Centre for Epidemiology & Biostatistics
Leeds Institute of Genetics, Health & Therapeutics

UNIVERSITY OF LEEDS

Investigating the association between ethnicity and survival from breast cancer using routinely collected health data: Challenges and potential solutions

A Downing[1], D Forman[1,2], RM West[1], J Thomas[1], G Lawrence[3], MS Gilthorpe[1]

[1]Centre for Epidemiology & Biostatistics, University of Leeds, UK; [2]Northern & Yorkshire Cancer Registry & Information Service, Leeds, UK;
[3]West Midlands Cancer Intelligence Unit, Birmingham, UK;

Background

- Information on ethnicity can be collected from surveys and face-to-face interviews, extracted using name recognition software or recorded in routine data, such as Hospital Episode Statistics (HES) data.
- Completeness of ethnicity recording in HES has improved over time, but it is not yet collected for 100% of patients and it is not known whether this information is more likely to be missing for some ethnic groups than for others.
- Patients with multiple hospital visits can have multiple ethnicities recorded. Methods of assigning patients a single ethnicity include using the 'most popular' or 'last recorded' ethnicity, but such methods do not make use of all of the available information. Another would be to use all of the patients' information and calculate proportions for each ethnicity for each patient.
- In addition imputation can be used to assign ethnicities to those episodes where the information is missing.

Aim

This study aims to investigate different methods of assigning ethnicity in order to assess the relationship between ethnicity and survival from breast cancer, using a registry-HES linked dataset for two English regions.

Data & methods

- Cases of female invasive breast cancer diagnosed in the Northern/Yorkshire & West Midlands regions during the period 01/01/1997-31/12/2003 were identified and linked to Hospital Episode Statistics (HES) data (*n=48,234*).
- Where multiple ethnicities were recorded for a patient (35% of cases), a single ethnicity was allocated according to the 'most popular' (highest frequency) and 'last recorded' (most recent) major ethnic group code.
- In addition, the data were expanded to include all available hospital episodes (and all ethnicity information) for each patient (452,061 'episode-level' records). Ethnicity proportions were then calculated for each patient.
- Due to small numbers in some of the ethnic groups the following categories were used; White, Asian, Black, Other (Mixed/Chinese/Other ethnic group combined). See Table 1 for the numbers in each group.
- Ethnicity was missing in 16.4% of the 'most popular' and 16.3% of the 'last recorded' patient-level records and 25.9% of the episode-level records.
- Multiple imputation (10 iterations) was undertaken using the ICE command in Stata with age, stage, IMD income domain and a census area measure of ethnicity (% White residents by super output area) as predictors.
- Stage was missing in 13.5% of cases and these data were also imputed.
- Survival analysis (with follow-up to 31/12/2006) was carried out using the imputed datasets (using the MIM command).

Table 1: Number and percentage of patients by ethnic group

	Most popular		Last recorded	
Major grouping	**N**	**%**	**N**	**%**
White	39,213	81.3	39,214	81.3
Asian	621	1.3	633	1.3
Black	306	0.6	326	0.7
Mixed	29	0.06	35	0.07
Chinese	34	0.07	35	0.07
Other	102	0.2	114	0.2
Unknown	7,929	16.4	7,877	16.3
Total	**48,234**	**100**	**48,234**	**100**

Results

White women were slightly older at diagnosis than the other groups (Figure 1), whilst Asian women had a higher proportion of early stage tumours, but these differences were not significant.

Figure 1: Box plot showing the age distribution of the patients by ethnic group

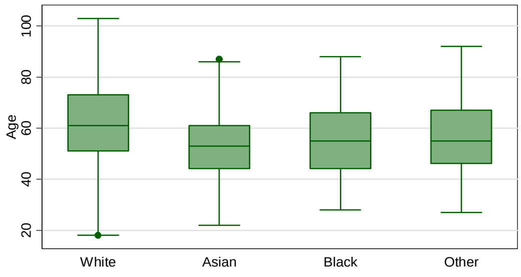

Table 2: Pre-imputation survival analysis results using the 'most popular' ethnicity

	Unadjusted		Adjusted	
	HR	**95% CI**	**HR**	**95% CI**
White	1.00		1.00	
Asian	0.83	0.69-1.00	0.91	0.76-1.09
Black	0.93	0.72-1.19	1.05	0.82-1.35
Other	0.78	0.53-1.14	0.82	0.56-1.20
Missing	1.04	0.99-1.10	1.07	1.02-1.13

- In the unadjusted analysis, the Asian group had improved survival compared to the White group (borderline significance), whilst there were no differences for the Black and Other groups (Table 2).
- After adjustment for age and stage, there were no significant differences in survival amongst the ethnic groups.
- The Missing group had worse survival compared to the White group and this became more marked after adjustment.
- This pattern was repeated when using the last recorded and episode-level data.

Table 3: Post-imputation survival analysis results (adjusted)

	Last recorded		Most popular		Episode-level	
	HR	**95% CI**	**HR**	**95% CI**	**HR**	**95% CI**
White	1.00		1.00		1.00	
Asian	0.98	0.82-1.16	0.96	0.81-1.15	0.99	0.82-1.19
Black	1.03	0.84-1.27	1.01	0.79-1.27	1.14	0.87-1.51
Other	0.74	0.51-1.06	0.87	0.61-1.25	0.87	0.58-1.29

- After imputation of the missing ethnicities, the results followed a similar pattern to those pre-imputation; the Asian group had improved survival in the unadjusted analyses but no differences were seen after adjustment for age and stage.
- The results were very similar for all three datasets (Table 3).

Table 4: Pre- and post-imputation ethnic group distribution (%)

	Last recorded		Most popular		Episode-level	
	Pre-imp	**Post-imp**	**Pre-imp**	**Post-imp**	**Pre-imp**	**Post-imp**
White	97.2	97.3	97.3	97.4	95.9	96.1
Asian	1.6	1.5	1.5	1.5	2.4	2.3
Black	0.8	0.8	0.8	0.7	1.3	1.2
Other	0.5	0.5	0.4	0.4	0.4	0.4

- Before and after imputation the proportions of the population in each ethnic group were very similar, suggesting that the missing ethnicities were not biased towards a particular group (Table 4).
- Approximately 97% of the population were White, 1.5% were Asian, 0.8% were Black and 0.5% were made up of other ethnic groups.
- The figures for the episode-level data were slightly different, with 96% White, 2.3% Asian, 1.2% Black and 0.4% Other.
- The analyses were repeated after 20 and 50 imputation iterations and the ethnic group distributions were very similar to those presented above.

Conclusions

- Our results show that whilst there is improved survival for Asian breast cancer patients in the unadjusted results, this survival difference disappears after adjustment for age and stage.
- The results are similar before and after imputation of the missing ethnicity information and for the three methods of assigning ethnicity.
- The missing data appears to be relatively evenly spread across the four ethnic groups; however, the worse survival for the Missing group requires further investigation.
- Ideally, we would like to look at the survival differences and patterns of missingness for the minor census ethnic groups (e.g. Asian split in to Indian, Pakistani & Bangladeshi) but the numbers are too small to give reliable results.
- The results should be validated in other regions of the UK.
- Assessment of the association between cancer survival and ethnicity presents many challenges. Failure to address the issues of missing data and multiple ethnicities may lead to biased results.

CANCER RESEARCH UK