How accurate is NHS ethnicity data?

A paper to support the COVID-19 response

Richard Webber, Trevor Phillips and Dr Marc Farr

April 2021

1. Intro

One element of the COVID-19 pandemic which has been consistently observed for the past year has been its disproportionately high impact on BAME (Black and Minority Ethnic) groups.

A range of factors may account for this. These include the relative deprivation of many BAME groups – which can mean higher population density and greater exposure to conditions within which the virus is able to spread. There may also be cultural factors. Living with elderly relatives is more common among South Asian communities, for example. And, to an extent to which we are not yet clear, genetic variations may result in different levels of vulnerability within certain ethnic groups as occurs for example with diabetes or sickle cell anaemia.

In acknowledgement of this there were calls from many quarters for BAME groups to be prioritised for vaccine rollout. The government announced in December that "Good vaccine coverage in BAME groups will be the most important factor within a vaccine programme in reducing inequalities for this group."[1]

Meanwhile, there is significant data suggesting that BAME groups are more likely to be sceptical about receiving the vaccine, with uptake among many minority groups much lower. A poll just before Christmas reported 57% of BAME respondents saying they would receive the jab, for example – compared with 79% of white British respondents.[2] There has been some discussion among health stakeholders about which approaches might best reassure BAME communities.

All of these questions reflect an important commitment by the NHS and other agencies to think seriously about data, ethnicity and health – acknowledging variations in behaviour and impact. Yet the ability to address these situations properly is predicated on having accurate information in the first place. If we want this debate to be serious and informed then getting the numbers right is critical.

Indeed, the one recommendation in this year's Commission on Race and Ethnic Disparities paper was that the term 'BAME' itself be retired. "It is demeaning to be categorised in relation to what we are not, rather than what we are," the report states. "The BAME acronym also disguises huge differences in outcomes between ethnic groups."[3]

We have used the term 'BAME' in a descriptive sense within this analysis, as it is the framework upon which data is still frequently collected; it needs to be used in order to be critiqued. But our findings very much reinforce the Commission on Race and Ethnic Disparities' findings when it comes to this term. The acronym BAME is not, according to our report, up to the task of understanding the complex relationships between ethnicity and health outcomes.

Written on the basis of Webber Phillips' analysis and that of an example NHS Trust, this paper looks at three respects in which current methods of data collection for health and ethnicity may be giving a misleading picture.

2. Hypothesis

National statistics on COVID-19 are constructed by adding together reported data from each of England's 151 hospital trusts. As a consequence, these statistics are only as reliable as the data each trust uploads. How accurate, detailed and consistent are they?

In this report we look at limitations which consistently appear in collection processes. Whilst not fully invalidating the accuracy of existing NHS statistics for ethnicity, we believe these limitations introduce biases – meaning the impact of the virus on BAME groups is consistently under-estimated. This is significant when assessing COVID-19's impact on different minorities at a national level, calling into question the reliability of the evidence on which public health policies and communications are based.

In particular, we have been concerned by three potential sources of error.

The first involves the size and composition of those whose ethnicity is unrecorded – and who, as a result, are categorised in the statistics as 'Other', 'Unknown' or 'Not stated'. Can we be sure that the makeup of these (comparatively large) groups are the same as those whose ethnicity is recorded? If it is not then how do we measure the resulting bias? The risk is that some BAME groups are more likely to misallocated into these categories than others, and that this has implications not just for the virus but also for other health policies. It would mean we are not getting the full picture when it comes to COVID-19-related racial inequalities, or to accurate evidence of BAME access to and uptake of vaccination programmes.

The second potential source of error involves the increasingly broad 'white other' category. This grouping is used to categorise people who self-identify neither as British nor black or Asian. But it includes many patients who, by many objective measures, would be considered as BAME and who, as a result, may experience the same systemic challenges as other BAME groups. The number of patients who fall into this category has grown substantially in Britain in the past two decades. (There are now about as many people of Polish descent as there are of African Caribbean ancestry, for example). And this growth of the 'white other' category means, in turn, that it has become increasingly heterogeneous. There are likely to be significant differences between people of German, Polish, Turkish and Bosnian heritage whether in terms of access to health and other services or appropriateness of different forms of communication. Likewise, when it comes to trust of institutions and cultural practices. Yet all might be classified as 'white other'.

The third potential error relates to the misclassification of some members of BAME groups as ethnically 'British'. This reflects the conflation on many government agency forms of 'Britishness' and 'Whiteness'. Our hypothesis is that many members of BAME groups currently self-describe as British – either for reasons of identity or due to anxieties regarding citizenship status – and that this inflates the 'White', 'British' and 'White British' ethnicity categories. It may mean ethnic vulnerabilities to the pandemic are not captured in the data, or that support for BAME groups is not being delivered in full.

Working in partnership with the Chief Analytical Officer at one of the country's 151 NHS Foundation Trusts, we have been able to address these questions in turn. We have done this using a complete set of patient records for the area the NHS Trust covers.

3. Methodology

Data Source

This is a joint project, carried out by the example NHS Foundation Trust in question and Webber Phillips. It involves a study of all patient visits between January 2018 and April 2020, to hospitals covered by the trust. The example NHS Trust serves a catchment area of over half a million adults. And it captures information on around 400,000 in-patients each year. This means that, in total, we were able to access to nearly a million records for the purposes of this project.

Patient ethnicity is, of course, one of the attributes which the Trust seeks to collect. Depending on the circumstances, this information may be collected by medical professionals assigning an ethnic category to a patient or by the patient themselves selecting a category from a list. What this means is that, for all 938,384 records analysed, there is both an ethnicity code provided by the hospital and a full name.

Use of Origins

Origins is a name recognition tool, developed by Webber Phillips. The tool is used across the public, private and third sectors, including by a number of government agencies. It is based on a database of around 1,300,000 unique forenames and about 4,000,000 unique surnames, collected worldwide, and is able to identify the regions of the world from which forebears bearing this name are most likely to have originated.

The attribution of names varies between cultures – sometimes occurring based on national borders, sometimes by faith and sometimes by language. Origins identifies 11 core categories (e.g. East European names), 50 sub-categories of names (e.g. Baltic names) and about 200 specific classifications (e.g. Lithuanian names).

Origins analyses both forenames and surnames, attributing a confidence score for each and weighting certain categories of names. It has been tested extensively and is able to assess, with a high degree of accuracy and granularity, the ethno-cultural make-up of a given group of names.[4] The larger the sample, the higher the degree of accuracy.

The advantage of the tool is that it is able to offer greater coverage, consistency and ease than regular sampling. Clearly no predictive system is likely to be able to predict the ethnicity of an adult with 100% accuracy from their name. However at the level of detail such as ‘South Asian’, ‘Black African’, ‘Eastern European’ or ‘Turkish’ the accuracy is very high.

This report cross-references the outcomes of the NHS Trust’s manual data collection against the Origins tool, looking at the three areas of potential error mentioned in Section 2.

4. Findings

The Trust associates each patient with one of the following 21 ethnic categories. These are listed in the table below, alongside the proportion of patients who fall into each one.

Code	Category	
99	NOT KNOWN	1.20
A	BRITISH	85.83
B	IRISH	0.33
C	ANY OTHER WHITE BACKGROUND	4.05
D	WHITE AND BLACK CARIBBEAN	0.14
E	WHITE AND BLACK AFRICAN	0.11
F	WHITE AND ASIAN	0.19
G	ANY OTHER MIXED BACKGROUND	0.62
H	INDIAN	0.31
J	PAKISTANI	0.06
K	BANGLADESHI	0.08
L	ANY OTHER ASIAN BACKGROUND	0.46
M	CARIBBEAN	0.07
N	AFRICAN	0.21
P	ANY OTHER BLACK BACKGROUND	0.21
R	CHINESE	0.12
S	ANY OTHER ETHNIC GROUP	1.08
T	ARAB	0.00
U	GYPSY OR IRISH TRAVELLER	0.00
V	NEPALESE	0.03
Z	NOT STATED (PATIENT ASKED BUT DECLINED)	4.90
ALL PATIENTS		100.00

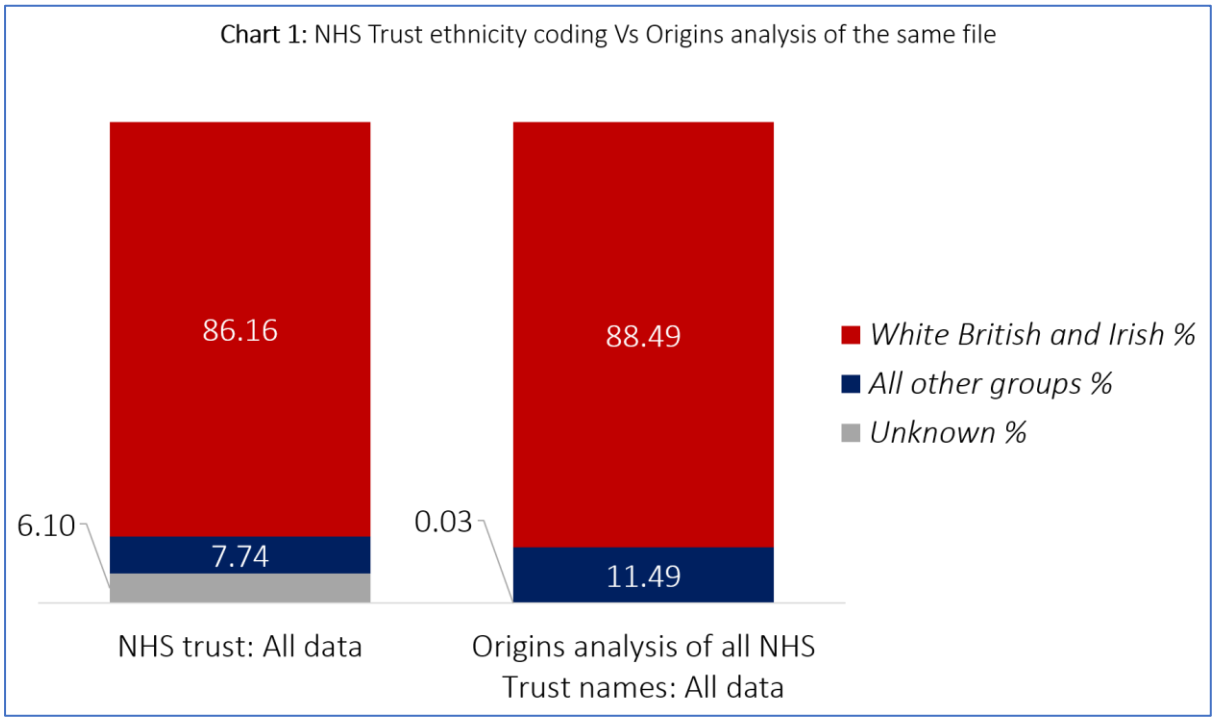
Table 1: *codes and categories recorded for each patient*

'Not known' and 'Not Stated'

Our first task was to look at the effect on ethnicity statistics of not recording or stating ethnicity. This impacts on the statistics which the NHS Trust – and, subsequently, Public Health England (PHE) – use to understand the impact of COVID-19 on the BAME community.

Of the Trust's 21 codes, two indicate that ethnic information is missing: code 99 ('Not Known') and code Z ('Not Stated (Patient asked but declined)'). In total 6.10% of patients are assigned one or the other of these two codes. That some codes are missing is unavoidable. But it constitutes a problem for the overall statistics if there is a relationship between the missing data and ethnicity itself.

When Webber Phillips used the Origins tool to examine the full name of each patient we found that there was, indeed, such a relationship. Missing ethnicity data was distributed in a far from random manner. Chart 1 shows the proportion within the sample who are coded as British or Irish (codes A or B), as Unknown (codes 99 or Z) and who belong to all other groups (codes C-V). The chart compares this with an Origins analysis of all 938,384 names within the same sample. The latter method allows a much greater level of coverage – with only 0.03% of names unclassifiable by Origins, compared to 6.10% unknown in the Trust database.



By looking at the two breakdowns, side by side, we can see that the 'All other groups' category grows proportionally larger when you use Origins to remove the vast majority of unknown names. In short, White British or Irish names were significantly underrepresented within the 6.10% which were 'Not known' or 'Not stated', compared to among the set of patients as a whole.

This is significant. It means that, over the past two and a half years, the NHS Trust has been reporting that 7.74% of patients were *not* of White British or Irish heritage, when in fact the figure was closer to 11.48% – a difference of 3.74% (or 35,095 individuals).

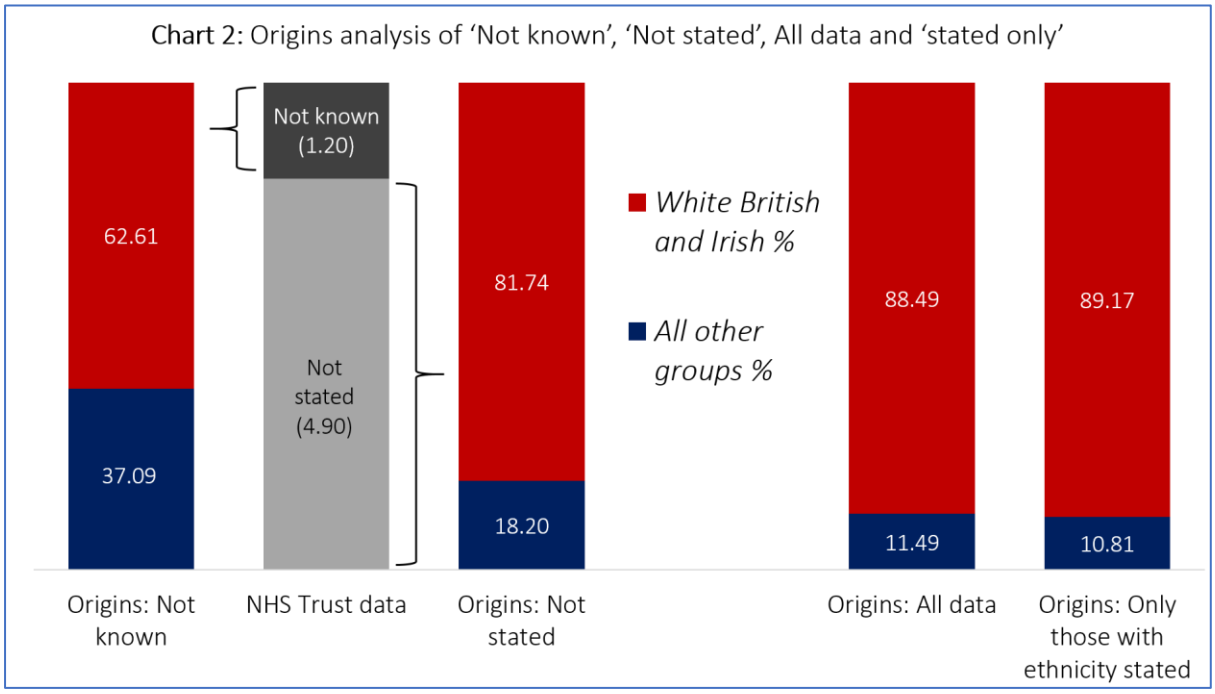


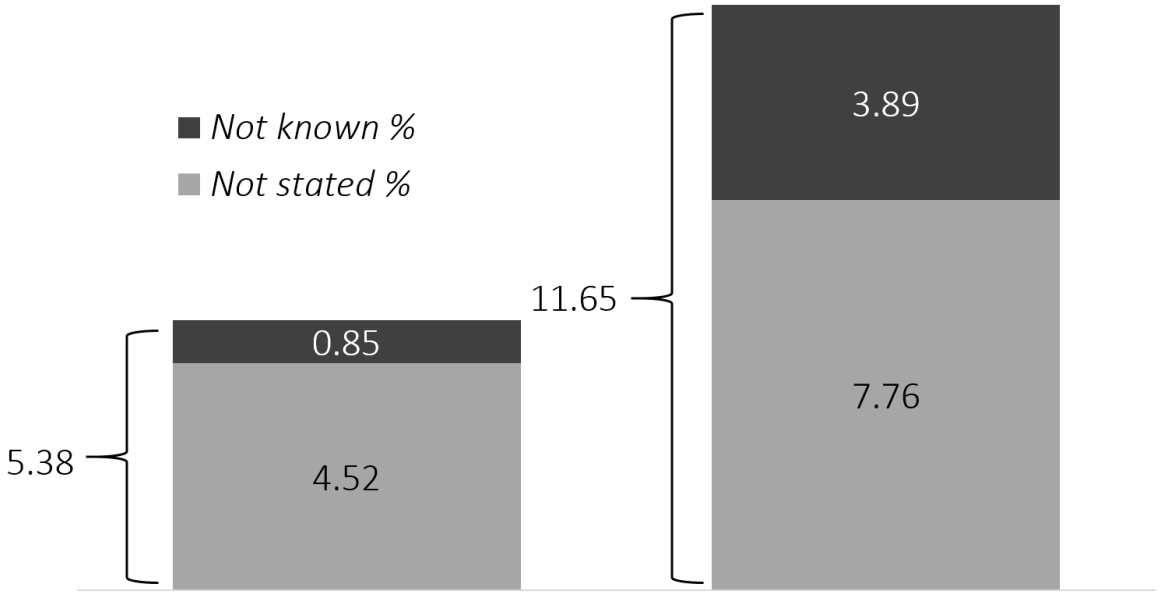
Chart 2 breaks down the 6.10% of records that are unknown (57,251 total), looking at the proportion which, according to Origins, are of white British or Irish heritage within each. Among the 4.90% that

are 'Not stated' – those cases where the patient declined to give their ethnicity – we can see that the proportion in 'all other groups' is significantly higher than for the database as a whole. (Figures for the database as a whole are shown, by way of comparison, in the second column in from the right). 18.20% of records coded as 'Not stated' do not have White British or Irish names, compared to 11.48% for the database as a whole.

When we look at 'Not known', meanwhile – those cases where a medical professional was coding, rather than the method being self-identification – we can see that this is even starker. 37.09% of those where the ethnicity is 'Not known' have neither White British nor Irish names. This is more than three times the proportion for the database as a whole. (NB: The column furthest the right in Chart 2 shows the Origins breakdown including only those where the names are stated, showing an even heavier slant than the database as a whole towards White British and Irish names).

Chart 3 looks at these findings in a slightly different aspect, focusing on likelihood, across groups, of ethnicity being unknown. It depicts the proportion that are 'Not known' or 'Not stated' among those who have White British and Irish names, and the proportion among those who do not. It shows that, whereas data on ethnicity was missing for just 5.38% of patients with White British or Irish names, the corresponding proportion for people with other categories of name was over twice as high, at 11.65%.

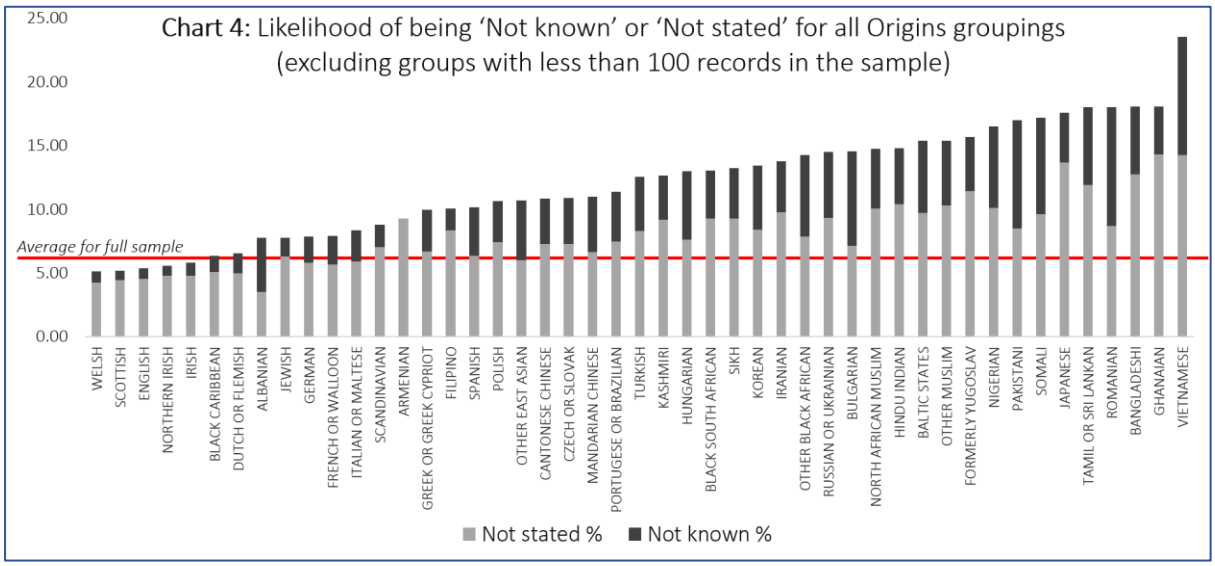
Chart 3: Likelihood of being 'Not known' or 'Not stated' – White British and Irish names compared with all other names



Meanwhile, by using Origins to look at this in a more granular way, we can also see that not all minorities were equally unlikely to have their ethnicity unrecorded. Chart 4 shows the percentage which were categorised as 'Not known' or 'Not stated' within each of the 50 Origins sub-groups (excluding groups with fewer than 100 members in the sample).

Ethnicity codes are missing for 15.35% of people with names from the Baltic States, for example, suggesting that those of Baltic heritage are three times as likely as those with English or Welsh names

to have their ethnicity go unrecorded. The figures are 16.51% for people with Nigerian names, 18.02% for those with Romanian names, 18.04% for those with Bangladeshi names and 23.51% among the 302 patients with Vietnamese names. We estimate that the proportion of Trust patients who are of Vietnamese ethnicity is being underestimated by some 25% in the statistics recorded centrally.



Overall, every Origins groups apart from those with White British and Irish names were more likely than the sample as a whole to be coded as 'Not known' or 'Not stated'. And populations that were more likely to have arrived recently appeared especially unwilling to state their ethnicity – particularly those with names from less economically developed countries. Trust of the state, fears about the migration system and language barriers may all be factors here.

There are also clear differences, demonstrated by variations in the 'Not known' figure, in how well medical professionals are able to identify different ethnicities. It would be interesting to look at this sort of dataset in other parts of the UK, to see whether hospital trusts with more diverse populations and more contact with minority groups are more confident in assigning codes.

The 'white other' code

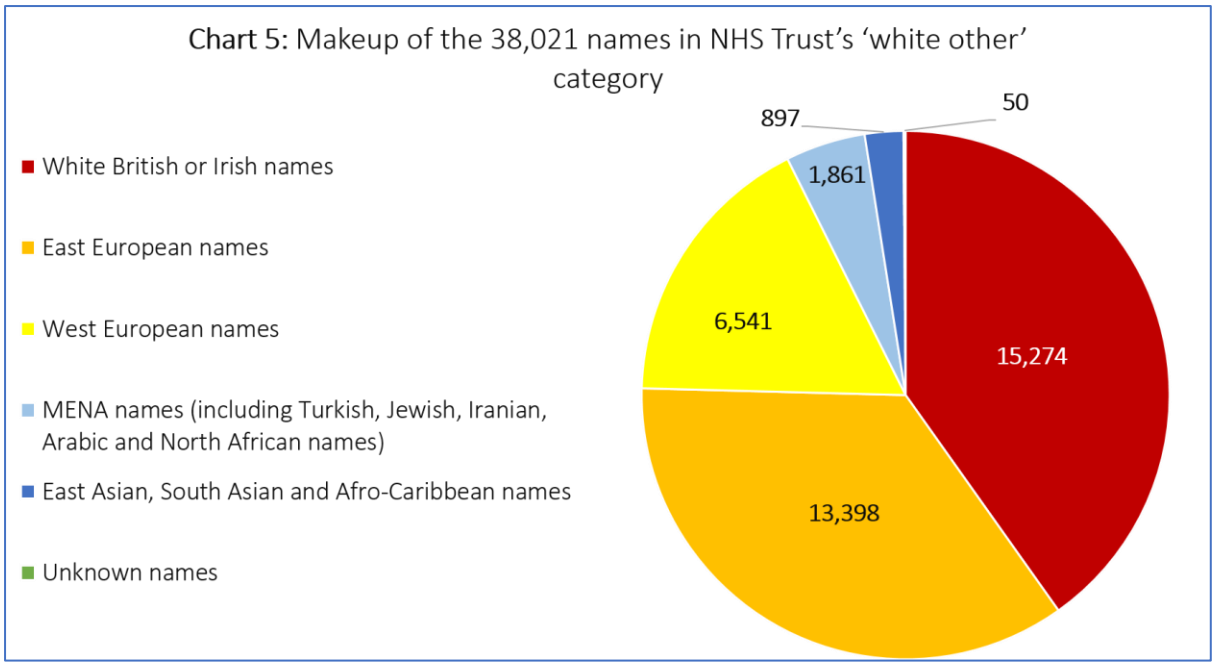
Our second concern with the NHS's ethnicity statistics lies with the categorisation system itself. We decided to look at the composition of the very large group of patients falling into the category 'Any other white background' (code C in Table 1). This group represented 4.05% of the overall sample, making it larger than all the other recorded categories apart from British put together.

Code	Category / categories	%
A	British	85.83
99 or Z	Not known or Not stated	6.10
C	Any other white background	4.05
All others	All other categories	4.02

Table 2: *Size of the 'Any other white background' category*

Who are the groups falling into this 'white other' category? How many members of the grouping are at a higher risk of COVID-19 or of other health issues? And what proportion are subject to the same systemic inequalities as BAME groups?

To understand this we again used the Origins tool, by analysing the names of all the 38,021 patients who were recorded as being of 'any other white background'. The results are shown in Chart 5, with the names grouped into broad categories.



To begin with, there are a large proportion – over a third – with White British or Irish names. It is hard to say whether this is caused by errors in completing forms, by anomalies (i.e. German nationals with British names) or by White British people who identify more by ethnicity than by nationality. But it confuses this category somewhat.

Turning to the other groups, we can see that in the area covered by the NHS Trust – although not necessarily elsewhere – the largest sub-groups within the 'white other' category are predominantly those of Eastern European heritage. Polish names are the most common, followed by Romanian names, Russian or Ukrainian names, Czech names and those from the Baltic States.

West European names also feature prominently – with significant numbers of German, Italian or Maltese, and Spanish names within the 'white other' sample. These groups are probably less likely to consider themselves BAME or to be as high risk when it comes to health.

However, there was also a significant proportion within the sample who we have grouped together as Middle East and North African (MENA). The largest groups within this are those with Turkish, Arabic and Iranian ancestry, according to Origins. Turkish names alone accounted for 962 'white other' records. These MENA categories are groups who are considered BAME according to some definitions, but whose risk is varied and difficult to gauge.

Lastly, we see 897 records whose heritage is from majority-BAME countries – including a significant number of South Asian and Black African names.

By exploring the composition of the 'white other' category using Origins we were therefore able to identify, in a much more granular way, the diverse group of minorities included within it. It is a category which, like the 'Not known'/'Not stated' question, potentially reveals significant under-reporting of the size of the BAME population.

But, more importantly, it poses big questions about what we mean by BAME. If we are looking at ethnicity or culture then Iranian, Turkish or Arabic names might be considered BAME. If we are looking at societal factors such as access to healthcare and housing, meanwhile, there could be an argument that Romanian and Bulgarian groups could be included within the category. These groups may face similar levels of precariousness to poorer BAME groups, such as living in close proximity to one another or facing language barriers.

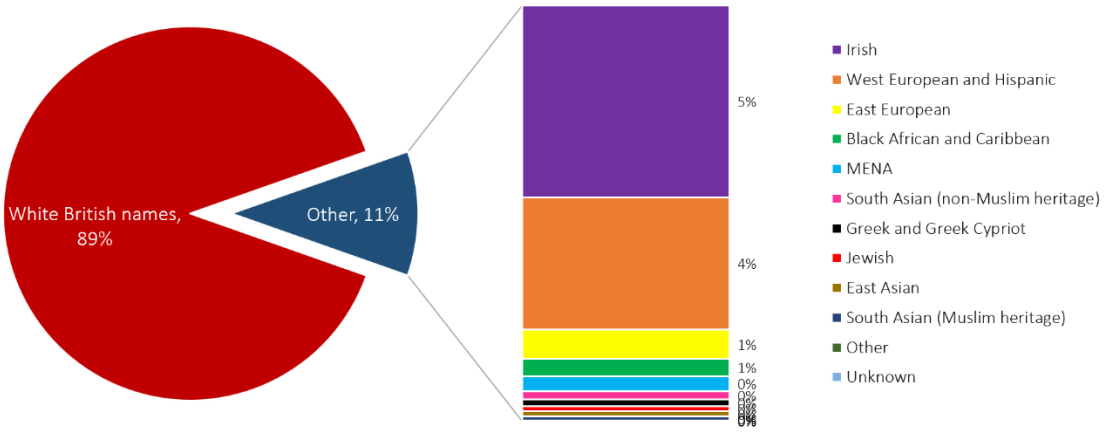
BAME groups and the British category

The number of BAME and White British names within the 'white other' category points to our third question, relating to 'misclassification'. This area of potential error relates to the over-representation of people of BAME heritage in 'British' or 'White British' categories.

Chart 6 examines the makeup of the 805,376 names categorised by the Trust as British (code A). This category is presumably aiming to capture British *ethnicity* rather than British *citizenship* – given that a number of other codes exist for British nationals from different backgrounds. And the term 'British' rather than 'White British' may be part of the issue here.

As the chart shows, our Origins analysis reveals a significant number with non-British names falling into this group. 11% of those coded as British have names which are not of White British heritage. This includes large numbers of Irish and West European names. But it also includes 6,158 East European names and 3,574 Afro-Caribbean or Black African names.

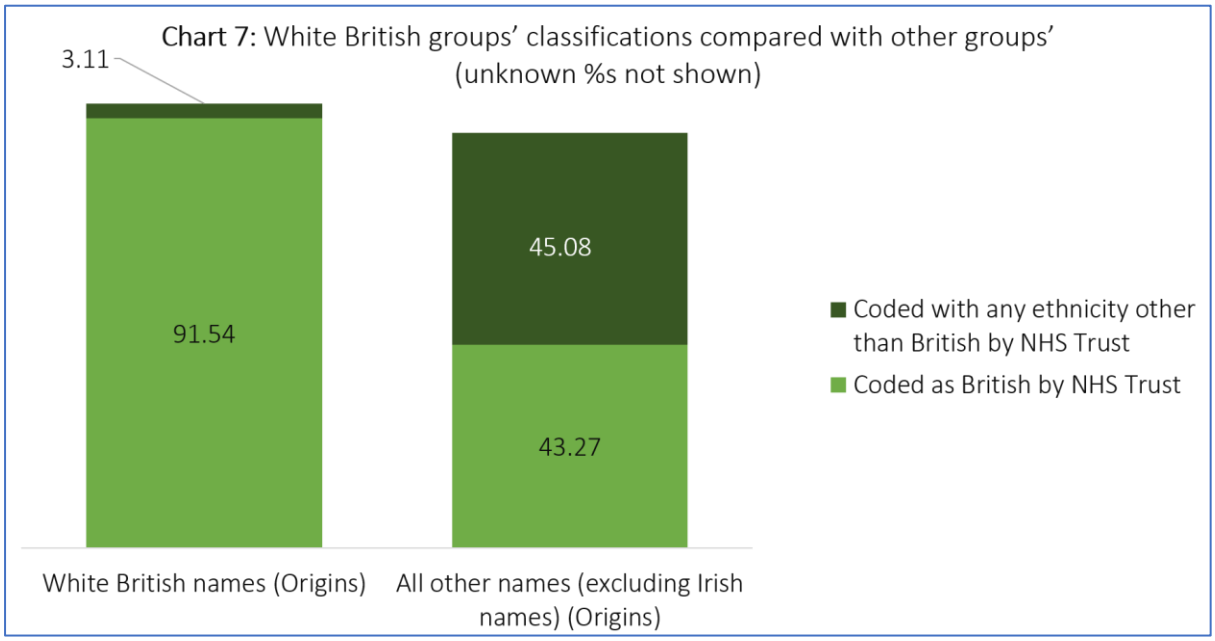
Chart 6: Origins analysis of the 805,376 names coded British



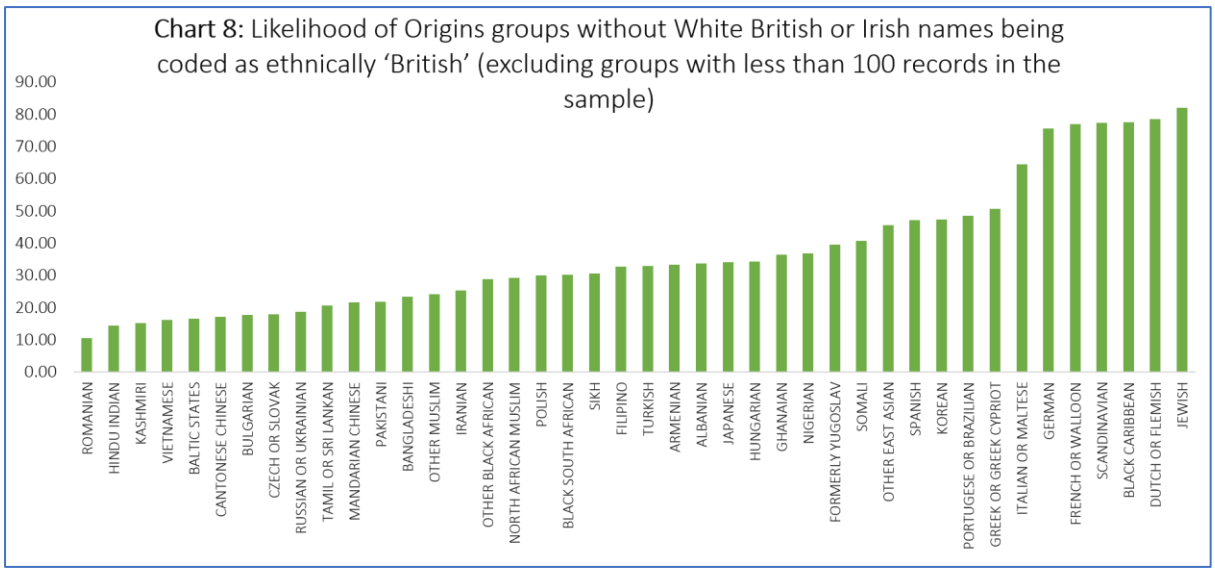
Obviously, within any sample like this, there will be an element of human error – the proportion 'ticking the wrong box'. Likewise, there will be genuine anomalies or questions around identity (which the Origins approach would obviously not seek to second guess). For example, there may be people with traditionally Anglo-Saxon forenames and surnames who are, in fact, German, or people with Germanic forenames and surnames who are, in fact, of White British ancestry. But issues arise if these misclassifications occur more in one direction than in the other, suggesting a bias in the data.

Chart 7 looks at the above data in a different light. It shows the *likelihood* that those with White British names (according to Origins) will be coded as British, compared with the same likelihood among all other names. (NB: In doing this we excluded Irish names from both categories, as these tended to skew the data whichever category they were included as).

When looking at this the findings are stark. Within the 785,145-strong sample with White British names, just 24,415 people (3.11%) were coded by the NHS Trust with a category other than British. Yet when looking at the 107,789 patients without White British (or Irish) names, the picture was completely different. 46,639 were coded as British (43.27%) – i.e. with an ethnic coding different to their name. Once you remove the unknowns, this means that very nearly half of those with names not of White British heritage were nevertheless coded as having British ethnicity according to the data collection processes used by the NHS Trust.



This is a remarkable finding, and there are several factors which may account for it. The first is to do with labelling. The fact that the category was called 'British' rather than 'White British' may have meant it was used to describe nationality rather than ethnic background. How many other NHS hospital trusts use 'British' as an ethnicity code is not known.



The second factor relates to citizenship status or country of residence. As we can see from Chart 8 – which looks at the likelihood of being coded as British among all Origins groups apart from White British and Irish – 39.60% of those names originating from Former Yugoslavian countries were coded

as British. Likewise, 40.63% of those with Somali names. That these groups code themselves as British in terms of ethnicity may again reflect anxieties about citizenship status.

Lastly, individual identity is clearly also a major factor here. According to our best estimates 77.53% of those with Black Caribbean names are coded as British, for example (a category we have carefully re-weighted because of the overlap with White British names). The same is true of Greek Cypriot names (50.62% of which were coded as British), Nigerian names (36.75%) and Turkish names (32.85%).

There are some interesting underlying distinctions here – for example, the difference between those with Sikh and Hindu Indian names or between Hungarian and Romanian. But the frequency of identifications as British among Origins groups with names from South Asia or West Africa is nevertheless very striking.

None of these patients will appear in a BAME category when data is reported centrally. Based on this data we estimate a further 10% under-recording of the number of patients of BAME origin, whether in this or in other hospital trusts.

This raises major questions. The ways in which people self-identify are very important. But they do not always conform to the categories used to develop or justify public policy in relation to BAME groups, when we are looking at the connection between COVID-19 and ethnic background.

5. Conclusions

The opportunity to analyse a database of a nearly one million records containing both the patient names and the hospital's record of their ethnicity has enabled us to draw a number of conclusions. Each of these have important implications for health statistics.

The first is that patients whose ethnicity codes are missing are not representative of the entire patient population. In the NHS Trust area we looked at this results in an underestimate of the proportion of patients who are not white British.

A second conclusion is that ethnicity needs to be measured with much greater granularity than categories such as 'white other' allow. People of Finnish, Colombian, Iranian, Turkish, Albanian or Cypriot backgrounds are not a cohesive group. It is not obvious which of these groups experience particular vulnerability to COVID-19 and which do not. This issue extends to the question of which groupings should be included within statistics for BAME outcomes.

The third conclusion is that the size of BAME populations is significantly underestimated thanks to those with non-British names selecting 'British' or even 'white' as their primary identity. As things stand self-identification and ethnic heritage are not precisely aligned. Over time they will drift apart even further. It does not detract from the former to collect data about the latter. This is a factor that needs to be taken far more seriously – and one where empirical analysis should be undertaken to understand outcomes and behaviours.

These three issues compromise decision-makers' ability to establish an accurate picture of the relationship between health and ethnicity. They limit the confidence with which we can identify the extent and nature of the racial inequalities exposed by the pandemic. And they restrict the ability to respond in an evidence-based way when addressing issues relating to vaccinations and treatment.

Table 3 provides our best numerical estimate of the level of undercounting of BAME group in the NHS Trust area (NB: we have explained the methodology behind the deductions within the end-notes, one-by-one).

Challenge	Consequence in Trust area
'Not known' and 'Not stated' code	• 5,979 records of those that are not white British are missing, thanks to the over-representation of certain groups in the 'Not known' and 'Not stated' categories (this includes 3,059 names of BAME origin)[5]
The 'white other' code	• 2,018 BAME names are included as 'white other' – including 1,574 names of Arab, North African, Iranian or Turkish descent[6] • 'White other' also includes 15,099 names from backgrounds that might face similar discrimination/ economic barriers to BAME groups[7]
The 'British' code	• 43,287 records with non-White British names are classified as having 'British' ethnicity (this includes 9,135 names of BAME origin)[8]

Table 3: *Consequences of each issue for top-line data*

If the same issues are occurring across all NHS trusts then the levels of under-reporting could be very large indeed. For instance, the UK's adult population as a whole was coded in the same way then upwards of 400,000 non white British records could be being wrongly classified as White British, just thanks to the question of 'Not known' and 'Not stated' codes.[9] Well over 300,000 of these individuals are likely to have names with BAME heritage (to have an ancestry that is Caribbean, African or from Asia and the Middle East) – as these groups are the most likely to fall into unknown codes.[10]

Meanwhile, if we assume that those of North African, Iranian, Turkish and Arab descent are within the category that is intended by the BAME label, then we find further issues. At a national level there could be upward of 100,000 people within these BAME groups cast as 'white other'.[11]

Finally, if the issue of BAME groups featuring under the 'British' ethnicity code was occurring at a national scale then we could be looking at 1.3 million individuals having their ethnicity recorded wrongly, thanks to issues ranging from identity and trust to unclear forms or lack of confidence among staff.[12] If we expanded this to include all non White British names then the figure could double or even treble.

These estimates involve a number of assumptions. But taken together they could account for over 1.5 million individuals with BAME names being miscounted – if other NHS Trusts had the same biases in data collection methods. This is before we include the many other white groups not of British descent – such as those from Eastern Europe – who are also being undercounted.

These findings corroborate the idea that broad and inexact terms like BAME are wholly insufficient when looking at health outcomes. They also demonstrate how imprecise the present data collection methods often are.

In the future, improving the accuracy and reliability of ethnicity figures could have implications for many other NHS questions beyond COVID-19. It is relevant to the procurement of the appropriate medicines for patients, the provision of staff training about groups served, the representativeness of organisations' workforces compared to those they care for, and many other questions besides.

Endnotes

[1] Joint Committee on Vaccination and Immunisation: advice on priority groups for COVID-19 vaccination, December 2020.
[2] New poll finds ethnic minority groups less likely to want COVID vaccine, Royal Society for Public Health, December 2020.
[3] Commission on Race and Ethnic Disparities: The Report, Commission on Race and Ethnic Disparities, March 2021. P.32.
[4] In order to comply with GDPR regulations and current COPI protocols Origins codes are only held against anonymised records. In other words no inferred ethnicity data is held against recognisable individuals.
[5] 11.4867% of all names within our sample are not of White British origins. This means that, of the 57,251 names categorised as 'Not known' and 'Not stated', 6,576 'should' be of non White British descent – if the category were representative. Instead, 12,555 are. The difference between these two figures is **5,979**. (The BAME figure is worked out on the same basis. 3.9184% of all names in the sample have an ancestry that is Caribbean, African or from Asia and the Middle East, meaning that 2,243 of those with BAME names 'should' be filed as 'Not known' and 'Not stated' – if it were a representative cross-section. Instead, 5,302 appear in these categories – a difference of **3,059**).
[6] 1.62769186% of Anglo-Saxon and Celtic names are coded as 'white other', and we have used this as a yardstick for the proportion of names that are likely to be outliers or errors within any sample. If BAME groups were similarly likely to identify as 'white other', then we might expect 599 to do so (36,770 BAME names x 1.62769186%). Instead 2,617 do – a difference of 2,018. (The same deductions occur for the Turkish/ Iranian/ Arab/ North African sub-group, for which there 'should' be 172 (10,538 Turkish/ Iranian/ Arab/ North African names x 1.62769186%). Instead there are 1,746 – a difference of **1,574**).
[7] Names included as potentially facing similar barriers include: Hispanic and Portuguese, Jewish and Armenian, Greek and Greek Cypriot, and all East European groups. We have not adjusted this figure for anomalies/ errors, as 'white other' is often the most accurate category for those with these names. The question is not whether groups are disproportionately mis-categorised, so much as whether the category itself is too broad.
[8] 3.11% of British names (English, Scottish, Welsh and Northern Irish) are coded with an ethnicity other than British (or Irish) according to our findings. We might assume that a similar number of anomalies and errors would occur in the other direction – meaning 3,352 of those with non White British names being coded as British (107,789 non White British names x 3.11%). Instead there are 46,639 coded as having British ethnicity – a very large difference, of **43,287**. (The BAME figure is worked out on the same basis. We might assume that 1,144 of those with BAME names would be coded as British ethnicity (36,770 BAME names x 3.11%). Instead, 10,279 are – a difference of **9,135**).
[9] According to mid-2019 ONS population estimates, there are 45,756,018 people in the UK aged 18 and above. Origins analysis of the UK population finds that 18.7% are not of White British heritage (i.e. 8,556,375 adults) and that 81.3% are White British (37,195,067). Our analysis in this paper, meanwhile, revealed that 5.38% of those with White British names were coded as 'Not known' and 'Not stated', compared with 11.65% of those with non White British names. Were the ratios from our NHS trust to play out nationally, therefore 996,817 people with non White British names would be categorised as 'Not known' or 'Not stated' (8,556,375 adults with non White British names x 11.65%). Meanwhile, 2,001,095 of those with White British names would be coded as 'Not known' or 'Not stated' (37,195,067 with White British names x 5.38%). This would mean 2,997,912 names in total being coded as 'Not known' or 'Not stated'. Without knowing anything about the records within this grouping, it would have to be assumed that the nearly 3 million names in this category were representative of the UK as a whole – i.e. that 560,610 were not White British (2,997,912 unknown names x 18.7% non-White British). The difference between this number and the actual figure would be **436,207**.
[10] According to Origins, 10.49% of UK names have an ancestry that is Caribbean, African or from Asia and the Middle East – 4,799,806 people in total (45,756,018 residents x 10.49%). Our analysis finds that 14.41936361163992% of those with an ancestry that is Caribbean, African or from Asia and the Middle East were 'Not known' or 'Not stated'. If a similar ratio fell into these categories across the UK then 692,101 would fall into the two unknown codes (4,799,806 individuals with BAME heritage x 14.41936361163992%). However, the assumption made would have to be that 314,480 were of BAME heritage (2,997,912 unknown names, as deduced above, x 10.49% of the UK with names from the Caribbean, Africa, Asia or the Middle East). The difference between these two figures is **377,621**.
[11] Of the 45,756,018 people in the UK aged 18 and above, 2.07% have North African, Turkish, Iranian or Arab names according to Origins – 947,150 people. Our analysis of the Foundation Trust sample, meanwhile, found

that 16.56860884418296% of individuals in this group are coded as 'white other'. If this were extrapolated to the UK as a whole, we might assume that 156,929 of those in these four groups would be coded 'white other' (947,150 people with North African, Turkish, Iranian or Arab names x 16.56860884418296%). Of course, we might also assume that a small proportion of errors and outliers would occur with any group. Within our sample 1.62769186% of Anglo-Saxon and Celtic names are coded as 'white other', for example; the same figure, if applied to the North African, Turkish, Iranian and Arab sub-group, would mean 15,416 people with this heritage being classified as 'white other' thanks to errors, outliers etc (947,150 people with North African, Turkish, Iranian or Arab names x 1.62769186%). The difference between this and the actual figure is **141,513** (156,929 minus 15,416).

[12] 27.96% of patients with BAME names (i.e. those from is the Caribbean, Africa, Asia and the Middle East) were classified as ethnically 'British' within our sample. According to Origins, meanwhile, 10.49% of UK names have an ancestry that is Caribbean, African or from Asia and the Middle East – 4,799,806 people in total (45,756,018 residents x 10.49%). If 27.96% of these people code themselves as ethnically British then there would be **1,342,026** individuals of BAME heritage that have been undercounted. (It is worth noting that, if we were to apply this to all non White British names – rather than just those with BAME ancestry – then we would be looking at over 3 million mis-coded records).