International Journal of Population Data Science

Journal Website: www.ijpds.org

How Ethnic Name Characteristics Influence String Similarities and Cause Bias when Linking Large Population Databases

Yashithi Dharmawimala[1], Peter Christen[1,2,*], Sumayya Ziyad[1], Rainer Schnell[3], Joseph Lam[4], Anushka Vidanage[1], and Thilina Ranbaduge[1,5]

Submission History

Submitted:	09/05/2026
Accepted:	22/08/2026
Published:	12/11/2026

[1] Australian National University, Canberra, 2600, Australia
[2] SCADR, University of Edinburgh, Edinburgh, EH3 9EF, United Kingdom
[3] University of Duisburg-Essen, Duisburg, 47057, Germany
[4] Great Ormond Street Institute of Child Health, University College London, London, WC1N 1EH, United Kingdom
[5] Data61, CSIRO, Canberra, 2601, Australia

Abstract

Comparing personal names is at the core of most record linkage systems, where it is common for names to be recorded with variations, either erroneous or intentional, and individuals can change their names over time. Therefore, string comparison functions are commonly employed to calculate similarities between names. Popular comparison functions include Jaro-Winkler, edit distance, and q-gram based techniques. Although the question of which functions are more suitable for a given record linkage application has received little attention, the selection of a suitable string comparison function can be crucial in the context of bias and fairness. This is particularly important when linking databases that contain ethnically diverse names. In this work, we investigate how different string comparison functions can result in various levels of bias when names from diverse ethnic groups are compared. We conduct extensive experiments on a large public population database where for each record an ethnicity is available. By comparing name variations of the same person using multiple string comparison functions, we show that different functions exhibit diverse similarity distributions for names from different ethnic groups. Moreover, the similarity score for the same name pair can vary greatly depending on the comparison function employed. Our results provide important insights into how practical record linkage of population-level databases should be conducted to limit bias with regard to ethnic groups.

Keywords
Record linkage, Data linkage, Entity resolution, Ethnicity, String comparison, Bias, Fairness

*Corresponding Author:
Email Address: peter.christen@anu.edu.au (Peter Christen)

https://doi.org/10.23889/ijpds.v11i1.3767
November 2026 © The Authors. Open Access under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.en)

Introduction

Record linkage aims to identify and match records from different data sources that refer to the same real-world entities, such as individuals. Comparing personal names[1] across records is a crucial operation in many record linkage systems [1]. However, when capturing names, they can be prone to variations due to typographical errors, phonetic spelling, transliteration, cultural naming conventions, scanning artefacts, as well as genuine name changes of individuals over time [2]. As a result, exact matching of names is rarely sufficient when aiming to identify records that refer to the same person. Therefore, string comparison functions are widely used to quantify a similarity between names [3].

Over the years, many string comparison functions have been proposed, such as edit distance, q-gram based (Dice and Jaccard coefficients), Editex, the Jaro-Winkler comparator, and various others [3–5]. However, the choice of which comparison function to use is often made pragmatically and with limited consideration of bias and fairness with regard to ethnic characteristics of the names being compared.[2] While ethnic bias in record linkage is a known problem [7–10] (as it is in data science more generally [11]), only limited work has explored how to recognise and address such bias with regard to any sub-populations when generating linked data sets.

In this study, we investigate whether commonly used string comparison functions behave unevenly when applied to names from different ethnic groups, raising concerns around bias and fairness in record linkage systems. Rather than focusing on linkage accuracy, we examine the distributions and skewness of the similarity values calculated by different string comparison functions, and how these vary across names from different ethnic groups.

Name changes and variations can occur because of a wide range of reasons, such as due to marriage (where a person changes their surname to that of their spouse, or surnames are compounded, like *Taylor-Thomas*) or assimilation processes (for example, the German name *Josef Käser* might be changed to *Joe Kaeser* after immigration to the US). Errors and assumptions during data capturing and data processing can also lead to variations of names [2].

In this work, we conduct an observational study to show that names from different ethnic groups differ in their similarities when different string comparison functions are employed. We are not considering the underlying reasons for these differences because these reasons can be manifold, and in a given record linkage project one is generally limited to working with the personal information available in the databases being linked. We therefore only assume that each record includes a value that indicates an ethnic group, and we do not make any assumptions on why name changes or variations occur for true matching record pairs.

We use pairs of names where the two names differ, and where we know both names occur in records of the same person (the record pair is therefore a true match [3]). This allows us to calculate recall (sensitivity) [12] of how many name pairs would be correctly considered as a match for different string similarity thresholds. Based on these results, we discuss how understanding the differences in similarities is essential to minimise ethnic bias and improve linkage quality in record linkage systems.

[1]For the remainder of this paper we simplify 'personal name' to 'name'.

[2]The continuous and multidimensional diversity of humans can be divided into categories, which are shaped by historical, political, and social forces. Therefore, for a person, no ethnic group can be identified by an objective method (Kaufman, 2025 [6], p. 31). We use 'ethnic group' as a global label to denote a proxy characteristic, which might be used to differentiate the naming practices of language groups. However, this label might be named by others in their work as 'race', 'ethnicity' or 'minority'.

Related Work

Despite record linkage being an active area of research for over eight decades [13–15], limited work has been conducted to investigate how record linkage methods can influence linkage quality and bias when records with names from diverse ethnic groups are being compared.

In early work, De Brou and Olson [16] highlighted that name variations across different languages and ethnic groups are a significant challenge for automated record linkage. This is because language-specific phonetic systems, such as Soundex [17] and FONEM [18], have been shown to be inadequate for linking multi-ethnic population data. As a remedy of this problem, De Brou and Olson [16] proposed the Guth algorithm, a string comparison function that is based on letter differences but is language agnostic. Similarly, Taft (1970), the principle developer of the NYSIIS (New York State Identification and Intelligence System) encoding and search system, observed: "*Existing techniques were developed to handle 'Anglo-Saxon' names in a high accuracy environment while actual search experience indicates a high proportion of 'Latin' names and very poor accuracy in the reporting process*." [19].

Bohensky et al. [20] conducted a systematic review of over thirty studies examining patient characteristics associated with incomplete record linkage, finding that several characteristics were consistently linked to uneven match rates: age (72% of studies), sex (50%), ethnic groups (64%), socio-economic status (82%), and health status (72%). Of particular relevance, seven studies found that minority groups had lower linkage rates, attributed to treatment at facilities with poorer data quality and lower rates of having key identifiers recorded. This work [20] showed that linkage errors are not random, but cluster around specific groups in a population. This confirms that record linkage methods that employ string comparison functions might fail more for some ethnic groups than others.

DuVall et al. [21] discussed that Hispanic names and names in other minority groups were overrepresented among identified matches when analysing duplicate records in a healthcare data warehouse. Some key factors affecting this were that minority names are being less familiar to clerks and thereby more prone to transcription errors, and non-western naming conventions that do not adhere to database designs built around first-middle-last name structures.

In a general discussion on linkage bias, Bohensky [8] described that different linkage rates for names from diverse ethnic groups have been reported by several studies. Reasons for such bias included that data entry personnel unfamiliar with such names can lead to an increase in misspellings, and that ethnic name structures which do not follow the Western first name and last name assumption can lead to name values

being entered into the wrong field leading to miscalculated similarities.

Randall et al. [22] investigated if errors in linkage quality are distributed evenly across socio-demographic groups by de-duplicating four large Australian administrative health data sets, and comparing results against a gold standard benchmark. Their findings showed that linkage quality varied across groups, where poorer quality was observed for younger individuals across all four data sets, and for those in remote areas across three of the four data sets. The study found that differences in linkage quality across socio-economic categories were highly data dependent. While this study did not explore ethnic bias in string comparison functions, it showed that linking without considering different error rates across groups can systematically create bias in research outcomes.

Mc Grath-Lone et al. [9] discussed that ethnic bias in record linkage can arise from structural factors such as ethnic minorities having missing or inaccurate records and poor overall data quality. Another aspect identified was that record linkage algorithms are generally designed around Western naming conventions and therefore they fail to accommodate alternative name structures.

Barboi et al. [23] explored whether the accuracy of probabilistic record linkage techniques applied on patient records varies across the socio-demo-graphic categories age, sex, race, and ethnic groups, using four health data sets from the US state of Indiana and a manually generated gold standard. Applying a modified Fellegi-Sunter probabilistic linkage algorithm [15] under a missing at random assumption [24], they found that matching performance was not demographically uniform, with F1-scores [25] ranging from 0.85 to 0.99 across different ethnic groups. These results indicate substantial variations in linkage quality for individuals from different ethnic groups, and also depending on the data set.

Lam et al. [26] investigated how heterogeneity in error mechanism (how names are changing) and inequality in error exposure contributes to ethnic differences in name-based record linkage. Using the North Carolina Voter Registration (NCVR) data set [27] (which we also use in our evaluation), error profiles were empirically generated by tracking name changes for the same individuals across ten snapshots of NCVR to parametrise realistic corruptions. Then, five replicate corrupted data sets were generated under three synthetic corruption settings and evaluated using three linkage models: (1) unadjusted Jaro-Winkler, (2) Term-Frequency (TF) adjusted Jaro-Winkler (which down-weights similarities for common names and up-weights similarities for rare names), and (3) a novel clustering-based name embedding model. These three settings were tested against the false match rate (false positive rate) and missed match rate (false negative rate) [12] outcomes across the different ethnicities that occur in NCVR. The results showed that some ethnic groups were more prone to recording errors and inconsistencies in names, leading to more biased linkage outcomes across ethnic groups than differences in name structure alone. Hispanic voters were up-to 36% more likely to be missed than Non-Hispanic voters, and up-to 60% more Indigenous or Pacific Islander voters were missed. While the TF adjusted Jaro-Winkler comparison reduced the effects of ethnic bias in linkage quality relative to unadjusted string similarities, it did not eliminate them. This indicates that penalising common names alone is insufficient to address the structural differences in how string comparison functions are affected by ethnically diverse name distributions. This work showed that the choice of string comparison function and its configuration directly influences the ethnic inequalities in linkage outcomes.

Besides these challenges with the ethnic differences of names in the context of record linkage discussed so far, name ethnicity has been considered in the analysis of (administrative) data more generally. Mateos [28] conducted a systematic review of thirteen name-based ethnic group classification studies. These showed that while name-based ethnicity classifiers are an alternative when self-reported ethnicity is not available, their accuracy varies considerably across groups, with sensitivity [12] ranging from 0.67 to 0.95 across studies. The review identified spelling variations, transliteration differences, and culturally distinct naming conventions as key sources of misclassification.

Schnell et al. [29] showed that sampling migrants based on names can introduce what they describe as *assimilation bias*. In their analysis, migrants who are more assimilated are more likely to be missed by name-based classifiers, as their names more likely resemble names that already occur in the host population and are therefore not flagged as foreign. Using a trigram-based Bayesian classifier applied to a large German survey data set, the authors showed [29] that false negative rates are not uniform, but rather they vary systematically with variables associated with assimilation. These results highlighted that string-based name classification methods can perform unequally across sub-populations.

Ye et al. [30] introduced *NamePrism*, a nationality classifier that learns name embeddings. The resulting embeddings were clustered by ethnic groups and nationality, demonstrating that names carry ethnically structured information at the character and pattern level. Similarly, Hafner et al. [31] evaluated three name ethnicity classifiers showing differences in their accuracy across ethnic groups. Their findings showed substantial variation in classification performance, with ranges of sensitivity from 33% for African names to 95% for East Asian names. Based on results from three surveys conducted in the US, where over 44,000 names were evaluated, Crabtree et al. [32] found that it is possible to assign socio-economic attributes such as income, education, and citizenship to names, however even the most specific ethnic names exhibit ambiguity with these attributes.

These studies reviewed showed that string comparison functions do not perform uniformly across ethnic groups. As discussed, name variations across ethnic groups seem to be a key factor for biased linkage results [8]. Furthermore, work on trying to infer ethnicity from names alone [30, 31, 33] demonstrated that name-based classification methods exhibit systematic differences in performance across ethnic groups. Detailed analyses revealed that bias can arise from factors like naming patterns, transliteration, and spelling variations [9, 16, 28]. Empirical work also demonstrated that the choice of string comparison function can systematically affect false negative rates across different ethnic groups [23, 26, 29].

Foxcroft et al. [34] proposed *Name2Vec*, a *Doc2Vec*-based method for calculating name similarity by creating name embeddings where each name is treated as a document and each letter in the name is considered a word. This allows

embeddings to be learned directly from the internal character structure of names. Using two data sets containing surnames, the authors showed that Cosine distance between the resulting embeddings can be used to classify a name pair as a match or non-match with high accuracy. This work did however not consider the ethnicity of names.

To the best of our knowledge, no systematic analysis has been conducted on how different string comparison functions perform across ethnic groups when applied to the same data set. Existing works either evaluated a single comparison function, focused on linkage outcomes rather than the comparison function itself, or only demonstrated ethnic inequalities as a finding rather than investigating their impacts. Our work addresses these gaps by directly comparing how different string comparison functions result in diverse distributions of similarities when names from different ethnic groups are being compared.

Methodology

We now describe our approach to investigate how different string comparison functions can result in diverse similarity distributions for names which are predominantly used by individuals with different ethnic backgrounds. We first discuss the data concepts we used in our work, then describe how we associated names to ethnic groups, and finally, how we conducted the comparison of name pairs to calculate their similarities. In the Experimental Evaluation section we then apply our methodology on a large real-world population data set, where in the Results and Discussion section we present and discuss our findings.

Notation and Data Concepts

We denote a database containing records of individuals with $\boldsymbol{D}$, and an individual record as $r_i \in \boldsymbol{D}$, where $1 \leq i \leq m$, and $m = |\boldsymbol{D}|$ is the number of records in $\boldsymbol{D}$. We assume each record contains a set of quasi-identifier (QID) attributes, such as the names, addresses, and dates of birth of individuals [3]. For our work, we require each record to include at least one name attribute (such as first name or last name), as well as an ethnicity or race attribute (we use the term *ethnic group* for the remainder of our work).

We denote these two QID attributes of record r_i as n_i (for name) and e_i (for ethnic group). Names are assumed to be textual strings while ethnicities are categories (such as *Asian*, *Black*, *White*, and so on). We denote the set of all unique names in a database with $\boldsymbol{n}$, and the set of all ethnic categories with $\boldsymbol{e}$. We assume $\boldsymbol{n}$ to be large (in the tens to hundreds of thousands of unique names) while $\boldsymbol{e}$ will contain a small number of ethnic categories (possibly below ten).

We assume that for each record it is known to which real-world entity it refers to, where we denote two records, r_i and r_j, that refer to the same entity with $r_i \equiv r_j$. If two databases, $\boldsymbol{D}_A$ and $\boldsymbol{D}_B$, are being linked, then the set of true matches, $\boldsymbol{M}$, are all record pairs (r_i, r_j) across the two databases where it holds that $\boldsymbol{M} = \{\forall (r_i, r_j) | r_i \equiv r_j \wedge r_i \in \boldsymbol{D}_A \wedge r_j \in \boldsymbol{D}_B\}$. The set of true non-matches, $\boldsymbol{U}$, are all pairs (r_i, r_j) where it holds that $\boldsymbol{U} = \{\forall (r_i, r_j) \mid r_i \neq r_j \wedge r_i \in \boldsymbol{D}_A \wedge r_j \in \boldsymbol{D}_B\}$. Overall, we have $|\boldsymbol{M}| + |\boldsymbol{U}| = |\boldsymbol{D}_A| \cdot |\boldsymbol{D}_B|$.

We compare the name values in each record pair in $\boldsymbol{M}$ using multiple of the many available string comparison functions, such as edit distance or the Jaro-Winkler comparator [3, 35], where we are interested in how such functions lead to different similarity distributions for names from diverse ethnic groups. We denote a string comparison function with f, where $s_{i,j} = f(n_i, n_j)$ and $0 \leq s_{i,j} \leq 1$ is the similarity value calculated by f for the pair of names n_i and n_j. For all string comparison functions we investigate experimentally in the Experimental Evaluation and Discussion sections, we assume that a similarity of 1 means two name strings are the same ($n_i = n_j$), while a value of $0 \leq s_{i,j} < 1$ means $n_i \neq n_j$. We expect that two names which are totally different, for example do not share a single letter (such as *David* and *Peter*), will result in a similarity of $s_{i,j} = 0$ for all string comparison functions [3].

Assigning Names to Ethnic Groups

The first step in our approach is to determine how frequently names occur in different ethnic groups in the databases being linked. Based on this, we then extract the set of names that predominantly occur with a single ethnic group. For example, *John* might occur commonly with records that have *White*, *Asian* and *Black* as their ethnicities, while *Priya* might only occur with records that have the *Asian* ethnic group. Names that are more common in a population are more likely to occur in records with different ethnicities, while names that are rarer in a population more likely occur with a single ethnic group (a name that occurs in a single record will only occur with one ethnic group), as we discuss further in the Experimental Evaluation section.

We first obtain the name pairs from all true matching record pairs in $\boldsymbol{M}$ and their corresponding ethnicities. For each unique name, we then count in how many record pairs it appears for each ethnic group. To hold these counts, we define the matrix $\boldsymbol{E}$ with $|\boldsymbol{n}|$ rows (one per unique name $n_x \in \boldsymbol{n}$) and $|\boldsymbol{e}|$ columns (one per category, $e_y \in \boldsymbol{e}$), with $1 \leq x \leq |\boldsymbol{n}|$ and $1 \leq y \leq |\boldsymbol{e}|$. The element $\boldsymbol{E}[x, y]$ holds the number of times when one of the records r_i or r_j in a pair $(r_i, r_j) \in \boldsymbol{M}$ had the name n_x and the corresponding ethnic group was e_y:

$$\boldsymbol{E}[x, y] = |\{\forall (r_i, r_j) \in \boldsymbol{M} \mid ((n_i = n_x) \wedge (e_i = e_y)) \vee ((n_j = n_x) \wedge (e_j = e_y))\}|$$

This means each name in a record pair is counted once, and each pair contributes to two elements in $\boldsymbol{E}$ (if $n_i \neq n_j$ or $e_i \neq e_j$), or it contributes twice to a single element (if $n_i = n_j$ and $e_i = e_j$). After all record pairs in $\boldsymbol{M}$ have been analysed with regard to their names and ethnic groups, it will hold that $2 \cdot |\boldsymbol{M}| = \sum_{x,y} \boldsymbol{E}[x, y]$. Each row x in $\boldsymbol{E}$ contains the number of times the name n_x occurred in a record with one of the different ethnic groups, while each column y in $\boldsymbol{E}$ contains the number of times each name was in a record with the given group e_y.

To identify names which predominantly occur with a certain ethnic group, we now find for each name $n_x \in \mathbf{n}$ the group it occurs most frequently with (for which the name n_x has the highest count in row $\boldsymbol{E}[x, :]$), denoted by p_x (predominance), where $p_x \in \boldsymbol{e}$:

$$p_x = argmax_y(n_x \in \boldsymbol{n} | max(\boldsymbol{E}[x, y]))$$

To select names based on the distribution of their ethnicities, we assume a threshold t_e, with $0 \leq t_e \leq 1$, which we use to decide if a name occurs predominantly in records with one ethnic group, or if it occurs commonly in records with multiple ethnicities. For each ethnic group, $e_y \in \boldsymbol{e}$, we extract the set of names $\boldsymbol{n}_y^{t_e}$ which predominantly occur with this ethnic group based on the threshold t_e as:

$$\boldsymbol{n}_y^{t_e} = \{n_x \in \boldsymbol{n} | (p_x = e_y) \wedge (max(\boldsymbol{E}[x,:])/\sum(\boldsymbol{E}[x,:]) \geq t_e)\}$$

For each name $n_x \in \boldsymbol{n}$, we assign this name to the set $\boldsymbol{n}_y^{t_e}$ of names of ethnic group e_y if the proportion of times the name occurs in records with ethnic group e_y is at least t_e based on the total number of times the name occurs in records with any ethnic group. Each name is only added to one such set of ethnic names, and many names will not be added to any set depending upon how often a name occurs with the different ethnic group categories and the value of the threshold t_e. The higher t_e is, the smaller the sets $\boldsymbol{n}_y$ become.

For each ethnic group $e_y \in \boldsymbol{e}$, we end up with a set of names which we use to decide which name pairs to compare when we calculate the similarities between names, as we discuss next.

Identifying Name Changes and Calculating Similarities

Since we are interested in name variations by the same person (and therefore we do not include exact name matches), and their corresponding similarities, we only consider true matching record pairs in the set $\boldsymbol{M}$, where both records have the same ethnic group but different name values. Formally, for each ethnic group $e_y \in \boldsymbol{e}$, we generate the set of name pairs $\boldsymbol{N}_y^{t_e}$ (for a given threshold t_e) as follows:

$$\boldsymbol{N}_y^{t_e} = \{\forall(r_i, r_j) \in \boldsymbol{M} | (n_i \neq n_j) \wedge (n_i \in \boldsymbol{n}_y^{t_e}) \wedge (n_j \in \boldsymbol{n}_y^{t_e})\}$$

For each ethnic group $e_y \in \boldsymbol{e}$, comparing its name pairs in $\boldsymbol{N}_y^{t_e}$ using a string comparison function f results in the list $\boldsymbol{S}_y^f$ of name similarities:

$$\boldsymbol{S}_y^f = \left[f(n_i, n_j) | \forall(r_i, r_j) \in \boldsymbol{N}_y^{t_e}\right]$$

$\boldsymbol{S}_y^f$ is a list and not a set because more than one name pair in $\boldsymbol{N}_y^{t_e}$ can have the same similarity value. Our aim is to analyse the similarity distributions of these lists for (1) different ethnic groups and (2) for different string comparison functions, as we discuss in the following section.

Experimental Evaluation

We now discuss how we applied our method to explore similarity distributions for names that predominantly occur with a certain ethnic group. We first describe the data set we used for our experiments, and then discuss how we set-up the comparison of name pairs. Finally, we discuss how we visualise the results.

Data Set and Experimental Set-Up

For our experiments, we used the North Carolina Voter Registration (NCVR) data set,[3] which is publicly available and is commonly used for record linkage research [27], including studies that examine the role of ethnic groups in record linkage [26]. We specifically obtain all records by the same voter using their *North Carolina IDentification* (NCID) number, a unique identifier for each voter that is stable over time. As the NCID uniquely identifies a single individual across all snapshots, every record pair derived from the same NCID is a true match. This ensures that the names in each record pair reflect genuine variation in how one individual's name was recorded over time, rather than differences between distinct individuals' names. We therefore restrict our analysis to true matching record pairs and do not include true non-matching pairs. For each voter, we track changes in their *First Name* (FN) and *Last Name* (LN) attributes over 61 snapshots of NCVR spanning from 2016 to 2026, containing a total of 368,791,415 records of active voters. In total, we obtain 66,963 pairs of FNs and 305,516 pairs of LNs where the two names in a record pair (r_i, r_j) are not the same $(n_i \neq n_j)$ but the two records have the same ethnic group $(e_i = e_j)$.

We consider the following grouping of ethnicities from the categories provided in the NCVR attribute *Race code*: *White*, *Black*, *Asian*, and *Grouped Others* (Grp Oth). The latter category consists of the ethnicities *Mixed*, *Alaska Native or American Indian*, and *Other*. We group these three categories because individually their corresponding numbers of name pairs were too low compared to the other categories, and some of these categories were also not available over the full eleven-year time period of all snapshots. Following our notation from the Methodology section, the resulting set of categories we use is $\boldsymbol{e} = \{$Asian, Black, White, Grouped Others$\}$. Table 1 shows the total numbers of records with these categories in the NCVR data set, and the corresponding percentages of each category. As can be seen, the majority of voters identify themselves as being *White*, followed by *Black*, *Grouped Others* and *Asian*. The remaining 25,727,565 records have no value (a missing ethnic group) or a value of 'unknown'. We note that ethnicities in the NCVR data set are self-reported, and therefore the same voter might have a changing ethnicity over time [42].

In Table 2 we provide a summary of the NCVR data set characteristics across each of these four ethnic categories. For each category, the table shows the total number of name pairs obtained (where the two names are different), the number of unique names in these pairs, and the average and maximum frequencies of names in these pairs. From this table we can see that the *White* ethnic category contains the largest number of name pairs as well as unique names, particularly for last names, and also exhibits the overall highest name frequencies. In comparison, the other categories have smaller numbers of name pairs, and correspondingly lower average and maximum name frequencies.

To gain further insight into the characteristics of the names within each ethnic category, we identified the names (both first and last names) which occur both commonly in the NCVR data set as well as predominantly occur with one category only. As we describe in detail in Appendix A, for each name we calculated a score that combines the frequency of the name

[3] See: https://dl.ncsbe.gov/?prefix=data/.

Table 1: Numbers of records and corresponding percentages for the four ethnic categories across all snapshots in the NCVR data set

	White	Black	Asian	Grouped others
Records	248,791,460	73,451,910	5,395,065	15,425,415
Percentage	67.5%	19.9%	1.5%	4.2%

Note that 6.9% of records (25,727,565) had either a missing or 'unknown' ethnic category.

Table 2: Characteristics of first name (FN) and last name (LN) pairs by ethnic category (EC) as obtained from true matching record pairs in the NCVR data set

EC	Name pairs		Unique names		Avg. frequency		Max. frequency	
	FN	LN	FN	LN	FN	LN	FN	LN
Asian	3,427	2,595	5,467	3,828	1.25	1.36	52	63
White	39,383	227,241	15,928	90,212	4.95	5.04	1,952	5,016
Black	20,713	60,581	22,281	29,282	1.86	4.14	254	2,126
Grp Oth	3,440	15,099	5,007	16,598	1.37	1.82	66	381

Table 3: Top ten first names by ethnic group and their group-specific frequency scores, as further discussed in Appendix A

Asian	Black	White	Grouped others
thanh (0.001)	willie (0.004)	james (0.015)	jose (0.008)
mai (0.001)	jasmine (0.002)	john (0.014)	juan (0.004)
phuong (0.001)	reginald (0.002)	william (0.014)	luis (0.004)
hong (0.001)	annie (0.002)	michael (0.013)	carlos (0.002)
young (0.001)	latoya (0.002)	robert (0.013)	jesus (0.002)
thuy (0.001)	antonio (0.002)	david (0.012)	jorge (0.002)
anh (0.001)	marcus (0.001)	mary (0.008)	ana (0.002)
muhammad (0.001)	derrick (0.001)	thomas (0.007)	miguel (0.002)
thao (0.001)	ebony (0.001)	richard (0.007)	francisco (0.001)
huong (0.001)	tyrone (0.001)	christopher (0.007)	hector (0.001)

Table 4: Top ten last names by ethnic group and their group-specific frequency scores, as further discussed in Appendix A

Asian	Black	White	Grouped others
nguyen (0.023)	williams (0.011)	smith (0.008)	locklear (0.018)
patel (0.021)	jackson (0.003)	johnson (0.004)	oxendine (0.006)
yang (0.009)	robinson (0.003)	miller (0.004)	rodriguez (0.004)
kim (0.009)	alston (0.003)	jones (0.003)	hernandez (0.004)
tran (0.008)	washington (0.002)	davis (0.003)	garcia (0.003)
le (0.007)	richardson (0.002)	brown (0.003)	martinez (0.003)
chen (0.006)	battle (0.001)	wilson (0.002)	gonzalez (0.003)
wang (0.006)	graham (0.001)	taylor (0.002)	lopez (0.003)
vang (0.005)	mckoy (0.001)	martin (0.002)	perez (0.002)
li (0.004)	hairston (0.001)	moore (0.002)	chavis (0.002)

with how specifically it occurs in its most predominant ethnic group.

The results obtained from calculating this score for first and last names are presented in Tables 3 and 4, respectively, where we show the top ten names with the highest scores per ethnic category. An interesting observation is that for both first and last names the *Grouped Others* names seem to be mostly Hispanic or Latino names. However, we retain the *Grouped Other* label rather than *Hispanic* since this category also includes non-Hispanic names from the constituent groups such as *Mixed* and *Other*.

The opposite of names occurring predominantly in one ethnic group category are names that occur with a (close to) uniform frequency across all ethnicities. In Table 5 we show the top ten such first and last names ranked by their uniformity score, calculated as described in Appendix A. As can be seen, these names seem to come from diverse ethnicities, which indicates that they are not strongly associated with a single ethnic group. The distribution of these names shows patterns of usage of names across cultures, which can be due to religious or migration influences (such as for *habib* and *rashid*).

Table 5: Top ten first and last names ranked by their uniformity score (calculated as described in Appendix A), and the number of records in the NCVR data set that contain these names

First Name (Asian, Black, White, Grouped Others)	Last Name (Asian, Black, White, Grouped Others)
iwan (60, 60, 60, 60)	vaz (428, 418, 432, 385)
isalene (60, 64, 64, 56)	rashid (1599, 1461, 1439, 1730)
farina (71, 60, 60, 60)	sika (88, 81, 92, 70)
habib (469, 501, 573, 485)	labador (64, 60, 77, 78)
saeed (635, 809, 665, 780)	baluch (275, 202, 249, 285)
malai (58, 70, 60, 75)	diokno (60, 40, 60, 54)
neki (57, 75, 60, 62)	tacla (40, 60, 60, 60)
clarita (279, 306, 373, 303)	mana (107, 75, 104, 78)
midori (321, 240, 279, 326)	carrazco (6, 9, 69, 82)
dorotea (88, 116, 120, 128)	kamal (556, 400, 451, 641)

Table 6: The numbers of record pairs with the same or different ethnic categories (EC), Where the first record, r_1, occurs in an earlier NCVR snapshot and the second record, r_2, in a later snapshot, and their corresponding name values are different

	EC of r_2							
	First name				Last name			
EC of r_1	Asian	White	Black	Grp oth	Asian	White	Black	Grp oth
Asian	3,427	18	7	266	2,595	74	24	327
White	30	39,383	48	311	84	227,241	159	2,337
Black	6	40	20,713	212	20	160	60,581	805
Grp Oth	336	363	238	3,440	280	2,114	815	15,099

Table 7: The numbers of name pairs selected into ethnic group sets for varying values of the threshold t_e

	First name				Last name			
t_e	Asian	White	Black	Grp oth	Asian	White	Black	Grp oth
1.0	746	607	4,111	281	240	8,787	2,407	1,510
0.9	820	3,679	6,333	294	273	42,942	3,190	1,590
0.8	1,129	14,541	7,433	320	435	71,664	3,935	1,798
0.7	1,581	25,682	8,268	386	639	111,005	4,768	2,008
0.6	1,929	31,597	9,124	538	760	157,451	5,982	2,420
0.5	2,137	34,700	9,847	649	889	195,654	9,092	5,319

Table 6 then shows the numbers of record pairs for all possible pairs of ethnic categories for first and last names. As can be seen from this table, for the majority of pairs both records have the same ethnic category. The *White* ethnic category has the largest number of record pairs for both name components, followed by the number of record pairs with the *Black* category. Both the *Asian* and *Grouped Others* ethnic categories occur in much lower numbers of record pairs, where especially for last names the number of *Asian* record pairs is substantially smaller than all other ethnic categories. Note also that *Grouped Others* is the category with the largest number of changes with any of the other ethnicities for both name components. In all our experiments, we will only consider the pairs where both ethnic categories are the same.

As described in the Methodology section, we generated the matrix $\boldsymbol{E}$ of counts for all names and how often they occur with the different ethnic categories, where we created one separate such matrix for first names and last names. In Table 7 we show how many name pairs will be included into the sets of ethnic name pairs, $\boldsymbol{N}_y^{t_e}$, with $e_y \in \boldsymbol{e}$, when the threshold t_e is set to different values ranging from 0.5 to 1. The number of pairs increases monotonically as t_e is being lowered. If we set $t_e = 0.5$ then a name pair which in half of its occurrences has one ethnic category will be included into the corresponding set of ethnic name pairs, independently of the occurrence of the pair with the other three ethnicities.

For the following analysis we set $t_e = 0.8$, because this ensures that a name pair predominantly (80% of the time) occurs in one specific ethnic group only. At the same time, with this threshold value we do not end up with too few name pairs that can be compared in all ethnic categories, as Table 7 shows. While higher thresholds would select name pairs that occur more exclusively with a single ethnic group, at the same time it would also limit the number of name pairs, especially for the *Grouped Others* (for first names) and *Asian* (for last names) categories given the smaller number of names in these groups.

Table 8: The average and maximum Frequencies of names included in Pairs in Table 7 for Different Similarity Thresholds, t_e. We Omit Minimum Frequencies as These Were All 1

$\mathbf{t}_e$	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
			First name			
Avg	110,130.7	116,227.5	124,587.5	81,557.0	32,514.7	74.5
Max	6,133,295	6,133,295	6,133,295	2,488,749	972,615	8,528
			Last name			
Avg	120,449.1	124,547.3	36,138.5	6,890.6	4,997.3	344.7
Max	3,815,281	3,815,281	874,567	207,231	207,231	19,297

Furthermore, the name pairs selected with higher threshold values are more likely to be rarer names, because many commonly used names occur across multiple ethnic groups (due to reasons such as migration or marriage). In Table 8 we show the average and maximum frequencies of the names (the numbers of records in the NCVR data set containing these names) in the name pairs that were selected for different values of the threshold t_e. As can be seen in this table, the numbers become very small, especially for $t_e = 1$.

As such, based on the availability of name pairs for analysis based on Tables 7 and 8, for $t_e = 0.8$ we obtain the smallest number of name pairs with the *Grouped Others* category (320 pairs) for first name and with the *Asian* category for last name (435 pairs). To ensure a fair analysis of the distribution of name similarities (as we present in the Results and Discussion section), for all other ethnic categories we randomly sample this corresponding minimum number of pairs. This means for first name we use a sample of 320 name pairs for each ethnic category, while for last name we use a sample of 435 pairs for each category.

Calculating Name Similarities

Considering only name pairs in the sets $\boldsymbol{N}_y^{t_e}$ (based on names included in the sets $\boldsymbol{n}_y$ when using the threshold $t_e = 0.8$), we then calculated the similarities between all name pairs in these sets. For this, we used the following four string comparison functions [3]:

- Edit distance (ED): This comparison function calculates the minimum number of single character operations (insertions, deletions, and substitutions) required to transform one string into another [36, 37].
- Editex (EX): This comparison function extends edit distance by grouping letters with similar pronunciation. Substitution costs are then reduced for phonetically similar characters in the same group, and duplicate consecutive letters are ignored. This approach will lead to higher similarities for names which have phonetic variations compared to typing or OCR errors [3, 38].
- Q-gram Dice (QD): This comparison function first converts the two strings being compared into one set each of overlapping substrings of length q characters (q-grams). As commonly used in record linkage [3], we set $q = 2$ (bigrams) in our experiments. The similarity is then calculated using the Dice coefficient [3] based on the intersection of the two sets and the sum of the numbers of q-grams in the two sets being compared.
- Jaro-Winkler (JW): This comparison function combines ideas from the edit distance and q-gram based string comparison functions by calculating a similarity based on the number of characters that are in common as well as the number of transposed characters [39, 40]. A similarity is further increased if the characters at the beginning are matching [41].

We define the set $\boldsymbol{F} = \{f_{ED}, f_{EX}, f_{QD}, f_{JW}\}$ with these four string comparison functions we investigate. We selected these four functions because they are comparing strings in fundamentally different ways, and where three of these functions (edit distance, q-gram based Dice coefficient, and the Jaro-Winkler comparator) are widely used in both practical record linkage systems as well as research studies [3]. Edit distance captures typographical differences based on character-level operations, and Editex extends this idea by incorporating phonetic information which puts more importance to handling spelling variations that sound similar. The Jaro-Winkler comparator gives additional weight to common prefixes by focusing on matching characters and their relative positions and ordering. In contrast, the q-gram based comparison function captures the similarity through the overlap of substring tokens, while not considering the overall sequence of characters.

For example, when comparing the names *john* and *jonh*, a higher similarity is calculated using Jaro-Winkler (0.94) and Editex (0.63) due to the ordering of characters and their phonetic similarity. Whereas edit distance (0.5) produces a comparatively lower similarity value because two character substitutions are required (in two names of four letter length). In contrast, the q-gram based Dice similarity (0.33) is very low due to the limited commonality of q-grams (only *jo*, assuming $q = 2$) as a result of the transposition.

In Tables 9 and 10 we provide further examples of names from the different ethnic categories that result in highly different similarities when compared with the four string comparison functions in $\boldsymbol{F}$. Note the unexpected similarity of 1.0 for some name pairs for the Editex comparison function, which is because of the phonetic adjustment by Editex (such as replacing *nn* with *n* and *ii* with *i*).

The examples in Tables 9 and 10 highlight how different string comparison functions can result in very different similarities for the same name pair. As such, analysing the correlations between these functions with regard to how they perform for names from different ethnic categories will enable us to identify if some of these string comparison functions introduce bias for names from different ethnic categories.

Table 9: Examples of first name pairs with the highest absolute differences between their similarity scores for the different string comparison functions and ethnic categories

Str. comp.	Asian	Black	White	Grouped others
JW/ED	naw	mendeecees	evan	sosa
	naw nay ni win	tamiko	jacob	yris
	(0.79/0.21)	(0.48/0.0)	(0.54/0.0)	(0.55/0.0)
JW/QD	noyumah	javon	frances	serjoine
	nurul amin	jovan	renae	serojnie
	(0.66/0.0)	(0.86/0.0)	(0.74/0.0)	(0.94/0.29)
JW/EX	mular	abdou	kay	deysi
	mular lakminaqaphna	abdou salami	kay-frances	deysi guadalupe
	(0.83/0.27)	(0.86/0.41)	(0.81/0.25)	(0.84/0.33)
ED/QD	khop	shamira	jessica	yika
	khup	shariva	joshua	ylka
	(0.75/0.33)	(0.71/0.33)	(0.43/0.0)	(0.75/0.33)
EX/QD	krshna maria	akaun	frances	neisi
	maria krshna	akuan	renae	nelci
	(0.15/0.91)	(0.71/0.25)	(0.48/0.0)	(0.67/0.25)
EX/ED	me me	towana	glen	gnanapragasam
	meemee	towanna	glenn	mariasoosai
	(0.86/0.67)	(1.0/0.86)	(1.0/0.8)	(0.41/0.23)

Table 10: Examples of last name pairs with the highest absolute differences between their similarity scores for the different string comparison functions and ethnic categories

Str. comp.	Asian	Black	White	Grouped others
JW/ED	kha	carthins	aldridge	almengor-vargas
	yang charron	deloatch	leonhardt	vargas almengor
	(0.58/0.0)	(0.63/0.0)	(0.61/0.0)	(0.7/0.07)
JW/QD	laui	agu	keirn	yanruw
	luai	mbadugha	kerin	yarnuw
	(0.93/0.0)	(0.68/0.0)	(0.95/0.25)	(0.96/0.4)
JW/EX	thao	deason	kluttz	locklear
	yang	mcneely	puckett	willaughby
	(0.55/0.0)	(0.54/0.0)	(0.68/0.06)	(0.6 /0.04)
ED/QD	gao	alford butler	beck	hernandez-magana
	guo	butler-alford	sheek	magana hernandez
	(0.67/0.0)	(0.0/0.83)	(0.4/0.0)	(0.18/0.87)
EX/QD	zov	freeman-thompson	shive	espitia toledo
	zuv	thompson freeman	wike	toledo espitia
	(0.78/0.0)	(0.2/0.87)	(0.43/0.0)	(0.15/0.85)
EX/ED	vii	ajasa	currin	tiimtim
	vu	ojukwu	oquinn	timtim
	(0.67/0.33)	(0.41/0.17)	(0.47/0.17)	(1.0/0.86)

Result Visualisation

We now describe the visualisation methods we use to analyse and identify the distributions of name similarity scores, and their correlations across ethnic categories and string comparison functions. In the Results and Discussion section we then present and discuss these results.

Cumulative Distribution Plots

We generated cumulative distribution function (CDF) plots for the similarities obtained for each string comparison function and each ethnic category, as represented in the lists $\boldsymbol{S}_y^f$ described in the Methodology section. In these plots (Figures 1 to 4), the x-axis represents the similarity scores in the range

$[0, 1]$, and the y-axis represents the cumulative proportion of name pairs at or below that similarity.

For first name and last name similarity scores we generated separate CDF plots, and for each string comparison function (ED, QD, JW, EX), showing the distributions of similarity scores within each ethnic category (*Asian*, *Black*, *White*, and *Grouped Others*); and for each ethnic category showing the distributions of similarity scores for each string comparison function. These CDF plots provide insight into the overall similarity distributions across string similarity functions and ethnic categories, how these differ, and how they can bias the linkage outcomes for different ethnic groups depending upon which string comparison function is being used.

In Table 11 we then present the percentage of true matching name pairs whose similarity score meets or exceeds a given similarity threshold, t_s, for different values of t_s, ethnic category, and string comparison function. Because we only consider true matches in our analysis, these percentages correspond to the recall [12] of how many name pairs would be correctly considered as a true match.

Scatter Plots for Pairs of String Comparison Functions

We generated scatter plots to visualise how pairs of string comparison functions calculate similarities for the same name pairs for each of the different ethnic categories. This allows us to explore how switching from one comparison function to another in a record linkage system might impact the resulting similarity distributions across the ethnic categories. These pairwise scatter plots (shown in Figures 5 and 6) include polynomial trendlines (calculated using polynomial fitting based on regression analysis) of degree 2 fitted to the data, where these lines capture any (linear or non-linear) relationships between the similarity distributions.

Specifically, for each ethnic category $e_y \in \boldsymbol{e}$, and each pair of string comparison function $(f_a, f_b) \in \boldsymbol{F}$, we generated a scatter plot where each point represents a name pair $(n_1, n_2) \in \boldsymbol{N}_y^{t_n}$, with its x-coordinate produced by $f_a(n_1, n_2)$ and its y-coordinate produced by $f_b(n_1, n_2)$. The six pairs for string similarities shown in Figures 5 and 6 are ED versus QD, EX versus ED, EX versus QD, JW versus EX, JW versus QD, and JW versus ED.

Results and Discussion

Figures 1 to 4 present the cumulative distribution function (CDF) results, where the plots in Figures 1 and 2 show name similarities grouped by string comparison function for first and last names, respectively, and Figures 3 and 4 presents the CDF results grouped by ethnic category for first and last names, respectively. For each comparison function, the CDF curves for different categories are visibly different, indicating that the same comparison function produces different similarity score patterns depending on the ethnic categorisation of names, even when all pairs represent true matches where the names are different.

As can be seen from Figures 1 and 2, across all four comparison functions, the ethnic category *White* shows a CDF curve that is shifted to the left when compared to the other ethnic groups. This shows that white name pairs (for both first and last names) tend to achieve lower similarity scores overall. In contrast, for last names, the *Grouped Others* ethnic category leads to CDF curves shifted to the right, indicating that name pairs in this category tend to be more similar to one another across all comparison functions.

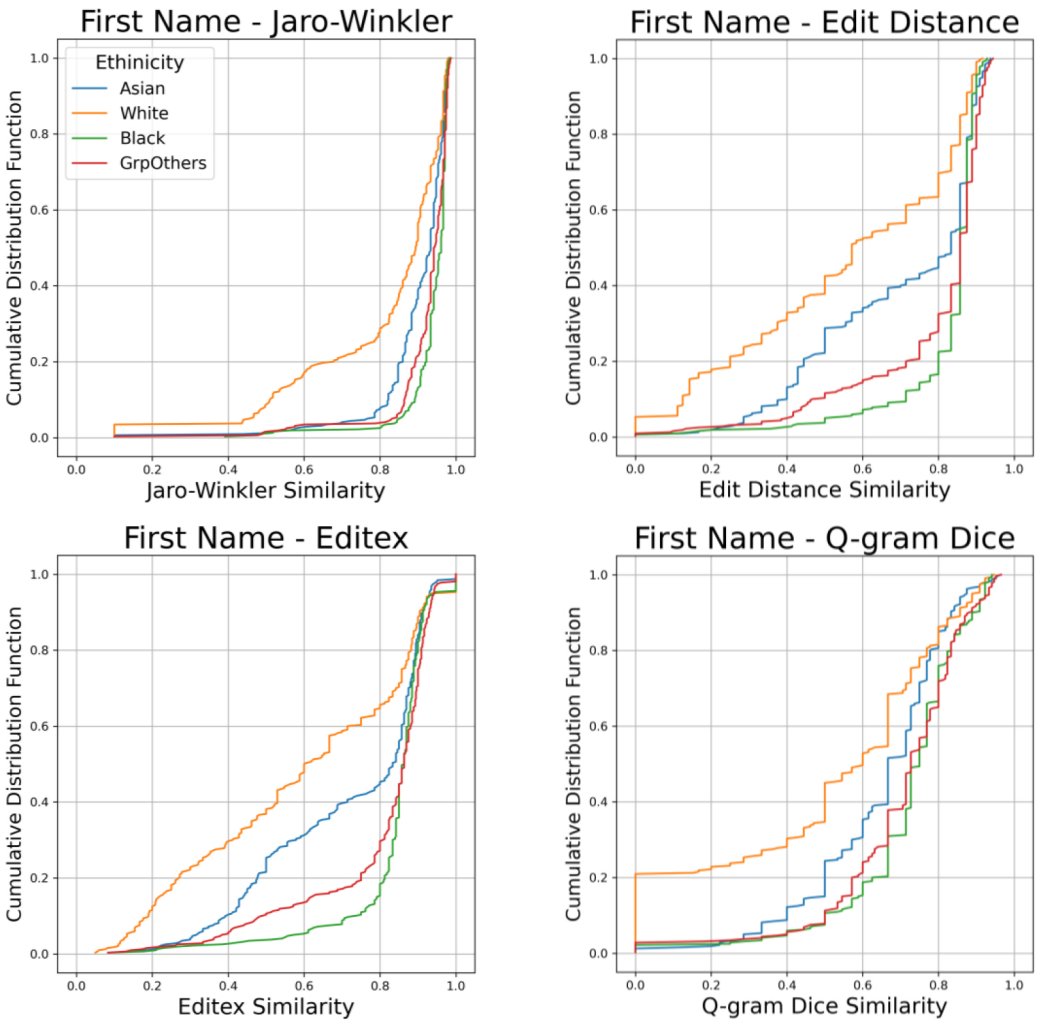
Furthermore, Figures 3 and 4 reveal that the choice of string comparison function substantially affects the overall shape of the CDF curves. Jaro-Winkler consistently produces the highest similarity scores across all ethnic categories for both first and last names, with CDF curves concentrated toward the upper end of the similarity range. Edit distance and the q-gram based Dice coefficient, on the other hand, produce more spread out similarity distributions, while Editex produces distributions that closely mirror those of edit distance, reflecting their calculations being done based on character-level edit operations. This demonstrates that Jaro-Winkler is more sensitive to any setting of a similarity threshold in a record linkage system.

The Jaro-Winkler string comparison function, as originally developed by the US Census Bureau for the comparison of US names [39–41] seems to work best for white names where it results in a more spread distribution of similarities compared to the other ethnic categories. On the other hand, q-gram based Dice similarity can result in very low similarities (or even 0) for short names, thereby affecting especially some types of Asian names.

As a result of these observations, we can conclude that a single fixed similarity threshold used to classify record pairs from all ethnic categories as matches and non-matches would not affect all ethnic categories equally, as the proportion of pairs with a similarity at least that threshold varies by ethnic category and by similarity function. This effect is observable for both first and last names, as summarised in Table 11. As the results in this table show, there are substantial results in the percentages of name pairs being considered as matches, across the different ethnicities, string comparison functions, and for all similarity thresholds. As highlighted in bold font in this table, for the *Black* ethnic category at threshold $t_s = 0.9$, Jaro-Winkler would consider 87.2% of true matches while edit distance only 9.4%. Changing this threshold to $t_s = 0.8$, edit distance would increase to 83.8%. Considering another instance with a threshold of $t_s = 0.7$ and Jaro-Winkler for last names, only 20.9% of *White* name pairs are considered a match compared to 90.6% of names from the *Grouped Others* category. These results clearly indicate that using a single similarity threshold can result in significant linkage bias when the ethnic categories of names are not considered in a record linkage system.

It is further observable that the CDF curves do not extend to a similarity of zero for the Jaro-Winkler comparison function. This is due to the nature of the function as a similarity of 0 is assigned only when the two strings share no characters within their matching windows. This is a rare occurrence for true matching name pairs, as even substantially different names of the same individual can retain at least one common character. However, in contrast, edit distance and the q-gram Dice comparison functions can both result in similarities of 0 more easily for short names or name pairs with non-common characters characters or q-grams.

Figure 1: CDF plots for first name similarities, grouped by string comparison function



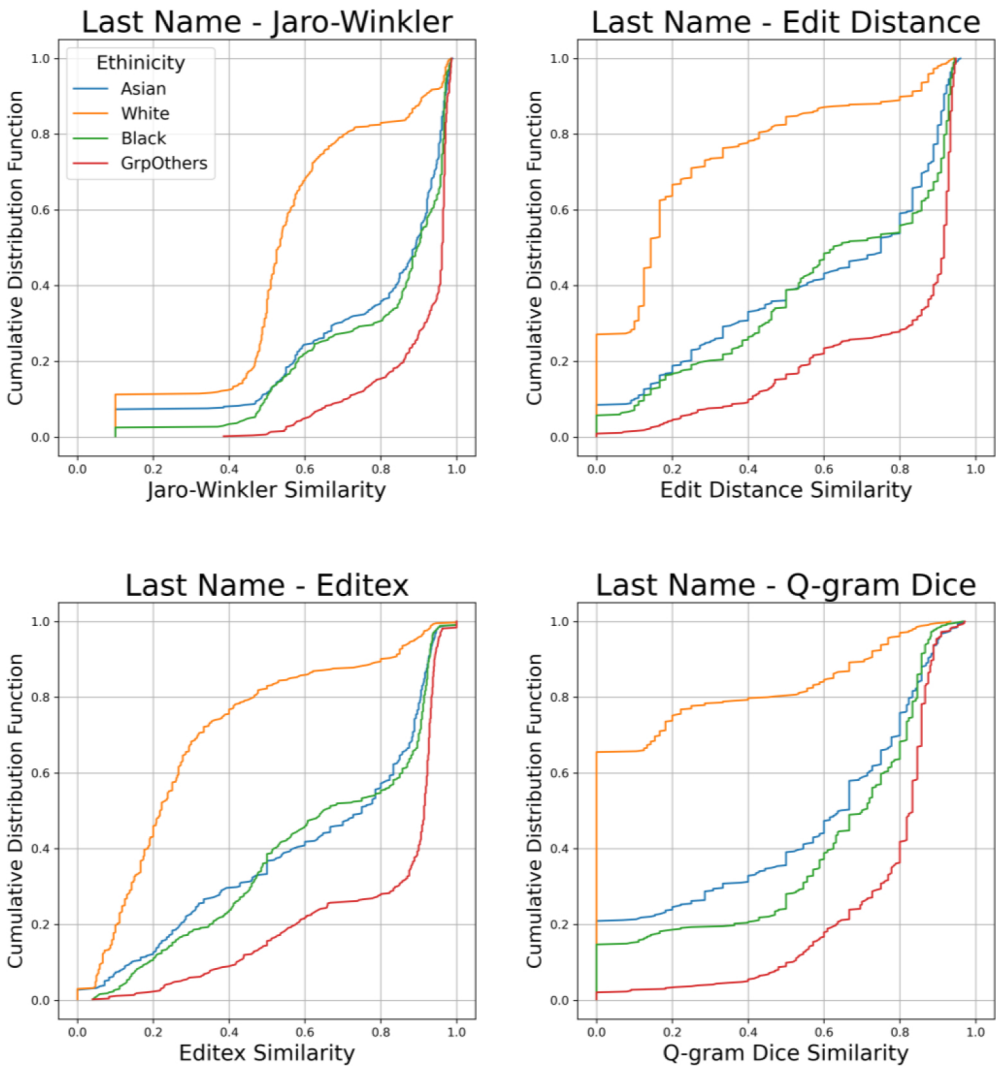
In Figures 5 and 6 we present pairwise scatter plots with degree 2 polynomial trendlines for first and last names, respectively, across all four ethnic categories. Across the scatter plots in these two figures, the general trends of similarities exhibit differences for the four ethnic categories. Some categories have most name pairs with similarities that are clustered tightly around the fitted trendlines, while others show more diversity with many name pairs substantially off the diagonal. This indicates that when name pairs from different ethnic groups are compared with different string similarity functions, they can result in very diverse similarity values. From these results we can conclude that replacing one comparison function with another in a record linkage system could introduce bias across ethnic groups. The only exemption of this would be if one replaced edit distance with Editex, or vice versa, as this would not result in large changes in the calculated similarities between names (when using the NCVR data set).

In contrast, when comparing Jaro-Winkler similarities to the similarities of any of the other three string comparison functions, we can see much larger divergences in their trendlines, particularly for first names. The trendlines in Figure 5 also show that the curves for different ethnicities diverge substantially.

While these plots reveal that the relationships between string comparison functions are not uniform across the similarity score range, they also reveal that the direction of bias between comparison functions can shift at certain threshold points. This means that a given comparison function which disadvantages one ethnic group in the low similarity range may actually disadvantage a different ethnic group in the high similarity range. This shift in bias is a key finding, as it demonstrates that ethnic bias in string similarities is not a fixed property of one comparison function but it can change depending on the similarity threshold being used.

The CDF plots in Figures 1 to 4 also support the finding that the direction of bias between ethnic groups can shift at certain similarity threshold values. For example, in the q-gram Dice plot in Figure 1 the CDF curves for the *Asian* and *White* ethnic categories cross at around a similarity of 0.8. This means that below this threshold a lower proportion of *White* name pairs would be considered as matches compared to *Asian* name pairs, while above this threshold this pattern is reversed. This confirms that ethnic bias in record linkage is not a fixed property of a given string comparison function.

Figure 2: CDF Plots for last name similarities, grouped by string comparison function



Discussion, Limitations and Recommendations

As our work has shown, different string comparison functions clearly can result in different similarity distributions when names from diverse ethnic categories are being compared. This can potentially lead to ethnic bias when population databases are being linked.

As an example, if the recall (sensitivity) for the four ethnic categories and their distributions are 0.95 for *White* name pairs (with 70% of pairs in this category), 0.9 for *Black* name pairs (with 10% of pairs in this category), 0.8 for *Asian* name pairs (with 10% of pairs in this category), and 0.75 for *Grouped Others* name pairs (with 10% of pairs in this category), then the overall recall for such a linkage would still be 0.91. This might be an acceptable overall linkage result, however it would substantially under-represent linked records from the smaller ethnic categories, especially of individuals in the *Grouped Others* category. This example highlights the need for a careful investigation of potential bias due to the way names are compared in record linkage systems.

Our work has several limitations, which we discuss first. Then we provide a set of recommendations for the practical linkage of (large) population databases that contain individuals from diverse ethnic backgrounds, which should lead to reduced ethnic bias in linkage results.

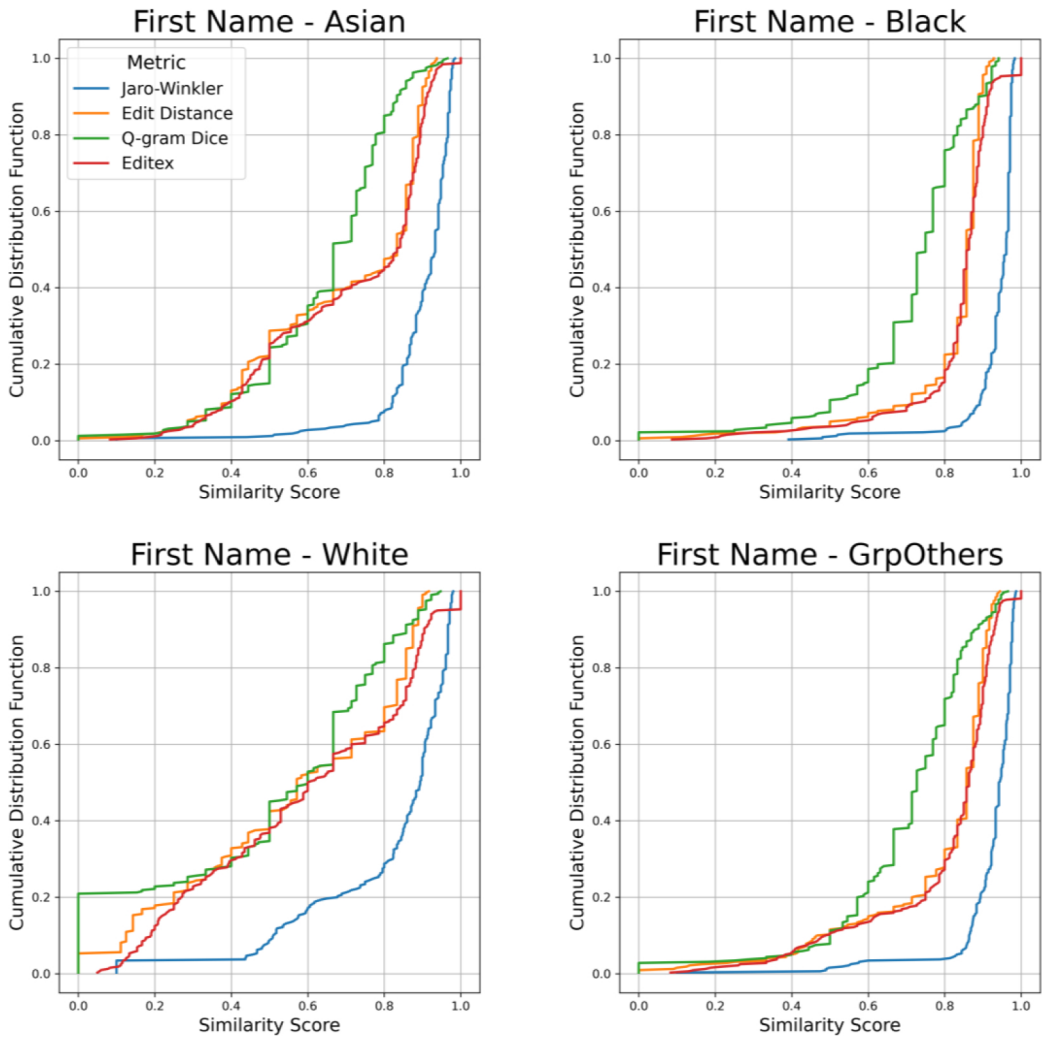
Limitations

A key limitation of this work is that we only used one data set (even though large and covering a diverse population over more than a decade) with self-reported ethnic categories (by voters via their voter registration forms) [42]. With many (administrative) population databases not containing such an ethnic attribute, the ethnic group of names might have to be inferred by using either an ethnicity classifier [28, 30, 31] or from attributes such as country of birth.

The self-reporting aspect of the ethnicity attribute by voters might furthermore have affected how names were selected into ethnic groups, as we discussed in the Methodology section. As previous research has shown [42], individuals are known to change the ethnic group they belong to over time.

From the very large number of records available over all snapshots in the NCVR data set, only a small fraction resulted in pairs where names were different. This is due to this data

Figure 3: CDF plots for first name similarities, grouped by ethnic category



set being of high quality, and name corrections and changes occurring on a rare basis. If this data set would be linked, for example over time and across two or more snapshots, for most voters exact matching record pairs would be identified. However, our work was focusing on actual name changes by the same individuals, and even with the small numbers of pairs used in our analysis we could see substantial differences as similarity distributions across ethnic categories.

Another limitation of this study arises from inconsistencies within the NCVR data set itself, where we used the NCID attribute to uniquely identify individuals. However, previous analysis of this data set [43] has identified situations in which records belonging to different individuals share the same NCID value. While the extent of such inconsistencies are not quantified, the number of records with incorrect NCID assignments is presumed to be very small (given the nature of this data set), and therefore unlikely to substantially alter the results from the analysis we conducted in this work.

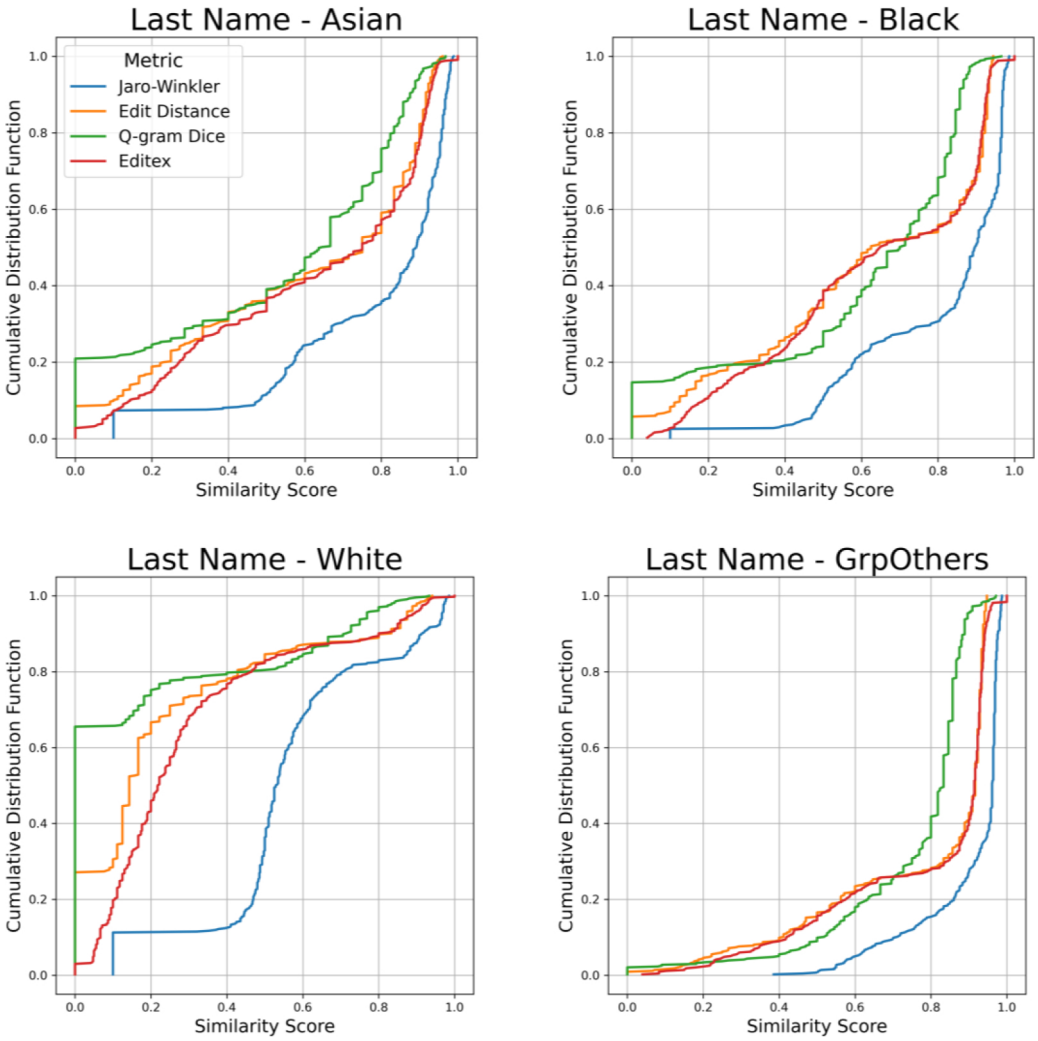
Another limitation of this work relates to the grouping of ethnic categories. The four categories we use for this study are derived from the *Race code* attribute available in the NCVR data set. *Asian, Black, White, and Grouped Others* (Combining *Mixed, Alaska Native or American Indian,* and *Other*) are broad groupings that may not align closely with the actual diversity of naming conventions found within each category. For example, the Asian category likely contains names with very different structural characteristics, such as Chinese, Japanese, Vietnamese, Thai, Indian, and Sri Lankan naming systems, among many others. Names within such ethnic subgroups can differ substantially due to name length, transliteration conventions, and so on. As a result, the similarity distributions for each of these ethnic categories may mask finer variations within the subgroups.

Recommendations for the Practical Linkage of Population Databases

Based on our work, we make the following general recommendations for record linkage projects that aim to link population databases that likely contain names from diverse ethnic categories.

1. A first step is to investigate if a database does indeed contain records with individuals from diverse ethnic groups. If no suitable attribute is available, then applying a name ethnicity classifier [28, 30, 31] (acknowledging these will not always be correct) or using an attribute such as country of birth [44] (if available), might be an option. Given many countries increasingly have multi-cultural populations, it is likely that names from diverse ethnic groups will be identified in any large database.

Figure 4: CDF Plots for last name similarities, grouped by ethnic category



2. None of the investigated string comparison functions can be seen as 'the best' when applied to names from different ethnic groups. However, the Jaro-Winkler string comparison function (developed for name comparison in the US census [39–41]) leads to generally very high similarities for many non-white name pairs. As a result, using this comparison function might result in linkage bias of individuals with non-white names as more records of such individuals might be linked due to their higher similarities. On the other hand, string comparison functions that are based on character q-grams seem to be less suitable for shorter names, such as those common in some Asian countries, because even a single character difference can lead to a very low or even zero similarity.
3. Exploring how changes in last names occur (such as after marriage) across different ethnic groups is important as this can also lead to linkage bias. In the NCVR data set we used in our experimental evaluation, individuals with names from the *White* category seem to have a larger proportion of name changes that lead to much lower similarities, as compared to individuals from other ethnic categories.
4. If ground truth data in the form of known true matches and true non-matches, including their ethnic categories, are available, then it might be possible to adjust and calibrate similarities across these ethnic categories to reduce linkage bias [45].
5. To assess linkage quality in a fair way, a stratified sampling of record pairs for manual assessment (such as via clerical review) across any sub-populations is required. This will provide information on how linkage quality might differ within different sub-populations. In addition, gender and age (or correspondingly year of birth) should be investigated for potential bias. Gender bias can occur due to different patterns in last name changes when people get married, while age bias can be due to different mobility patterns of people at different stages of their lives (for example, students compared to settled families with children).

We hope these recommendations are starting points that allow record linkage practitioners to consider and hopefully improve their linkage projects, with the ultimate aim of reducing linkage bias due to the way how strings such as names are being compared.

Conclusions

This work provides an observational study of how commonly used string comparison functions result in diverse similarity

Table 11: Percentage of true matching name pairs with a similarity score of at least the similarity threshold, t_s, for first names and last names, grouped by ethnic category (EC) and string comparison function

t_s	EC	First name				Last name			
		JW	ED	QG	EX	JW	ED	QG	EX
0.9	Asian	61.2	12.5	3.4	17.2	47.1	22.8	6.0	23.4
	Black	**87.2**	**9.4**	10.0	20.9	49.4	32.6	1.8	31.7
	White	43.4	4.4	5.0	13.1	12.6	2.8	0.5	4.4
	Grp Oth	78.4	24.1	8.4	29.7	72.9	59.3	4.6	60.7
0.8	Asian	92.2	55.6	19.7	55.6	64.4	46.4	30.3	44.1
	Black	97.5	**83.8**	33.8	85.0	69.4	46.2	36.6	45.5
	White	71.9	36.9	18.8	35.6	17.0	11.3	4.1	10.8
	Grp Oth	96.3	72.5	35.3	73.1	84.6	72.4	63.9	72.4
0.7	Asian	96.3	60.6	48.4	60.6	69.7	53.3	41.4	54.0
	Black	98.1	90.9	69.1	92.5	72.6	48.3	50.6	48.1
	White	79.1	43.8	31.6	41.9	**20.9**	12.2	10.8	12.4
	Grp Oth	96.6	81.9	62.2	83.1	**90.6**	74.3	74.9	74.3
0.6	Asian	97.2	67.2	69.7	69.1	75.6	58.4	56.1	59.3
	Black	98.1	94.1	85.0	95.0	77.9	53.3	63.0	54.5
	White	83.4	48.1	50.6	52.5	31.7	13.1	16.1	14.3
	Grp Oth	96.6	85.9	79.1	86.6	95.2	78.2	83.5	78.6

Results are shown for the four ECs Asian, Black, White, and Grouped Others. These percentage values correspond to the recall of true matches at different thresholds. We highlight in bold font examples with large differences.

distributions when applied to names from different ethnic groups. The study highlights that comparison functions can produce substantially different similarity ranges and distributions depending on the characteristics of names.

The findings demonstrate that the choice of the string comparison function can influence linkage outcomes in ways that disproportionately affect certain ethnic groups. These results show the importance of selecting and evaluating string comparison functions carefully when linking large, heterogeneous population databases, and they provide practical guidance for designing fairer and less biased record linkage systems.

A direction to extend this work is to explore ethnicity-aware name similarity methods that automatically learn similarities and that can be adapted to capture ethnicity-specific naming conventions. Foxcroft et al. [34] proposed a Doc2Vec-based approach to name matching, where each name is treated as a "document" and each constituent letter as a "word". This allows embeddings to be learned directly from the internal character structure of names. Moreover, Garg et al. [46] showed that embeddings trained on large data sets can capture ethnic associations, because common surnames from different ethnic groups are located closer to different words in the embedding space. Together, these works suggest an extension for future work in name similarity research that would allow name similarity functions to be adapted to specific ethnic populations, which could support more approaches to combining name matching with ethnicity detection.

Acknowledgements

This work was undertaken with support from the ANU University Research Scholarship awarded to Yashithi Dharmawimala. Sumayya Ziyad gratefully acknowledges the Australian Government Research Training Program. Joseph Lam was funded by the Wellcome Trust [212953/Z/18/Z]. Peter Christen gratefully acknowledges his funding by ADR Scotland (through grant numbers ES/W010321/1 and UKRI3324), which receives funding from the UK Economic and Social Research Council (ESRC) through their investment of £168 million awarded to ADR UK until March 2031.

Statement on Conflicts of Interest

The authors have no conflicts of interest.

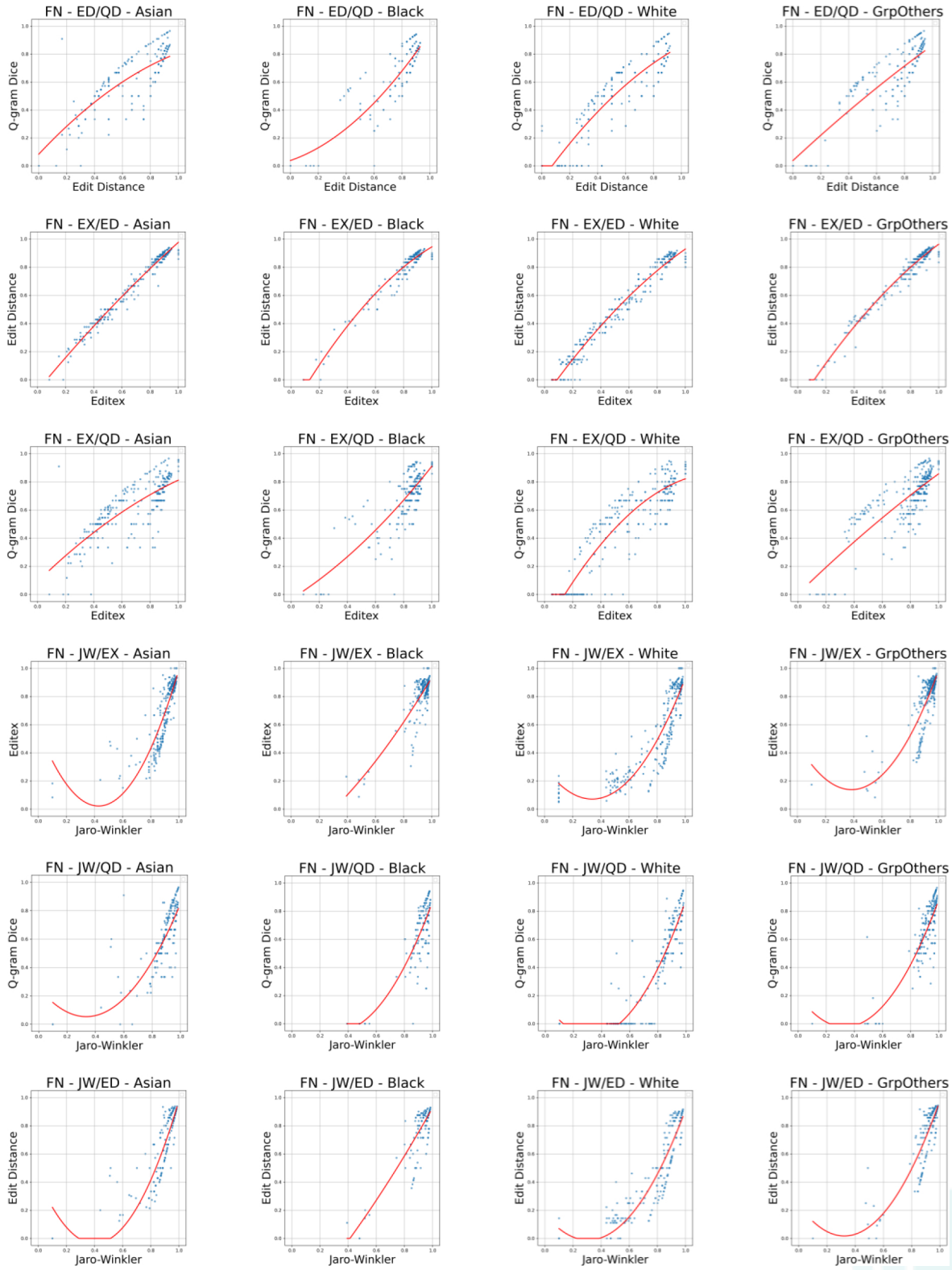
Ethics Approval and Data Availability

The use of the North Carolina Voter Registry (NCVR) database for this project has been approved by the Australian National University Human Research Ethics Committee (Protocol H/2026/0354).

AI Statement

The authors declare that no generative AI tools were used, neither for this research nor for the preparation of this manuscript.

Figure 5: Scatter plots with polynomial trendlines of degree 2 across string comparison functions and ethnic categories for first names

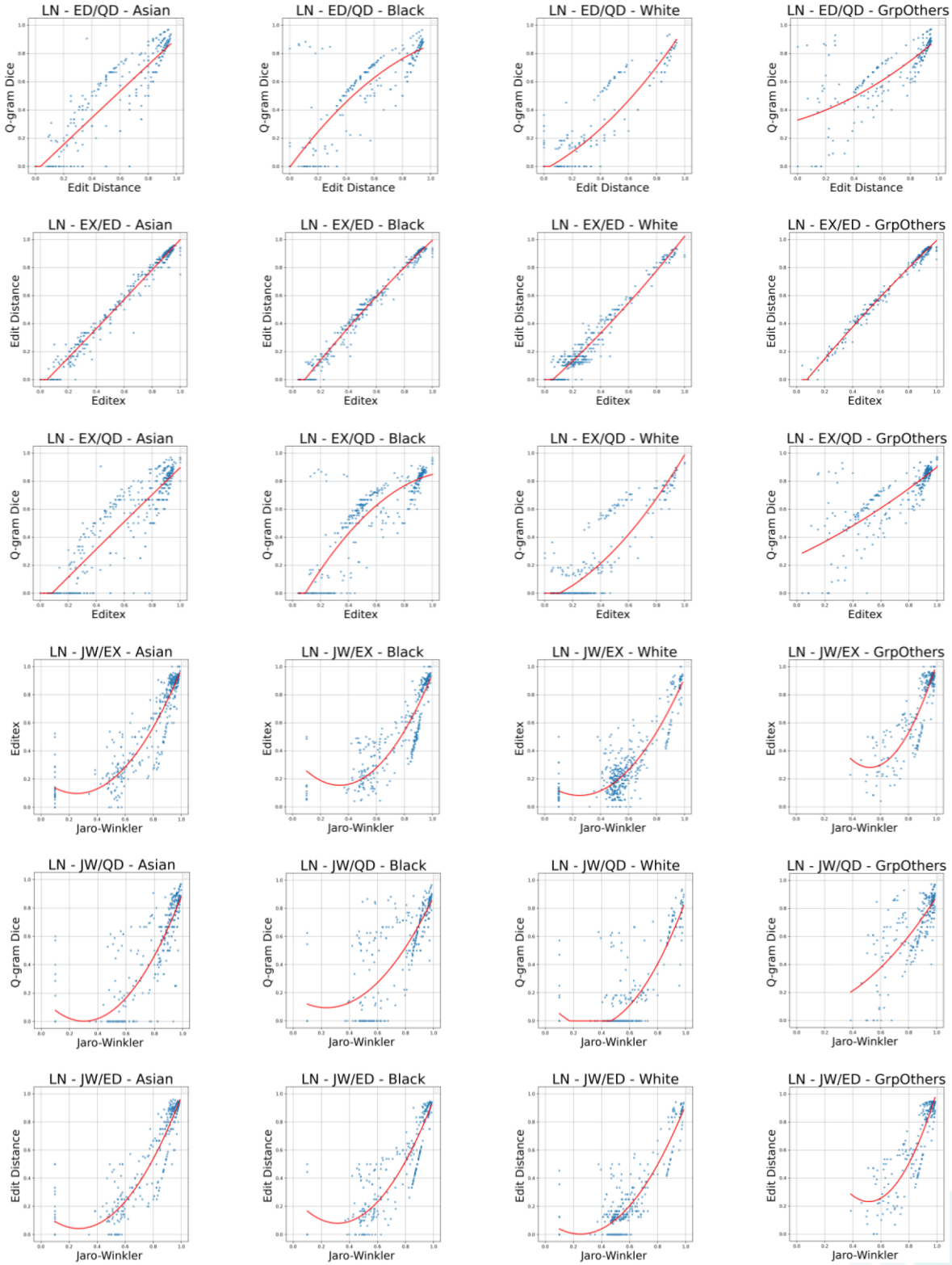


References

1. P. Christen, R. Schnell, T. Ranbaduge, Linking sensitive data – Methods and techniques for practical privacy-preserving information sharing, Springer, 2020. https://doi.org/10.1007/978-3-030-59706-1

2. P. Christen, R. Schnell, Thirty-three myths and misconceptions about population data: from data capture and processing to linkage, International Journal of Population Data Science 8 (1) (2023). https://doi.org/10.23889/ijpds.v8i1.2115

Figure 6: Scatter plots with polynomial trendlines of degree 2 across string comparison functions and ethnic categories for last names



3. P. Christen, Data matching – Concepts and techniques for record linkage, entity resolution, and duplicate detection, Data-Centric Systems and Applications, Springer, 2012. https://doi.org/10.1007/978-3-642-31164-2

4. W. W. Cohen, P. Ravikumar, S. Fienberg, A comparison of string distance metrics for name-matching tasks, in: Workshop on Information Integration on the Web, held at IJCAI, Acapulco, 2003, pp. 73–78.

5. P. Jokinen, J. Tarhio, E. Ukkonen, A comparison of approximate string matching algorithms, Software – Practice and Experience 26 (12) (1996)

1439–1458. https://doi.org/10.1002/(SICI)1097-024X(199612)26:12<1439::AID-SPE71>3.0.CO;2-1

6. J. S. Kaufman, Statistical aspects of racial and ethnic health disparities, Annual Review of Statistics and Its Application 13 (2025). https://doi.org/10.1146/annurev-statistics-042324-061403

7. M. DeBell, J. A. Krosnick, K. Gera, D. S. Yeager, M. P. McDonald, The turnout gap in surveys: Explanations and solutions, Sociological Methods and Research (2018). https://doi.org/10.1177/0049124118769085

8. M. Bohensky, Bias in data linkage studies, in: K. Harron, H. Goldstein, C. Dibben (Eds.), Methodological developments in data linkage, Wiley Online Library, 2015, pp. 63–82. https://doi.org/10.1002/9781119072454

9. L. Mc Grath-Lone, N. Libuy, D. Etoori, R. Blackburn, R. Gilbert, K. Harron, Ethnic bias in data linkage, The Lancet Digital Health 3 (6) (2021) e339.30 https://doi.org/10.1016/S2589-7500(21)00081-9

10. J. Lam, M. Cortina-Borja, R. Aldridge, R. Blackburn, K. Harron, Data note: Alternative name encodings-using Jyutping or Pinyin as tonal representations of Chinese names for data linkage, International Journal of Population Data Science 8 (5) (2025) 2935. https://doi.org/10.23889/ijpds.v8i5.2935

11. H. E. Knight, S. R. Deeny, K. Dreyer, J. Engmann, M. Mackintosh, S. Raza, M. Stafford, R. Tesfaye, A. Steventon, Challenging racism in the use of health data, The Lancet Digital Health 3 (3) (2021) e144–e146. https://doi.org/10.1016/S2589-7500(21)00019-4

12. D. J. Hand, P. Christen, S. Ziyad, Selecting a classification performance measure: matching the measure to the problem, arXiv preprint arXiv:2409.12391 (2024). https://doi.org/10.48550/arXiv.2409.12391

13. H. Dunn, Record linkage, American Journal of Public Health 36 (12) (1946) 1412.

14. H. Newcombe, J. Kennedy, S. Axford, A. James, Automatic linkage of vital records, Science 130 (3381) (1959) 954–959. https://doi.org/10.1126/science.130.3381.954

15. I. P. Fellegi, A. B. Sunter, A theory for record linkage, Journal of the American Statistical Association 64 (328) (1969) 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

16. D. De Brou, M. Olsen, The Guth algorithm and the nominal record linkage of multi-ethnic populations, Historical Methods: A Journal of Quantitative and Interdisciplinary History (1986). https://doi.org/10.1080/01615440.1986.10594163

17. R. Russell, The Soundex coding system, US patent 1261167 (1918).

18. G. Bouchard, P. Brard, Y. Lavoie, FONEM: Un code de transcription phonétique pour la reconstitution automatique des familles Saguenayennes, Population (french edition) (1981) 1085–1103. https://doi.org/10.2307/1532326

19. R. L. Taft, Name Search Techniques, Tech. Rep. NCJ 1743, Bureau of Systems Development (1970).

20. M. A. Bohensky, D. Jolley, V. Sundararajan, S. Evans, D. V. Pilcher, I. Scott, C. A. Brand, Data linkage: a powerful research tool with potential problems, BMC Health Services Research 10 (346) (2010) 1–7. https://doi.org/10.1186/1472-6963-10-346

21. S. L. DuVall, A. M. Fraser, R. A. Kerber, G. P. Mineau, A. Thomas, The impact of a growing minority population on identification of duplicate 31 records in an enterprise data warehouse, in: MEDINFO 2010, IOS Press, 2010, pp. 1122–1126. https://doi.org/10.3233/978-1-60750-588-4-1122

22. S. Randall, A. Brown, J. Boyd, R. Schnell, C. Borgs, A. Ferrante, Sociodemographic differences in linkage error: an examination of four large-scale datasets, BMC Health Services Research 18 (1) (2018) 678. https://doi.org/10.1186/s12913-018-3495-x

23. C. Barboi, F. Ouyang, L. Lembcke, A. Martin, A. Griffith, K. S. Allen, X. Li, H. Xu, S. J. Grannis, Evaluation of the accuracy of probabilistic record linkage across sociodemographic categories in 4 databases: Exploratory study, JMIR Formative Research 10 (1) (2026) e78622. https://doi.org/10.2196/78622

24. T. C. Ong, M. V. Mannino, L. M. Schilling, M. G. Kahn, Improving record linkage performance in the presence of missing linkage data, Journal of Biomedical Informatics 52 (2014) 43–54. https://doi.org/10.1016/j.jbi.2014.01.016

25. P. Christen, D. J. Hand, N. Kirielle, A review of the F-measure: Its history, properties, criticism, and alternatives, Computing Surveys 56 (3) (2023) 1–24. https://doi.org/10.1145/3606367

26. J. Lam, M. Cortina-Borja, R. Aldridge, R. Blackburn, K. Harron, Cluster-based name embeddings reduce ethnic disparities in record linkage quality under realistic name corruption: evidence from the North Carolina Voter Registry, arXiv preprint arXiv:2601.07693 (2026). https://doi.org/10.48550/arXiv.2601.07693

27. P. Christen, Preparation of a real temporal voter data set for record linkage and duplicate detection research, Australian National University (2014).

28. P. Mateos, A review of name-based ethnicity classification methods and their potential in population studies, Population, Space and Place 13 (4) (2007) 243–263. https://doi.org/10.1002/psp.457

29. R. Schnell, M. Trappmann, T. Gramlich, A study of assimilation bias in name-based sampling of migrants,

Journal of Official Statistics 30 (2) (2014) 231–249. https://doi.org/10.2478/jos-2014-0015

30. J. Ye, S. Han, Y. Hu, B. Coskun, M. Liu, H. Qin, S. Skiena, Nationality classification using name embeddings, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 1897–1906. 32 https://doi.org/10.1145/3132847.3133008

31. L. Hafner, T. P. Peifer, F. S. Hafner, Equal accuracy for Andrew and Abubakar—Detecting and mitigating bias in name-ethnicity classification algorithms, AI & Society 39 (4) (2024) 1605–1629. https://doi.org/10.1007/s00146-022-01619-4

32. C. Crabtree, J. Y. Kim, S. M. Gaddis, J. B. Holbein, C. Guage, W. W. Marx, Validated names for experimental studies on race and ethnicity, Scientific Data 10 (1) (2023) 130. https://doi.org/10.1038/s41597-023-01947-0

33. P. Mateos, R. Webber, P. Longley, The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names, University College London, Centre for Advanced Spatial Analysis (CASA Working Paper 116) (2007).

34. J. Foxcroft, A. d'Alessandro, L. Antonie, Name2vec: Personal names embeddings, in: Canadian Conference on Artificial Intelligence, Springer, 2019, pp. 505–510. https://doi.org/10.1007/978-3-030-18305-9_52

35. P. Konda, S. Das, P. S. G. C, A. Doan, A. Ardalan, J. R. Ballard, H. Li, F. Panahi, H. Zhang, J. Naughton, et al., Magellan: toward building entity matching management systems over data science stacks, Proceedings of the VLDB Endowment 9 (13) (2016) 1581–1584. https://doi.org/10.14778/2994509.2994535

36. F. J. Damerau, A technique for computer detection and correction of spelling errors, Communications of the ACM 7 (3) (1964) 171–176. https://doi.org/10.1145/363958.363994

37. V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet Physics Doklady 10 (8) (1966) 707–710.

38. J. Zobel, P. Dart, Phonetic string matching: Lessons from information retrieval, in: ACM Conference on Research and Development in Information Retrieval, Zürich, 1996, pp. 166–172. https://doi.org/10.1145/243199.243258

39. M. A. Jaro, Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida, Journal of the American Statistical Association 84 (1989) 414–420. https://doi.org/10.1080/01621459.1989.10478785

40. E. H. Porter, W. E. Winkler, Approximate string comparison and its effect on an advanced record linkage system, Tech. Rep. RR97/02, US Bureau of the Census (1997). 33

41. W. Winkler, String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage, in: Proceedings of the Section on Survey Research Methods, American Statistical Association, 1990, pp. 354–359.

42. S. Ziyad, J. Lam, P. Christen, R. Schnell, C. Nanayakkara, Self-reported ethnicity changes over time: A population level study using public voter databases, International Journal of Population Data Science 9 (5) (2024). https://doi.org/10.23889/ijpds.v9i5.2591

43. F. Panse, A. Düjon, W. Wingerath, B. Wollmer, Generating realistic test datasets for duplicate detection at scale using historical voter data, in: International Conference on Extending Database Technology, Open Proceedings, Nicosia, 2021, pp. 570–581. https://doi.org/10.5441/002/edbt.2021.67

44. F. Stanaway, L. Zhu, B. McDonald, J. Ravulo, M. Dickson, N. Nassar, M. L. Yap, L. Jorm, S. Aitken, L. Kritharides, et al., Using linked census ancestry data to examine all-cause mortality by ethnicity in Australia, International Journal of Population Data Science 10 (1) (2025) 2476. https://doi.org/10.23889/ijpds.v10i1.2476

45. M. H. Moslemi, M. Milani, Threshold-independent fair matching through score calibration, in: Conference on Governance, Understanding and Integration of Data for Effective and Responsible AI, Santiago, Chile, 2024, pp. 40–44. https://doi.org/10.1145/3665601.3669845

46. N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, Proceedings of the National Academy of Sciences 115 (16) (2018) E3635–E3644. https://doi.org/10.1073/pnas.1720347115

47. B. S. Everitt, A. Skrondal, et al., The Cambridge dictionary of statistics, Vol. 140, Cambridge University Press, Cambridge, 2002.

Abbreviations

CDF:	Cumulative Distribution Function
cv:	Coefficient of Variation
ED:	Edit Distance
efs:	Ethnic-Specific Frequency Score
EX:	Editex
FN:	First Name
Grp Oth:	Grouped Others (ethnic group)
JW:	Jaro-Winkler
LN:	Last Name
NCID:	North Carolina Identification
NCVR:	North Carolina Voter Registration
NYSIIS:	New York State Identification and Intelligence System
OCR:	Optical Character Recognition
QD:	Q-gram Dice
QID:	Quasi-Identifier
TF:	Term-Frequency

Appendix A. Identifying Frequent Ethnic Group Specific Names

In this appendix we describe how we calculated the scores to identify those names which predominantly occur with a single ethnic category, as discussed in the Experimental Evaluation section and shown in Tables 3 and 4. We also discuss the uniformity score calculation we used to rank the names shown in Table 5.

We used the matrix $\boldsymbol{E}$, as we defined in the Methodology section, where for each name in the data set we have the number of its occurrences associated with each ethnic category, which indicate how strongly a name is associated with a specific ethnic category. By using this information, we can identify the most frequent names across ethnicities. However, rating names based on how exclusively a name belongs to a specific ethnic category alone does not capture how frequent a name is within that category. For example, the first name *Nam Woong* might only occur with the *Asian* ethnic category, however, it might only occur in two records. As a result, *Nam Woong* might be ranked at the top of the *Asian* ethnic category (with many other uniquely Asian names) even though it is a rare name (in the NCVR data set which contains several million records).

To address this issue, we define a ranking score which jointly considers both the relative frequency of a name within its predominant ethnic category, and its exclusivity to that category. For each name $n_x \in \boldsymbol{n}$ and its predominant ethnic category, $e_d \in \boldsymbol{e}$, we calculate its ethnic-specific frequency score, *efs*, as follows:

$$efs(n_x, e_d) = \left(\frac{\boldsymbol{E}[x,d]}{\sum_x \boldsymbol{E}[x,d]}\right) \times \left(\frac{\boldsymbol{E}[x,d]}{\sum_y \boldsymbol{E}[x,y]}\right),$$

where $\boldsymbol{E}[x,d]$ is the count of name n_y within its predominant ethnic category e_d, $\sum_x \boldsymbol{E}[x,d]$ is the total count of the occurrence all names within the category e_d, and $\sum_y \boldsymbol{E}[x,y]$ is the total count of occurrence of name n_x across all ethnic categories. The first term in this equation provides the relative frequency of a specific name within its predominant ethnic category, while the second term captures how strongly the name is associated with that category.

The results obtained by calculating this score for all first names and last names are presented in Tables 3 and 4, respectively, together with their *efs* scores.

To identify names that occur (close to) uniformly across ethnic categories, we calculate a uniformity score, u for each name $n_x \in \boldsymbol{n}$ based on the name frequencies across the four ethnic categories. For each name, we first calculate the mean $f'(n_x)$ and standard deviation $\sigma(n_x)$ of its frequency counts across all ethnic categories. The uniformity score of a name is then calculated as the ratio of its mean frequency divided by its standard deviation:

$$u(n_x) = \frac{f'(n_x)}{\sigma(n_x)}.$$

This uniformity score is equivalent to the inverse of the coefficient of variation (cv) [47], $u = \frac{1}{cv}$. A higher value of u indicates lower relative variability in a name's frequency across the ethnic groups. As such, names with higher uniformity scores have a more balanced distribution across the ethnic categories than names with lower uniformity scores. In Table 5 we show the top ten first and last names ranked by their uniformity score to provide examples of names that occur consistently across ethnic groups.