POPULATION, SPACE AND PLACE
Popul. Space Place 13 (2007) *in press*

# *A Review of Name-based Ethnicity Classification Methods and their Potential in Population Studies.*

**Pablo Mateos ***
Department of Geography, University College London

* Department of Geography, University College London, Gower Street, London WC1E 6BT Tel: +44 (0) 20 7679 0500 / Fax: +44 (0) 20 7679 0565, e-mail: p.mateos@ucl.ac.uk

**Abstract:**

*BACKGROUND*: Several approaches have been proposed to classify populations into ethnic groups using people's names, as an alternative to ethnicity self-identification information when this is not available. These methodologies have been developed, primarily in the Public Health and Population Genetics literature in different countries, in isolation from and with little participation from demographers or social scientists.

*OBJECTIVE*: To bring together these isolated efforts and provide a coherent comparison, a common methodology and terminology in order to foster new research and applications in this promising and multi-disciplinary field.

*METHODS*: A systematic review of the most representative studies that develop new name-based ethnicity classifications has been conducted, extracting methodological commonalities, achievements and shortcomings.

*FINDINGS*: 13 studies met the inclusion criteria and all followed a very similar methodology to create a name reference list with which to classify populations into a a few most common ethnic groups. The different classifications' *sensitivity* varies

between 0.67 and 0.95, their *specificity* between 0.80 and 1, their *positive predicted value* between 0.70 and 0.96, and their *negative predicted value* between 0.96 and 1.

*CONCLUSION*: Name-based ethnicity classification systems have a great potential to overcome data scarcity issues in a wide variety of key topics in population studies, as have been proved by the 13 papers analysed. Their current limitations are mainly due to a restricted number of names and a partial spatio-temporal coverage of the reference population datasets used to produce name reference lists.

*RECOMMENDATION*: Improved classifications with extensive population coverage and higher classification accuracy levels will be achieved by using population registers with wider spatio-temporal coverage. Furthermore, there is a requirement for such new classifications to include all of the potential ethnic groups present in a society, and not just one or a few of them.

**Keywords**:

Name origins, ethnicity classifications, identity measurement, inter-disciplinary methods, surnames

## 1. Introduction

Since the last decade and a half, there has been an explosion of interest in issues of ethnicity, nationalism, race and religion, around a renewed preoccupation with the question of defining and asserting collective identities in an increasingly globalised world (Castells, 1997). Governments and social scientists have struggled to keep track of the reality of rapidly changing populations that are constantly re-defining their self-perceptions of their collective identities (Skerry, 2000). Although highly contested, the practice of classifying the population into discrete groups according to race, ethnicity or religion has made a strong re-appearance in many countries' recent national censuses (Howard and Hopkins, 2005, Kertzer and Arel, 2002, Nobles, 2000). Such questions in the censuses not only quantify the size and geographical extent of collectively pre-perceived racial, ethnic and religious groups, but more interestingly helps to reinforce the self-identity of those groups or accelerate the emergence of new identities (Christopher, 2002) by solidifying transient labels (Howard and Hopkins, 2005).

Due to the subjective nature of collective identities, its categorization process, that is, the problematic definition of ethnic groups' boundaries and labels, has been a significant issue in social science (Peach, 1999). Following an impassioned debate around the essentialism of ethnicity labels (Modood, 2005), there seems to be a consensus, at least in the demographic and public health literature, that the classification of populations into ethnic groups has proven useful to fight discrimination and entrenched health and social inequalities (Bhopal, 2004, Mitchell et al, 2000). There is a vast literature that demonstrates the persistence of stark inequalities between ethnic groups, specially in health outcomes, access to housing and labour markets, educational

outcomes and socioeconomic status (for a review in Britain see Mason, 2003). As long as these inequalities between population subgroups persist, no matter how these are defined or perceived, the use of ethnic group definitions and labels will be useful to identify them and combat their causes. However, several of the current ethnicity classification practices have proved inappropriate to uncover the true nature of specific factors of ethnic minorities' inequalities. This paper summarises these issues, before reviewing an alternative methodology of classifying populations into ethnic groups using the origins of people's names.

The basic hypothesis of this methodology is that the classification of surnames and forenames into ancestral groups of origin provides a viable alternative to subdivision of populations or classifications of neighbourhoods into groups of common origin. This is of particular importance when ethnicity, linguistic or religious data are not available at appropriate temporal, spatial or nominal (number of categories) resolutions. The paper reviews the different theoretical and methodological approaches that have developed independently in the fields of public health/epidemiology, population genetics, linguistics, and statistics. The purpose is to bring together these isolated efforts from very different research angles, so far reduced to the study of a small number of ethnic groups in a few migration destination countries, and provide a coherent comparison, a common methodology and terminology in order to foster new research and applications in this promising and multi-disciplinary field.

## 2. Defining and Measuring Ethnicity and Race

The term 'ethnicity' is derived from the Greek '*ethnos*' meaning 'nation', and thus is closely related to the concept of 'peoples' that share a perceived common ancestry or descent (Weber, 1997[1922]). Therefore, at the core of the concept of ethnicity is the question of an individual's identity, which is defined by the characteristics of the ethnic group he or she considers herself to belong to, usually understood in a contextual rather than in an essentialist way (Peach, 1996). Ethnicity is a multi-faceted concept comprised of the different dimensions that makes a person's identity, usually summarized as kinship, religion, language, shared territory, nationality, and physical appearance (Bulmer, 1996).

Due to the subjective, multi-faceted and changing nature of ethnic identification and because there is not a clear consensus on what constitutes an 'ethnic group' (Coleman and Salt, 1996, Office for National Statistics, 2003) the measurement of ethnicity is even more contentious than its definition. This paper will not go any further in the dense debate over the definition of Ethnicity and Race and its measurement as scientific research variables. Literature reviews are available in public health (Bhopal, 2004, Senior and Bhopal, 1994), genetics (Cavalli-Sforza, 1997), geography (Coleman and Salt, 1996), sociology (Banton, 1998, Brubaker, 2004), and anthropology (Eriksen, 2002). For a review of how ethnicity has been measured in 141 countries' recent population Censuses see Morning (forthcoming).

There are three major problems with the way ethnicity is currently officially measured in most developed countries. Firstly, ethnicity is usually measured as a single variable, that of an 'ethnic group' into which the individual self-assigns his or herself from a classification of a reduced number of classes, what restricts its ability to represent the characteristics of the multi-faceted nature of self-identity exposed above. This problem has been partially addressed in the U.S. 2000 Census in which respondents were able to choose from more than one 'race/ethnic group'.

A second problem is that pre-conceived ethnic group classifications are used, as opposed to just an open question whose responses are then arranged according to the more meaningful common identities. This is of course justified with the need to facilitate the reproduction and comparison of the resulting statistics over time and between different sources (Office for National Statistics, 2003). However, these categories have proved not to reflect the complex heterogeneity found within each group (Connolly and Gardener, 2005, Rankin and Bhopal, 1999), for example 'Black African' (Agyemang et al, 2005), 'Asian' (Aspinall, 2003), 'White' (Peach, 2000), or 'Hispanic' (Choi and Sakamoto, 2005). Efforts made to reach a consensus between the major stakeholders of government statistics on a set of meaningful ethnic categories comprises a highly contested issue in the arena of identity politics (Skerry, 2000). Furthermore, such categories are always contextual to a country and moment in time (Peach, 2000), according to each society's response to their own particular historical processes of ethnogenesis (Eriksen, 2002).

A third problem comes with the current consensus in the method of self-assessment of ethnicity (Bhopal, 2004), as opposed to it being assigned by a third person or a computer. As a result of this, the classification of the same person can vary in time and space, since perceptions of individual and social identity changes over time (Aspinall, 2000) and are influenced by the type of ethnicity question asked (Arday et al, 2000), the definitions of categories offered (Olson, 2002), the country and method of data collection, and the time or generations passed since migration (degree of 'assimilation').

In addition to these three major issues, there is a recognised problem of lack of routine collection of ethnicity data in most government or public service datasets, which is especially critical in population registers, such as birth, death, electoral and health general practice registrations (London Health Observatory, 2003, Nanchahal et al, 2001). Even when ethnicity information is routinely collected, such as in U.K. Hospital Admissions, its quality, consistency and coverage is very poor (London Health Observatory, 2005), despite its critical importance in public policy decisions (Department of Health, 2005). As a result, the only major trustworthy source of ethnicity information is usually censuses of population, which are generally only carried out every ten years and results disseminated only in aggregated form.

Taken together, the issues of lack of reflection upon the multi-dimensional nature of ethnicity, the use of a limited range of pre-defined coarse categories, the variability of self-assignment of ethnicity, and the lack of routine collection of ethnicity information present a major shortcoming for researchers and public policy decision makers. As a consequence of these issues, they are frustrated in measuring socioeconomic

inequalities, equity of access to and uptake of public services, and demonstration of compliance with anti-discrimination and equal opportunities legislation. These are each important issues in increasingly multicultural populations.

Due to these issues other proxies, such as country of birth, have been used to ascribe a person's ethnicity when it is not appropriately known for the purpose of analysis (Marmot et al, 1984, Wild and McKeigue, 1997). Despite the utility of country of birth to classify migrants' origins, with growing numbers of second generation ethnic minorities born in the 'destination' or 'host' country (e.g. 50% of ethnic minority members in the U.K. 2001 Census), the proportion of people of the 'majority ethnicity' born abroad, and migrants born in 'intermediate' countries (e.g. East African Indians), this method has become increasingly inappropriate (Gill et al, 2005, Harding et al, 1999). In some countries, such as Spain or France, an alternative variable used is nationality, which is not recorded in many countries (such as in the U.K. Census of Population). This proxy is also problematic since it can change over time, there are people with more than one nationality, and usually second or third generation migrants acquire the host country's nationality. A third option is the analysis of name origins (surname and forename), which in particular has been used to identify South Asian, Chinese and Hispanic populations, with a relatively high degree of accuracy; this will be the focus of the rest of this paper.

As already mentioned, ethnicity is a multidimensional concept reflecting kinship, religion, language, shared territory, nationality, and physical appearance (Bulmer, 1996). In principle one could accurately classify a person into an ethnic group if these

six dimensions were to be measured separately, which is the preferred way forward proposed by several researchers in health and ethnicity (Bhopal, 2004, Gerrish, 2000, McAuley et al, 1996), although physical appearance seems to be a much more sensitive aspect to ask about and even more to classify, than the other five dimensions. Name origin analysis has the potential to provide embedded information about several of these dimensions of a person's origins, when no other ethnicity information is available, since names are usually unique to a language, a religion, a geographical area, a cultural tradition, a group of kin, a migration flow, etc. Although name analysis does not completely overcome the three major problems with the way ethnicity is currently officially measured, mentioned in this section, it does have the potential to substantially improve the situation at a fraction of the cost of other alternatives, as it will be explained through this paper.

## 3. Languages, Names, Genes and Human Origins

Charles Darwin's 'On the Origins of the Species' (1859) included a parallelism between the evolution of languages and humans, suggesting that the genealogical arrangement of the 'races of man' necessarily had to follow a taxonomy of languages.

> *'It may be worth while to illustrate this view of classification, by taking the case of languages. If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world'* (Darwin, 1859, 422).

With the subsequent advances in modern genetic techniques, population geneticists have demonstrated the existence of such a relationship in human evolution, mapping human origins, gene evolution, and geographical spread and intermixing through the

planet and comparing it to language evolution and the archaeological record (Cavalli-Sforza and Cavalli-Sforza, 1995, Piazza et al, 1987).

Moreover, in order to analyse the genetic linkages between human groups, the Human Genome Diversity Project has defined those human groups, called 'populations', by the common mother language of the subjects to be studied (M'charek, 2005), avoiding cases where there is known to have been a historic language replacement (e.g. Spanish imposed to Native Americans, or Finno-Urgic language to Hungarians: Cavalli-Sforza, 1997). They then compare the genetic linkages between such populations (i.e. an evolutionary tree) with the language taxonomy most widely accepted, that of Greenberg and Ruhlen (Ruhlen, 1987), to corroborate the geographical spread or explain the differences with historical data (Cavalli-Sforza et al, 1988).

Furthermore, due to a known relationship between surnames distribution and population structure (Piazza et al, 1987), surnames have been used since the 19th century to understand the relationships between population subgroups (Darwin, 1875) at regional or national levels (for a review see Lasker, 1985). Today, surnames have been demonstrated to correlate well with Y-chromosomes, since both are patrilinearly inherited (Jobling, 2001, McEvoy and Bradley, 2006), and this is opening up a new era of genetic genealogy (Shriver and Kittles, 2004). Moreover, a recent extensive study of the surname distribution of the total population of eight European countries has concluded that the present surname structure of Western Europe is strictly linked to local languages (Scapoli et al, forthcoming).

The combined facts that; first, surnames correlate well with Y-chromosomes at the regional and national level; second, several genetic markers also significantly correlate with languages at a continental and global scale; and, third, there is an obvious link between names and the languages from which they originate, indicate that analysis of people's names can offer a reliable method to ascribe individuals to common human groups, where such groups are defined as having a common linguistic, geographic and ethnic origin. There is a vast literature on surnames and genetics, which has made great advances in disentangling ancestral human movements and distant historic settlement and migrations, as well as to study populations genetic structure, endogamy, and cultural evolution (for a full review see Colantonio et al, 2003, Lasker, 1985). This paper will not cover these aspects but will only focus on name origin analysis to classify contemporary populations according to recent migrations (their own or that of their ancestors to three or four generations back).

The use of people's names' origins to subdivide contemporary populations into ethnic groups has been applied to population studies in the U.S. at least since the beginning of the 20$^{th}$ century, (Rossiter, 1909). Initial applications were primarily focussed on the purpose of calculating immigration quotas, which were set according to the estimated ethnic composition of the "original national stock" of the population of the United States in the 1790 Census (American Council of Learned Societies, 1932, US Senate, 1928). However, name origin analysis has been more widely applied and independently validated in the fields of public health and genetics, in studies since the 1950s (Lasker, 1985, U.S. Bureau of the Census, 1953, Winnie, 1960). The application of such techniques has grown very rapidly through the past 20 years, following increasing

interest in research in international migration, improvements in computer processing power, and (most importantly) with the wider availability of digital name datasets covering entire populations at individual person level. Given this interest in name-based techniques, and the known limitations about their accuracy (Choi et al, 1993), a few studies have concentrated upon measuring the accuracy of different name-based ethnicity classification methods, a stream of research opened by Nicoll et al (1986) and with growing interest and relevance today (Nanchahal et al, 2001).

Hereinafter two types of personal names will be distinguished and named as follows; surnames (also known as family names or last names), which normally correspond to the components of a person's name inherited from his or her family, and forenames (also known as first names, given names, or Christian names), which refer to the proper name given to a person usually at birth.

## 4. A review of Name-based Ethnicity Classification methods

A literature search has been carried out to identify the most representative research papers that specifically deal with the problem of classifying lists of names of individuals into ethnic groups and provide a full evaluation of their accuracy. This section presents a summary of this review, the main characteristics of the studies evaluated, and the results of the comparison.

### 4.1 Search Strategy

The literature search was carried out using three databases of scholarly publications; PubMed Medline, ISI Web of Knowledge (CrossSearch), and Google Scholar. The keywords and search string used to search these databases were:

1) [ethnic* OR race OR racial OR minorit* OR migrant* OR immigrant*]; in the title, keywords or abstract of the publication (abstract not used for Google Scholar)

AND

2) [name* OR surname* OR forename*]; only in the title or keywords of the publication (due to the common use of the word "name" in abstracts)

This search retrieved 186 unique publications at the time (January 2006)

The inclusion criterion was to select any study that developed or used a name-based ethnicity classification method to subdivide contemporary populations at the individual level, and evaluated its accuracy. On the other hand, the exclusion criteria were; a) studies that neither offered a new method of name-based ethnicity classification, nor evaluated a previously developed method that had not been tested before; b) studies that did not validate the classification using an alternative ethnicity information source (i.e. non-name-based); c) studies that provided insufficient detail of their research process and results as to support this systematic review, for which at least the methods' sensitivity and specificity needed to be explicit, and d) studies that were not published in English.

The 186 publications retrieved by the search where filtered through a three tier process. First, potentially relevant publications where evaluated against the inclusion criteria, using solely the information offered in their title, with non-relevant publications being rejected, most of them using surnames in the genetic domain to study ancient migrations or isonomy. In case of doubt, the publication was left included in this phase. This reduced the number of publications to 129. Secondly, these were then evaluated against the exclusion criteria using the information provided in their abstract, what reduced the number of selected publications to 37. Finally, the full text of these 37 publications was analysed against the exclusion criteria, ending up with 11 publications that met all the selection criteria. These 11 publications where analysed in-depth, and all of their references were retrieved and also checked against the inclusion and exclusion criteria. This last step contributed with two additional publications that were not found by the original search, one of them because the word "name" or its equivalents did not appear neither in the title nor in the keywords (Sheth et al, 1999), and the second one because it is a government report only published on-line (Word and Perkins, 1996).

The final selection of publications consisted in 13 papers representing five countries (Canada, Germany, Netherlands, U.K., and the U.S.), and most of them from the field of public health. Table 1 shows the key characteristics of these studies, whose findings will be analysed in the following sections. The subset of ethnic minorities studied represent the biggest and most recently arrived groups in each country: a) South Asians (Indian, Pakistanis, Bangladeshis, Sri Lankans), b) Chinese, c) other East and South-east Asians (Vietnamese, Japanese, Korean, and Filipino), d) Hispanics, e) Turks, and f) Moroccans (see Table 1 for the correspondence between these groups and each study).

*Insert Table 1 about here*

**Table 1: Summary of the general characteristics of the 13 studies reviewed**

Amongst the publications excluded in the last phase of the selection strategy (n=26) there were some other interesting research papers in which an independent name-based approach was developed, although not explicitly explained nor independently evaluated. However, some of these studies are worth mentioning, since they typically used telephone directories to select names from a particular ethnic group as a sampling strategy for their surveys, showing the usefulness of the name-based approach to classify Vietnamese (Hinton et al, 1998, Rahman et al, 2005), Korean (Hofstetter et al, 2004), Cambodian (Tu et al, 2002), Chinese (Hage et al, 1990, Lai, 2004), South Asian (Chaudhry et al, 2003), Japanese (Kitano et al, 1988), Irish (Abbotts et al, 1999), Jewish (Himmelfarb et al, 1983) Iranian (Yavari et al, 2005) and Lebanese (Rissel et al, 1999) names, in the U.S., Canada, U.K. and Australia.

### 4.2 Structure of the selected studies

The 13 finally selected papers aimed to demonstrate a satisfactory accuracy rate in separating individuals of either one, or just a few, ethnic minority groups from the rest of the resident population in some developed countries. None of them tried to classify the whole population into all of the potential ethnic groups in a country, something that remains a research gap. The studies differ substantially in the sizes of the target populations to be classified (from 137 to 1.9 million people), the numbers of unique forenames or surnames in the reference list used in the search (from fewer than 100 to 27,000 names), and hence the method to allocate them (manual vs. automatic

classification). However, each of the studies includes a number of common methodological processes and research components: firstly a name *reference list* is independently built or sourced from another study or from 'an expert'; secondly a separate *target population* is manually or automatically classified into ethnic groups; and thirdly the *accuracy* of the method is *evaluated* against a previously known '*gold standard'* for ethnicity in the target population. These common structure and processes are summarised through a flow chart in Figure 1.

Insert Figure 1 about here

**Figure 1: Structure and processes of Name Classifications Evaluated**

### 4.3 Source data, reference and target populations

The primary source material for each of the studies is datasets of individuals' personal data that are usually sourced from population administrative files, health registers, or surveys.. *Target population* is the term given to the list of individuals to be classified into ethnic groups using their names, either manually or automatically. Automatic classification methods require an independent *reference list* of surnames or forenames with their pre-determined ethnic origin, that is used to perform the computerized search and allocation of ethnicity for each individual in the target population (in the manual methods the equivalent to the *reference list* is embedded in the expert's knowledge). This distinction between *reference* and *target* lists of names is key to the understanding of the methodologies here analysed.

### 4.4 Building Reference Lists

The first step thus involves building *reference lists* or borrowing them from previous studies, that would finally include several hundreds or thousands of surnames, each of

one of them with a pre-assigned ethnic group (e.g. Nguyen – Vietnamese; Chang – Chinese, etc). The characteristics of how the reference lists in the eight studies that used automatic classification were developed are further detailed in Table 2. Two of these studies used a software application already developed to identify South Asian names in the U.K., *Nam Pehchan* (Cummins et al, 1999, Harding et al, 1999), which contains 2,995 unique South Asian surnames, and was derived from the Linguistic Minorities Project (1985). Another study, Nanchahal et al (2001), develop a similar software called *SANGRA,* but do not offer sufficient information about how they built their reference list of 9,422 South Asian names. In the remaining five studies a purpose-built reference list was constructed, containing from 427 to 25,276 unique surnames. These *reference lists* were typically built from an independent source to the *target population*, a second population generally described as *reference population* (see the left half of Table 2), except in Choi et al (1993) and Coldman et al (1988), with important consequences for their results, as will be mentioned later.

*Insert Table 2 about here*

**Table 2: Characteristics of Reference Populations and Reference Lists in Automatic Methods**

Despite the big differences in the sizes of the reference populations, the methods employed to derive the name reference lists were broadly similar. Generally, they all used some type of 'ethnic origin information' in the reference population, such as self-reported ethnicity, country of birth, or nationality, to classify individuals into ethnic groups, and they then aggregated them by surname and produced a frequency count for each surname and ethnic group combination (and the same for forenames when available). Each surname or forename was then assigned to the ethnic group with the

highest frequency, using a series of rules or thresholds in some cases (Lauderdale and Kestenbaum, 2000, Word and Perkins, 1996), producing the final *reference list*.

In general, there are four factors affecting the accuracy and coverage of the *reference list* as will be explained in the accuracy evaluation section: the independence between reference and target populations, the size of the reference population, its spatio-temporal coverage (the countries and regions where it was sourced and the time period covered), and the method to ascribe ethnicity (using proxies vs. self-reported ethnicity). Therefore, the desired qualities of the reference list is to be large enough as to maximise coverage in the target population, and accurate enough as to minimise misclassifications (Coldman et al, 1988, Nanchahal et al, 2001). These two qualities are usually mutually exclusive, and there is a trade-off to be made between marginal extra coverage of a larger number of names and marginal extra accuracy of the classification, as each extra name tends to be rarer than the last. The final decision will depend on each specific type of application. A similar issue arises regarding the nominal resolution of the ethnic group categorizations used: the finer the groups are defined (e.g. Hindu, Bengali, Tamil, Urdu, Gujarati, Punjabi, etc vs. 'Indian' or 'South Asian'), the less accurate the name classification becomes and vice versa.

### 4.5 Minimum size of the reference list

As per calculating the ideal size of the reference population, the best attempt has been proposed by Cook et al (1972: 40) using the following formula:

$$n \geq \frac{\log(1-x)}{\log y}$$

Where $n$ is the required minimum size of the reference population, $x$ is the desirable level of confidence for the allocation of an individual to his or her appropriate ethnic group, and $y$ is the required level of confidence that a particular surname will perform as desired. For example, for $x$= 80% and $y$=95% the minimum size of the reference population required will be $n \geq 13.4$, meaning that for every surname to be classified a list of at least 13.4 individuals with that surname and their ethnicity is required. The minimum value of $n$ (in the above example equal to 13.4) refers to the unlikely situation that all individuals with the same surname in the reference population had the same ethnicity, and hence the size would have to be extended in proportion to the 'noise' found in each specific reference population. Cook et al (1972) propose to multiply $n$ by a factor of 4 to obtain a workable reference population size. The actual reference population sizes used in the five studies evaluated here, that built their own reference lists, has been compared against these two 'Cook et al criteria': *first criterion*; n=13.4 people per surname, and *expanded criterion*; n=13.4 x 4= 53.6 people per surname. It is surprising to find out that only two of the five studies' reference populations satisfy the first 'Cook first criterion' (Lauderdale and Kestenbaum, 2000, Word and Perkins, 1996), with the remaining three below 75% of the required size. Moreover, only one satisfies the 'Cook expanded criterion' (Lauderdale and Kestenbaum, 2000), with the rest below 45% of the required minimum reference population size.

### 4.6 Classification of Target Populations

The second step in the 13 studies analysed consisted in classifying the target population into ethnic groups, using either a manual (by an expert) or an automatic method (through computer algorithms). The characteristics of the target populations selected in each of the 13 studies are summarised in Table 3 ('Target Population' section).

Manual methods have the advantage of not requiring a name reference list and also to include a rich number of 'fuzzy rules' that the experts performing the classification can apply in order to decide the group into which an individual should be assigned. However, the manual method has a series of major limitations, the main one being that it is cumbersome and time-consuming (Bouwhuis and Moll, 2003) and this seriously constrains the size of the target population to be coded. In order to increment the number of individuals to be coded, additional experts need to be recruited, which causes inconsistency in the subjective decisions taken by different human subjects. Additionally, most of the manual classification studies focus on a two-group classification problem, that only requires a simple binary decision on whether the individual belongs to a specific ethnic minority group or not, but when more groups are introduced, several experts from different cultural backgrounds are required, and hence the number or misclassifications quickly rises, especially between similar ethnic groups when names overlap between groups (Martineau and White, 1998). For these reasons, not further specific attention will be given here to those studies which used manual methods (last four papers in Table 3).

On the other hand, automatic methods to classify the target population rely on the availability of an appropriate name reference list. The studies analysed here applied an automated algorithm to search the name of each individual in the target population against the reference list, and then assign the pre-coded ethnic group for that name to the individual. One of the main differences between the studies is whether they used only one name component of the individual (surname) or more (forename and surname,

or even middle name) (see last column of Table 1 for details). *Nam Pehchan* includes a set of rules that use name stems if the name has no match in the reference list (Cummins et al, 1999), but this is avoided by *SANGRA* since it is deemed to derive an unacceptable number of false positives (Nanchahal et al, 2001).

A second difference between studies is whether one or several ethnic groups are to be classified. It must be emphasized that almost all of the studies that used automatic classification were designed to classify individuals with a binary taxonomy in mind that seeks to identify members of a particular minority group or macro group (i.e. South Asians) from a general population. The exception is Lauderdale and Kestenbaum (2000) classifying six substantially different Asian ethnic groups (Chinese, Vietnamese, Japanese, Korean, and Filipinos). A third difference, is the use of certain name scores or thresholds related to the strength of the association between each name and ethnic group of origin (e.g. Heavily Spanish, Moderate Spanish, etc.), to the final user's advantage when fine-tuning the classification to their specific target population and purpose. Only two studies use such thresholds (Lauderdale and Kestenbaum, 2000, Word and Perkins, 1996).

*Insert Table 3 about here*

**Table 3: Summary of Target Population characteristics and results of the evaluation of classification accuracy in the 13 papers reviewed**

## 5. Evaluating Name Classifications

All of the studies measure the accuracy of the name-based classification, by comparing it to a 'gold standard' for the ethnicity of the individuals in the target population, that had to be previously known through an independent source (the exception is Word and

Perkins, 1996, but another study that evaluates their method is used here: Stewart et al 1999). This 'gold standard' is either the person's ethnicity (self-reported, by a next-of-kin, or by a third party), or a proxy for it such as country of birth or nationality (of the person or of his/her parents), all of which are assumed to represent the individuals 'true ethnicity'. However, we have to bear in mind that an objective entity such as the 'true ethnicity' does not exist, and hence *'there can be no such thing as a completely correct method of classifying individuals into ethnic groups'* (Cook et al, 1972 : 39), but to a certain extent a more appropriate one.

### 5.1 Accuracy Evaluation

The studies reviewed here self-evaluated their accuracy using the epidemiological measures of *sensitivity*, *specificity*, *positive predictive value* (PPV), and *negative predicted value* (NPV). *Sensitivity*, is the proportion of members of 'Ethnic Group X' (gold standard) who were correctly classified as such; *specificity*, the proportion of members of 'Other Ethnic Groups'(gold standard) who were correctly classified as such; *Positive Predictive Value* (PPV), is the proportion of persons classified as 'Ethnic Group X' (predicted) who were actually from 'Ethnic Group X'; *Negative Predictive Value* (NPV), is the proportion of persons classified as 'Other Ethnic Groups' (predicted) who were actually from 'Other Ethnic Groups'. These concepts are better explained in Table 4 in a more visual fashion. Any classification's objective is to maximize the number of correct classifications across the diagonal ('a' and 'd') and to minimise the number of misclassifications ('b' and 'c').

*Insert Table 4 about here*

**Table 4: Explanation of measures of classification accuracy: Sensitivity, Specificity, PPV and NPV**

The results for these four variables in the 13 studies are offered in Table 3 ('Method Evaluation' section) and a range of values is offered where the study evaluated different populations, or made separate evaluations for subpopulations (e.g. by gender). If certain isolated outliers are excluded, the *sensitivity* varies between 0.67 and 0.95, the *specificity* between 0.8 and 1, the *PPV* between 0.7 and 0.96, and the *NPV* between 0.96 and 1 (only reported in four studies).

It is striking to notice that there are no substantial differences between the accuracy of the manual (bottom four in Table 3) and automatic classification methods, removing the theoretical advantage, in accuracy terms, of the former over the latter. In general the studies tend to reach a high specificity and NPV (near to 1), in detriment of a slightly lower sensitivity and PPV (see for example Razum et al, 2001), a fact linked to the mentioned trade-off between the marginal extra coverage of a classification and its marginal extra accuracy. The differences between the statistics of the 13 studies do not seem to imply substantial differences in the quality of the methods adopted. Rather more, they reflect variations between the degree of distinctiveness of each subpopulation's names in the particular context of the general population studied, as well as constraints imposed by the characteristics of the datasets utilised.

All authors read in these results a validation of the name-based classification method to ascribe ethnicity, when other data sources are not available, giving further details of their advantages and the limitations found which will be both discussed in the next two

sections. However, one could argue as well the issue of publication bias, in which studies which did not achieve satisfactory results may have not been published.

### 5.2 Limitations found in the methodology

The 13 studies list a series of issues and limitations, many of them common between them, and that have been summarised here complementing them with other studies (Jobling, 2001, Senior and Bhopal, 1994) under the following eight major themes:

a) *Temporal differences in name distribution between the reference and the target populations; because of different migration waves and variations in the geographic distribution patterns through time, which introduces misclassification and low coverage in the classification. For example, Lauderdale and Kestenbaum (2000) use a list of people born in Asia before 1941, which might misrepresent today's common Asian names in the U.S., and a similar problem is present in Coldman et al (1988) with Chinese names in Canada.*

b) *Regional differences* in the frequency distribution of names, whether these are *between* the origin and the host country, *within* either of them, or between *different host countries*, due to differential geo-historical processes and migration flows. If this heterogeneity in name distribution is ignored when sampling the reference population, the subsequent name reference lists will be biased and names from a single region might not represent well the names present in other regions. Some examples found are: different Pakistani names present in the North of England, compared to the South East of England (Cummins et al, 1999), Turkish names between Germany and Turkey (Razum et al, 2001) , or Chinese migrant names between Australia and Canada (Choi et al, 1993)

*c) Differences in the average ratio of people per surname; between the ethnic minority (higher) and the host population (lower), and the ethnic minority in the host country (higher) and in the origin country (lower) (see* Table 2*: 'E.M. People / Surname'). This asymmetry is caused by a combination of a phenomenon of 'family autocorrelation' in the data (Lasker, 1997), and the uneven initial distribution of migrant names due to selective migration (a few initial names that can be rare in the origin country grow rapidly because of intra-group marriages in the host country). This causes the false assumption that a common name in the host country might also be common in the origin country, which together with item b) above make a strong case for a sourcing of name reference lists from the whole population of both origin and host countries.*

*d) Name Normalisation issues;* data entry misspellings, forename and surname inversions, and name corruptions, all need to be normalised both in the reference and target populations in order to cleanse the datasets, but making the difficult decision to keep the ones that might be accepted as official names, even for several generations (Lasker, 1985). This could be due to different *transcriptions* of a name into a different language's alphabet and/or pronunciation (called *transliteration*); what creates name duplications and long lists of name variants, presenting a barrier to the accuracy of the reference lists. This problem is linked to other processes of name change, the 'acculturation of a name' in a host country, and the degree of inter-marriages between groups, which are all well documented for 'older' immigrant groups in the U.S. such as Norwegians (Kimmerle, 1942), Finnish (Kolehmainen, 1939), Italian (Fucilla, 1943) or Polish (Lyra, 1966).

*e)* Names usually *only reflect patrilineal heritage*; and thus, the methodology assumes a high degree of group endogamy, and is incapable of identifying mixed ethnicity or women's ethnicity in mixed marriages (when women maiden name is not available) (Harland et al, 1997). If exogamy increases, as is anticipated in the near future, the method's discriminatory ability may decline. This has already happened in highly mixed populations such as the U.S. or Argentina, where more than three generations have passed since immigration of the traditional European migrant groups, their populations are assimilated into the general population, and the male surnames that are passed on do not normally reflect a perceived ethnic identity (Petersen, 2001).

*f)* There is a *different history of name* adoption, naming conventions and surname change that varies from country to country (e.g. Caribbeans have British surnames, Spanish women do not change surname at marriage), leading to the overlapping of certain names between ethnic groups (Martineau and White, 1998) which is difficult to accommodate in a single classification.

All of the above issues result in *differences in the strength of association* of a particular name with an ethnic group, measured by the proportion of people with a name ascribed to a certain ethnic group that actually consider themselves from that ethnic group.The effects of issues a) b) and c) can be mitigated by sourcing broad reference populations from both the origin and host country and from a wide enough time period, using the Cook et al (1972) formula mentioned before to calculate its minimum size. This would

ensure that the name reference list would reflect all of the potential names and true frequencies from the regions of the origin and host countries in equal probability than the methods analysed here have. Moreover, when aggregating the reference population by household surname, the issue of family autocorrelation can be avoided (Word and Perkins, 1996). The effects of issues d) to h) can be ameliorated by the use of 'name scores' to measure the strength of the association between a name and its ethnic group (Lauderdale and Kestenbaum, 2000), and use them in different ways sensitive to other context information (e.g. such as address of residence, which can be linked to Census information on the distribution of ethnic groups in an area).

### 5.3 Advantages of the methodology

According to the authors of the studies analysed here, name-based ethnicity classification methods present a valid alternative technique to ascribe individuals to ethnic groups through their name origins, when self-identification is not available. The criterion for such validity is that the methodology makes it possible to subdivide populations to a sufficient degree of accuracy at the ethnic group aggregate level, and not necessarily at the individual level (i.e. it produces relatively accurate total figures and orders of magnitude). In general, there is a consensus in the literature that although this methodology cannot entirely replace self-assigned ethnicity information, it provides a sufficient level of classification confidence to be used in the measurement of inequalities and in the design and delivery of services that meet the needs of ethnic minorities. In predicting these types of outcomes, name-based classifications have proved a very cost effective method compared to conventional collection of self-assigned ethnicity information (e.g. projects aiming to collect all patients' self-reported

ethnicity in the U.K. have had an average response rate of 56%: Adebayo and Mitchell, 2005).

Some of the methods evaluated here also provide a degree of strength in the assignment of an ethnic group to each name (Lauderdale and Kestenbaum, 2000, Word and Perkins, 1996), and others offer the probable religion and language associated with each group of names (those using *Nam Pehchan* or *SANGRA*). These efforts have produced three computerised name classification systems, *Nam Pehchan* (Cummins et al, 1999) and *SANGRA* (Nanchahal et al, 2001), designed to classify South Asian names in the U.K., and GUESS (Generally Useful Ethnicity Search System) (Buechley, 1976) which identifies Hispanic names in the U.S.. These computer systems have been used in a wide variety of studies in Public Health, having proved very useful to identify areas of inequality and health needs within populations (Coronado et al, 2002, Honer, 2004).

Furthermore, name-based methods have been successfully applied to sample members of particular ethnic groups using electoral registers or telephone directories (see discarded studies listed in section 4.1), presenting significant cost advantages over other alternatives (Cook et al, 1972). Moreover, this methodology has also proved useful in combination with conventional ethnicity classification information (Coronado et al, 2002). When some degree of ethnicity information is already available for a population, name-based classification can provide complementary information to detect errors, fill missing data, or correct bias introduced by proxies of ethnicity used, such us country of birth (e.g. second generation migrants).

Despite having found some inconsistencies between *Nam Pehchan* and *SANGRA*, when trying to classify the entire U.K. population (using the electoral roll), Peach and Owen (2004) conclude that name-based methods have a potential value to health organisations, local authorities, commerce and academics, but further research to improve the classifications is needed. Furthermore, a similar conclusion is reached by Bhopal et al (2004), who also used *Nam Pehchan* and *SANGRA* in an extensive study linking Census and health data in Scotland, highlighting that name-based methods are valuable in the absence of alternative information sources, and more crucially, they produce important information at relatively low costs (Bhopal et al, 2004).

## 6. Promising developments in name-based classifications

The 13 research studies reviewed here have demonstrated the advantages of name-based methods as well as their current main limitations. From the latter, three general needs for improvement arise, as justified in the previous section: a) a need for a reference population with high spatio-temporal coverage including name frequency data sourced both in the host and origins countries, b) the need to use name scores to measure the probability of a name being associated with a particular ethnic group, and c) the need for a system that classifies the whole population into all of the potential ethnic groups, and not just one or a few.

These tasks are made much easier today by the use of population registers that cover most of the population, such as electoral registers or telephone directories, providing very valuable name frequency information, name spelling variants, linkages between surnames and forenames, precise addresses, etc. A few of the studies analysed make use

of some of these resources, although they only cover parts of a country, or use manual methods such as counting names in a paper telephone directory. Electronic versions of such registers can today be accessed through special requests or purchased from data providers, making this type of analyses much simpler.

### 6.1 The Cultural Ethnic Language Group (CELG) technique

However, such directories or registers do not obviously contain any ethnicity information associated with people's names. To be able to develop a name reference list from such datasets an alternative method has been recently proposed in the onomastics field by Tucker (2005), who pre-classified over 70,000 surnames into 44 'Cultural Ethnic and Linguistic' groups (CELG) for the Oxford Dictionary of American Family Names (DAFN) (Hanks, 2003). Tucker (2005) developed a technique that termed *Cultural-Ethnic-Language Group (CELG)* in which a database of individuals with both forenames and surnames is required. To do this he uses the U.S. telephone directory with 88 million subscribers.

Firstly, a set of 'diagnostic forenames' (good predictors of ethnicity) is manually classified into cultural-ethnic-linguistic groups (CELG) by onomastic experts (8,000 forenames: Hanks and Tucker, 2000). Secondly, this diagnostic list of forenames is applied to classify the forenames of all the individuals in the telephone directory by CELG. Thirdly, for each surname in the database (1.75 million) the following calculation is done:

Surname X; % Forenames of CELG-1, % Forenames CELG-2 …etc.

(E.g. a fictitious surname being: 72% English, 17% Polish, 4% Spanish, and 3% Jewish)

That is, the relative frequency of people bearing that surname in each of the ethnic groups assigned to their forenames in the previous step. Finally, the surname is assigned to the group of highest frequency other than 'English', due to a 'host-country' assimilation effect (the previous example resulting in the surname being classified as Polish). This technique can be repeated iteratively to increase the number of diagnostic forenames classified and then the number of surnames and so forth. The performance of the CELG technique is deemed to have an accuracy of 88-94 % tucker (Tucker, 2005).

This method is very efficient because it leverages on the difference in the asymmetry of the name frequency distribution between that of forenames (extremely positively skewed) and surnames (largely positively skewed). To illustrate this with an example, 10% of the surnames in the U.S. are sufficient to cover 91% of the population, while 1% of forenames is sufficicient to cover 95% of the population. There are 1.25 million unique forenames in the U.S., so concentrating just in 1% of them (12,500 forenames) one can code the forenames' ethnicity of 95% of the U.S. population, and hence their surnames' ethnicity (Tucker, 2001). Furthermore, by applying the CELG technique this population coverage can be increased to nearly 100%, while improving the overall accuracy of the names classified. This is further eased by the use of etymology dictionaries of forenames origins to code 'diagnostic forenames', with larger coverage and availability than surname dictionaries.

## 6.2 Towards a total population multi-ethnicity classification method based on names.

The CELG technique has not been used in any of the studies reviewed in this paper but it has a great potential for efficiently classifying hundreds of thousands of names into all of the potential ethnic groups present in a given population. Furthermore, it makes it possible to create the desired 'surname scores', measuring the degree of association between a surname and an ethnic group by setting thresholds to the ethnicity distribution of its bearers forenames (as in the Polish example mentioned above). This approach is being followed by the team of researchers at University College London to which the author belongs, what has provided promising developments that at the time of writing are being evaluated in the same way as other studies have (see Mateos et al, 2007 for initial results). .

Finally, in order to create an ethnicity classification covering all of the potential ethnic groups present in a population, the name reference list has to be created using reference populations originated in a large number of countries, what is made possible today through the use of electronic telephone directories, population registers and a growing realm of genealogical internet resources (Hanks, 2003). Furthermore, a set of alternative classification techniques, such as census area information, text pattern mining, etc. which are discussed in detail in Mateos et al (2007) can be brought to the effort of improving name classification methods currently available.

## 7. Conclusion

Name-based Ethnicity Classification Methods have been successfully applied, primarily in public health applications, to subdivide populations into groups of common origin, although they clearly present room for improvement. Moreover, these methods present a high potential to be applied in broader population studies about ethnicity, such as: in ethnic group population forecasting by small area (Large and Ghosh, 2006), monitoring migration (Stillwell and Duke-Williams, 2005), detecting Census undercount (Graham and Waterman, 2005), measuring residential segregation (Simpson, 2004), analysing the geography of ethnic inequalities (Dorling and Rees, 2003) or of mortality and morbidity (Boyle, 2004), evaluating equal opportunity policies (Johnston et al, 2004) and political empowerment processes (Clark and Morrison, 1995), and improving public and private services to ethnic minorities (Van Ryn and Fu, 2003). All of these research and public policy areas present a lack of appropriate timely and detailed data on ethnicity, a problem that is increasing as the last round of Census data age and new migration flows are changing the composition and demands for public services. Improved methods in these areas are thus of key policy importance in today's multi-cultural society.

The name-based ethnicity classification methodology evaluated here through 13 representative studies, offers a few advantages over traditional information sources such as the Censuses of population. Amongst them, it can develop a more detailed and meaningful classification of people's origins categories (finer categories based on a very large number of languages versus just 10 to 20 ethnic groups in the Census), offers improved updating (annually through registers with substantial population coverage; e.g. electoral or patient registers), it better accommodates changing perceptions of

identity than ethnicity self-classification (through independent assignment of ethnicity and or cultural origins according to name) and is made available, subject to confidentiality safeguards, at the individual or household level (rather than an aggregated Census area). Moreover, according to the literature its main advantage remains its capability to provide an ethnicity classification when self-reported ethnicity is not available, which is the case in most population registers and datasets about individuals, and at a fraction of the cost of alternative methods. However, this advantage will tend to disappear with time, as the recording of self-reported ethnicity becomes a routine practice, and data linkage methods allow that this information is only recorded once throughout population registers (Bhopal et al, 2004, Blakely et al, 2000)

However, this review has also revealed a series of limitations that remain mostly unsolved, and have hindered the wider adoption of name-based classifications. The comparative approach taken here has enabled to group the common causes of these issues and propose a few improvements to overcome them. These issues are; spatio-temporal differences in the frequency distribution of names, the selective process of migration, family autocorrelation, differences in the strength of association between a name and an ethnic group, name spelling errors and name normalisation issues, different transcriptions or transliteration of a name into a different alphabet or pronunciation, names usually only reflecting patrilineal heritage, different histories of name adoption, naming conventions and surname change, and that they currently only classify a few ethnic groups.

In order to overcome or ameliorate these issues, future name-based classifications will have to use large enough reference populations with wide spatio-temporal coverage sourced both in the host and origins countries. They will also necessarily require the development of name-to-ethnicity probability scores, and they will need to be able to classify complete populations into all of the potential ethnic groups present in a society at any given time and place. Two of the studies analysed here stand out from the rest in that they manage to gather some of these qualities (Lauderdale and Kestenbaum, 2000, Word and Perkins, 1996), and while the rest present important shortcomings, they have all demonstrated their value and sufficient accuracy in classifying ethnicity in the context for which they were designed for. There is an important potential for future research on improvements to this methodology (Bhopal et al, 2004, Peach and Owen, 2004), and key advancements have already been proposed by Hanks and Tucker (2000) and Tucker (2005), which are being adopted into new classifications aiming for complete population coverage (Mateos et al, 2007). Finally, there are certain uncertainties regarding the ethical and legal implications of using names in this manner, which also need to be assessed and clarified.

There is evidence today that names are unfortunately still being used to discriminate people in access to the labour, housing, and credit market (Carpusor and Loges, 2006, Williams, 2003), due to the prejudices that some still have about people's ancestry, language, religion, culture, or skin colour. Using the same weapons as the 'enemy', in the 'The Causes and Consequences of Distinctively Black Names' Fyer and Levitt (2004) present a crude picture of ethnic inequalities and discrimination in the U.S. through an innovative analysis using forenames. A golden opportunity would be missed

if researchers in population studies do not be creative enough to find alternative ways to reduce persistent discrimination and inequalities between ethnic groups in today's ever increasingly multi-cultural cities.

**Acknowledgements for support:**

I am grateful to two anonymous referees who have provided useful feedback on aspects of this paper. This research has been sponsored by a Knowledge Transfer Partnership between ESRC and Camden Primary Care Trust (Dti KTP-037). This work was undertaken with the partial support of Camden Primary Care Trust who received a proportion of funding from the NHS Executive. The views expressed in this publication are those of the author and not necessarily of the NHS Executive nor of the ESRC.

## 8. References


Abbotts J, Williams R, Smith GD. 1999. Association of medical, physiological, behavioural and socio-economic factors with elevated mortality in men of Irish heritage in West Scotland. *Journal Of Public Health Medicine* **21**(1): 46-54

Adebayo C and Mitchell P. 2005. Patient Profiling. Presented at *GEONom*, London, 25 May. Available at: www.casa.ucl.ac.uk/geonom/Initial_meeting Accessed: 12/05/2006.

Agyemang C, Bhopal R, Bruijnzeels M. 2005. Negro, Black, Black African, African Caribbean, African American or what? Labelling African origin populations in the health arena in the 21st century. *Journal Of Epidemiology And Community Health* **59**(12): 1014-1018

American Council of Learned Societies. 1932. *Report of Committee on Linguistic and National Stocks in the Population of the United States.* Annual Report for the Year 1931. American Historical Association. Washington, D.C.

Arday SL, Arday DR, Monroe S, Zhang J. 2000. HCFA's Racial and Ethnic Data: Current Accuracy and Recent Improvements. *Health Care Financing Review* **21**(4)

Aspinall PJ. 2000. 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? *Ethnicity and Health* **5**(1): 33 - 40

Aspinall PJ. 2003. Who is Asian? A category that remains contested in population and health research. *Journal Of Public Health Medicine* **25**(2): 91-97

Banton M. 1998. *Racial Theories*. Cambridge: Cambridge University Press.

Bhopal R. 2004. Glossary of terms relating to ethnicity and race: for reflection and debate. *Journal Of Epidemiology And Community Health* **58**(6): 441-445

Bhopal R, Fischbacher C, Steiner M, Chalmers J, Povey C, et al. 2004. *Ethnicity and health in Scotland: can we fill the information gap?*, Centre for Public Health and Primary Care Research. University of Edinburg. Available at: http://www.chs.med.ed.ac.uk/phs/research/Retrocoding%20final%20report.pdf

Accessed: 22/11/2005.

Blakely T, Woodward A, Salmond C. 2000. Anonymous linkage of New Zealand mortality and Census data. *Australian and New Zealand Journal of Public Health* **24**(1): 92

Bouwhuis CB and Moll HA. 2003. Determination of ethnicity in children in the Netherlands: Two methods compared. *European Journal of Epidemiology* **18**(5): 385

Boyle P. 2004. Population geography: migration and inequalities in mortality and morbidity. *Progress in Human Geography* **28**(6): 767-776

Brubaker R. 2004. *Ethnicity without groups*. London: Harvard University Press.

Buechley RW. 1976. *Generally useful ethnic search system: GUESS* (Mimeo) Cancer Research and Treatment Center. University of New Mexico. Albuquerque.

Bulmer M. 1996. The ethnic group question in the 1991 Census of Population. In *Ethnicity in the 1991 Census. Volume 1. Demographic characterisitics of the ethnic minority populations*, Coleman D, Salt J (eds.), Office for National Statistics, HMSO: London: xi -xxix.

Carpusor AG and Loges WE. 2006. Rental Discrimination and Ethnicity in Names. *Journal of Applied Social Psychology* **36**(4): 934-952

Castells M. 1997. *The power of identity - Information age: economy, society and culture - Vol. 2*. Oxford: Blackwell.

Cavalli-Sforza LL. 1997. Genes, Peoples, and Languages. *Proceedings of the National Academy of Sciences* **94**(15): 7719-7724

Cavalli-Sforza LL and Cavalli-Sforza F. 1995. *The Great Human Diasporas*. Reading, Massachusetts: Addison-Wesley.

Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J. 1988. Reconstruction of Human Evolution: Bringing Together Genetic, Archeological and Linguistic Data. *Proceedings of the National Academy of Sciences* **85**: 6002-6006

Chaudhry S, Fink A, Gelberg L, Brook R. 2003. Utilization of papanicolaou smears by South Asian women living in the United States. *Journal of General Internal Medicine* **18**(5): 377-384

Choi BCK, Hanley AJ, Holowaty EJ, Dale D. 1993. Use of surnames to identify individuals of Chinese ancestry. *American Journal of Epidemiology* **138**: 723-734

Choi KH and Sakamoto A. 2005. *Who is Hispanic? Hispanic ethnic identity among African Americans, Asian Americans, and Whites.* PRC Working Paper Series. Rep. No. 04-05-07, Population Research Centre. University of Texas at Austin. Available at: http://www.prc.utexas.edu/working_papers/wp_pdf/04-05-07.pdf Accessed: 22/02/2005.

Christopher AJ. 2002. "To define the indefinable": population classification and the census in South Africa. *Area* **34**(4): 401-408

Clark WAV and Morrison PA. 1995. Demographic Foundations of Political Empowerment in Multiminority Cities. *Demography* **32**(2): 183-201

Colantonio SE, Lasker GW, Kaplan BA, Fuster V. 2003. Use of surname models in human population biology: a review of recent developments. *Human Biology* **75**(6): 785-807

Coldman AJ, Braun T, Gallagher RP. 1988. The classification of ethnic status using name information. *Journal Of Epidemiology And Community Health* **42**(4): 390-395

Coleman D and Salt J, eds. 1996. *Ethnicity in the 1991 Census. Volume 1. Demographic characterisitics of the ethnic minority populations*. Office for National Statistics, HMSO London

Connolly H and Gardener D. 2005. *Who are the 'Other' ethnic groups?* Social and Welfare reports. Office for National Satistics. London. Available at: http://www.statistics.gov.uk/articles/nojournal/other_ethnicgroups.pdf. Accessed: 27/01/2006.

Cook D, Hewitt D, Milner J. 1972. Uses of the surname in epidemiologic research. *American Journal of Epidemiology* **95**: 38-45

Coronado GD, Koepsell TD, Thompson B, Schwartz SM, Wharton RS, et al. 2002. Assessing cervical cancer risk in Hispanics. *Cancer Epidemiology Biomarkers and Prevention* **11**(10 Pt 1): 979-984

Cummins C, Winter H, Cheng K-K, Maric R, Silcocks P, et al. 1999. An assessment of the Nam Pehchan computer program for the identification of names of south Asian ethnic origin. *Journal Of Public Health Medicine* **2**(4): 401-406

Darwin C. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray.

Darwin GH. 1875. Marriages between first cousins in England and their effects. *Journal of the Statistical Society* **38**: 153-184

Department of Health. 2005. *A Practical Guide to Ethnic Monitoring in the NHS and Social Care*

Available at: http://www.dh.gov.uk/assetRoot/04/11/68/43/04116843.pdf. Accessed: 23/09/2005.

Dorling D and Rees P. 2003. A nation still dividing: the British census and social polarisation 1971 - 2001. *Environment And Planning A* **35**(7): 1287-1313

Eriksen TH. 2002. *Ethnicity and Nationalism*. London: Pluto Press.

Fryer RG and Levitt SD. 2004. The Causes and Consequences of Distinctively Black Names. *The Quarterly Journal of Economics* **119**(3): 767-805

Fucilla JG. 1943. The Anglicization of Italian Surnames in the United States. *American Speech* **18**(1): 26-32

Gerrish K. 2000. Researching ethnic diversity in the British NHS: methodological and practical concerns. *Journal of Advanced Nursing* **31**: 918-925

Gill P, Bhopal R, Wild S, Kai J. 2005. Limitations and potential of country of birth as proxy for ethnic group. *British Medical Journal* **330**(7484): 196

Graham D and Waterman S. 2005. Underenumeration of the Jewish population in the UK 2001 Census. *Population, Space and Place* **11**(2): 89-102

Hage BH, Oliver RG, Powles JW, Wahlqvist ML. 1990. Telephone directory listings of presumptive Chinese surnames: an appropriate sampling frame for a dispersed population with characteristic surnames. *Epidemiology* **1**(5): 405-408

Hanks P. 2003. *Dictionary of American Family Names* New York: Oxford University Press.

Hanks P and Tucker DK. 2000. A Diagnostic Database of American Personal Names. *Names* **48**(1): 59-69

Harding S, Dews H, Simpson S. 1999. The potential to identify South Asians using a computerised algorithm to classify names. *Population Trends* **97**: 46-50

Harland JO, White M, Bhopal RS. 1997. Identifying Chinese populations in the UK for epidemiological research experience of a name analysis of the FHSA register. Family Health Services Authority. *Public Health* **111**: 331-337

Himmelfarb HS, Loar RM, Mott SH. 1983. Sampling by ethnic surnames: The case of American Jews. *Public Opinion Quarterly* **47**: 247-260

Hinton L, Jenkins CN, McPhee S, Wong C, Lai KQ, et al. 1998. A survey of depressive symptoms among Vietnamese-American men in three locales: prevalence and correlates. *The Journal of Nervous and Mental Disease* **186**(11): 677-683

Hofstetter CR, Hovell MF, Lee J, Zakarian J, Park H, et al. 2004. Tobacco use and acculturation among Californians of Korean descent: a behavioral epidemiological analysis. *Nicotine and Tobacco Research* **6**(3): 481-489

Honer D. 2004. *Identifying Ethnicity: A comparison of two computer programmes designed to identify names of South Asian ethnic origin.* Uk Centre for Evidence in Ethnicity Health & Diversity. University of Warwick. Available at: http://www2.warwick.ac.uk/fac/med/research/csri/ethnicityhealth/aspects_diversity/identifying_ethnicity/. Accessed: 22/06/2006.

Howard D and Hopkins PE. 2005. Editorial: race, religion and the census. *Population, Space and Place* **11**(2): 69-74

Jobling MA. 2001. In the name of the father: surnames and genetics. *Trends in Genetics* **17**(6): 353-357

Johnston R, Wilson D, Burgess S. 2004. School Segregation in Multiethnic England. *Ethnicities* **4**(2): 237-265

Kertzer DI and Arel D. 2002. *Census and Identity. The Politics of Race, Ethnicity, and Language in National Censuses*. Cambridge: Cambridge University Press.

Kimmerle MM. 1942. Norwegian-American Surnames in Transition. *American Speech* **17**(3): 158-165

Kitano HH, Lubben JE, Chi I. 1988. Predicting Japanese American drinking behavior. *The International Journal of the Addictions* **23**(4): 417-428

Kolehmainen JI. 1939. Finnish Surnames in America. *American Speech* **14**(1): 33-38

Lai DW. 2004. Impact of culture on depressive symptoms of elderly Chinese immigrants. *Canadian Journal of Psychiatry* **49**(12): 820-827

Large P and Ghosh K. 2006. A methodology for estimating the population by ethnic group for areas within England. *Population Trends* **123**: 21-31

Lasker G. 1997. Census versus sample data in isonymy studies: relationship at short distances. *Human Biology* **69**(5): 733-738

Lasker GW. 1985. *Surnames and genetic structure*. Cambridge: Cambridge University Press.

Lauderdale D and Kestenbaum B. 2000. Asian American ethnic identification by surname. *Population Research and Policy Review* **19**(3): 283-300

Linguistic Minorities Project. 1985. *The Other Languages of England.* London: Routledge & Kegan Paul.

London Health Observatory. 2003. *Missing Record: The Case For Recording Ethnicity At Birth And Death Registration.* LHO Reports. Available at: http://www.lho.org.uk/viewResource.aspx?id=7954. Accessed: 01/09/2006.

London Health Observatory. 2005. *Using Routine Data to Measure Ethnic Differentials in Access to Revascularisation in London.* Available at: http://www.lho.org.uk/viewResource.aspx?id=9732. Accessed: 20/07/2006.

Lyra F. 1966. Polish Surnames in the United States. *American Speech* **41**(1): 39-44

M'charek A. 2005. *The Human Genome Diversity Project*. Cambridge: Cambridge University Press.

Marmot M, Adelstein A, Bulusu L. 1984. *Immigrant Mortality in England and Wales 1970-78: Causes of Death by Country of Birth.* Opcs. Her Majesty's Stationery Office. London.

Martineau A and White M. 1998. What's not in a name. The accuracy of using names to ascribe religious and geographical origin in a British population. *Journal Of Epidemiology And Community Health* **52**(5): 336-337

Mason D. 2003. *Explaining ethnic differences: changing patterns of disadvantage in Britain*. Bristol: Policy Press.

Mateos P, Webber R, Longley PA. 2007. *Using Names to Classify People and Neighbourhoods by their Name Origins.* CASA Working Papers. Centre for Advanced Spatial Analysis. London. Available at: http://www.casa.ucl.ac.uk/publications/workingpapers.asp. Accessed: 19/02/2007.

McAuley J, De Souza L, Sharma V, Robinson I, Main CJ, et al. 1996. Self defined ethnicity is unhelpful. *British Medical Journal* **313**(7054): 425b-426

McEvoy B and Bradley DG. 2006. Y-chromosomes and the extent of patrilineal ancestry in Irish surnames. *Human Genetics* **119**(1-2): 212-219

Mitchell R, Shaw M, Dorling D. 2000. *Inequalities in life and death: what if Britain were more equal?* Bristol: Policy Press.

Modood T. 2005. *Multicultural Politics: Racism, Ethnicity and Muslims in Britain*. Edimburg: Edimburg University Press.

Morning A. forthcoming. Ethnic Classification in Global Perspective: A Cross-National Survey of the 2000 Census Round. *Population Research and Policy Review*

Nanchahal K, Mangtani P, Alston M, dos Santos Silva I. 2001. Development and validation of a computerized South Asian Names and Group Recognition Algorithm (SANGRA) for use in British Health-related studies. *Journal Of Public Health Medicine* **23**(4): 278-285

Nicoll A, Bassett K, Ulijaszek SJ. 1986. What's in a name? Accuracy of using surnames and forenames in ascribing Asian ethnic identity in English populations. *Journal Of Epidemiology And Community Health* **40**(4): 364-368

Nobles M. 2000. *Shades of Citizenship: Race and the Census in Modern Politics*. Stanford: Stanford University Press.

Office for National Statistics. 2003. *Ethnic group statistics: A guide for the collection and classification of data.* Available at: http://www.statistics.gov.uk/about/ethnic_group_statistics/downloads/ethnic_group_statistics.pdf. Accessed: 13/02/2006.

Olson S. 2002. *Mapping human history: genes, race, and our common origins*. New York: First Mariner Books.

Peach C. 1996. *Ethnicity in the 1991 Census. Volume 2. The ethnic minorities of Great Britain*. London: Office for National Statistics, HMSO.

Peach C. 1999. Social Geography. *Progress in Human Geography* **23**(2): 282-288

Peach C. 2000. Discovering white ethnicity and parachuting plurality. *Progress in Human Geography* **24**(4): 620-626

Peach C and Owen D. 2004. *Social Geography of British South Asian Muslim, Sikh and Hindu Sub-Communities.* ESRC End of Project Full Report R-000239765. Available at: http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/ (search for "R-000239765"). Accessed: 15/08/2006.

Petersen W. 2001. Surnames in US Population Records. *Population and Development Review* **27**(2): 315

Piazza A, Rendine S, Zei G, Moroni A, Cavalli-Sforza LL. 1987. Migration rates of human populations from surname distribution. *Nature* **329**: 714 - 716

Rahman MM, Luong NT, Divan HA, Jesser C, Golz SD, et al. 2005. Prevalence and predictors of smoking behavior among Vietnamese men living in California. *Nicotine and Tobacco Research* **7**(1): 103-109

Rankin J and Bhopal R. 1999. Current census categories are not a good match for identity. *British Medical Journal* **318**(7199): 1696

Razum O, Zeeb H, Akgun S. 2001. How useful is a name-based algorithm in health research among Turkish migrants in Germany? *Tropical Medicine and International Health* **6**(8): 654-661

Rissel C, Ward JE, Jorm L. 1999. Estimates of smoking and related behaviour in an immigrant Lebanese community: does survey method matter? *Australian and New Zealand Journal of Public Health* **23**(5): 534-537

Rossiter WS. 1909. *A Century of Population Growth, from the First Census of the United States to the Twelfth, 1790-1900.* Government Printing Office. Washington DC.

Ruhlen M. 1987. *A guide to the World's Languages*. Standford, CA: Stanford University Press.

Scapoli C, Mamolini E, Carrieri A, Rodriguez-Larralde A, Barrai I. forthcoming. Surnames in Western Europe: A comparison of the subcontinental populations through isonymy. *Theoretical Population Biology*: In Press, Accepted Manuscript

Senior PA and Bhopal R. 1994. Ethnicity as a variable in epidemiological research. *British Medical Journal* **309**(6950): 327-330

Sheth T, Nair C, Nargundkar M, Anand S, Yusuf S. 1999. Cardiovascular and cancer mortality among Canadians of European, south Asian and Chinese origin from 1979 to 1993: an analysis of 1.2 million deaths. *Canadian Medical Association Journal* **161**(2): 132

Shriver MD and Kittles RA. 2004. Genetic ancestry and the search for personalized genetic histories. *Nature Reviews Genetics* **5**(8): 611-618

Simpson L. 2004. Statistics of racial segregation: measures, evidence and policy. *Urban Studies* **41**: 661-681

Skerry P. 2000. *Counting on the Census? Race, Group Identity, and the Evasion of Politics*. Washington: Brookings Institution Press.

Stillwell J and Duke-Williams O. 2005. Ethnic population distribution, immigration and internal migration in Britain. What evidence of linkage at the district scale. Presented at *British Society for Population Studies Annual Conference*, University of Kent at Canterbury 12-14 September. Available at: http://www.lse.ac.uk/collections/BSPS/pdfs/Stillwell_ethnicpopdist_2005.pdf Accessed: 20/06/2006.

Tu SP, Yasui Y, Kuniyuki A, Schwartz SM, Jackson JC, et al. 2002. Breast cancer screening: stages of adoption among Cambodian American women. *Cancer Detection and Prevention* **26**(1): 33-41

Tucker DK. 2001. Distribution of Forenames, Surnames, and Forename-Surname Pairs in the United States. *Names* **49**: 69-96

Tucker DK. 2005. The cultural-ethnic-language group technique as used in the Dictionary of American Family Names (DAFN). *Onomastica Canadiana* **87**(2): 71-84

U.S. Bureau of the Census. 1953. *Persons of Spanish Surname.* Washington D.C.: U.S. Government Printing Office.

US Senate. 1928. *Immigration quotas on the basis of national origin.* Rep. Miscellaneous Documents 8870 vol.1 nr 65, 70th Congress 1st Session. Washington, DC.

Van Ryn M and Fu SS. 2003. Paved With Good Intentions: Do Public Health and Human Service Providers Contribute to Racial/Ethnic Disparities in Health? *American Journal of Public Health* **93**(2): 248-255

Weber M. 1997[1922]. What is an Ethnic Group. In *The Ethnicity Reader. Nationalism, Multiculturalism and Migration*, Guibernau M, Rex J (eds.), Polity Press: Cambridge: 15-32.

Wild S and McKeigue P. 1997. Cross sectional analysis of mortality by country of birth in England and Wales, 1970-92. *British Medical Journal* **314**(7082): 705-

Williams A. 2003. Who will be hired: Stacey or Shakisha? *Journal of the National Medical Association* **95**(2): 109-110

Winnie WW, Jr. 1960. The Spanish Surname Criterion for Identifying Hispanos in the Southwestern United States: A Preliminary Evaluation. *Social Forces* **38**(4): 363-366

Word DL and Perkins RC. 1996. *Building a Spanish surname list for the 1990s a new approach to an old problem.* Technical Working Paper 13. US Census Bureau, Population Division. Washington DC. Available at: http://www.census.gov/population/documentation/twpno13.pdf. Accessed: 29/05/2005.

Yavari P, Hislop TG, Abanto Z. 2005. Methodology to identify Iranian immigrants for epidemiological studies. *Asian Pac J Cancer Prev* **6**(4): 455-457

| Paper Reference | Geographical area of study<br>*Country and (Region)* | Ethnic Minorities (E.M.) classified | Name to Ethnicity Assignment | |
|---|---|---|---|---|
| | | | ***Method***<br>---<br>***A**utomatic*<br>***M**anual* | ***Name components***<br>---<br>***S**urname*<br>***F**orename*<br>***M**iddle name* |
| **Choi, *et al* (1993)** | Canada (Ontario) | Chinese | A | S |
| **Coldman, Braun & Gallagher (1988)** | Canada (British Columbia) | Chinese | A | F, S, M |
| **Lauderdale & Kestenbaum (2000)** | U.S. (National) | Chinese, Japanese, Filipino, Korean, Indian, & Vietnamese | A | S |
| **Razum, Zeeb, & Akgun (2001)** | Germany (Rhineland-Palatinate & Saarland) | Turkish | A | F, S |
| **Word & Perkins (1996) / Stewart et al (1999)** | U.S. (National) | Hispanic | A | S |
| **Harding, Dews, & Simpson (1999)** | U.K. (Bradford & Coventry) | South Asian + Hindu, Muslim & Sikh | A | F, S |
| **Cummins, *et al* (1999)** | U.K. (Thames, Trent, W.Midlands & Yorkshire) | South Asian | A | F, S |
| **Nanchahal, *et al* (2001)** | U.K. (London, W.Midlands, Glasgow) | South Asian | A | F, S, M |
| **Sheth, *et al* (1997)** | Canada (National) | South Asian and Chinese | A/M | S |
| **Martineau & White (1998)** | U.K. (Newcastle; 4 General Practices) | Bangladeshi, Pakistani, Indian Muslims, Non-South Asian Muslims, Sikh, Hindu, White, Other | M | F, S and Gender |
| **Bouwhuis &. Moll (2003)** | Netherlands (Rotterdam; 1 Hospital) | Turkish, Moroccan, Surinamese | M | F, S |
| **Nicoll, Bassett, & Ulijaszek (1986)** | U.K. (Selected areas) | South Asian | M | F, S |
| **Harland, White & Bhopal (1997)** | U.K. (Newcastle) | Chinese | M | F, S |

**Table 1: Summary of the general characteristics of the 13 studies reviewed.** Method of name to ethnicity assignment: 'A' = Automatic, 'M' = Manual. Name components used in the classification; 'S'= Surname, 'F'= Forename, 'M'= Middle Name.

| Paper Reference | Reference Population | | | | | Reference List | | |
|---|---|---|---|---|---|---|---|---|
| | **Total Population** | **E.M. population identified** | **% E.M.** | **Source** | **Dates** | **Production Method** | **Nr. Unique E.M. Surnames** | **E.M. people / Surname** |
| **Choi, *et al* (1993)** | 270,139 | 1,899 | 0.7% | Mortality database | 1982-1989 | Country of Birth + Manual cleansing | 427 | 4.4 (Chinese) |
| **Coldman, Braun & Gallagher (1988)** | 203,354 | 5,430 | 2.7% | Death registrations | 1950-1964 | Ethnicity (family) | 544 | 16 (Chinese) // 1.7 (Non-Chinese) |
| **Lauderdale & Kestenbaum (2000)** | 1,765,422 | 1,609,679 | 91.2 % | Social Security Card Applications (MBR) | Born <1941 | Country of Birth | 27,000 | 59.6 (avg.) |
| **Razum, Zeeb, & Akgun (2001)** | 4,000,000 | 108,500 | 2.7% | Rhineland-Palatinate Population Register | c.2000 | Nationality + Manual cleansing | 12,188 | 12.8 (in Germany) / 3.1 (in Turkey) |
| **Sheth, *et al* (1997)** | 2,782,000 (estimated) | N/K | N/K | Canadian Mortality Data Base (CMBD) | 1979-1993 | Country of Birth (deceased & parents) | 4,271 | N/K |
| **Word & Perkins (1996) / Stewart et al (1999)** | 5,609,592 people; 1,868,781 households. | 597,533 | 10.7 % | 1990 US Census Post-enumeration Sample | U.S. Census Day 1990 | Ethnicity (self-assigned) | 25,276 | 23.6 (avg.) |
| **Harding, Dews, & Simpson (1999)** | List of 2,995 surnames in *Nam Pehchan* program | | | *Nam Pehchan* program | 1981-1998 | Experts' knowledge | 2,995 | N/A |
| **Cummins, *et al* (1999)** | List of 2,995 surnames in *Nam Pehchan* program | | | *Nam Pehchan* program | 1981-1998 | Experts' knowledge | 2,995 | N/A |
| **Nanchahal, *et al* (2001)** | List of 9,422 surnames in *SANGRA* program | | | Surveys and Hospital Records | 1995-1999 | From list of voluntary organisations and ONS | 9,422 | N/A |

**Table 2: Characteristics of Reference Populations and Name Reference Lists in Automatic Methods.** (E.M. = Ethnic Minority, N/K= Not Known, N/A= Not Available). Reference Population: 'Total population' is the input dataset used, of which 'E.M. population identified' is the ethnic minority population identified within the 'total population'. Reference List: 'Production Method' is the technique or piece of ethnicity information in the reference population used to produce the reference list; 'Nr. Unique E.M. Surnames' is the final number of ethnic minority surnames present in the reference list. 'E.M. People / Surname' is the average number of people of the ethnic minority sharing the same surname (column 3 / column 8).

| Paper Reference | Division of Reference & Target Population | Target Population | | | | | | Method Evaluation *(single value or a range)* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Total Population | Nr. E.M. classified | % E.M. | Source | Dates | Ethnicity Gold Standard | Sensitivity | Specificity | PPV | NPV |
| **Choi, *et al* (1993)** | Random split | 270,138 | 1,910 | 0.7% | Same as Reference | 1982-1989 | Country of Birth | 0.73 | N/K | 0.81 - 0.84 | N/K |
| **Coldman, Braun & Gallagher (1988)** | Chronological split sample | 155,629 | 3,205 | 2.1% | Same as Reference | 1965-1973 | Ethnicity | 0.89-0.97 | 1.00 | N/K | N/K |
| **Lauderdale & Kestenbaum (2000)** | Different sources | 1,900,000 | N/K | N/K | 1990 US Census Sample | 1990 | Ethnicity | 0.55 - 0.70 | N/K | 0.76 - 0.83 | N/K |
| **Razum, Zeeb, & Akgun (2001)** | Different sources | NK | 192 | N/K | Saarland Population Register | c.2000 | Nationality | 0.40 - 0.84 | 0.99 | 0.14 - 0.98 | 1.00 |
| **Word & Perkins (1996) / Stewart et al (1999)** | Different research papers | 7,232 | 780 | 10.8% | Greater Bay Area Cancer Register | 1990 | Ethnicity (self-reported) | 0.61 | 0.98 | 0.70 | 0.96 |
| **Sheth, *et al* (1997)** | Different sources | 200 | 100 | 50% | Telephone survey | 1990s | Ethnicity (self-reported) | 0.96 | 0.95 | N/K | N/K |
| **Harding, Dews, & Simpson (1999)** | Different sources | 275,353 | 6,585 | 2.4% | a) Resident Survey, b) School Survey, c) Death Register, d) Census Longitudinal Study | 1981-1998 | Ethnicity [self-rep. (a)&(d) parents(b)], c)Visual inspection | 0.94 | 0.99 | 0.96 | N/K |
| **Cummins, *et al* (1999)** | Different sources | 356,555 | 3,845 | 1.1% | Thames, Trent, W. Midlands & Yorkshire Cancer registers | 1990-1992 | Visual inspection + computerised dictionary | 0.90 | N/K | 0.63 | N/K |
| **Nanchahal, *et al* (2001)** | Different sources | 130,993 | 15,390 | 11.7% | London and Midlands Hospital Admissions | 1995-1999 | Ethnicity (self-reported) | 0.89 - 0.96 | 0.94 - 0.98 | 0.80 - 0.89 | 0.98 - 0.99 |
| **Martineau & White (1998)** | N/A | 137 | 107 | 78.1% | Family Health Service Authority Register (FHSA) | Born Oct 93 - Sep 94 | Ethnicity (3rd party reported) | 0.87- 0.98 (outlier 0.5) | 0.60 - 0.97 | N/K | N/K |
| **Bouwhuis &. Moll (2003)** | N/A | 335 | 99 | 29.6% | Hospital Internal Survey to parents of children | Sep - Dec 99 | Parents' country of birth (COB) | 0.40 - 0.95 | 0.80 - 0.99 | 0.61 - 0.86 | N/K |
| **Nicoll, Bassett, & Ulijaszek (1986)** | N/A | 846 | 348 | 41.1% | (a)Child Register, (b)School Survey (c)Stillbirth Certificate | N/K | Ethnicity [(3rd pty. (a),parents (b)]; Mother COB (c) | 0.67-1.00 | 0.92 - 1.00 | 0.72-1.00 | 0.96-1.00 |
| **Harland, White & Bhopal (1997)** | N/A | 129,914 | 1,702 | 1.3% | Family Health Service Authority Register (FHSA) | 1991 | Individual contact | N/K | 1.00 | 0.95 | N/K |

**Table 3: Summary of Target Population characteristics and results of the evaluation of classification accuracy in the 13 papers reviewed.**
'E.M.' = Ethnic Minorities; COB = Country of Birth; 'N/K' = Not Known; 'PPV'= Positive Predictive Value; 'NPV'= Negative Predictive Value. A ranges of values is included here when a study reports several values of results for different subpopulations (e.g. by gender or ethnic group), or under different evaluation criteria.

| | **Gold Standard ('true' ethnicity)** | |
|---|---|---|
| **Classification (predicted ethnicity)** | Ethnic Group X | Other Ethnic Groups |
| Ethnic Group X | **a** | b |
| Other Ethnic groups | c | **d** |

Measures of classification accuracy:

*Sensitivity = a / (a + c)*

*Specificity = d / (b + d)*

*Positive Predictive Value (PPV) = a / (a + b)*

*Negative Predictive Value (NPV) = d/ (c + d)*

**Table 4: Explanation of measures of classification accuracy: Sensitivity, Specificity, PPV and NPV**

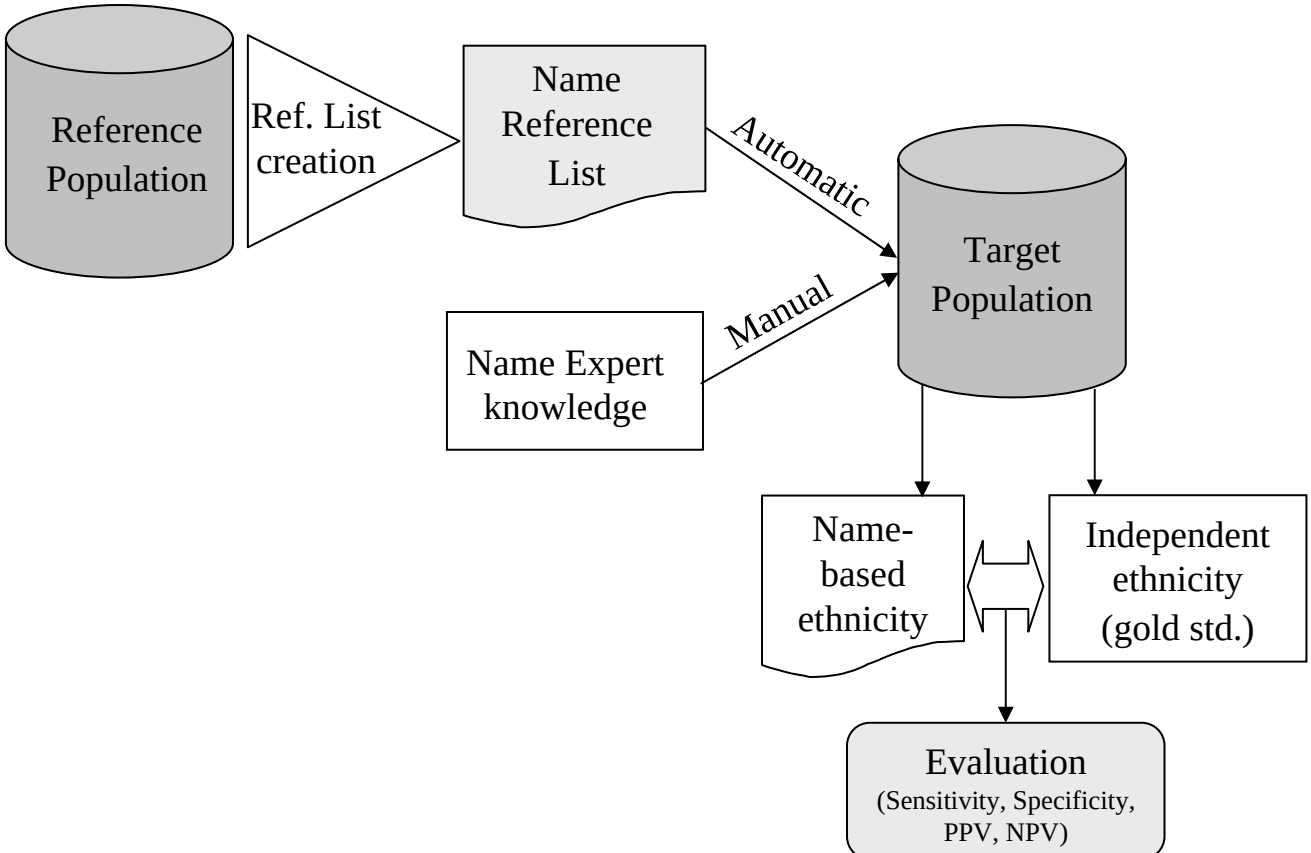


Figure 1: Structure and processes of Name Classifications Evaluated