# Names-based ethnicity enhancement of hospital admissions in England, 1999-2013

Jakob Petersen[1*], Jens Kandt[2], Paul A. Longley[1]

1 Consumer Data Research Centre (CDRC), Department of Geography, University College London (UCL), Gower Street, London, WC1E 6BT 2 The Bartlett Centre for Advanced Spatial Analysis (CASA), Gower Street, UCL, London, WC1E 6BT

*) Corresponding author: j.petersen@ucl.ac.uk (JP)

**Word count**

2,622

**Highlights**

- Ethnicity data were missing for more than half of patients admitted to English hospitals in 1990s.
- Name-based ethnicity classifications have merit for the predictions of many ethnic minorities.
- Prediction success of a names-based ethnicity classification tool has been quantified.

**Abstract**

*Background*

Accurate recording of ethnicity in electronic healthcare records is important for the monitoring of health inequalities. Yet until the late 1990s, ethnicity information was absent from more than half of records of patients who received inpatient care in England. In this study, we report on the usefulness of names-based ethnicity classification, Ethnicity Estimator (EE), for addressing this gap in the hospital records.

*Materials and methods*

Data on impatient hospital admission were obtained from Hospital Episode Statistics (HES) between April 1999 and March 2014. The data were enhanced with ethnicity coding of participants' surnames using the EE software. Only data on the first episode for each patient each year were included.

*Results*

A total of 111,231,653 patient-years were recorded between April 1999 and March 2014. The completeness of ethnicity records improved from 59.5% in 1999 to 90.5% in 2013. Biggest improvement was seen in the White British group, which increased from 55.4% in 1999 to 73.9% in 2013. The correct prediction of NHS-reported ethnicity varied by ethnic group (2013/14 figures): White British (89.8%), Pakistani (81.7%), Indian (74.6%), Chinese (72.9%), Bangladeshi (63.4%), Black African (57.3%), White Other (50.5%), White Irish (45.0%). For other ethnic groups the prediction success was low to none. Prediction success was above 70% in most areas outside London but fell below 40% in parts of London.

*Conclusion*

Studies of ethnic inequalities in hospital inpatient care in England are limited by incomplete data on patient ethnicity collected in the 1990s and 2000s. The prediction success of a names-based ethnicity classification tool has been quantified in HES for the first time and the results can be used to inform decisions around the optimal analysis of ethnic groups using this data source.

**Keywords:** Electronic health records; Health services research; Public health

## 1. **Introduction**

Ethnicity is defined as a sensitive personal characteristic under European Union (2016) General Data Protection Regulation (GDPR) [1]. It is often considered to be inherently subjective [2] and may not always be collected for reasons of statute [3,4]. This can handicap the conduct of equality audits, analysis of corporate governance [5] and, most recently, monitoring of hospital admissions and outcomes during the COVID-19 pandemic [6,7].

Provision has been made for Hospital Episode Statistics (HES) to include patient-reported ethnicity since 1995 by drawing on a central National Health Service (NHS) patient register [5]. Yet until the late 1990s, ethnicity information was absent from more than half of records of patients who received inpatient care. General practitioners were financially incentivised to record patient ethnicity through the Quality Outcomes Framework (QOF) between 2006-2012 with a resultant increase in completeness of inpatient ethnicity data to more than 80% during this time [5].

The problem of missing ethnicity data in NHS datasets has previously been studied [8,9]; although not in the full range of ethnic groups in a national study over several years. Ryan et al. (2012) who used Onomap and Nam Pehchan to impute the ethnicity of White, South Asian, Black and Other groups in the UK's West Midlands [8]. Ryan et al. 2012 used a multiple imputation strategy with characteristics of the individual patients, their care, and the ethnic composition of their neighbourhoods: they reported that the sensitivity of the multiple imputation was above 90% for White and South Asian ethnicities but was very low for other groups. Smith et al. 2017 used the Onomap software to assign children and young people with cancers to either White, South Asian, or Other groups in a Yorkshire study, concluding that combining different data sources including names-based ones increased the representation of ethnic minorities, albeit with some ambiguity [9]. Both studies concluded that there is no perfect substitute for more complete self-reported ethnicity data.

Personal names are commonly used to impute ethnicity information when self-reported ethnicity data are not collected systematically or available through linkage [10,11]. An early example of names-based ethnicity classification exploiting large scale data sets is the work by Mateos et al. (2011) [11]. The applied cluster analysis to data on personal names and residential codes from telephone directories and other administrative data from 17 different countries. Kandt & Longley (2018) used cluster analysis to define more detailed clusters for the UK making use of data on names and country of origin in the Census 2011 microdata [10]. In this paper we report on the use of names-based ethnicity classifications to address incomplete ethnicity information in inpatient hospital records. It is a national study covering the whole of England over fifteen years (1999/00-2013/14). The study quantifies the prediction success of the complete range of ethnic groups – nationally and regionally – against self-reported, NHS-recorded ethnicity. A freely available software, Ethnicity Estimator (EE), was used [10]. EE was developed by the Consumer Data Research Centre (CDRC: cdrc.ac.uk) in partnership with the Office for National Statistics (ONS) and using enhanced algorithmic procedures [10,12]. The results of this study can be used to inform decisions around analysis of ethnicity in HES.

## 2. Materials and methods

Hospital inpatient admission records were obtained from NHS England HES for the period April 1999-March 2014 (financial years referred by the first year only from here onwards). The ethnicity information was coded on patient forename and surname separately using an enhanced version of the Ethnicity Estimator (EE) software [10]. Where a patient changed surname, e.g. due to marriage, the ethnicity category of the earliest name was used. To retain full anonymity, the coding was carried out in an air-gapped, secure data facility by NHS Digital linking name information in the Patient

Demographic Service to HES. The main 12 EE categories map onto the Census 2011 ethnicity categories except for mixed ethnicity, which is not predicted by EE. All NHS-recorded mixed ethnicities were combined into a single mixed category. Not-Stated and Missing were combined, and Black Other was combined with Other. It should be noted that the NHS used a simpler coding frame in 1999-2001, which did not include categories for mixed, White Irish or Asian Other.

Using the EE software, we developed three different ways of estimating ethnicity; each of which we compared to the benchmark of self-reported ethnicity as recorded by the NHS.

1. NHS-recorded ethnicity with additional ethnicity estimation based on patient surname where data were missing (in the following: supplementary estimation)
2. Ethnicity estimation based on patient surname alone (surname-based estimation)
3. Ethnicity estimation based on patient forename and surname; selecting only those records where the estimated ethnic groups were identical for forename and surname (full name-based estimation), e.g. a record would only be classified as Pakistani where both forename and surname were estimated as Pakistani by the EE software, etc.

The completeness of ethnicity records was defined and measured as the proportion (%) of non-missing ethnicity records out of all records. Annual estimates of the prediction success or sensitivity (proportion of true positives among the sum of true positives and false negatives) were calculated as the percentage of correctly predicted records based on surname; overall and by gender and region. Only data on the first episode of care for each patient in each financial year were used, i.e. the unit of analysis was patient-years. Confidence intervals around point estimates were calculated, but too small be visible on the graphs due to the considerable sample size of the study.

The geography of correctly predicted ethnicities based on patient surname was mapped at Local Authority level. The specificity (proportion of true negatives among the sum of true negatives and false positives) of the estimator was also calculated.

Ethical approval was obtained from Bromley Research Ethics Committee (Reference: 13/LO/1355) for analyses of patient-level HES data. The HES data licence reference was DARS-NIC-28051-Q3K7L.

## 3. Results

A total of 111,231,653 patient-years were recorded between 1999 and 2013. The completeness of ethnicity records improved from 59.5% in 1999 to 89.2% in 2009 and peaked at 90.5% in 2013 (Figure 1). The biggest absolute improvement was seen in the White British group, which increased from 55.4% in 1999 to 73.9% in 2013. Figure 2 shows increased representation for other ethnic groups.

The sensitivity analysis comparing EE estimates with NHS-recorded ethnic group, in 2013, suggested that the accuracy of prediction was highest for White British individuals (89.8%) followed by those of Pakistani (81.7%), Indian (74.6%), Chinese (72.9%) and Bangladeshi (63.4%) extraction. Lower levels of success were recorded for Black African (57.3%), White Other (50.5%) and White Irish (45.0%) groups (Table S1, supplementary materials). For other ethnic groups the sensitivity was very low and none at all for mixed ethnic groups. The sensitivity increased for the White Other group from 10.5% in 1999 to 50.5% in 2013, whereas it remained more stable for other ethnic groups over time (Figure 3). The confusion matrix for NHS-recorded ethnicity against surname predicted ethnicity can be found in Table 1. The sensitivity and specificity of the EE prediction by ethnic group each year can be found in Table S1 (supplementary materials).

The prediction success within ethnic groups were similar for males and females (Figure 3). The prediction success was however higher for females than males among Bangladeshis. The prediction success of the full name-based classification was consistently lower than when using patient surname alone for all ethnic groups (Figure S1, supplementary materials). The prediction success of the surname-based estimation for the different ethnic groups across regions, in 2013, were relatively similar except for Asian others, White Other, White Irish, and Indian (Figure 4).

The proportion of patient ethnicities predicted by the software, in 2013, was calculated and mapped for English Local Authorities (Figure 5). Prediction success was above 70% in most areas outside London but fell below 40% in parts of London.

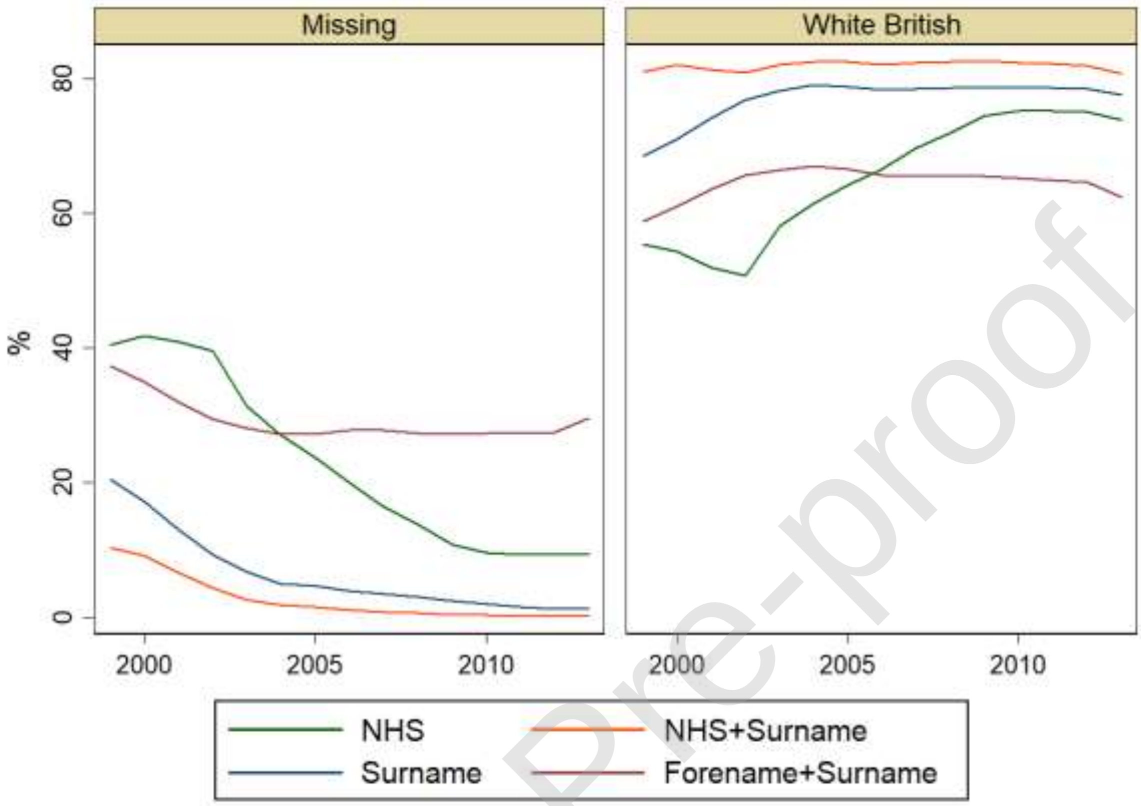


*Figure 1 Proportion of patients for missing and White British ethnicity over time.*

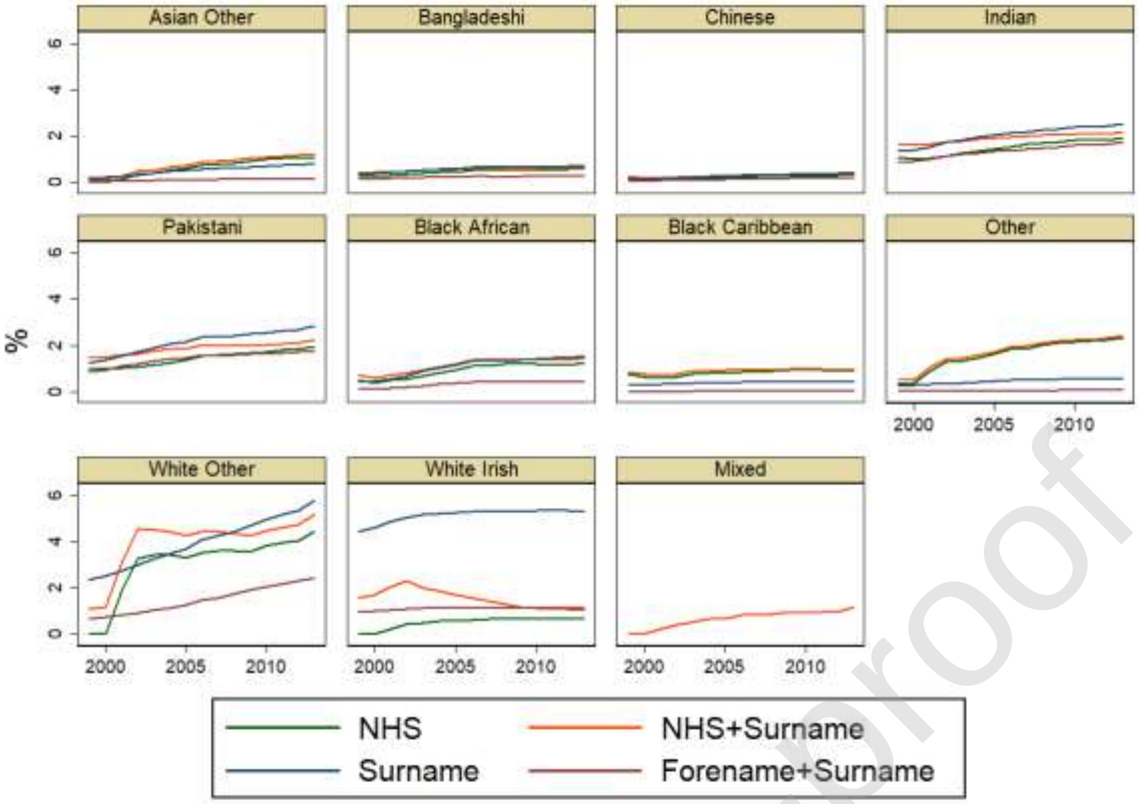


*Figure 2 Proportion of patients for each ethnic minority group over time.*

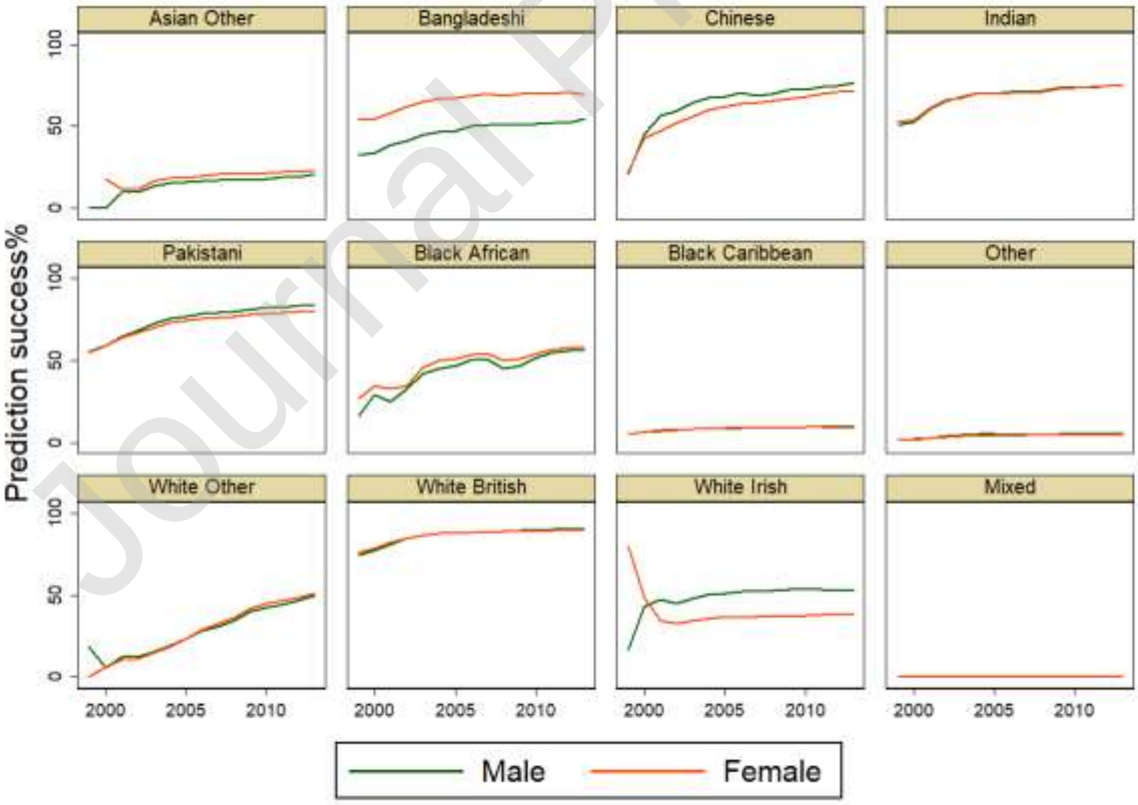

*Figure 3 Sensitivity% of Ethnicity Estimator (EE) software (Kandt & Longley, 2018) in predicting NHS-recorded ethnicity by ethnic group and gender.*

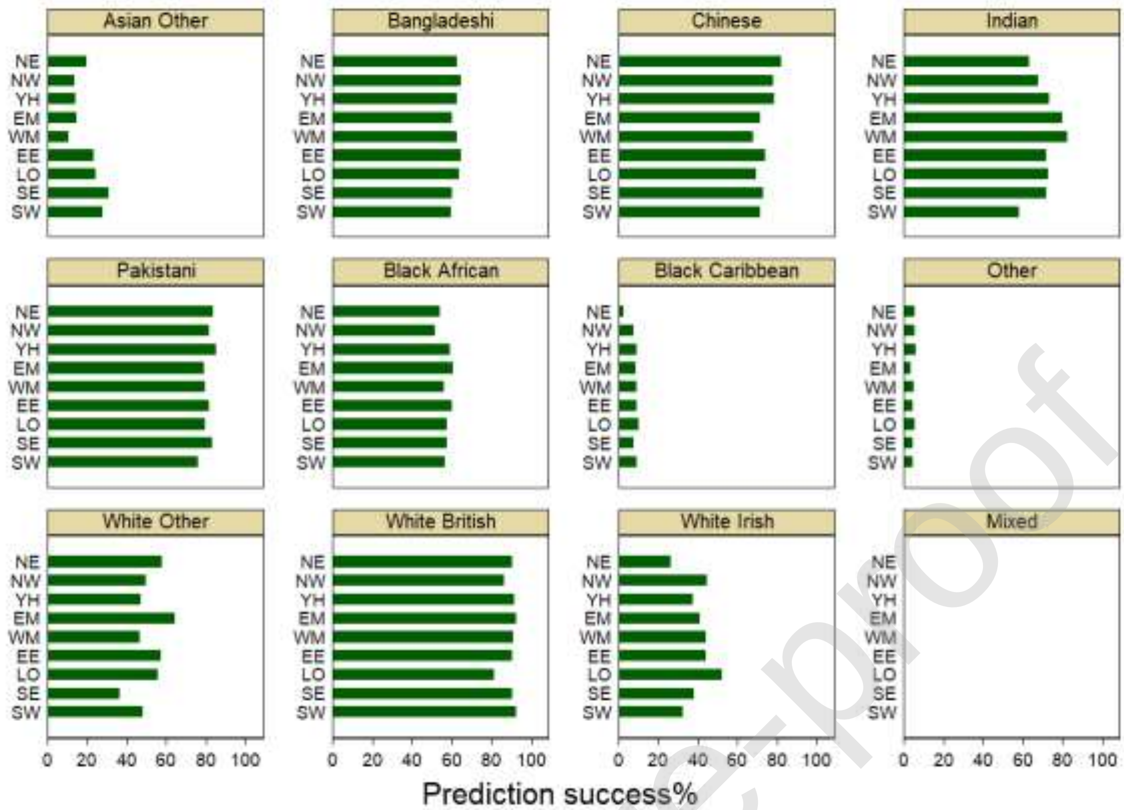


*Figure 4 Prediction success (Sensitivity) of Ethnicity Estimator (EE) software (Kandt & Longley, 2018) in predicting NHS ethnic group from patient surname in 2013 by region. Abbreviations: North East (NE), North West (NW), Yorkshire & The Humber (YH), East Midlands (EM), West Midlands (WM), East of England (EE), Greater London (LO), South East (SE), South West (SW)*

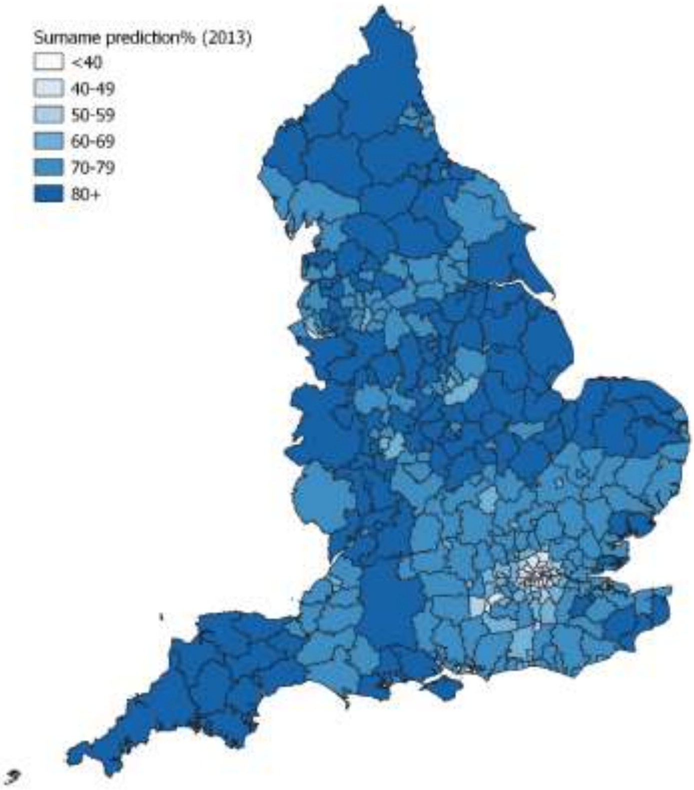


*Figure 5 Proportion of patient ethnicities correctly predicted on surname (%) in England, 2013, by Local Authority, using the EE software (Kandt & Longley, 2018).*

**Table 1 Confusion matrix of NHS-recorded versus EE surname predicted ethnicity (Kandt & Longley 2018) of HES patients in 2013.**

| | Surname prediction (EE) | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NHS-recorded | Asian other | Bangla deshi | Chinese | Indian | Pakistani | Black African | Black Caribbean | Missing | Other | White Other | White British | White Irish | Total |
| Asian Other | 21,200 | 4,119 | 2,068 | 18,513 | 18,750 | 3,142 | 266 | 5,803 | 4,242 | 6,897 | 10,495 | 409 | 95,904 |
| Bangladeshi | 685 | 32,388 | 21 | 1,486 | 12,805 | 600 | 17 | 1,272 | 392 | 482 | 901 | 51 | 51,100 |
| Chinese | 1,621 | 21 | 18,623 | 171 | 104 | 106 | 20 | 1,017 | 115 | 584 | 3,016 | 139 | 25,537 |
| Indian | 5,399 | 2,418 | 125 | 122,042 | 12,799 | 2,198 | 212 | 3,116 | 1,620 | 3,658 | 9,772 | 320 | 163,679 |
| Pakistani | 1,770 | 11,213 | 29 | 8,171 | 137,807 | 2,137 | 46 | 2,224 | 1,329 | 1,084 | 2,825 | 96 | 168,731 |
| Black African | 1,736 | 331 | 153 | 1,786 | 4,945 | 62,144 | 661 | 6,887 | 3,982 | 6,249 | 19,019 | 568 | 108,461 |
| Black Caribbean | 303 | 59 | 153 | 1,078 | 334 | 2,939 | 7,655 | 1,916 | 1,154 | 2,570 | 58,790 | 1,858 | 78,809 |
| Missing | 10,478 | 5,753 | 4,744 | 24,459 | 24,363 | 16,206 | 3,679 | 21,563 | 7,073 | 63,538 | 596,372 | 39,381 | 817,609 |
| Other | 13,803 | 2,518 | 2,397 | 9,680 | 14,441 | 24,353 | 3,184 | 15,336 | 10,183 | 42,466 | 60,324 | 3,711 | 202,396 |
| White Other | 4,981 | 721 | 790 | 4,073 | 3,573 | 4,876 | 906 | 21,113 | 6,200 | 192,404 | 131,211 | 10,050 | 380,898 |
| White British | 5,946 | 2,509 | 3,067 | 21,425 | 10,530 | 9,010 | 17,365 | 33,579 | 12,200 | 163,615 | 5,715,973 | 372,119 | 6,367,338 |
| White Irish | 72 | 29 | 25 | 353 | 105 | 101 | 214 | 749 | 128 | 1,390 | 27,928 | 25,403 | 56,497 |
| Mixed | 3,387 | 977 | 1,351 | 5,606 | 5,967 | 7,253 | 2,346 | 4,160 | 3,215 | 11,823 | 50,965 | 3,512 | 100,562 |
| Total | 71,381 | 63,056 | 33,546 | 218,843 | 246,523 | 135,065 | 36,571 | 118,735 | 51,833 | 496,760 | 6,687,591 | 457,617 | 8,617,521 |

## 4. Discussion

We found that the completeness of ethnicity data for hospital patients in England improved from 59.5% in 1999 to 90.5% in 2013. The biggest improvement was seen in the White British group, which increased from 55.4% in 1999 to 73.9% in 2013. The correct prediction of NHS-reported ethnicity varied by ethnic group (2013/14 figures): White British (89.8%), Pakistani (81.7%), Indian (74.6%), Chinese (72.9%), Bangladeshi (63.4%), Black African (57.3%), White Other (50.5%), White Irish (45.0%). For other ethnic groups the prediction success was low to none. Prediction success was above 70% in most areas outside London but fell below 40% in parts of London.

The coronavirus disease 2019 pandemic (COVID-19) has re-emphasised the importance of a better understanding of the many factors causing ethnic inequalities such as poorer living and working conditions as well as co-morbidities exacerbating infection and survival [13]. A review on ethnic inequalities in health in the UK published in 2020 reported evidence of not only persistent ethnic inequalities, but also that ethnic minorities rated the healthcare services lower than the White majority in terms of the experience and overall care [14]. One survey alarmingly found that patients from ethnic minorities had to see their family doctors several more times than White majority patients before being referred to a cancer specialist [14]. It is therefore timely to reassess the completeness and quality of ethnicity information in electronic healthcare records and propose methods to estimate ethnicity in a secure and robust manner.

Rates of NHS recording of ethnicity in HES have improved over the period of this study, especially for the White British group between 1999 and 2009. The gaps in the ethnicity records in the 1990s and early 2000s, are however likely to limit studies of ethnic inequalities. The availability of patient names was more complete than NHS-recorded ethnicity during the entire study period. There are therefore good reasons to consider alternative ways to enhance the ethnicity records either by linkage or by using name-based ethnicity classification softwares. We report on the latter approach.

NHS-recorded ethnicity supplemented with surname-based ethnicity yielded the highest completeness across the years with a few exceptions. The main exception was that using surnames alone would assign eight times more patients to White Irish background than recorded by the NHS. An Irish surname alone is in other words not a very strong predictor of individuals perceiving themselves as Irish. This is likely due to the long migration history of people from Ireland to Great Britain. The regional data showed that the prediction success for the White Irish group was higher in London than other regions, which may reflect that London has more first generation Irish migrants who still perceive themselves as Irish [10].

We would expect that the full name-based estimation would lead to higher sensitivity in the prediction. Empirically, however, we found the opposite, surname-based estimation would outperform full name-based estimation in predicting self-reported, NHS-recorded ethnicity. In the subsequent analyses, we therefore focused on the surname-based estimation. Kandt and Longley (2018) came to similar conclusions finding that forenames added little to ethnicity estimations that were based only on surnames [10]. It should in this context be mentioned that classifications based on groups of closely associated forenames and surnames are also available, i.e. the methodology used for creating the related Onomap software [10,15]. Onomap was validated against the Scottish birth registration database for 2004-2008, with slightly higher sensitivity for White, South Asian, and Chinese names than found in this study [15]. The reported sensitivity for Black African names was however as low as 25% (compared with nearly 50% in this study). It should be noted that is not possible to make a direct comparison in this case as the name-based classification methodology (surname vs forename-surname groups), study population (England HES vs Scottish birth register), and study period (1999-2011 vs. 2004-2008) were not identical.

The sensitivity of the EE software in correctly predicting NHS-recorded ethnicity was stable over time, except for the White Other group where success rates improved over time, possibly following successive EU enlargements. The sensitivity, in 2013, was nearly 90% for the White British group followed by those of Pakistani, Indian, Chinese, or Bangladeshi extraction (63%-82%) and was close to 50% for the Black African, White Other, and White Irish groups. For other ethnic groups the sensitivity was very low and none at all for mixed ethnic groups. The surname prediction results are broadly comparable with those reported in an analysis of Census 2011 microdata for England and Wales using the same name-based classification [10]. For the comparison, it should be noted that the HES only covers England and that the HES patient population is skewed towards older individuals, whereas the Census is designed to cover the entire residential population. The geographical coverage (England vs. England and Wales) and time period (1999-2013 vs. 2011) of the two sources were not identical but overlapping.

In many cultures women change surname upon marriage and this can lead to lower prediction success. The prediction success within ethnic groups were however similar in HES for males and females (Figure 3). This may in part be because we used the earliest name for name-based coding if there had been any changes over time. The prediction success was notably higher for females than males among Bangladeshis. This was also found in a study of 2011 Census microdata [10].

A main reason for the high prediction success rates among women may be that inter-ethnic marriages are still fairly rare in England [16]. Only 9% of relationships are inter-ethnic, which was even lower in the earlier time of our study period. Out of those relationships, only half were marriages that may have induced a name change. Individuals of Bangladeshi, Pakistani and Indian descent were least likely to be in inter-ethnic relationships, while those of Chinese descent showed considerable gender differences. As a result, there may only be a few cases in which individuals have adopted a surname from a different ethnic group prior to the first record of HES.

The problem of imputing ethnicity in NHS databases has previously been considered by other authors. Ryan et al. (2012) who used Onomap and Nam Pehchan to impute the ethnicity of White, South Asian, Black and Other groups in the UK’s West Midlands [8]. Nam Pehchan is based on distinctive surnames and Onomap is based on clusters of closely associated forenames and surnames. Ryan et al. 2012 used a multiple imputation strategy with characteristics of the individual patients, their care, and the ethnic composition of their neighbourhoods: they reported that the sensitivity of the multiple imputation was above 90% for White and South Asian ethnicities but was very low for other groups. Smith et al. 2017 used the Onomap software to assign children and young people with cancers to either White, South Asian, or Other groups in a Yorkshire study, concluding that combining different data sources including names-based ones increased the representation of ethnic minorities, albeit with some ambiguity [9]. Both studies concluded that there is no perfect substitute for more complete self-reported ethnicity data.

The Scottish NHS presents a parallel case with many similarities [17]. Knox et al. (2019) imputed missingness in the Scottish hospital admission database in two ways [17]. First, by assigning last recorded ethnic group to previous records for each patient. Second, by assigning remaining patients to an ethnic group based on the distribution of the different ethnic groups by sex and 5-year age band under a missing-at-random assumption. Knox et al. were in this way able to increase the completeness of the ethnicity information from 76% to 100%. The unevenness of ethnicity recording for different groups over time in both England and Scotland does not support the missing-at-random assumption [5,17], which indicate that further work is required to either collect more accurate on ethnicity or develop more sophisticated methods of imputation.

When the sensitivity of EE was mapped it also showed that successful prediction was greatest outside London (Figure 5). This is likely to be a compositional effect with a significant presence of groups with relatively low prediction success. The reason might be a higher prevalence of Anglo-Saxon names that were re-imported from abroad, such as US, Australia, or the Caribbean. In this case, the use of given names may improve the prediction for London and other cosmopolitan places. Simultaneously, London acts as an arrival hub for new migrants, among whom ethnic identities may be firmer. This latter point is supported by the fact that the prediction success in London within most ethnic groups was among the highest in the country (Figure 4).

There is currently no good surname distinction of, e.g. patients of Black Caribbean background. Future work may consider involving geographical and demographic data to improve the prediction for these groups. This may however be challenging, especially, for rarer or more geographically dispersed ethnic groups [8,17]. For more common ethnic groups, however, previous work on neighbourhood classification in London by the authors encouragingly found data on ethnicity and housing tenure correlated with large scale groupings in the data [18]. Alternatively, ethnic composition of neighbourhoods can also act as ancillary data when neighbourhood definitions are sufficiently granular (such as Census Output Areas in the UK).

In summary, studies of ethnicity in HES, 1999-2013, are compounded by a number of caveats. The completeness of ethnicity data was below 60% in 1999. It improved in the 2000s and reached a plateau of 89-90% in 2009-2013. The completeness of patient surnames improved from 79% in 1999 to 93% in 2002; then gradually improved to 99% in 2013. If patient names are used for ethnicity estimation, it should be noted that the sensitivity (prediction success) varies by ethnic group, e.g. it is close to 90% for White British and approximately 50% for Black African. For White Other, the sensitivity notably increased from 10.5% in 1999 to 50.5% in 2013. We also found that the surname estimation inflated the White Irish group considerably relative to individuals reporting themselves as White Irish. The representation of different ethnic groups in HES could potentially be improved by retrospective linkage to the patient register or other data sources with better quality data. Names-based classification can however be a method for estimating ethnicity in studies where linkage is not feasible. And work on barriers in recording self-reported ethnicity [19] further highlights the importance of developing alternative approaches to capture and differentiate ethnicity in health care to assess multiple inequalities in health.

## 5. Limitations

As a limitation, it should be noted that ethnicity is a complex concept encompassing biological, cultural, and subjective aspects. Which aspect matters most depends on the kind of inequalities that are the object of the study and the related assumptions about disease aetiology. Variation in prediction success of name-based ethnicity classification can therefore arise for different reasons including individuals' sense of belonging and resulting choice of ethnic group, socio-cultural naming and name-change practices, distinctiveness of names across ethnic groups, and the extent to which the name-based classification covers different origins at a given time point, e.g. when later waves of immigration have widened the range of diasporic names in the host country since the creation of the software. As a result, a given surname can be found in different ethnic groups, although the vast majority of surnames concentrate significantly in one particular ethnic group at this level of ethnic classification [10]. More detailed analysis of secondary ethnic groups and non-matching ethnicity predictions can help to disentangle these different aspects but were outside the scope of the current study.

Moreover, it should be acknowledged that names-based ethnicity imputation is biased towards common over rarer names. It should also be noted that even though the NHS hospital services are free at the point of use to all, there are groups that underuse the services. These include very wealthy persons who exclusively prefer private medical care and groups avoiding contact with all authorities or the use of Western medicine It seems unlikely that this should create biases in the names material available for this study, but the possibility cannot be ruled out. Finally, there may also be individuals who have affinities for more than a one ethnic group due to multiple nationality, adoption, mixed parentage, or lifestyle choice.

**6. Conclusion**

Studies of ethnic inequalities in hospital inpatient care in England are limited by incomplete data on patient ethnicity in the 1990s and 2000s. Financial incentives for general practitioners to collect and report ethnicity to the central patient register between 2006 and 2012 have greatly improved completeness during this period. Personal names of patients remain an untapped source for closing this gap for the earlier years. As demonstrated in this - and other studies - name-based ethnicity classifications have merit for the predictions of many ethnic minorities. The case for name-based ethnicity classification is naturally stronger for databases where ethnicity is not collected systematically, e.g. accident and emergency department data [20] or more recently COVID-19 admissions in the Welsh hospital admission database [7]. The current work also highlights areas where name-based ethnicity classifications can be improved. There is currently no good surname distinction of, e.g. patients of Black Caribbean background. Future work may consider involving geographical and demographic data to improve the prediction for these groups.

**Source of funding**

The UK Economic and Social Research Council is acknowledged for its support for the UCL Consumer Data Research Centre (CDRC) enabling this research (Grant ES/L011840/1).

**Authors' contributions**

All authors made substantial contributions to the conception and design of the study, interpretation of data, revision, and final approval of the submitted version (JP, JK, PAL). JK and PAL contributed to the acquisition of data. JP contributed with analysis and drafting of the article.

**Statement of competing interests**

None.

**Acknowledgements**

None.

**Summary table**

*What is already known on this subject*

- Studies of ethnic inequalities in hospital inpatient care in England are limited by incomplete data on patient ethnicity collected in the 1990s and 2000s.

- Name-based ethnicity classifications have merit for the predictions of many ethnic minorities, yet there has to date not been any studies of the completeness of personal names in HES or the prediction success of name-based classifications over time.

*What this study added to our knowledge*

- The prediction success of a names-based ethnicity classification tool has been quantified in HES for the first time and the results can be used to inform decisions around the optimal analysis of ethnic groups using this data source.
- The work also highlights areas where name-based ethnicity classifications can be improved, e.g. for patients of Black Caribbean background.
- Future work may consider involving geographical and demographic data to improve the prediction for these groups.

**Appendix A. Supplementary materials**

Supplementary material related to this article can be found, in the online version.

**References**


[1] European Union, Regulation (EU) 2016/679 of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), (2016). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e40-1-1 (accessed August 20, 2020).

[2] B. Byrne, C. Alexander, O. Khan, J. Nazroo, W. Shankley, Ethnicity, Race and Inequality in the UK - State of the Nation, Policy Press, 2020. https://policy.bristoluniversitypress.co.uk/ethnicity-race-and-inequality-in-the-uk (accessed November 4, 2020).

[3] The Economist, American ideas about racism are influencing Europe, The Economist. (2020). https://www.economist.com/europe/2020/08/08/american-ideas-about-racism-are-influencing-europe (accessed August 19, 2020).

[4] The Economist, An edgy inquiry, The Economist. (2015). https://www.economist.com/europe/2015/04/04/an-edgy-inquiry (accessed August 19, 2020).

[5] R. Mathur, K. Bhaskaran, N. Chaturvedi, D.A. Leon, T. vanStaa, E. Grundy, L. Smeeth, Completeness and usability of ethnicity data in UK-based primary care and hospital databases, J. Public Health Oxf. Engl. 36 (2014) 684–692. https://doi.org/10.1093/pubmed/fdt116.

[6] R.W. Aldridge, D. Lewer, S.V. Katikireddi, R. Mathur, N. Pathak, R. Burns, E.B. Fragaszy, A.M. Johnson, D. Devakumar, I. Abubakar, A. Hayward, Black, Asian and Minority Ethnic groups in England are at increased risk of death from COVID-19: indirect standardisation of NHS mortality data, Wellcome Open Res. 5 (2020). https://doi.org/10.12688/wellcomeopenres.15922.2.

[7] D.R. Thomas, O. Orife, A. Plimmer, C. Williams, G. Karani, M.R. Evans, P.A. Longley, J. Janiec, R. Saltus, A.G. Shankar, Ethnic variation in outcome of people hospitalised with Covid-19 in Wales (UK): A rapid analysis of surveillance data using Onomap, a name-based ethnicity classification tool, (In prep.).

[8] R. Ryan, S. Vernon, G. Lawrence, S. Wilson, Use of name recognition software, census data and multiple imputation to predict missing data on ethnicity: application to cancer registry records, BMC Med. Inform. Decis. Mak. 12 (2012) 3. https://doi.org/10.1186/1472-6947-12-3.

[9] L. Smith, P. Norman, M. Kapetanstrataki, S. Fleming, L.K. Fraser, R.C. Parslow, R.G. Feltbower, Comparison of ethnic group classification using naming analysis and routinely collected data: application to cancer incidence trends in children and young people, BMJ Open. 7 (2017). https://doi.org/10.1136/bmjopen-2017-016332.

[10] J. Kandt, P.A. Longley, Ethnicity estimation using family naming practices, PLOS ONE. 13 (2018) e0201774. https://doi.org/10.1371/journal.pone.0201774.

[11] P. Mateos, P.A. Longley, D. O'Sullivan, Ethnicity and Population Structure in Personal Naming Networks, PLOS ONE. 6 (2011) e22943. https://doi.org/10.1371/journal.pone.0022943.

[12] J. Kandt, J. Van Dijk, P.A. Longley, Family name origins and inter-generational demographic change in Great Britain, Am. Geogr. Soc. 110 (2020) 1726–1742.

[13] N. Bhala, G. Curry, A.R. Martineau, C. Agyemang, R. Bhopal, Sharpening the global focus on ethnicity and race in the time of COVID-19, The Lancet. 395 (2020) 1673–1676. https://doi.org/10.1016/S0140-6736(20)31102-8.

[14] K. Chouhan, J. Nazroo, Health inequalities, in: Ethn. Race Inequal. UK - State Nation, Policy Press, Bristol, UK, 2020: pp. 73–92.
[15] F. Lakha, D.R. Gorman, P. Mateos, Name analysis to classify populations by ethnicity in public health: Validation of Onomap in Scotland, Public Health. 125 (2011) 688–696. https://doi.org/10.1016/j.puhe.2011.05.003.
[16] ONS, 2011 Census analysis: What does the 2011 Census tell us about Inter-ethnic Relationships?, 2014. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/marriagecohabitationandcivilpartnerships/articles/whatdoesthe2011censustellusaboutinterethnicrelationships/2014-07-03 (accessed January 22, 2021).
[17] S. Knox, R.S. Bhopal, C.S. Thomson, A. Millard, A. Fraser, L. Gruer, D. Buchanan, The challenge of using routinely collected data to compare hospital admission rates by ethnic group: a demonstration project in Scotland, J. Public Health. fdz175 (2019). https://doi.org/10.1093/pubmed/fdz175.
[18] J. Petersen, M. Gibin, P. Longley, P. Mateos, P. Atkinson, D. Ashby, Geodemographics as a tool for targeting neighbourhoods in public health campaigns, J. Geogr. Syst. 13 (2010) 173–192. https://doi.org/10.1007/s10109-010-0113-9.
[19] W.-C. Lee, S.P. Veeranki, H. Serag, K. Eschbach, K.D. Smith, Improving the Collection of Race, Ethnicity, and Language Data to Reduce Healthcare Disparities: A Case Study from an Academic Medical Center, Perspect. Health Inf. Manag. 13 (2016). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5075235/ (accessed January 21, 2021).
[20] J. Petersen, P. Longley, M. Gibin, P. Mateos, P. Atkinson, Names-based classification of accident and emergency department users, Health Place. 17 (2011) 1162–1169. https://doi.org/10.1016/j.healthplace.2010.09.010.