | 12 214 | <70 | 1 766 | <70 | 1 276 | 12 569 | **122 874** | <70 | <70 | <70 | 159 689 |
| **Southeast Asian** | 1 250 | 102 | 2 517 | 17 008 | <70 | <70 | 244 | <70 | <70 | **47 556** | <70 | <70 | 68 817 |
| **Korean** | 232 | <70 | <70 | 94 | <70 | <70 | <70 | <70 | <70 | <70 | **40 680** | <70 | 41 106 |
| **Japanese** | 174 | <70 | 166 | 329 | <70 | <70 | 217 | <70 | <70 | <70 | 98 | **5 506** | 6 585 |
| **Total** | 37 582 | 506 882 | 582 812 | 400 771 | 254 189 | 179 118 | 150 898 | 139 608 | 137 842 | 63 652 | 41 452 | 5 708 | 2 500 514 |

*Specific data for cells with value < 70, including cells with value 0, were suppressed to protect privacy. Row and column totals are the true sums, including the suppressed values.
†Values in cells along the diagonal are shown in bold to highlight similar classification by the 2 methods.

Table 3

**Frequency of immigrants to Ontario in each self-reported ethnic category (based on the Canadian Community Health Survey [CCHS]) and as classified using country of birth plus mother tongue (method B), based on data in the Citizenship and Immigration Canada Permanent Resident Database*†**

| | Method B (using country of birth plus mother tongue) | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Self-reported (CCHS) | Excluded | White | South Asian | Chinese | Black | Latin American | Filipino | West Asian | Arab | Southeast Asian | Korean | Japanese | Total |
| **White** | 40 | **2021** | <10 | <10 | 19 | 91 | <10 | 74 | 40 | <10 | <10 | <10 | 2300 |
| **South Asian** | 26 | <10 | **1105** | <10 | 60 | 66 | <10 | 13 | <10 | <10 | <10 | <10 | 1296 |
| **Chinese** | <10 | <10 | < 10 | **720** | <10 | <10 | 12 | <10 | <10 | 22 | <10 | <10 | 769 |
| **Black** | <10 | 12 | < 10 | <10 | **586** | 26 | <10 | <10 | 15 | <10 | <10 | <10 | 645 |
| **Latin American** | <10 | <10 | < 10 | <10 | 11 | **364** | <10 | <10 | <10 | <10 | <10 | <10 | 389 |
| **Filipino** | <10 | <10 | < 10 | <10 | <10 | <10 | **321** | <10 | <10 | 27 | <10 | <10 | 355 |
| **West Asian** | <10 | <10 | 10 | <10 | 16 | 11 | <10 | **155** | 16 | <10 | <10 | <10 | 224 |
| **Arab** | <10 | <10 | <10 | <10 | <10 | <10 | <10 | 12 | **185** | <10 | <10 | <10 | 212 |
| **Southeast Asian** | <10 | <10 | 96 | 12 | <10 | <10 | 12 | <10 | <10 | **61** | <10 | <10 | 205 |
| **Korean** | <10 | <10 | <10 | <10 | <10 | <10 | <10 | <10 | <10 | <10 | **80** | <10 | 83 |
| **Japanese** | <10 | <10 | <10 | <10 | <10 | <10 | <10 | <10 | <10 | <10 | <10 | **19** | 21 |
| **Total** | 96 | 2059 | 1227 | 735 | 715 | 562 | 354 | 259 | 262 | 128 | 82 | 20 | 6499 |

*Specific data for cells with value < 10, including cells with value 0, were suppressed to protect privacy. Row and column totals are the true sums, including the suppressed values.

†Values in cells along the diagonal are shown in bold to highlight similar classification by the 2 methods.

Table 4

**Validation of ethnic classification using country of birth plus mother tongue (method B), with self-reported ethnicity (from Canadian Community Health Survey) as reference (n = 6499)**

| Classification by method B | Sensitivity (95% CI) | Specificity (95% CI) | Positive predictive value (95% CI) | Negative predictive value (95% CI) |
|---|---|---|---|---|
| White | 0.87 (0.86–0.89) | 0.99 (0.98–0.99) | 0.98 (0.97–0.98) | 0.93 (0.92–0.94) |
| South Asian | 0.85 (0.83–0.87) | 0.97 (0.97–0.98) | 0.90 (0.88–0.91) | 0.96 (0.95–0.96) |
| Chinese | 0.93 (0.91–0.95) | 0.99 (0.99–0.99) | 0.97 (0.96–0.98) | 0.99 (0.98–0.99) |
| Black | 0.90 (0.88–0.92) | 0.97 (0.97–0.98) | 0.81 (0.78–0.84) | 0.98 (0.98–0.99) |
| Latin American | 0.93 (0.90–0.95) | 0.96 (0.96–0.97) | 0.64 (0.60–0.68) | 0.99 (0.99–0.99) |
| Filipino | 0.90 (0.86–0.93) | 0.99 (0.99–0.99) | 0.90 (0.87–0.93) | 0.99 (0.99–0.99) |
| West Asian | 0.69 (0.62–0.75) | 0.98 (0.98–0.98) | 0.59 (0.53–0.65) | 0.98 (0.98–0.99) |
| Arab | 0.87 (0.82–0.91) | 0.98 (0.98–0.99) | 0.70 (0.64–0.76) | 0.99 (0.99–0.99) |
| Southeast Asian | 0.29 (0.23–0.36) | 0.98 (0.98–0.99) | 0.47 (0.38–0.56) | 0.97 (0.97–0.98) |
| Korean | 0.96 (0.89–0.99) | 0.99 (0.99–1.00) | 0.97 (0.91–0.99) | 0.99 (0.99–0.99) |
| Japanese | 0.90 (0.69–0.98) | 0.99 (0.99–1.00) | 0.95 (0.75–0.99) | 0.99 (0.99–1.00) |

CI = confidence interval.

## Interpretation

We used 2 methods (country of birth alone or country of birth plus mother tongue) to classify Ontario immigrants in the CIC data set into 11 predefined ethnic groups. We found a high degree of agreement between self-reported ethnic groups from CCHS data and those assigned by our 2 classification methods. Compared with country-specific or world region–specific classifications used previously,[14] our classification by visible minority groups may be more practical for researchers and health policy-makers, as it is comparable to other important population statistics on visible minorities produced by Statistics Canada and other international organizations. Our methods may also prove useful (with local customization) in other countries where health-related information regarding self-reported ethnicity is not available or is not routinely collected but data on immigrants' country of birth and/or mother tongue are available.

Using country of birth to define ethnicity has been reported as a robust method for health care research in countries such as the Netherlands, where this variable was closely correlated with self-reported ethnicity.[13,28] Nevertheless, this method has been criticized,[13,29] because the definition of ethnicity is complex and may not always be determined by geography. Problems can arise with multi-ethnic countries (e.g., Australia, the United States, South Africa) or with individuals born to a family whose ethnicity is different from the predominant ethnic group of their country of residence. To further investigate this issue, we analyzed the CCHS self-reported ethnicity of a subset of the immigrants in the CIC data who had a CCHS-linked record and came from a large, multi-ethnic country (i.e., United Kingdom, United States, South Africa, or Australia, as defined by country of birth in CIC data). Among these CIC–CCHS linked records, 292 (94%) of the 310 immigrants born in the United Kingdom self-identified as white, as did 179 (87%) of the 206 immigrants born in the United States, 56 (89%) of the 63 born in South Africa, and 24 (96%) of the 25 born in Australia. These data support the validity of our ethnicity classification algorithm for immigrants from these countries. Classification methods that use additional information such as language and parents' country of birth have been shown to improve classification accuracy over methods based on country of birth alone.[13] In our study, adding mother tongue to country of birth resulted in only slight improvements in ethnicity classification, relative to country of birth alone. Methods using mother tongue alone to define ethnicity also have their limitations. Second- or third-generation immigrants in some countries (e.g., the United States) may not share the mother tongue of their ancestors. Moreover, native individuals may report their mother tongue to be a world language (originating from a predominantly white country) that is accepted as their birth country's official language (e.g., French in Congo, English in India). We controlled for the latter problem in our data set by recording the world languages spoken as official languages specific to such countries.

Despite unanimously high specificity values, the sensitivity of our methods to detect Southeast Asian immigrants was low relative to self-reported ethnicity. Among Southeast Asians there was considerable misclassification into the South Asian category. This result may be due to individuals' uncertainty about world geographic boundaries for South Asia and Southeast Asia (e.g., a South Asian might think that his or her country is located in Southeast Asia) or self-identification by country of residence rather than country of birth (e.g., a person born in India who lived in Malaysia for a long time before immigrating to Canada may self-identify as Southeast Asian). We also found relatively low positive predictive value for the Latin American group, despite the high sensitivity and specificity of our methods. Some immigrants from certain Latin American countries (e.g., Brazil and Argentina) are descendants of European immigrants and self-identify as white. Moreover, a large proportion of the population in Guyana, the Latin American country with the largest number of immigrants to Ontario, are members of the South Asian diaspora.

Some limitations may exist for the CCHS data that we used to validate our classification methods. First, the CCHS population linked to our Ontario CIC data set may not be a representative sample of Ontario immigrants. Second, the sample size for some visible minority groups (e.g., Japanese and Koreans) was limited, which can result in less reliable estimates. Third, self-reported ethnicity, although often considered a preferred method of ethnic group classification, has some shortcomings. Self-reported ethnicity may change over time and can be influenced by psychosocial factors, such as the feeling of pride that a person attaches to his or her ethnic or national identity, uncertainty about ethnic origin, or even concern related to disclosing one's ethnicity.[5,7,8,13,30,31] For instance, 5% of the immigrants who self-reported as white in the CCHS validation data set were classified as West Asian or Arab on the basis of country of birth and mother tongue. It is likely that these CCHS participants were West Asians or Arabs who reported their ethnicity as white on the basis of skin colour.

In conclusion, in a large data set of Ontario immigrants, we found close agreement between self-reported ethnic categories and ethnic categories based on country of birth alone or country of birth plus mother tongue. These findings suggest that the 2 methods of ethnic classification described are valid for categorizing most immigrants to Canada into the country's official visible minority groups. Use of a larger validation data set in future studies may further illuminate the external validity of these methods.

**Contributors:** Mohammad R. Rezai and Laura C. Maclagan contributed to the literature review, drafted the first version of the manuscript, and revised it critically for important intellectual content. Mohammad R. Rezai also classified the countries and languages and performed statistical and geographic analyses. Linda R. Donovan participated in the analysis and interpretation of the data and revised the manuscript critically for important intellectual content. Jack V. Tu conceived and designed the project,

contributed to data acquisition, revised the manuscript critically for important intellectual content, obtained funding, and supervised the entire project. All authors gave final approval for publication.

## References


1. *Visible minority population and population group reference guide, 2006 Census.* Ottawa (ON): Statistics Canada; 2008. Catalogue no. 97-562-GWE2006003. Available from: www12.statcan.gc.ca/census-recensement/2006/ref/rp-guides/ethnic-ethnique-eng.cfm (accessed 2012 Oct 17).
2. *Ethnic diversity and immigration.* Ottawa (ON): Statistics Canada; 2010. Available from: www41.statcan.gc.ca/2009/30000/cybac30000_000-eng.htm (accessed 2012 Dec 21).
3. *Table 051-0011: International migrants, by age group and sex, Canada, provinces, and territories, annual (persons).* Ottawa (ON): Statistics Canada; 2012. Available from: www.statcan.gc.ca/pub/91-214-x/2009000/related-connexes-eng.htm (accessed 2012 Oct 5).
4. Lin SS, Kelsey JL. Use of race and ethnicity in epidemiologic research: concepts, methodological issues, and suggestions for research. *Epidemiol Rev* 2000;22(2):187–202.
5. Mays VM, Ponce NA, Washington DL, Cochran SD. Classification of race and ethnicity: implications for public health. *Annu Rev Public Health* 2003;24:83–110.
6. Veenstra G. Racialized identity and health in Canada: results from a nationally representative survey. *Soc Sci Med* 2009;69(4):538–542.
7. Bhopal R. Glossary of terms relating to ethnicity and race: for reflection and debate. *J Epidemiol Community Health* 2004;58(6):441–445.
8. Kaplan JB, Bennett T. Use of race and ethnicity in biomedical publication. *JAMA* 2003;289(20):2709–2716. Erratum in: *JAMA* 2004;292(9):1022.
9. Elliott MN, Fremont A, Morrison PA, Pantoja P, Lurie N. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Serv Res* 2008;43(5 Pt 1):1722–1736.
10. Fiscella K, Fremont AM. Use of geocoding and surname analysis to estimate race and ethnicity. *Health Serv Res* 2006;41(4 Pt 1):1482–1500.
11. Mateos P. A review of name-based ethnicity classification methods and their potential in population studies. *Popul Space Place* 2007;13(4):243–263.
12. 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.
13. Stronks K, Kulu-Glasgow I, Agyemang C. The utility of 'country of birth' for the classification of ethnic groups in health research: the Dutch experience. *Ethn Health* 2009;14(3):255–269.
14. Creatore MI, Moineddin R, Booth G, Manuel DH, DesMeules M, McDermott S, et al. Age- and sex-related prevalence of diabetes mellitus among immigrants to Ontario, Canada. *CMAJ* 2010;182(8):781–789.
15. Guttmann A, Manuel D, Stukel TA, DesMeules M, Cernat G, Glazier RH. Immunization coverage among young children of urban immigrant mothers: findings from a universal health care system. *Ambul Pediatr* 2008;8(3):205–209.
16. Urquia ML, Ying I, Glazier RH, Berger H, De Souza LR, Ray JG. Serious preeclampsia among different immigrant groups. *J Obstet Gynaecol Can* 2012;34(4):348–352.
17. Walton-Roberts MW. Immigration, the university and the welcoming second tier city. *Int Migr Integr* 2011;12(4):453–473.
18. Lin Z. *Foreign-born vs native-born Canadians: a comparison of their inter-provincial labour mobility.* Ottawa (ON): Statistics Canada; 1998. Catalogue no. 11F0019MIE1998114.
19. Wang S, Lo L. Chinese immigrants in Canada: their changing composition and economic performance. *Int Migr* 2005;43(3):35–71.
20. Akbari AH, Lynch S, McDonald J, Rankaduwa W. *Socioeconomic and demographic profiles of immigrants in Atlantic Canada.* Halifax (NS): Atlantic Metropolis Centre; 2007.
21. *Employment Equity Act,* S.C. 1995, c. 44. Available from: http://laws-lois.justice.gc.ca/eng/acts/E-5.401/ (accessed 2013 Feb 12).
22. *Ethnocultural portrait of Canada highlight tables, 2006 Census.* Ottawa (ON): Statistics Canada; 2008. Catalogue no. 97-562-XWE2006002.
23. *Standard country or area codes for statistical use.* New York (NY): United Nations, Statistics Division; 2012. Available from: http://unstats.un.org/unsd/methods/m49/m49.htm (accessed 2012 Nov 20).
24. *How we classify countries.* Geneva (Switzerland): World Bank; 2012. Available from: http://data.worldbank.org/about/country-classifications (accessed 2012 Nov 20).
25. *The world factbook 2012–13.* Washington (DC): US Central Intelligence Agency; 2012. Available from: www.cia.gov/library/publications/the-world-factbook/index.html (accessed 2012 Nov 20).
26. Lewis MP, editor. *Ethnologue: languages of the world.* Dallas (TX): SIL International. Available from: www.ethnologue.com (accessed 2012 Jul 30).
27. Sandvik B. *World borders dataset.* Self-published; 2009. Available from: http://thematicmapping.org/downloads/world_borders.php (accessed 2012 Nov 21).
28. Haasnoot A, Koedijk FD, Op De Coul EL, Götz HM, van der Sande MA, Van DB IV; CSI Group. Comparing two definitions of ethnicity for identifying young persons at risk for chlamydia. *Epidemiol Infect* 2012;140(5):951–958.
29. Gill PS, Bhopal R, Wild S, Kai J. Limitations and potential of country of birth as proxy for ethnic group [letter]. *BMJ* 2005;330(7484):196.
30. Smith FD, Woo M, Austin SB. 'I didn't feel like any of those things were me': results of a qualitative pilot study of race/ethnicity survey items with minority ethnic adolescents in the USA. *Ethn Health* 2010;15(6):621–638.
31. Morning A. Ethnic classification in global perspective: a cross-national survey of the 2000 census round. *Popul Res Policy Rev* 2008;27(2):239–272.


**Published:** 1 October 2013

**Citation:** Rezai MR, Maclagan LC, Donovan LR, Tu JV. Classification of Canadian immigrants into visible minority groups using country of birth and mother tongue. *Open Med* 2013;7(4):e85–e93.

**Copyright:** Open Medicine applies the Creative Commons Attribution Share Alike License, which means that anyone is able to freely copy, download, reprint, reuse, distribute, display or perform this work and that authors retain copyright of their work. Any derivative use of this work must be distributed only under a license identical to this one and must be attributed to the authors. Any of these conditions can be waived with permission from the copyright holder. These conditions do not negate or supersede Fair Use laws in any country. For more information, please see http://creativecommons.org/licenses/by-sa/2.5/ca/.