ORIGINAL ARTICLE

Why poor quality of ethnicity data should not preclude its use for identifying disparities in health and healthcare

Peter J Aspinall, Bobbie Jacobson

Qual Saf Health Care 2007;**16**:176–180. doi: 10.1136/qshc.2006.019059

See end of article for authors' affiliations

Correspondence to: Dr P J Aspinall, London Health Observatory, 4th floor, Southside, 105 Victoria St, London, SW1E 6DT, UK; p.j.aspinall@kent.ac.uk

Accepted 1 March 2007

Background: Data of quality are needed to identify ethnic disparities in health and healthcare and to meet the challenges in governance of race relations. Yet concerns over completeness, accuracy and timeliness have been long-standing and inhibitive with respect to the analytical use of the data.
Aims: To identify incompleteness of ethnicity data across routine health and healthcare datasets and to investigate the utility of analytical strategies for using data that is of suboptimal quality.
Methods: An analysis by government office regions of ethnicity data incompleteness in routine datasets and a comprehensive review and evaluation of the literature on appropriate analytical strategies to address the use of such data.
Results: There is only limited availability of ethnically coded routine datasets on health and healthcare, with substantial variability in valid ethnic coding: although a few have high levels of completeness, the majority are poor (notably hospital episode statistics, drug treatment data and non-medical workforce). In addition, there is also a more than twofold regional difference in quality. Organisational factors seem to be the main contributor to the differentials, and these are amenable—yet, in practice, difficult—to change. This article discusses the strengths and limitations of a variety of analytical strategies for using data of suboptimal quality and explores how they may answer important unresolved questions in relation to ethnic inequalities.
Conclusions: Only by using the data, even when of suboptimal quality, and remaining close to it can healthcare organisations drive up quality.

Concern over the quality of ethnicity data, especially its completeness, has been long-standing, and has significantly contributed to the widespread lack of use of the substantial volumes of data now collected to identify ethnic disparities in health and healthcare.[1] However, the pursuit of equality agendas as a matter of governance has created an imperative to analyse the data. A recent compilation of public health indicators across the English government office regions[2] identified just 10 ethnically coded routine datasets that could be exploited (table 1), with a more than twofold overall regional difference in completeness. However, analysis of hospital episode statistics (HES) data suggests no structural (but the possibility of unmeasured) impediments to the achievement of high levels of ethnic coding: indeed, more than a quarter of National Health Service (NHS) trusts have achieved ⩾90% completeness in 2003/4, regardless of size[3] (fig 1).

ANALYTICAL STRATEGIES FOR USING SPARSE AND INCOMPLETE ETHNICITY DATA

When ethnicity data are of suboptimal quality because of incomplete coding or other quality problems, there are strategies that can be adopted to make use of it. They include methods such as donor imputation, record linkage and the use of distinctive naming algorithms for assigning—with varying degrees of success—an ethnic group to those records where it is missing (called item non-response) in population-based survey and administrative data. By so populating such records with ethnic group, analysis can be conducted on a greater number, thereby addressing possible non-response bias. Some of these methods can also be used when there is non-response to the survey itself (case non-response), if a subsequent coverage survey has been undertaken. A second set of strategies involve analytical methods that enable data to be used when the numerator is incomplete and/or compatible denominators are not available (proportional mortality ratios (PMRs)), or a clear picture to emerge from studies which, when examined alone, tend to be inconclusive in their findings (meta-analysis). An attempt is made to assess the scope of these approaches for answering important and currently unresolved questions in relation to ethnic inequalities.

DONOR IMPUTATION

Donor imputation is a method that can be used when a survey respondent does not answer a background question on a characteristic such as religion or ethnic group. In such instances, that characteristic is imputed from that of someone (the donor) for whom this has been given, who is geographically close (a nearest neighbour) and who matches the non-respondent on other selected characteristics. The judgement is that it is highly probable that the non-respondent will have the same characteristic as that of the donor, owing to their similarity. Clearly, imputation is an unreliable method when the sample size in surveys is small. Moreover, using the Office for National Statistics (ONS) Longitudinal Study (LS), research on the imputation of the ethnic group item in the 2001 census—that is, analysis of 1991 minority ethnic groups with an imputed ethnicity in 2001—showed it to be an unreliable approach for minority ethnic groups.[4] Of those from 1991 minority ethnic groups with ethnicity imputed in their 2001 census record, less than half were imputed to the same ethnicity as they used in their 1991 census response. Although 97.5% in the White ethnic group in 1991 were imputed to the same group, this fell to around 49% for Indians and Pakistanis, 37.5% for Bangladeshis, 29.9% for Black Caribbeans, 28.3% for Black Africans and just 10.0% for Chinese. The investigators concluded that census imputation

Abbreviations: HES, hospital episode statistics; LS, longitudinal study; NHS, National Health Service; ONS, Office for National Statistics; PM, proportional mortality; PMR, proportional mortality ratio

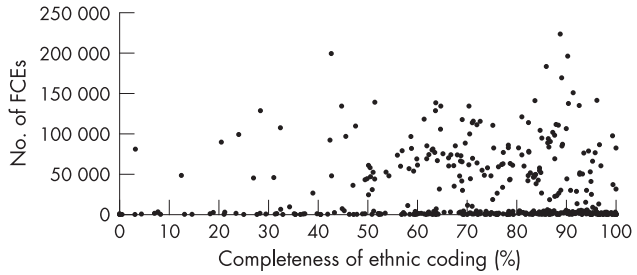

Figure 1 Completeness of ethnic coding, England, National Health Service (NHS) trusts, 2003/4. Source: Department of Health, Hospital Episode Statistics database. Based on 416 primary care trusts and hospital inpatient NHS trusts with ⩾100 finished consultant episodes (FCEs; excluding 65 trusts with ⩽100 FCEs). Overall completeness: 72.4%, based on 12 710 591 records; 59 trusts (14.2%) with ⩾95% completeness; 108 trusts (26.0%) with ⩾90% completeness.

of the ethnic group introduced inconsistency in the recording of the item, and that studies of individual-level LS data with 2001 ethnic group as an independent variable should consider omitting such cases. This method would therefore seem to have limitations in an ethnicity context.

RECORD LINKAGE: DATA FUSION AND OTHER METHODS

Record linkage can provide a more reliable method of restoring ethnic group when it is missing—so as to mitigate the effects of possible non-response bias—or of supplying it on a dataset when it is not collected. It can also be used to enrich data and supply additional variables, to remove duplicates from files, and that to establish the accuracy of data items common to the linked datasets. The Oxford Record Linkage Study[5] pioneered work on medical record linkage in the UK on the basis of discriminating personal identifying variables. The scope for record linkage has now increased substantially with the addition of unique identifiers—such as the NHS Number and National Insurance Number—to administrative datasets.

Several methods for matching records are available.[6] For deterministic matching, a unique number (such as the NHS Number) is needed on both files to be matched. However, such exact matching (data fusion) is frequently not possible, as the records may not contain such high-quality identifiers—especially when survey records are linked to administrative databases—or they may not be accessible, for reasons of confidentiality. A probabilistic matching method is then used, by which number of identifying variables from the two datasets inform the judgement whether the records are from the same person. Such algorithms often encompass variables such as name, address (or postcode or address coding), sex and date of birth. This method frequently leads to a combination of true matches, non-matches and possible matches, the last requiring resolution through clerical intervention or some rule-based computerised method.

Several examples are available from linkage studies of administrative records to surveys, or to the same or other administrative records where one or more datasets are ethnically coded. Perhaps the best known is the ONS LS, in which data from the 1971–2001 censuses have been linked together, along with information on events such as births, deaths and cancer registrations, for 1% of the population of England and Wales. The availability of ethnic group in the 1991 and 2001 censuses for sample members has enabled indices of stability and change in ethnic group assignment at the individual level to be systematically investigated for the first time.[4] Furthermore, as data accrue, the LS will be of increasing value as a source of data on mortality by ethnic group, currently largely unreported, as death registration records country of birth (of interest, of course, in its own right) rather than ethnic group. As an example of linked administrative data, the Pupil Level Annual School Census, which collects data on pupil ethnic group (as well as on other demographic domains, free school meals entitlement, special educational needs and others), has been linked with the National Pupil Database through the unique pupil number, enabling educational attainment to be stratified by entitlement to free school meals.[2] An ONS pilot study into the feasibility of linking maternity records from the HES system with the corresponding birth records held by ONS established that it was possible to achieve a 99% linkage using an iterative process of matching.[7] As the ONS record contains mother's country of birth and the HES record contains mother's ethnic group, this exercise enabled the ethnic group of mothers in the sample to be broken down by countries of birth. However, the analysis was incomplete, as only 50.5% of the matched-HES records had an ethnic group recorded. As ethnic coding on HES improves, such linkage would provide important information on fertility rates by ethnic group and improved statistics on parity and gestation. The Department of Health has commissioned the University of Oxford's Unit of Health-Care Epidemiology to prepare a national record linked file of HES and mortality data for England from 1998/9 to the present, and to analyse the national HES data linked as a person-based dataset.

An increasing number of surveys that collect data on ethnic group also now use record linkage methods. In the ethnically coded Millennium Cohort Study (relating to children born in 2000–1), details of mothers and children have been linked with birth registration and HES. The Avon Longitudinal Study of Parents and Children, which collected information on the ethnic origin of the mother but has a shortfall in ethnic minority mothers, makes extensive use of additional information from administrative data sources. Agreement has been obtained to link the survey records of the ethnically coded English Longitudinal Study of Ageing to a variety of health and administrative records, including HES, and Inland Revenue and Department for Work and Pensions records. However, few examples are available from the UK for the use of record linkage to specifically populate a database with ethnic group or to validate this field.

With respect to the former, perhaps the best known recent example of record linkage to address the paucity of ethnic coding in administrative records has been that undertaken in Scotland: the linking of the ethnic code in the Scottish 2001 Census to the Scottish NHS Community Health Index number as a unique identifier, using the linkage variables of names, dates of birth and addresses.[8] Overall, 94% of census records were matched (more than 85% in minority ethnic groups). Outcomes on the SMR01 database (the Scottish morbidity record linked to mortality) were linked to the census data via the community health index number, thereby creating an ethnically coded healthcare database of quality, albeit excluding persons added to the population after the 2001 census, and death outcomes before 2001. The exploitation of this database has shown important variations in mortality and morbidity rates, and survival, for coronary heart disease by ethnic group: in particular, the incidence of acute myocardial infarction in South Asians was about 60–70% higher than that in non-South Asians. There is no similar linked dataset in England and Wales, although ONS has recently proposed a UK-wide integrated population statistics system of linked census, survey and administrative data to be developed over the next decade or two.[9] Examples of the use of record linkage to validate ethnic group are available only from other countries. Stehr-Green

Table 1 Incompleteness of ethnicity data across government office regions

	Percentage (%) incompleteness and rank (R)* across government office regions										
	England	NE	NW	Y&H	EM	WM	EE	L	SE	SW	Quotient†
PLASC data, 2004											
Primary schools	2.3	3.0 (8)	1.7 (5)	1.4 (2)	1.3 (1)	1.5 (3)	2.8 (6)	1.6 (4)	4.4 (9)	2.9 (7)	3.4
Secondary schools	3.4	4.3 (7)	2.3 (4)	1.6 (1)	2.0 (2)	2.2 (3)	4.2 (6)	2.5 (5)	6.1 (9)	4.8 (8)	3.8
Educational attainment/PLASC 2003	5.7	6.1 (6)	4.5 (4)	2.9 (1)	4.7 (5)	3.8 (2)	6.7 (7)	3.9 (3)	9.7 (9)	8.1 (8)	3.3
Children in need, 2003											
Children looked after	1.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Children supported in families or independently	11.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
All children in need	8.0	5.0 (1)	10.0 (8)	6.0 (4)	6.0 (4)	9.0 (7)	15.0 (9)	8.0 (6)	6.0 (4)	8.0 (6)	3.0
Enhanced TB surveillance, 2000–2	6.6	1.8 (4)	1.2 (1)	2.7 (5)	43.4 (9)	1.3 (2)	3.0 (6)	5.0 (8)	3.8 (7)	1.5 (3)	36.2
AIDS/HIV: SOPHID data, 2003	3.0	3.3 (8)	1.8 (6)	0.7 (3)	0.7 (3)	1.2 (4)	1.6 (5)	4.0 (9)	2.4 (7)	0.5 (1)	5.8
Drug misuse: NDTMS data	15.6	15.3 (6)	11.3 (4)	17.1 (5)	3.3 (2)	2.2 (1)	25.9 (8)	9.5 (3)	30.2 (9)	22.9 (7)	15.6
Social services workforce, 2004	8.9	3.1 (1)	10.6 (7)	4.8 (3)	4.8 (3)	5.4 (4)	18.1 (9)	7.1 (5)	14.4 (8)	9.6 (6)	5.8
Non-medical workforce, 2004	11.7	8.6 (4)	7.8 (3)	6.9 (1)	7.3 (2)	10.7 (6)	17.7 (9)	16.8 (8)	15.8 (7)	9.0 (5)	2.6
Medical and dental workforce, 2004	2.0	1.1 (1)	1.4 (3)	1.6 (4)	1.1 (1)	2.0 (6)	3.0 (9)	1.9 (5)	2.6 (8)	2.5 (7)	2.7
HES, 2003/4	36.0	19.0 (1)	33.0 (4)	45.0 (8)	30.0 (3)	25.0 (2)	42.0 (7)	34.0 (5)	47.0 (9)	42.0 (7)	2.5
Smoking cessation	4.0	2.0 (3)	4.0 (6)	2.0 (3)	7.0 (8)	2.0 (3)	4.0 (6)	10.0 (9)	7.0 (8)	3.0 (4)	5.0
Summation of percentages and ranks		72.6 (50)	89.6 (55)	92.7 (40)	111.6 (43)	66.3 (43)	144.0 (87)	104.3 (70)	149.4 (94)	114.8 (69)	
Overall ranks*		2 (4)	3 (5)	4 (1)	6 (2)	1 (3)	8 (8)	5 (7)	9 (9)	7 (6)	

EE, East of England; EM, East Midlands; HES, hospital episode statistics; L, London; NA, not applicable; NE, North East; NDTMS, National Drug Treatment Monitoring System; NW, North West; PLASC, Pupil Level Annual School Census; SE, South East; SOPHID, Survey of Prevalent HIV Infections Diagnosed; SW, South West; TB, tuberculosis; WM, West Midlands; Y&H, Yorkshire & The Humber.
*Rankings: 9, worst on data completeness; 1, best.
†High–low rate ratio based on percentages.

et al[10] used probabilistic record linkage to match 1989–97 Washington state death files to the Northwest Tribal Registry. Of the matches for 2819 decedents, 414 (14.7%) had been misclassified on death certificates as non-American Indians and Alaskan natives.

Most methods of record linkage require particular attention to be accorded to issues of data protection and data confidentiality,[11 12] especially when unique personal identifiers such as the NHS Number are used. This means that record linkage on major administrative datasets can usually be undertaken only by the government and its agencies. Despite the complexity of the rules governing data-sharing within the government and with researchers, the strengths of record linkage lie in the robust and timely content of such administrative data, and the limitation of costs to data extraction, cleaning and the linkage process.

USE OF DISTINCTIVE NAMING ALGORITHMS

A further method for assigning ethnic group when it is missing is that of computerised name recognition algorithms. Such algorithms—to assign ethnic group on a probabilistic basis—can clearly be used only when the researcher has full access to name information. This is frequently not the case for routinely reported data, although healthcare organisations may have access to the full names of individuals on contract minimum dataset records. In those datasets for which ethnic group is missing on some records (eg, on many cancer registration datasets) or not collected, this method can be used to populate such records, but only for certain ethnic groups and with less than full accuracy. Datasets with both names and ethnic group can be used to validate name recognition methods. Name recognition algorithms, based on the distinctiveness of names, have been developed in different country settings (mainly North America and Britain) and for only a limited number of ethnic groups, including South Asians, Chinese, Vietnamese, Koreans, Hispanics and Jews. For some of these groups—notably South Asians—their use in an epidemiological and health services research context in the UK is now widespread. However, given the probabilistic nature of the allocation based on the distinctiveness of names, the method does not definitively assign ethnic group, and is clearly less satisfactory than the recommended gold standard of self-assignment.

Two such algorithms have been developed to assign South Asian ethnicity: the Nam Pehchan, and the South Asian Names and Group Recognition Algorithm. With respect to Nam Pehchan, the programme identified 36.8% of all South Asian cases (n = 5506) as false positives and 9.5% as false negatives, which, when compared with the reference standard, gave Nam Pehchan a sensitivity of 90.5% and a positive predictive value of 63.2%.[13] The investigators concluded that the programme alone was not an adequate single strategy. The South Asian Names and Group Recognition Algorithm was successful in recognising people of South Asian origin in reference datasets, with a sensitivity of 89–96%, a specificity of 94–98%, a positive predictive value of 80–89% and a negative predictive value of 98–99%.[14] In addition, religious origin was correctly assigned in the majority of cases. Other studies have reported less satisfactory findings, and there may be differences across the UK nations. For example, computer-based name search algorithms were found to be inaccurate in Scotland: Nam Pehchan, when subject to expert visual inspection, gave a positive predictive value of only 34.7%.[8] With some refinement or supplementation, then, for one of the largest pan-ethnic groups in Britain, South Asians, distinctive name algorithms do offer an alternative strategy for investigating ethnic disparities in health and healthcare, in datasets with information on full names but lacking or having incomplete ethnic coding. Such

methods are problematic for the Irish and Afro-Caribbean groups, and, apart from the Chinese, remain largely untested in other ethnic groups in the UK. Again, the contribution of such studies to our knowledge on ethnic disparities has been significant, as shown in the studies on cancer incidence and survival in South Asians[15 16] and the mortality of patients with insulin-treated diabetes mellitus.[17]

Finally, a second set of strategies involves the use of analytical methods to utilise data that are of suboptimal quality: PMRs and meta-analysis.

PROPORTIONAL MORTALITY (MORBIDITY) RATIOS

Proportional mortality (PM)—the number of deaths due to cause "x" divided by the total number of deaths—can be calculated by sex, age group or any other appropriate subdivision of the population: these figures can be compared between populations, places or time periods by calculating the PMR (in simpler terms, the ratio of PMs in the study and standard or comparison populations).[18] As Bhopal states, the PMR answers the question: is there a difference in the proportion of deaths attributable to disease "x" in one population compared with a second population? Either the overall proportion in the standard population can be applied to obtain the expected proportion in all ages (the actual or crude PMR) or the age-specific proportions can be applied. The denominator may also be cause-specific—for example, deaths from coronary heart disease could be examined as a proportion of deaths from stroke, or cancer, or accidents, rather than all causes.

The PMR—which can also be used to present data on hospital admissions by cause—is used for risk data presentation when reliable data are available only on numerators (cases) and compatible denominator data (on persons at risk) are not available or are inaccurate. With respect to denominators for country of birth and ethnic group, these are currently usable only for the 2 or 3 years on either side of the decennial census: otherwise, population sizes must be calculated by linear interpolation between data collected in the censuses. Currently, population estimates and projections by ethnic group for the intercensal years are not routinely available, except in London and a few local authorities, although work is in progress on a national set of population projections by ethnic group.

Given that the size of ethnic groups is difficult to estimate with accuracy several years beyond the census enumeration, the case for using PMRs or proportional admission ratios is strengthened. This is also the case when a denominator cannot be satisfactorily derived from the census—for example, when the numerator is a registered—rather than resident—population. For NHS hospital trusts, too, there are no accurate denominators, as their catchments are not discrete geographical areas and usually overlap with neighbouring trusts, although there are complex best-fit statistical procedures for estimating populations in hospital catchment areas. The alternative—of using a denominator based on a geographically resident population—requires information on all patients treated in that geographical area and not just those of the one provider. Finally, there may be some instances in which record systems use an ethnic coding system that is different from the one used in the census and cannot easily be mapped to it.

Many examples can be found in the ethnicity (including country of birth) literature of the application of proportional ratios, in the context of mortality,[19 20] hospital admissions[20 21] and medical school admissions.[22] However, there are few studies that provide a more critical discussion of PMRs and their limitations. Bhopal[18] is cautious, citing the well-established shortcomings of the method. The magnitude of the PM depends not only on the number of deaths from the cause under study but also on the number of deaths from other causes. Thus, in comparisons of the PM between populations, differences might arise from either differences in the disease under study or differences in other diseases. He cites the example of South Asians in whom cancers are less common than in the population as a whole: thus, a high PMR could be due to either a high level of coronary heart disease or a lower rate of cancer. He prefers to see the PMR as a preliminary, or corroborative, analysis tool, because its fundamental assumption—that the distribution of deaths from causes other than the one under study is the same in the two populations—is unlikely to hold.

A stronger recommendation for the use of the PMR is taken by Aveyard,[23] who argues that PMRs are a useful tool by which the NHS organisations can monitor the health of a population, that they should be more widely used for that purpose, and that the bias as a measure of risk is small and of no practical importance. Indeed, a recent study of revascularisations in London by ethnic group—undertaken on incomplete data—showed that similar results were obtained by proportional admission ratios, direct standardisation and indirect standardisation.[21] Earlier work in occupational studies—the examination of data from 30 randomly selected occupational units described by the UK census agency—revealed that age-standardised cause-specific standardised mortality ratios and PMRs had an almost constant relationship[24]; furthermore, around 70% of conditions with significantly high PMRs >200 had corresponding standardised mortality ratios that were also significantly high. Aveyard[23] argues that several techniques can be used to reduce bias in PMR studies, including the use of several controls, the use of positive and negative controls, and the use of only one type of death in the denominator, rather than all causes of death. For the NHS organisations, then, the PMR can be viewed as a simple, quick to calculate and potentially useful indicator, but, given its potential flaws, it does require cautious interpretation and should be used with other corroborative evidence.

META-ANALYSIS

Finally, the utility of meta-analysis as a statistical technique for combining the findings from a number of independent studies is considered. Meta-analyses are usually based on systematic reviews, a method which applies rigorous standards of study selection and assessment of design and execution to secondary research. The benefit of meta-analysis is that it overcomes the bias in unsystematic or narrative reviews. Small or medium-sized studies may have low statistical power: by drawing on patients in many studies, meta-analyses have more power to detect small but clinically significant effects. This may become important in specific subgroup analyses (eg, by ethnicity), in which patients in the subgroup of interest in individual studies may be too few for significant effects to be detected. Meta-analyses can provide a precise estimate of effect by giving due weight to the size of the studies included, and through aiming for complete coverage of all relevant studies. The likely presence of selection bias is assessed through funnel plots, and the robustness of findings through sensitivity analysis.

The yield from this kind of approach is illustrated by McDowell *et al*'s[25] recent meta-analysis of ethnic differences in risks of adverse reactions to drugs used in cardiovascular medicine. In all, 24 studies provided data for such reactions for at least two ethnic groups. Pooled analyses enabled the presentation of relative risks of angio-oedema from ACE inhibitors (black vs non-black patients), of cough from ACE inhibitors (East Asian vs white patients) and of intracranial haemorrhage from thrombolytic therapy (black vs non-black patients). However, this study exemplifies an inherent

difficulty in undertaking meta-analyses when the findings relate to ethnic groups. Subjects may be recruited to research studies within specific national contexts and on the basis of varying ethnicity criteria, such as self-identification, membership in groups and ethnic origin, perhaps combined with other dimensions such as country of birth or migration status. Frequently, such methods of assignment are poorly reported in studies. The pooling of findings on treatment effects for pan-ethnicities such as East Asian, black, non-black and white raises issues of validity, given the likely substantial heterogeneity within these collectivities. The investigators do, indeed, acknowledge the limitations that this imposed on comparisons between the different studies.

CONCLUSION

There are no valid reasons for eschewing analysis of ethnically coded datasets when the quality and completeness of the data are regarded as suboptimal for such purposes. Experience has shown that, by using the data and remaining close to it, healthcare organisations can drive up quality. What has been missing has been a framework of incentives to do this, including performance indicators beyond those tied to quality. The importance attached by the NHS to the implementation of policies and ethnic monitoring, at the expense of the analysis of the data, is unfortunate. Ultimately, discrimination can routinely and successfully be challenged only if organisations are able to demonstrate this in the analysis of their ethnically coded datasets. Collecting the data simply to report on its quality serves no useful purpose and continues to be wasteful of substantial resources. It is hoped, therefore, that this presentation of analytical strategies for using sparse and incomplete ethnicity data will encourage the greater use of the data already collected, and will also stimulate further research into ethnic disparities in health and healthcare.

Authors' affiliations

Peter J Aspinall, Centre for Health Services Studies, University of Kent, Canterbury, Kent, UK
Bobbie Jacobson, London Health Observatory, London, UK

Funding: This study was funded by the London Health Observatory to PJA (under part-time secondment).

Competing interests: None.

REFERENCES

1 **Aspinall PJ**. Secondary analysis of administrative, routine and research data sources: lessons from the UK. In: Nazroo JY, ed. *Health and social research in multiethnic societies*. London: Routledge, 2006:165–95.
2 **Fitzpatrick J**, Jacobson B, Aspinall PJ. *Indications of public health in the English regions. Vol 4. Ethnicity and health*. London: Association of Public Health Observatories, 2005.
3 HES Data Quality Indicator Report for data year 2003/2004. http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=452.
4 **Platt L**, Akinwale B, Simpson L. Stability and change in ethnic group in England and Wales. *Popul Trends* 2005;**121**:35–45.
5 **Goldacre MJ**. The Oxford Record Linkage Study: current position and future prospects. In: Howe GR, Spasoff RA, eds. *Proceedings of the Workshop on Computerised Record Linkage in Health Research*. Toronto: University of Toronto Press, 1986:106–29.
6 **Office for National Statistics**. *National Statistics Code of Practice. Protocol on data matching*. London: TSO, 2004.
7 **Abrahams C**, Davy K. Linking HES maternity records with ONS birth records. *Health Stat Q* 2002;**13**:22–30.
8 **Bhopal R**, Fischbacher CM, Steiner M, *et al. Ethnicity and health in Scotland: can we fill the information gap?* Edinburgh: Public Health Sciences, University of Edinburgh, 2005.
9 **Office for National Statistics**. *Proposals for an integrated Population Statistics System*, Discussion paper. London: ONS, 2003.
10 **Stehr-Green P**, Bettles J, Robertson LD. Effect of racial/ethnic misclassification of American Indians and Alaskan Natives on Washington State Death Certificates, 1989–1997. *Am J Public Health* 2002;**92**:443–4.
11 **Social Exclusion Unit**. *Better information. Report of Policy Action Team 18*. London: The Stationery Office, 2000.
12 **Cabinet Office**. *Privacy and data-sharing. The way forward for public services*. London: Performance and Innovation Unit, Cabinet Office, 2002.
13 **Cummins C**, Winter H, Cheng KK, *et al*. An assessment of the Nam Pehchan computer program for the identification of names of south Asian ethnic origin. *J Public Health Med* 1999;**21**:401–6.
14 **Nanchahal K**, Mangtani P, Alston M, *et al*. Development and validation of a computerized South Asian Names and Group Recognition Algorithm (SANGRA) for use in British health-related studies. *J Public Health Med* 2001;**23**:278–85.
15 **Winter H**, Cheng KK, Cummins C, *et al*. Cancer incidence in the south Asian population of England (1990–92). *Br J Cancer* 1999;**79**:645–54.
16 **dos Santos Silva I**, Mangtani P, De Stavola BL, *et al*. Survival from breast cancer among South Asian and non-South Asian women resident in South East England. *Br J Cancer* 2003;**89**:508–12.
17 **Swerdlow AJ**, Laing SP, Dos Santos Silva I, *et al*. Mortality of South Asian patients with insulin-treated diabetes mellitus in the United Kingdom: a cohort study. *Diabet Med* 2004;**21**:845–51.
18 **Bhopal RS**. *Concepts of epidemiology*. Oxford: Oxford University Press, 2002.
19 **Marmot MG**, Adelstein AM, Bulusu L. *Immigrant mortality in England and Wales 1970–78*. London: HMSO, 1984.
20 **Bardsley M**, Hamm J, Lowdell C, *et al. Developing health assessment for black and minority ethnic groups: analysing routine health information*. London: Health of Londoners Project, 2000.
21 **Mindell J**, Klodawski E, Fitzpatrick J. *Using routine data to measure ethnic differentials in access to revascularisation in London*, A technical report. London: LHO & APHO, 2005.
22 **Seyan K**, Greenhalgh T, Dorling D. The standardised admission ratio for measuring widening participation in medical schools: analysis of UK medical school admissions by ethnicity, socioeconomic status, and sex. *BMJ* 2004;**328**:1545–6.
23 **Aveyard P**. A fresh look at proportional mortality ratios. *Public Health* 1998;**112**:77–80.
24 **Roman E**, Beral V, Inskip H, *et al*. A comparison of standardized and proportional mortality ratios. *Stat Med* 1985;**3**:7–14.
25 **McDowell SE**, Coleman JJ, Ferner RE. Systematic review and meta-analysis of ethnic differences in risks of adverse reactions to drugs used in cardiovascular medicine. *BMJ* 2006;**332**:1177–81.