BMJ open

Accuracy of routinely recorded ethnic group information compared with self-reported ethnicity: evidence from the English Cancer Patient Experience survey

C L Saunders, G A Abel, A El Turabi, F Ahmed, G Lyratzopoulos

To cite: Saunders CL, Abel GA, El Turabi A, *et al*. Accuracy of routinely recorded ethnic group information compared with self-reported ethnicity: evidence from the English Cancer Patient Experience survey. *BMJ Open* 2013;**3**: e002882. doi:10.1136/bmjopen-2013-002882

► Prepublication history and additional material for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2013-002882).

Received 14 March 2013
Revised 2 May 2013
Accepted 15 May 2013

Cambridge Centre for Health Services Research, University of Cambridge, Institute of Public Health, Cambridge, UK

Correspondence to
Dr Catherine Saunders;
ks659@medschl.cam.ac.uk

ABSTRACT

Objective: To describe the accuracy of ethnicity coding in contemporary National Health Service (NHS) hospital records compared with the 'gold standard' of self-reported ethnicity.

Design: Secondary analysis of data from a cross-sectional survey (2011).

Setting: All NHS hospitals in England providing cancer treatment.

Participants: 58 721 patients with cancer for whom ethnicity information (Office for National Statistics 2001 16-group classification) was available from self-reports (considered to represent the 'gold standard') and their hospital record.

Methods: We calculated the sensitivity and positive predictive value (PPV) of hospital record ethnicity. Further, we used a logistic regression model to explore independent predictors of discordance between recorded and self-reported ethnicity.

Results: Overall, 4.9% (4.7–5.1%) of people had their self-reported ethnic group incorrectly recorded in their hospital records. Recorded White British ethnicity had high sensitivity (97.8% (97.7–98.0%)) and PPV (98.1% (98.0–98.2%)) for self-reported White British ethnicity. Recorded ethnicity information for the 15 other ethnic groups was substantially less accurate with 41.2% (39.7–42.7%) incorrect. Recorded 'Mixed' ethnicity had low sensitivity (12–31%) and PPVs (12–42%). Recorded 'Indian', 'Chinese', 'Black-Caribbean' and 'Black African' ethnic groups had intermediate levels of sensitivity (65–80%) and PPV (80–89%, respectively). In multivariable analysis, belonging to an ethnic minority group was the only independent predictor of discordant ethnicity information. There was strong evidence that the degree of discordance of ethnicity information varied substantially between different hospitals ($p<0.0001$).

Discussion: Current levels of accuracy of ethnicity information in NHS hospital records support valid profiling of White/non-White ethnic differences. However, profiling of ethnic differences in process or outcome measures for specific minority groups may contain a substantial and variable degree of misclassification error. These considerations should be taken into account when interpreting ethnic variation audits based on routine data and inform initiatives aimed at improving the accuracy of ethnicity information in hospital records.

ARTICLE SUMMARY

Article focus

- Accurate recording of ethnicity in administrative health data is a pre-requisite for efforts to ensure equality or reduce inequalities in healthcare.
- This paper describes the accuracy of ethnicity coding in English National Health Service hospital records compared with the 'gold standard' of self-reported ethnicity, and identifies areas where improvement is needed.

Key messages

- Hospital records will usually code the ethnicity of patients with self-reported White British ethnic group correctly, but the levels of incorrect coding of the ethnicity of all ethnic minority groups are high.
- Belonging to an ethnic minority group is the only independent predictor of having an incorrectly coded hospital record ethnicity; there is substantial variation in the quality of coding between hospitals.
- The probability look-up tables provided in this paper can be used for weighting of incidence or prevalence estimates where hospital record ethnicity is being used, or in regression analysis to improve estimation of ethnic variation in processes or outcomes of care.

Strengths and limitations of this study

- This was a unique opportunity to carry out an audit of the accuracy of hospital record coding of ethnicity in England, using data from a large national survey of recently treated patients.
- We were not able to account for the different processes by which ethnicity was ascertained in hospital records and the study population was skewed towards older ages, which may limit the generalisability of the findings.

BACKGROUND

Modern healthcare systems aspire to equal access and quality of care for patients of any ethnic group.[1 2] In England, there are nevertheless well-documented ethnic inequalities in the processes and experiences of care.[3 4] Good practice in measuring these inequalities is needed.[5] Therefore, the availability of complete and valid information on patients' ethnicity in routine National Health Service (NHS) data is a fundamental first step to enable equality audits to inform improvement actions. Further, there are variations in the incidence and prevalence of some conditions between different patient ethnic groups. Understanding such variations can be vital for service planning and required capacity estimates. NHS hospitals began to routinely record ethnicity information in their Patient Administration System (PAS) records in the mid-1990s.[6] (PAS records are a precursor to Hospital Episode Statistics (HES) data—hereafter, we refer to hospital records as 'HES records' for simplicity, as this is the more commonly used convention to denote patient administrative data in the NHS.) Although implementation of this measure has been slow,[7] HES data in recent years have high completeness of ethnic group information (typically exceeding 90%).[6 8] There is, however, little evidence about the accuracy of routinely recorded information on patients' ethnicity. In the past, audits of the quality of ethnicity information in NHS records have principally focused on data completeness, rather than accuracy,[6] and completeness of ethnicity coding information currently forms part of the Commissioning Outcomes Data Set,[6] a prescribed standard for data quality that is linked to hospital reimbursement.

Ethnicity can be classified by referring to a community of people who share the same culture and/or by referring to an ancestral population which comprises their self-identity.[9] Self-reported ethnicity captures both the shared experiences/culture of an individual and their self-identity. Methods such as surname recognition algorithms or geocoding (or combinations of both these approaches) have been developed to indirectly infer the ethnicity of individuals,[10–15] but both have limitations that do not apply to self-reported ethnicity. For example, geocoding methods rely on high levels of geographical (residential) segregation between different ethnic groups. Surname recognition methods rely on low levels of interethnic marriages and the existence of distinctive surname nomenclatures (which do not always exist for some ethnic groups, eg, Black populations in the USA). Further, by definition, indirect methods will misclassify the ethnicity of some patients and disproportionately do so for the ethnicity of people from ethnic minorities. For all the above reasons, self-report is currently considered the gold standard measure of ethnicity.[16 17]

Patient surveys, when linked to routine ethnicity data, provide opportunities to determine the accuracy of ethnic group information contained in routine health system records. As this linkage is rarely performed, few studies have cross-examined the accuracy of ethnic group information included in routine healthcare records against self-reported ethnicity information. Those that have examined this have been conducted in the USA.[18–20]

Against this background and using data from an English national patient survey, we aimed to examine the overall accuracy of recorded versus self-reported ethnicity; and whether discordance between recorded and self-reported ethnicity information varied between patients of different ethnic groups.

METHODS

Data

We used publicly available anonymous data from the 2010 Cancer Patient Experience Survey in England.[21] An unusual feature of this survey is that hospital-recorded ethnicity was collected when compiling the sample and linked to patient responses by the survey provider. All patients treated for cancer in an English NHS hospital during the first quarter of 2010 were invited to participate in the survey, with a response rate of 67%.[21] Of 67 713 respondents, 64 418 (94.7%) provided valid self-reported ethnicity information. As self-reported ethnicity was used as the gold standard against which the accuracy of HES-recorded ethnicity was compared, data on respondents with missing self-reported ethnicity were excluded from further analysis (see figure 1). Data on an additional 348 respondents with missing information on deprivation status were also excluded from further analysis, leaving 63 770 respondents. An additional 5049 records with missing HES-recorded ethnicity were excluded, leaving 58 721 records for analysis. For all respondents included in our analyses, completely observed data were available for patients' HES-recorded age, gender and socioeconomic status (using the Index of Multiple Deprivation score of lower super output area of residence). In both

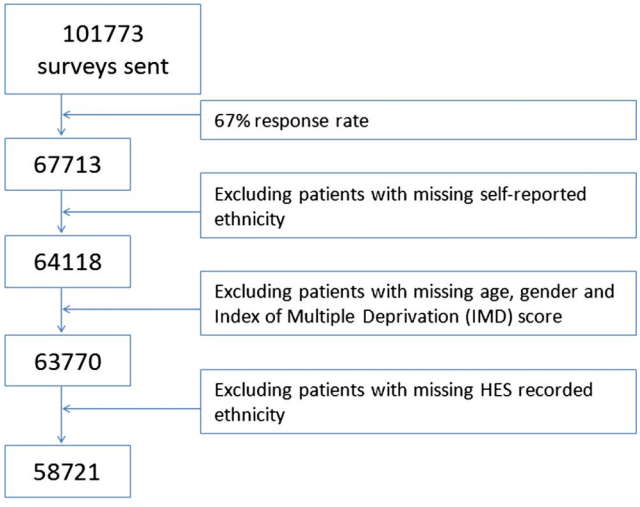

Figure 1 Survey responders and exclusions.

HES-records and patient reports, ethnicity was classified using the same 16-group categorisation (Office of National Statistics (ONS16) 2001; see online supplementary appendix 1).

Analysis

We first described the overall degree of discordance between HES-recorded and self-reported information on ethnicity using the Cohen's κ statistic. We further calculated the sensitivity and the positive predictive value (PPV) of HES-recorded ethnicity in respect of self-reported ethnicity. In this context, sensitivity denotes the proportion of patients with a given self-reported ethnicity with concordant ethnicity information in their HES record; and PPV denotes the probability that a patient with a particular HES-recorded ethnic group will self-report the same ethnic group.

Subsequently, we explored independent predictors of discordant ethnicity information by constructing a multivariable logistic regression model with concordant/discordant ethnicity status as the binary outcome variable and self-reported ethnicity as a covariate. Adjustment was also made for other patient sociodemographic characteristics (age in 10-year age groups, gender and postcode-linked area-based deprivation). For age and gender, we used HES-recorded information because the completeness of these variables among survey respondents was higher than self-reported age and gender, and as the degree of concordance between HES-recorded and self-reported age (based on year of birth) and gender were very high (>99.5% for both).

To explore whether clustering of patients in some groups in hospitals with higher or lower discordance levels could in part explain the findings, we subsequently repeated the regression model described above including a random effect for hospital—in addition, this model allows us to explore the degree of variation in the level of ethnicity information discordance between different hospitals. As less detailed classifications of ethnicity are often used in health research, we also carried out supplementary analysis looking at discordance when using six (as opposed to 16) ethnic groups (ie, White, Mixed, Asian or Asian British, Black or Black British, Chinese and other).

Lastly, we constructed a probability 'look-up' table indicating the probability that HES-recorded ethnicity represents true (self-reported) ethnicity for each of the 16 different ethnic groups. These probabilities can be used for weighting of incidence or prevalence estimates or used in regression analysis to improve estimation of ethnic variation in processes or outcomes of care.[14 22] For the calculation of the probability 'look-up' table only, we combined data from two surveys (2010 and 2011/2012) to improve the precision of the probabilities presented (increasing our sample size to 133 204). STATA V.11 was used for all analyses.

RESULTS

Overall, the frequency of discordance of HES-recorded and self-reported ethnicity information was 4.9% (4.7–5.1%). Patients who identified themselves as White British had the lowest frequency of discordant ethnicity information in their HES records (2.2% (2.0–2.3%)). In contrast, patients who identified themselves as belonging to any other ethnic group had a substantially higher frequency of discordant HES-recorded ethnicity (41.2% (39.7–42.7%)). Cohen's κ for HES-recorded and self-reported ethnicity was 0.64 overall and 0.54 if White British patients were excluded. The frequency of discordance was particularly high for patients who self-reported that they belonged to the Any Other Black Background (90.0% (55.5–99.7%)) and the Any Other Mixed Background (87.8% (78.2–97.3%)) groups (table 1).

HES-recorded 'White British' ethnicity had a high sensitivity of 97.8 (97.7–98.0) and PPV of 98.1 (98.0–98.2) for self-reported White British ethnicity (figure 2, estimates and CIs in online supplementary appendix table 2 for the 6 and 16 group classification). In contrast, HES-recorded 'Mixed' ethnicity had very low sensitivities (12–31%) and PPVs (12–42%). HES-recorded 'Indian', 'Pakistani', 'Bangladeshi', 'Chinese', 'Black-Caribbean' and 'Black African' ethnicity had intermediate levels of sensitivity (65–80%) and PPV (80–89%), respectively. HES-recorded 'White Irish' ethnicity had low sensitivity (47.8% (44.5–51.0%)) but high PPV (81.5% (77.9–84.6%)). This means that of all individuals who self-identify themselves as White Irish, only 48% would have their ethnic group recorded as such in their HES records; however, among patients whose HES records indicate that they are 'White Irish', 82% would identify themselves as belonging to this group too.

While there was some evidence that age, gender and deprivation are crudely associated with discordance of ethnicity information, in multivariable logistic regression analysis, adjusting for other patient characteristics, the sole independent predictor of discordance was self-reported ethnicity (table 1). Repeating this model with a random effect for hospital produced only trivial differences to associations of discordance with patient characteristics. This means that the association between discordance and self-reported ethnic minority group cannot be explained by ethnic minority patients attending hospitals that have poor levels of accuracy of ethnicity information overall. There was, however, strong evidence ($p<0.0001$) of variation in discordance between different hospitals (accuracy of coding of ethnicity across hospitals ranged from 67% to 100%). Specifically, if all hospitals were to be arranged in order of frequency of discordance of ethnicity information, and after accounting for differences in the proportion of ethnic minority patients attending the hospital, the OR of discordance between the hospital in the 97.5th and 2.5th centiles (the 95% reference range) will be about 13, indicating a 13-fold difference in the odds of

Table 1 Crude and adjusted predictors of discordant hospital record ethnicity coding

	Total respondents		Total discordant					
	n	Per cent	n	Per cent	Crude OR of discordance	Joint p value	Adjusted OR	Joint p value
Age								
16–24	399	0.7	30	7.5	1.77 (1.28–2.46)	<0.0001	0.90 (0.55–1.48)	0.54
25–34	950	1.6	93	9.8	2.37 (1.86–3.01)		0.94 (0.71–1.26)	
35–44	3139	5.3	231	7.4	1.73 (1.51–1.99)		1.02 (0.84–1.23)	
45–54	7763	13.2	484	6.2	1.45 (1.30–1.61)		1.12 (0.97–1.29)	
55–64	15 375	26.2	743	4.8	1.11 (0.99–1.23)		1.11 (0.98–1.25)	
65–74	18 366	31.3	805	4.4	reference		reference	
75–84	10 747	18.3	412	3.8	0.87 (0.76–0.99)		0.99 (0.86–1.14)	
85+	1982	3.4	80	4.0	0.92 (0.71–1.18)		1.08 (0.83–1.42)	
Gender								
Men	27 395	46.7	1268	4.6	reference	0.014	reference	0.69
Women	31 326	53.3	1610	5.1	1.12 (1.02–1.22)		1.02 (0.93–1.12)	
Deprivation								
Lowest	13 301	22.7	488	3.7	reference	<0.0001	reference	0.69
2	13 432	22.9	509	3.8	1.03 (0.91–1.18)		1.04 (0.90–1.2)	
3	12 498	21.3	559	4.5	1.23 (1.05–1.44)		0.99 (0.85–1.14)	
4	10 756	18.3	652	6.1	1.69 (1.36–2.11)		1.03 (0.89–1.2)	
Highest	8734	14.9	670	7.7	2.18 (1.68–2.84)		0.94 (0.80–1.11)	
Ethnic group								
British (White)	54 589	93.0	1177	2.2	reference	<0.0001	reference	<0.0001
Irish (White)	938	1.6	490	52.2	49.63 (32.72–75.28)		47.73 (41.00–55.55)	
Any other White background	1066	1.8	485	45.5	37.88 (24.95–57.52)		36.49 (31.52–42.25)	
White and Black Caribbean (Mixed)	72	0.1	50	69.4	103.14 (55.73–190.86)		109.97 (64.65–187.04)	
White and Black African (Mixed)	41	0.1	33	80.5	187.19 (90.23–388.35)		202.59 (90.28–454.64)	
White and Asian (Mixed)	77	0.1	65	84.4	245.81 (125.89–479.94)		283.62 (149.43–538.33)	
Any other Mixed background	49	0.1	43	87.8	325.22 (124.17–851.84)		399.68 (166.26–960.79)	
Indian (Asian or Asian British)	516	0.9	101	19.6	11.04 (7.16–17.04)		8.06 (6.33–10.27)	
Pakistani (Asian or Asian British)	206	0.4	46	22.3	13.05 (8.39–20.28)		10.26 (7.19–14.66)	
Bangladeshi (Asian or Asian British)	50	0.1	13	26.0	15.94 (7.44–34.18)		12.02 (6.17–23.41)	
Any other Asian background	129	0.2	57	44.2	35.93 (20.28–63.63)		30.42 (20.94–44.2)	
Caribbean (Black or Black British)	471	0.8	136	28.9	18.42 (12.64–26.85)		10.37 (8.21–13.10)	
African (Black or Black British)	319	0.5	111	34.8	24.22 (16.97–34.56)		15.54 (11.9–20.28)	
Any other Black background	10	0.0	9	90.0	408.42 (49.3–3383.53)		410.23 (49.73–3384.32)	
Chinese (other ethnic group)	124	0.2	30	24.2	14.48 (8.77–23.92)		10.66 (6.86–16.57)	
Any other ethnic group	64	0.1	32	50.0	45.38 (27.55–74.75)		34.04 (20.05–57.79)	
Hospital								
OR 95% reference range	158	–			30.48	<0.0001	12.85	<0.0001

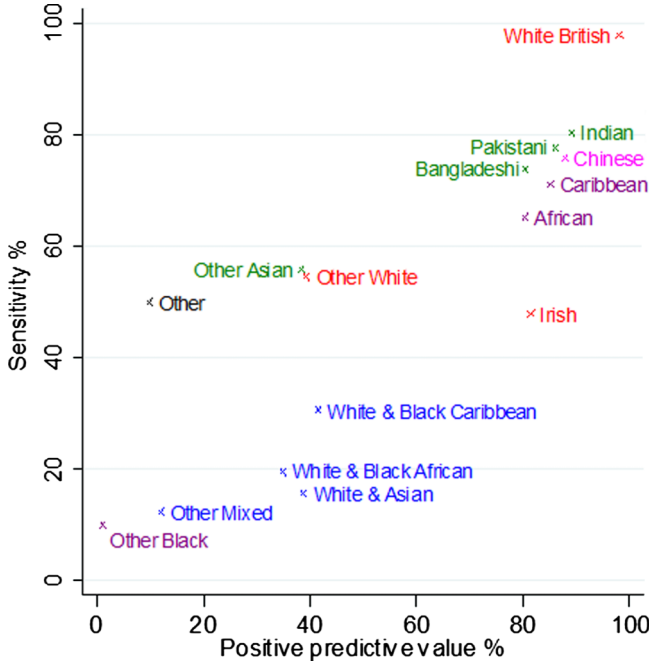

Figure 2 Sensitivity* and positive predictive value** of hospital record-recorded ethnicity compared with self-reported ethnicity as a gold standard. *If a patient self-reports that they belong to a particular ethnic group, then the sensitivity of the hospital record ethnicity coding is the probability that the hospital record will record the same (correct) ethnicity. **If a patient's hospital record states that they belong to a particular ethnic group, then the positive predictive value of the hospital record ethnicity code is the probability that this code has been recorded correctly and that the patient will self-report the same ethnicity.

discordance across hospitals. After accounting for differences in the proportion of ethnic minority patients attending each hospital, the hospital at the bottom 2.5th centile of coding accuracy had concordant self-reported and hospital record-recorded ethnicity codes in 90% of records.

Similar findings to those observed in the main analysis were observed when using a six-group classification (table 2). The total numbers of respondents with discordant ethnicity decreased when using a six-group classification to 796 participants (1.4%) from 2878 (4.9%), indicating that much of the discordance was within the cruder six-group classification. However, self-reporting a non-White ethnic background remained as strongly associated with having an incorrectly coded hospital-record ethnic group as for the 16-group classification. Details of the ethnic groups to which people from each self-reported ethnic background were misclassified are given in table 3, showing the variation between and within the 6-group and 16-group coding.

People who self-report mixed ethnic backgrounds are particularly likely to have an incorrectly coded ethnic group in their hospital records, with high numbers using both the 16 (79.9%) and 6 (74.9%) group classifications. Looking at table 3, we see that the self-reported ethnicity falls into one of three groups—the concordant mixed category, white (either British or Other white) or the corresponding ethnic minority category—over 90% of the time, explaining why for this group the broader six-category classification does not improve the accuracy of hospital record ethnicity coding.

Finally, the probability table (table 3) can be used in estimating ethnic group variations in incidence, prevalence or measures care quality when using HES data. Examples of such applications in the context of US health research have been reported previously.[14 22]

DISCUSSION

Using data from a recent national survey of hospital patients with cancer, we explored the accuracy of ethnicity information in HES data. Overall, we found that the level of discordance of ethnicity information between HES records and patient self-reports is low, particularly for the majority White British patients. There is, however, a substantial degree of inaccuracy in the recorded ethnicity of patients who self-report themselves as belonging to ethnic minority groups. For many major ethnic groups ('Indian', 'Pakistani', 'Bangladeshi', 'Chinese', 'Black-Caribbean' and 'Black African'), routine hospital data will miscode between 20% and 35% of all patients who self-report that they belong to these ethnic groups (sensitivity 65–80%). Further, up to 20% of patients recorded as belonging to some major ethnic groups will self-report that they actually belong to other ethnic groups (PPV 80–89%, respectively). For patients who self-report being of mixed ethnic groups, HES records are usually discordant. We provide probability tables that can be used in re-estimating ethnic group variations when investigating ethnic variation in processes or quality of care from the UK hospital records in order to improve the precision of such estimates.

The study explored a unique opportunity provided by patient survey data to explore the accuracy of ethnicity information in the UK-routine healthcare data. Previous evidence on the accuracy of ethnicity information in administrative datasets relates to US settings.[18–20] Our findings concord with previous US literature which also indicates that the accuracy of ethnicity information tends to be highest for the majority white population, lower for major ethnic minority groups (like African American or Hispanic) and lowest for smaller ethnic groups (such as American Indian/Alaskan Natives) and Mixed or Other racial/ethnic groups.[18 19] Other strengths of the study include its large sample size, enabling the profiling of discordance for small ethnic groups; and the use of regression analyses to explore independent predictors of discordance and to examine variation between different hospitals.

A limitation is that the study population included patients who attended hospital for cancer treatment (most of whom are aged 65 years of age or older). Therefore, in principle, the generalisability of the

Table 2 Crude and adjusted predictors of discordant hospital record ethnicity coding (ONS 6-group classification)

	Total	Per cent	N discordant (ONS6)	Per cent	Crude OR of discordance	Joint p value	Adjusted OR of discordance	Joint p value
Age								
16–24	399	0.7	17	4.3	4.15 (2.47–6.96)	<0.0001	0.87 (0.43–1.79)	0.10
25–34	950	1.6	37	3.9	3.78 (2.63–5.42)		1.23 (0.77–1.94)	
35–44	3139	5.3	83	2.6	2.53 (1.99–3.22)		1.32 (0.96–1.82)	
45–54	7763	13.2	166	2.1	2.04 (1.68–2.46)		1.26 (0.98–1.62)	
55–64	15 375	26.2	200	1.3	1.23 (1.05–1.44)		1.16 (0.93–1.46)	
65–74	18 366	31.3	195	1.1	reference		reference	
75–84	10 747	18.3	86	0.8	0.75 (0.58–0.98)		0.84 (0.63–1.12)	
85+	1982	3.4	12	0.6	0.57 (0.30–1.09)		0.73 (0.39–1.39)	
Gender								
Men	27 395	46.7	321	1.2	reference	0.0003	reference	0.49
Women	31 326	53.3	475	1.5	1.30 (1.13–1.49)		1.06 (0.90–1.26)	
Deprivation								
Lowest	13 301	22.7	119	0.9	reference	<0.0001	reference	0.30
2	13 432	22.9	120	0.9	1.00 (0.75–1.34)		1.13 (0.84–1.52)	
3	12 498	21.3	143	1.1	1.28 (0.92–1.78)		1.23 (0.92–1.64)	
4	10 756	18.3	191	1.8	2.00 (1.44–2.79)		1.36 (1.02–1.80)	
Highest	8734	14.9	223	2.6	2.9 (2.14–3.93)		1.27 (0.94–1.70)	
Ethnic group								
White	56 593	96.4	332	0.6	reference	<0.0001	reference	<0.0001
Mixed	239	0.4	179	74.9	505.56 (333.91–765.45)		564.17 (399.47–796.78)	
Asian	901	1.5	100	11.1	21.16 (14.50–30.88)		14.30 (10.97–18.65)	
Black	800	1.4	123	15.4	30.79 (19.65–48.23)		17.51 (13.35–22.96)	
Chinese	124	0.2	30	24.2	54.08 (32.35–90.42)		38.62 (24.39–61.14)	
Other	64	0.1	32	50.0	169.46 (103.63–277.12)		109.64 (63.49–189.34)	
Hospital								
OR 95% reference range	158	–		73.97		<0.0001	22.96	<0.0001

ONS, Office for National Statistics.

findings (particularly regarding younger patients) might be limited; equally, the need for accurate recording of ethnicity among cancer patients has been identified.[23] We were also unable to explore potential inaccuracies in ethnicity categorisation resulting from longitudinal person-level discordance among patients with more than one hospital care episode. Previous research indicates that up to 3% of patients with more than one episode of care have longitudinally discordant ethnic group information.[24] Evolving sociocultural trends or changes in Census methodology (such as the introduction of Mixed ethnic groups to the UK census in 2001) could contribute to changes in a person's self-identified ethnic group over their lifetime. Lastly, we were not able to account for the process by which ethnicity was ascertained in hospital records. It is quite likely that this process involves a combination of self-reports, reports by relatives or carers (eg, in the case of infirm patients, or patients with language or other communication difficulties) or even guesswork by hospital staff (eg, when there are language barriers, or in clinical emergencies).[16]

The 2001 Census has expanded the previous ONS classification of ethnicity (originally containing 10 groups) to the 16-group classification also used in this survey,[2] a change which was subsequently reflected into HES coding. The impact of this secular change may explain some of the discordance observed. For example, the discordance in self-reported Irish White ethnicity may reflect the fact that this ethnic group was only included in the 2001 classification. The 2011 Census included two new ethnic groups; 'Gypsy or Irish Traveller' and 'Arab'.[25] It remains to be seen if this change will be reflected into the routinely collected health data.

While this study considers the accuracy of recorded ethnicity, there are issues with defining ethnicity using any simple classification; for example, the potential for 'concealed heterogeneity' within each of the ethnic groups.[16] It is possible, for example, that within the Indian ethnic group there are certain social, linguistic or religious subgroups which are more or less likely to have a discordant ethnicity. Indeed, evidence indicates that ethnic minority patients with discordant ethnicity information may be systematically different from other patients from the same ethnic group with concordant ethnicity.[20] Within-ethnic group heterogeneity is a complex issue inherent in any type of ethnicity research, and the nature of our study did not allow us to address this.

Table 3 The probability (%) of 'true' ethnic group for each hospital record-recorded ethnic group (ie, the probability that a given person with a specific HES-recorded ethnicity belongs to any (self-reported) group

	Probability (%) of true self-reported ethnic group for each HES code															
HES code	**British (White)**	**Irish (White)**	**Any other White background**	**White and Black Caribbean (Mixed)**	**White and Black African (Mixed)**	**White and Asian (Mixed)**	**Any other Mixed background**	**Indian (Asian or Asian British)**	**Pakistani (Asian or Asian British)**	**Bangladeshi (Asian or Asian British)**	**Any other Asian background**	**Caribbean (Black or Black British)**	**African (Black or Black British)**	**Any other Black background**	**Chinese (other ethnic group)**	**Any other ethnic group**
A=British (White)	98.1	0.9	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
B=Irish (White)	20.1	78.7	1.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C=Any other White background	53.1	1.2	42.1	0.2	0.1	0.5	0.7	0.2	0.0	0.0	0.5	0.1	0.3	0.1	0.2	0.8
D=White and Black Caribbean (Mixed)	11.2	0.0	2.8	38.3	0.0	0.0	2.8	0.0	0.0	0.0	0.0	41.1	2.8	0.9	0.0	0.0
E=White and Black African (Mixed)	12.2	0.0	6.1	4.1	26.5	0.0	8.2	0.0	0.0	0.0	0.0	8.2	34.7	0.0	0.0	0.0
F=White and Asian (Mixed)	25.0	0.0	2.8	0.0	0.0	40.3	2.8	8.3	1.4	1.4	13.9	1.4	0.0	0.0	1.4	1.4
G=Any other Mixed background	30.6	0.9	10.8	4.5	4.5	7.2	12.6	4.5	0.9	0.9	3.6	5.4	2.7	0.9	4.5	5.4
H=Indian (Asian or Asian British)	1.9	0.0	0.4	0.0	0.1	0.6	0.1	88.4	2.8	0.4	4.6	0.2	0.1	0.1	0.0	0.1
J=Pakistani (Asian or Asian British)	2.3	0.0	0.8	0.0	0.0	0.5	0.3	5.9	87.7	1.0	1.0	0.0	0.0	0.0	0.0	0.5
K=Bangladeshi (Asian or Asian British)	2.1	0.0	0.0	0.0	0.0	1.1	1.1	5.3	2.1	82.1	4.2	0.0	1.1	0.0	1.1	0.0
L=Any other Asian background	4.4	0.0	4.2	0.2	0.5	3.7	3.0	21.5	7.6	0.9	41.0	0.5	0.2	0.7	3.5	8.1
M=Caribbean (Black or Black British)	4.2	0.0	0.1	3.6	0.5	0.0	0.6	0.0	0.0	0.0	0.4	84.8	3.6	2.0	0.0	0.1
N=African (Black or Black British)	5.0	0.2	1.4	0.6	2.6	0.0	0.6	0.0	0.0	0.0	0.2	6.6	81.3	1.4	0.0	0.2
P=Any other Black background	3.1	0.0	0.9	3.6	2.7	0.9	1.8	0.4	0.0	0.0	1.3	38.1	42.6	4.0	0.4	0.0
R=Chinese (other ethnic group)	4.2	0.0	0.8	0.0	0.0	0.8	0.0	0.4	0.0	0.0	4.2	0.0	0.0	0.0	83.2	6.3
S=Any other ethnic group	43.5	0.6	17.3	0.4	0.9	1.6	1.8	3.5	1.1	0.6	9.1	2.5	3.4	0.9	1.6	11.3
Z=Not stated	90.6	1.5	2.8	0.1	0.1	0.3	0.2	0.9	0.5	0.0	0.7	1.0	0.6	0.1	0.3	0.3
X or Missing	91.3	1.3	2.5	0.1	0.1	0.2	0.2	1.3	0.4	0.1	0.5	0.7	0.7	0.1	0.4	0.2

HES, Hospital Episode Statistics.

Another issue that we have not addressed in this paper is the completeness of ethnicity information, which has the potential to bias estimates of ethnic variation in addition to the misclassification detailed here. We performed similar analysis to those discussed here for predictors of missing ethnicity (not shown). We found only small variations by self-reported ethnic group but a much larger degree of variation between hospitals than that seen for discordance, implying that the issues with missing ethnicity are primarily driven by hospital-level processes.

Our findings have implications for policy and future research. First, although HES data currently have a high level of completeness of ethnic information, if the aim of any audit is to compare outcomes between White and non-White groups, the current classification system will perform well, but a degree of caution is required when interpreting more detailed evidence on ethnic inequalities in care quality or disease incidence and prevalence that is based solely on HES data, particularly for minority ethnic groups found to have higher discordance rates. Misclassification of ethnicity in HES data could result in either an underestimation of ethnic variation or an inability to detect such variation when it exists ('type 1' error).

Second, we provide a list of probabilities that can be used to improve estimates of ethnic variation in healthcare in a UK setting.[14 22] These probabilities can be used both to improve the precision of prevalence estimates and in statistical models where hospital record ethnic group is a predictor.[14]

Third, as the completeness of ethnicity information in hospital records is currently high, more attention needs to be given to the accuracy of recorded information. The substantial variation between hospitals in the accuracy of ethnicity information indicates that there is great potential for improving the quality of ethnicity information in poorly performing hospitals. Improvement in the quality of HES data (which is generally desirable[26]) should also encompass improvements in the quality of ethnicity coding. Qualitative studies have found a willingness among ethnic minority groups to provide this information,[27] and future research should explore optimal ways for efficiently obtaining current self-reported information on ethnicity in patient records.

Acknowledgements We thank the UK Data Archive for access to the anonymous survey data (UKDA study number: 69488), the Department of Health as the depositor and principal investigator of the Cancer Patient Experience Survey 2010, Quality Health as the data collector; and all NHS Acute Trusts in England for provision of data samples.

Contributors All authors had substantial input into the study concept, design, analysis and access to the data. GL had the original idea for the study, which was further conceptualised in discussions between all authors; methods were principally developed by CLS and further amplified by GAA. All authors communicated frequently during the course of the project, including through face-to-face meetings. CLS and GL wrote the first draft, which was edited by all authors over multiple versions. All authors saw and approved the final manuscript.

Funding The paper is independent research arising from a Post-Doctoral Fellowship award to GL supported by the National Institute for Health Research (PDF-2011-04-047). AET is funded by an Academic Clinical Fellowship award from the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. US Department of Health and Human Services Agency for Healthcare Research and Quality. *2009 National Healthcare Disparities and Quality Reports*. 2009. http://www.ahrq.gov/qual/qrdr09.htm (accessed 29 Nov 2012).
2. Department of Health. *Equity and excellence: liberating the NHS* (White Paper). 2010. http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_117353 (accessed 29 Nov 2012).
3. Cancer Incidence and Survival By Major Ethnic Group, England, 2002–2006. 2009. http://www.ncin.org.uk/publications/reports/default.aspx
4. Mindell J, Klodawski E, Fitzpatrick J. Using routine data to measure ethnic differentials in access to coronary revascularization. *J Public Health* 2008;30:45–53.
5. Mir G, Salway S, Kai J, *et al*. Principles for research on ethnicity and health: the Leeds Consensus Statement. *Eur J Public Health* 2013;23:504–10.
6. The NHS Information Centre. How good is HES ethnic coding and where do the problems lie? 2011. http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=171 (accessed 7 Feb 2013).
7. Szczepura A. Access to health care for ethnic minority populations. *Postgrad Med J* 2005;81:141–7.
8. Jack RH, Linklater KM, Hofman D, *et al*. Ethnicity coding in a regional cancer registry and in Hospital Episode Statistics. *BMC Public Health* 2006;6:281.
9. Isajiw WW. Definition and Dimensions of Ethnicity: a theoretical framework. In: *Challenges of Measuring an Ethnic World: Science, politics and reality: Proceedings of the Joint Canada-United States Conference on the Measurement of Ethnicity April 1–3, 1992*. Washington, DC: U.S. Government Printing Office, 407–27.
10. Lyratzopoulos G, McElduff P, Heller RF, *et al*. Comparative levels and time trends in blood pressure, total cholesterol, body mass index and smoking among Caucasian and South-Asian participants of a UK primary-care based cardiovascular risk factor screening programme. *BMC Public Health* 2005;5:125.
11. Shah BR, Chiu M, Amin S, *et al*. Surname lists to identify South Asian and Chinese ethnicity from secondary data in Ontario, Canada: a validation study. *BMC Med Res Methodol* 2010;10:42.
12. Maringe C, Mangtani P, Rachet B, *et al*. Cancer incidence in South Asian migrants to England, 1986–2004: Unraveling ethnic from socioeconomic differentials. *Int J Cancer* 2013;132:1886–94.
13. Quan H, Wang F, Schopflocher D, *et al*. Development and validation of a surname list to define Chinese ethnicity. *MedCare* 2006;44:328–33.
14. Elliott MN, Fremont A, Morrison PA, *et al*. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Serv Res* 2008;43:1722–36.
15. Nitsch D, Kadalayil L, Mangtani P, *et al*. Validation and utility of a computerized South Asian names and group recognition algorithm in ascertaining South Asian ethnicity in the national renal registry. *QJM* 2009;102:865–72.
16. Aspinall PJ. The utility and validity for public health of ethnicity categorization in the 1991, 2001 and 2011 British Censuses. *Public Health* 2011;125:680–7.
17. Mays VM, Ponce NA, Washington DL, *et al*. Classification of race and ethnicity: implications for public health. *Annu Rev Public Health* 2003;24:83–110.
18. Arday SL, Arday DR, Monroe S, *et al*. HCFA's racial and ethnic data: current accuracy and recent improvements. *Health Care Financ Rev* 2000;21:107–16.
19. Lee LM, Lehman JS, Bindman AB, *et al*. Validation of race/ethnicity and transmission mode in the US HIV/AIDS reporting system. *Am JPublicHealth* 2003;93:914–17.

20. Zaslavsky AM, Ayanian JZ, Zaborski LB. The validity of race and ethnicity in enrollment data for Medicare beneficiaries. *Health Serv Res* 2012;47:1300–21.
21. Department of Health. National Cancer Patient Experience Survey 2010 [computer file]. UKDA-SN-6742-1. Colchester, Essex: UK Data Archive [distributor]. http://dx.doi.org/10.5255/UKDA-SN-6742-1
22. McCaffrey DF, Elliott MN. Power of tests for a dichotomous independent variable measured with error. *Health Serv Res* 2008;43:1085–101.
23. Iqbal G, Gumber A, Szczepura A, *et al*. Improving ethnic data collection for statistics of cancer incidence, management, mortality and survival in the UK. http://www2.warwick.ac.uk/fac/med/research/csri/ethnicityhealth/research/crc.pdf (accessed 25 Apr 2013).
24. Georghiou T, Thorlby R. Quality of ethnicity coding in Hospital Episode Statistics: beyond completeness. Paper presented at Healthcare Commission seminar 'Measuring what matters: measuring ethnicity in health data' 12/02/2007. 2007.
25. Office of National Statistics. 2011 Census. http://www.ons.gov.uk/ons/guide-method/census/2011/index.html (accessed 8 Mar 2013).
26. Spencer SA, Davies MP. Hospital episode statistics: improving the quality and value of hospital data: a national internet e-survey of hospital consultants. *BMJ Open* 2012;2:e001651.
27. Iqbal G, Johnson MR, Szczepura A, *et al*. UK ethnicity data collection for healthcare statistics: the South Asian perspective. *BMC Public Health* 2012;12:243.