Iranian J Publl Health, Vol. 37, No.4, 2008, pp.26-31

Original Article

The Evaluation of Matching in a Case-Control Study of Colorectal Cancer Using General Practice Lists

**M Movahedi [1], T Bishop [2], JH Barrett [2], GR Law [3]*

[1] *Dept. of Epidemiology, School of Public Health, Shaheed Beheshti University of Medical Sciences, Tehran, Iran*
[2] *Genetic Epidemiology Unit, Cancer Research UK, St.James Hospital, Beckett Street, LS9 7FT, Leeds, UK*
[3] *Centre for Epidemiology and Biostatistics, University of Leeds, 30 Hyde Terrace, Leeds, LS2 9LN, UK*

(Received 15 Apr 2008; accepted 24 Oct 2008)

Abstract

Background: A crucial part of a case-control study is the selection of a sample of controls that represent the base-population from which cases were drawn. Controls may be matched to cases by one or more potentially important confounding variables, such as socioeconomic status. In the United Kingdom, one method for control selection has been based on the patient list of the General Practice with whom the cases were registered, which we refer to as GP-matching. We aimed to explore whether GP-matching adequately control for the potential confounding effect of socioeconomic status.

Methods: The Townsend index of deprivation was calculated for different two national census geography levels of Electoral ward/Postcode Sector and Enumeration District/Output area for the three study areas of Dundee, Leeds and York. Conditional logistic regression was used to estimate the association of cases with deprivation (based on the Townsend index) compared with that of matched controls for the two geographical scales.

Results: At the largest geographical level (Electoral ward/Postcode Sector) there was no evidence of a difference in the distribution of deprivation scores between cases and controls. However, analysis at the smallest level (Enumeration District/Output area) showed that, despite GP matching, cases were more likely to live in deprived areas than matched controls.

Conclusion: Using General Practice lists for the selection of controls for controlling the confounding effect of socioeconomic status might not be an appropriate method for case-control studies conducted in the United Kingdom.

Keywords: *Case-control study, Confounding, Townsend deprivation index*

Introduction

Valid inference about disease aetiology from a case-control study requires the controls to represent the population from which cases were selected. An unrepresentative sample of controls might lead to a biased estimation of the association between exposure and risk of disease, known as selection bias (1). Controls are often matched to cases by one or more potentially important confounding variables, which results in reducing the effect of those variables on the estimated risk of disease for the target exposure.

Matching is much used in case control studies. There are some sources for control series: Population based study, the cases are a representative sample of all cases in a defined and identified population and controls should be randomly selected from the disease free members of the same population. If we can not identify the source population, simple random sampling is not possible and as a result it is better to use other source of control selection (1). Neighbourhood controls method selects controls through sampling of residences but people living in the same area are likely to be similar in many respects. Moreover some times this method is not easy to use because usually all geographic address of residences is not available. In this condition, one can decide to sample controls that are individually matched to the cases from the same neighbourhood. The third source of control selection is hospital-based controls. When using hospital-based cases, it may not be possible to determine source population, because not all cases refer to the hospital and those

***Corresponding to author:** Tel: +98 912 5049644, Fax: +98 21 22432037, E-mail: movahed20@gmail.com

referred might be selected regarding some specific criteria. In this situation based population controls are the better choice. On the advantages of this method is that the effect of many selective factors that bring people to hospitals such as financial standing area of residence, ethnicity will be reduced. Moreover, they reply to questions more accurately because of having hospitalisation and illness. On the other hand the most advantage of this kind of control selection is that some their illness may share relating exposure with the study disease. It means they may have a lower, higher exposure prevalence compared to the population from which the cases arise (1).

For nominal variables such as neighbourhood, which contains a wide range of environmental factors certain variables, if matching were not applied in the design part of the study, there would not be a sufficient number of individuals in the study groups who were alike with respect to these confounding factors to allow for any type of controlling in the analysis.

Having said that, in our case-control study (2-3), which conducted to investigate the association between environmental factors and genetic polymorphisms on the risk of colorectal cancer, we selected controls through the patient's General Practice (GP) list. The rationale was to select controls with a similar socioeconomic status (SES) and geographical area of residence to cases. It was assumed that patients registered to the same GP were from the same geographical area with some similarity in socioeconomic profiles.

Thus, the main research question in this study was to show how successful the GP-matching approach was to control for the potential confounding effect of SES.

Materials and Methods

Study design

The data for this analysis were taken from a multicentre case-control study, which was conducted to investigate the association between environmental factors and genetic polymorphisms and the risk of colorectal cancer. The study was approved by the local research Ethics Committees, and signed informed consent was obtained from all participants. Further details of study design are given elsewhere (2-3). Briefly, cases were between 45 and 80 yr of age when diagnosed with colorectal cancer between 1997 and 2001 from hospitals in Leeds, Dundee and York. Healthy population-based controls (GP controls), with no history of previous cancer, were recruited from the patient's GP practice list. An age (within one year), and sex matched control was identified for each consented case. Multiple GP controls were recruited for some cases, while no matching control was available for a minority of cases. Contact was made initially by post, including a standard letter of invitation by, or behalf of, their GP. The controls who most closely matched their case were approached. They could be divided into two main groups: a) those who agreed to be interviewed and were interviewed (Interviewed Controls= 397); b) those who were approached but refused or failed to reply to invitation letter (Refused Controls= 57). The refused controls were then replaced by the alternative controls.

Linkage to the census of Great Britain

In Great Britain the census is a count of all households and persons, which is carried out every 10 yr. The 1991 census contained questions on housing, ethnic group, car ownership, and occupation (4). In order to maintain confidentiality census information at household level aggregated to an areal level. The smallest census geography in England and Wales is the Enumeration District (ED) and in Scotland is the Output Area (OA). These areas can be aggregated into Electoral Wards (EW) in England and Wales and Postcode Sectors (PS) in Scotland. The postcode for the residential address of cases and controls was checked using a postal address directory available from the Royal Mail. Cases and matched controls were then linked through their residential postcodes to the 1991 census.

Townsend deprivation index

The Townsend deprivation index (5), a measure of SES, was calculated for each ED/OA and EW/PS level using data for the three populations of the study-area (Dundee, Leeds, and York) combined. The Townsend score was based on four pieces of information: the percentages of (I) unemployed economically active persons (age 16 and over), (II) households with more than one person per room, (III) households not owner occupied and (IV) households without a car. Unemployment and Overcrowding were transformed using a logarithmic transformation, allowing the variables to be symmetrically distributed. Each of the four percentages was standardized to a mean of zero with a standard deviation of one in order to make all four factors contribute equal weight to the Townsend score. These standardised scores were added to obtain the Townsend score. A high positive value represents an area with high deprivation and a high negative value represents an affluent area.

Statistical analyses

The evaluation of GP matching for ED/OA and EW/PS levels was performed first based on the comparison of the mean deprivation index of cases with those of matched controls using the paired t-tests. The Townsend scores were then divided into five equally-sized groups according to study-area population quintiles of the score. In next step being case or control was chosen as dependent variable and deprivation quintile as independent variable and then using conditional logistic regression, the association of disease risk with deprivation was estimated for the two geographical scales (ED/OA and EW/PS). The least deprived quintile was considered as baseline for this analysis. The analyses were carried out using STATA (Stata Statistical Software: Release 7.0. College Station, TX: Stata Corporation).

Results

In total, 500 cases and 742 controls were recruited to the original study, of whom 484 cases and 738 controls were interviewed. Overall 461 matched case-control pairs (922 individuals) were available for this secondary study. A valid postcode was identified for 454 pairs (seven pairs were excluded because the correct postcodes for 13 consented participants were unknown). In Leeds centre, 57 first choice controls (6), eligible but who did not consent for interview were replaced by 42 second choice and 15 third, fourth and fifth controls.

Evaluation of matching at the ED/OA level

The paired t-tests showed a significant difference between the Townsend score in cases and controls matched by GP at the ED/OA level (difference= 0.64, 95% CI: 0.30, 0.99) (Table 1). The distribution of cases and matched controls at ED/OA level by deprivation quintile, compared with the distribution of the study-area population was shown in Table 2. In a representative sample from this population, 20% of the participants would be in each quintile. There was a marked difference in deprivation distribution of the cases and controls and also for all study participants and the study-area population with over-representation of those in the least deprived quintile for both groups but particularly for controls (30.5%) compared with the distribution of the study-area. Cases were significantly more likely to belong to the most deprived fifth than controls compared to the least deprived fifth (OR=1.80, 95% CI: 1.18-1.50).

Evaluation of matching at the EW/PS level

The results of the paired t-tests for the Townsend score at EW/PS level showed no significant difference between the mean index in cases and their matched controls (difference: 0.16, CI: -0.13,0.44) (Table1). There was no major difference in distribution of deprivation quintile between cases and controls at EW/PS level (Table 2). Cases and controls were similarly likely to live in the most deprived compared to the least deprived areas (OR= 1.05, 95% CI: 0.73-1.50). A marked gradient was observed in cases and controls with over-representation of those in the least deprived quintiles (39.1% and 38.7% respectively) compared with the distribution of the study-area population (Table 2).

Table 1: Summary of deprivation index for cases and controls and the result of paired *t-test* at ED/OA and EW/PS level

	Observation	**Mean**	**Standard Deviation**	**95% Confidence Interval**	**Minimum**	**Maximum**
ED/OA level						
Case	454	- 0.34	3.31	-0.65, -0.04	-7.66	7.17
Control	454	- 0.99	3.14	-1.28, -0.7	-7.03	7.21
Difference	454	0.64	3.74	0.30, 0.99	-11.15	11.05
EW/PS level						
Case	454	1.00	3.66	0.66, 1.33	-7.26	8.41
Control	454	0.84	3.56	0.50, 1.17	--6.75	8.41
Difference	454	0.16	0.15	-0.13, 0.44	-10.99	10.99

Table 2: Distribution of cases and matched controls, and odds ratios, by Townsend deprivation index fifth based on the study-area population distribution shown at ED/OA and EW/PS level

Deprivation fifth	**Cases** **n (%)**	**Controls** **n (%)**	**Odds ratio** [2]	**95%Confidence Interval** [2]
ED/OA level [1]				
1(least deprived)	107 (23.6)	138 (30.5)	1.00	-
2	109 (24.0)	96 (21.0)	1.46	1.01-2.13
3	72 (15.9)	102 (22.5)	0.91	0.61-1.35
4	84 (18.5)	59 (13.0)	1.84	1.21-2.79
5(most deprived)	82 (18.0)	59 (13.0)	1.80	1.18-2.73
EW/PS level [1]				
1(least deprived)	178 (39.1)	176 (38.7)	1.00	-
2	44 (9.70)	59 (13.0)	0.74	0.74-1.15
3	60 (13.2)	53 (11.7)	1.12	0.73-1.71
4	81(17.8)	80 (17.6)	1.00	0.69-1.45
5(most deprived)	92 (20.2)	86 (19.0)	1.05	0.73-1.50
Total	454 (100)	454 (100)		

[1] ED/OA Enumeration District/Output area and EW/PS Electoral Ward/ Postal Sector- see text for further details.
[2] Conditional logistic regressions

Table 3: Mean difference of Townsend deprivation index between cases and two types of controls at ED/OA scale in Leeds centre

Mean of Townsend deprivation index				
Cases	**Controls**		**Mean difference**	***P***
	Including second , third and forth choices	**Including the first choices**		
-0.08	-0.59	-	-0.50	0.02
-0.08	-	-0.42	-0.35	0.10

Discussion

Our findings show that at the ED/OA level, the smallest census geography available compared to the cases, the controls that were selected from the same GP's list had a much more skewed distribution in the least deprived quintile.

The findings also identify and provide a warning against, a particular sort of methodological "bad practice". However, one would hope that social researchers in the UK would typically have a sufficiently awareness of the possible impact of underlying socioeconomic differences on research findings to avoid this form of bad practice. It is possible that medical researchers are less alert to this kind of issue, though one would imagine that a practical awareness of the nature GP practices would lead to recognition of the flaws in such an approach.

Regarding the differential participation of cases and controls, where a proportion of the first choice controls in this study refused to participate, selection bias may be a possible explanation for this significant difference between cases and controls. For example, in the Leeds centre, when 57 second and other next choice controls were replaced by non-participating first choice controls there was no significant difference between cases and controls in terms of deprivation at ED/OA scale (Table 3). It shows that those controls living in the most deprived areas were less likely to participate in this study.

In contrast, when the Townsend deprivation index at EW/PS scale was used, the lack of homogeneity in the area covered by a GP was not evident. The lack of significant difference between cases and controls in terms of deprivation distribution at this level demonstrates that general practices cover a false homogenous area with respect to SES. Since the aim of matching is to make the distribution of potential confounding factors similar between cases and controls, it is necessary that cases be matched to controls as accurately as possible for the relevant confounders.

Using the Carstairs score as a deprivation index at the ED level, a case-control study of heart attacks in young women of GB found similar results to these (7). It was shown that cases had a skewed deprivation distribution, with more than 35% in the most deprived quintile compared with the distribution of the British population. The controls who were selected from the same general practices showed a much less skewed distribution. Cases and matched controls with different SES might have similar Townsend scores at the EW/PS level while their scores at ED/OA level differ. Electoral Ward and Postcode Sector usually cover around 2000 households with possibly wide variation in SES, while Enumeration District and Output Area cover 200 and 50 households respectively (8) with more homogeneity in SES than households covered by EW/PS and GP.

In terms of the role of participation bias in the validity and generalisability of the study, in a case-control study of acute leukemia in England it was shown that the controls who participated differed markedly from those who did not (9). Those who could not be contacted tended to live in the most deprived areas followed by those whose GP refused contact and those who were contacted but declined to participate.

We understand that the ecological fallacy might be a limitation for our study as a result of working on aggregated data. The SES of cases and controls was estimated based on aggregated level in place of individual level (ecological fallacy) which might have distorted the findings.

In conclusion, the area covered by a general practice is likely to cover a population with heterogeneous SES and as a result using a general practice might not be a very appropriate approach as to matching characteristics between cases and controls. However, participation bias should be considered as a possible explanation for different deprivation distribution of cases and controls in case-control studies conducted in the UK.

Acknowledgments

We gratefully acknowledge the help of all study participants, clinicians and interviewers who contributed and supervised this project. We also acknowledge financial support from the Food Standards Agency and from Cancer Research UK for initial study.

The authors declare that they have no conflict of interests.

References

1. Hennekens C, Buring J (1987). Case-control studies. In: *Epidemiology in Medicine*. Ed, Mayrent. 1st ed, Lippincott Williams and Wilkins. Philadelphia, USA, pp. 132-52.
2. Barrett JH, Smith G, Waxman R, Gooderham N, Lightfoot T, Garner RC, et al. (2003). Investigation of interaction between N-acetyltransferase 2 and heterocyclic amines as potential risk factors for colorectal cancer. *Carcinogenesis*, 24(2): 275-82.
3. Sachse C, Smith G, Wilkie MJ, Barrett JH, Waxman R, Sullivan F, et al. (2002). A pharmacogenetic study to investigate the role of dietary carcinogens in the aetiology of colorectal cancer. *Carcinogenesis*, 23(11): 1839-49.
4. Majeed FA, Cook DG, Poloniecki J, Martin D (1995). Using data from the 1991 census. *BMJ*, 310(6993): 1511-14.
5. Townsend P, Phillimore P, Beattie A (1988). *Health and deprivation: inequality and the North*. Croom Helm. London. Law GR, Smith AG, Roman E, on behalf of the UK Childhood Cancer Study
6. Investigators (2002). The importance of full participation: Lessons from a national case-control study. *British Journal of Cancer*, 86(3): 350-55.
7. Thorogood M, Arscott A, Walls P, Dunn N, Mann R (2002). Matched controls in a case-control study. Does matching by doctor's list mean matching by relative deprivation. *International Journal of Social Research Methodology,* 5(2): 165-72.
8. Cole K (1993). The 1991 Local Base and Small Area Statistics. In: *The 1991 Census user's Guide*. Eds, Dale and Marsh. 1st ed, HMSO Publications. London, pp. 201-48.
9. Smith AG, Fear NT, Law GR, Roman E (2004). Representativeness of samples from general practice lists in epidemiological studies: case-control study. *BMJ*, 328(7445): 932-36.