# Secondary matching: a method for selecting controls in case-control studies on environmental risk factors

Antonio Agudo and Carlos A González

**Background** The problem of control selection was considered in a population-based case-control study on pleural mesothelioma in Spain. Random sampling from the population was discarded because of potential selection bias due to low participation. Selection of hospital controls by matching them to cases by hospital seemed unsuitable for investigating environmental exposures, as the choice of hospital may be related to the place of residence; controlling for residence may avoid bias but could produce overmatching.

**Methods** A three-step procedure was proposed. First, a random sample of primary controls from the population census of the province of Barcelona was obtained. Second, the hospital closest to the residence of the primary control was identified as the control hospital. Third a secondary control was chosen among patients admitted to the hospital matched to the primary control by sex, age and municipality.

**Results** An overall participation rate of 85% was achieved. The hospital control group showed a distribution of residences similar to that of the general population, and independent of the distribution of cases.

**Conclusions** This procedure may be considered as an alternative for control selection when studying environmental factors or, generally, when matching cases and controls by hospital is to be avoided. Its validity was assessed according to the principles of comparability with cases regarding the study base, accuracy of information, deconfounding and efficiency.

**Keywords** Case-control studies, epidemiological methods, overmatching, selection bias, study design

**Accepted** 22 June 1999

The issue of control selection has been regarded as crucial to validity of case-control studies, generating controversy and discussion among epidemiologists, usually concerning comparisons between the two main sources of controls, population or hospital. Cole[1] identified the selection of the control group as 'the unique and truly large problem of the case-control study'. The subject was also addressed by Miettinen[2] within the theoretical framework of case-referent studies, as a problem of valid selection of subjects. The primary challenge in a case-control study is the identification of the appropriate study base (a population's experience over time). The guideline to valid selection of subjects is that the case and control series should be representative of the same base. Wacholder *et al.*[3] took comparability as the starting point to assess the selection of controls. They proposed and summarized four basic principles: study base, deconfounding, comparable accuracy and efficiency. Following similar principles, Miller[4] examined the issue of the source of controls, hospital versus population. Both[3,4] concluded that it may be difficult to satisfy all principles in a study and they need to be balanced. On the other hand, the appropriate source often depends on the question being addressed. The implications of alternative approaches need to be considered carefully. Thus, detailed examination of the problem will sometimes dictate the solution, but each study must be evaluated on its own merits.

Recently we undertook the design of a case-control study of malignant pleural mesothelioma in Spain as a part of a European multicentric study,[5] with the aim of assessing the risk of environmental and occupational exposure to asbestos. All incident cases from a geographical area during the study period were identified. Two ways of selecting controls were considered: first, random sampling from the general population; second, selection from patients admitted to the same hospital where each case

Institute of Epidemiology and Clinical Research (IREC), Mataró, Spain.

Reprint requests to: Antonio Agudo, IREC (Institute of Epidemiology and Clinical Research), c. Sant Pelegri 3, 08301 Mataró, Spain. E-mail: agudo@csm.scs.es

was diagnosed. Neither of these approaches seemed entirely satisfactory, and an alternative procedure was proposed. The purpose of this paper is to present the rationale and a detailed description of the method, as well as assessing its validity according to the general principles of case-control study design.

## Methods

### The study design; valid selection of controls

The theoretical framework of the study was one of 'primary base', as defined by Miettinen:[2] a geographical area over a period of time demarcated *a priori*. The study base consisted of residents in the provinces of Barcelona (4.6 million inhabitants) and Cádiz (1.1 million inhabitants) between 1 January 1993 and 31 December 1996. Although procedures and study design were identical, all data presented in this paper for illustrative purposes are restricted to the province of Barcelona.

With a primary base the main challenge is the identification of all cases. Potential cases were all subjects from the study base newly diagnosed with primary malignant mesothelioma of the pleura and histologically confirmed. Cases could come from the 24 hospitals with pathology departments.

Regarding selection of controls, it follows that the first option should be a random sample of the population, for which census data were available. However, in the pilot phase a low participation rate (below 50%) was achieved; non-participation is considered as one of the biggest threats to validity. Furthermore, the rate of participation was lower in urban than rural areas, and the place of residence may in turn be related to environmental exposures. A condition for valid selection is that the probability of being sampled is independent of the exposure. This rule could have been violated in our study given the observed pattern of response.

Since cases were actually identified in hospitals, an obvious alternative would be to assemble a hospital control group. It is well known that hospitalized patients are readily available and cooperative.[1,3] The usual procedure is to match controls to cases, selecting one or more patients from the same hospital where the case was diagnosed. However, in Spain, hospitalization is mainly determined on the basis of proximity, rather than other factors such as social class or type of disease. Thus admission to a hospital is highly associated with the place of residence, and exposure to environmental factors can be closely related to the proximity of available workplaces and industrial activities. As a consequence, hospitalization in a particular centre can become a correlate of the exposure being measured. Matching by a factor related to the exposure but unrelated to the disease is known as overmatching, which is more a problem of efficiency than of validity.[6] Although appropriate analysis would produce valid estimates of the effect, they would be imprecise and there would be little confidence in the estimates obtained.

### 'Secondary matching': a method for selecting hospital controls

Higher response rates made hospital controls preferable to population controls but the problem remained of how to choose the hospital where each control should be selected, independent of cases. We followed a three-step procedure. First, a random sample of subjects from the population census of the province of Barcelona (the study base) was obtained, stratified by age and sex according to the expected age-sex distribution of cases, with two controls per case. For this group of subjects (primary controls) we compiled a list with information on sex, date of birth and address. Second, the hospital closest to the residence of the primary control was identified as the control hospital. Third, a hospitalized patient (secondary control) was selected from the control hospital's admission list and matched to his/her primary control by sex, age and municipality, with appropriate exclusions. For residents in small municipalities (<5000 inhabitants) it was sometimes difficult to find a suitable (secondary) hospital control the first time round. In this case relaxation of the procedure was allowed in two ways: the patient could be a resident in an adjacent municipality, or a resident in the same municipality could be selected from a neighbour hospital. Patients admitted because of known or suspected conditions related to asbestos were excluded as potential controls. These diagnoses were asbestosis, lung cancer, cancer of the larynx and colon cancer. Usually matched controls are selected according to some characteristics of their matched cases, but here a hospital control group ('secondary controls') is matched with a population series ('primary controls'), which is why we have called the procedure 'secondary matching'.

## Results

According to mortality rates for mesothelioma and the incidence/mortality ratio in Barcelona,[7] the expected number of cases was 119, while the final number of cases included in the study was 117. By using the procedure described above, 267 potential controls were selected in Barcelona, of which 227 were interviewed (overall participation of 85%). Only 8 (3%) subjects actively refused to participate. Of the remaining subjects, 18 were not interviewed because they were too ill or had a condition that made the interview impossible and 14 because they had been discharged by the time of the interview. Of the 227 interviewed, 187 were selected by strict application of the procedure; 32 subjects had to be found in an adjacent municipality, eight were from the same municipality but selected in a second hospital, and for two subjects both circumstances concurred.

Cases were identified in only 14 of the participating hospitals, while controls were selected from 22 out of the 24 participating centres. The distribution of cases and controls by hospital is shown in Table 1. There are 308 municipalities in the province of Barcelona; residences of cases were distributed between 35 municipalities, and those of controls from 53. Within the province, municipalities are grouped into 11 administrative areas ('comarques'); in order to give an idea of the geographical origin of cases and controls, their distribution according to these areas is presented in Table 2, together with the distribution of the general population. The proportion of males was 77% for cases and 78% for controls, with a median age of 68 years (range 35–92). Cases and controls were similarly distributed for educational level and smoking habits as compared to general population (results not shown).

## Discussion

In the design of the study there were two main issues to be dealt with. The first one led us to select hospital controls instead of a

**Table 1** Distribution of cases and controls from a study of malignant pleural mesothelioma in the province of Barcelona (Spain) according to the hospital where they were identified (cases) or selected (controls)

| | Cases | | Controls | |
|---|---|---|---|---|
| **Hospital** | n | (%) | n | (%) |
| Hospital 1 | 12 | (10.3) | 34 | (15.0) |
| Hospital 2 | 16 | (13.7) | 20 | (8.8) |
| Hospital 3 | 19 | (16.2) | 24 | (10.6) |
| Hospital 4 | | | 20 | (8.8) |
| Hospital 5 | 3 | (2.6) | | |
| Hospital 6 | 11 | (9.4) | 19 | (8.4) |
| Hospital 7 | 1 | (0.9) | 5 | (2.2) |
| Hospital 8 | 10 | (8.5) | 10 | (4.4) |
| Hospital 9 | | | 10 | (4.4) |
| Hospital 10 | 3 | (2.6) | 11 | (4.8) |
| Hospital 11 | | | 4 | (1.8) |
| Hospital 12 | 5 | (4.3) | 7 | (3.1) |
| Hospital 13 | 8 | (6.8) | | |
| Hospital 14 | 10 | (8.5) | 15 | (6.6) |
| Hospital 15 | 17 | (14.5) | 17 | (7.5) |
| Hospital 16 | 1 | (0.9) | 6 | (2.6) |
| Hospital 17 | | | 5 | (2.2) |
| Hospital 18 | | | 1 | (0.4) |
| Hospital 19 | | | 3 | (1.3) |
| Hospital 20 | | | 3 | (1.3) |
| Hospital 21 | | | 2 | (0.9) |
| Hospital 22 | | | 4 | (1.8) |
| Hospital 23 | | | 1 | (0.4) |
| Hospital 24 | 1 | (0.9) | 6 | (2.6) |
| **Total** | 117 | (100) | 227 | (100) |

**Table 2** Distribution of population (age 35–94 years), and controls and cases from a study of malignant pleural mesothelioma in Barcelona (Spain), according to the area of residence[a]

| | | Controls | | Cases | |
|---|---|---|---|---|---|
| **Geographical area[a]** | **Population (%) 35–94 years** | n | (%) | n | (%) |
| Area 1 | 1.44 | 3 | (1.32) | | |
| Area 2 | 1.71 | 2 | (0.88) | | |
| Area 3 | 3.46 | 10 | (4.41) | 4 | (3.42) |
| Area 4 | 10.11 | 25 | (11.01) | 8 | (6.84) |
| Area 5 | 55.68 | 124 | (54.63) | 57 | (48.72) |
| Area 6 | 1.01 | 3 | (1.32) | 1 | (0.85) |
| Area 7 | 1.54 | 6 | (2.64) | 1 | (0.85) |
| Area 8 | 5.54 | 11 | (4.85) | 4 | (3.42) |
| Area 9 | 2.49 | 4 | (1.76) | 1 | (0.85) |
| Area 10 | 12.50 | 31 | (13.66) | 31 | (26.50) |
| Area 11 | 4.51 | 8 | (3.52) | 10 | (8.55) |
| **Total** | (100) | 227 | (100) | 117 | (100) |

[a] Geographical areas correspond to the eleven administrative units ('comarca') in which the province is divided, each one including a variable number of municipalities. Specific identification of these areas are: (1) Alt Penedès, (2) Anoia, (3) Bages, (4) Baix Llobregat, (5) Barcelonès, (6) Berguedà, (7) Garraf, (8) Maresme, (9) Osona, (10) Vallès Occidental, and (11) Vallès Oriental.

population-based control series. The second determined that these controls should be distributed by hospitals independently of cases. The first decision was largely based on high non-response from population controls. This is a relatively common situation.[4] Most epidemiologists acknowledge that non-response is an important threat to validity but there does not seem to be a clear cutoff point between acceptable and unacceptable non-response. We decided that the response rate of lower than 50% found in population controls during the pilot study was too low, while the 85% observed in the hospital control series was acceptable. Furthermore, the different level of participation by place of residence, and consequently by exposure, was determinant in our decision to reject population controls to avoid selection bias.

The second issue may be illustrated by Tables 1 and 2. The distribution of cases and controls by hospitals is quite different, while the geographical distribution of hospital controls is similar to that of the general population. Selecting each control in the same hospital as the case would have produced overmatching by making cases and controls have a similar distribution of residence, a correlate of exposure not independently associated with the disease. It has been shown that matching and conditional analysis by an indicator of exposure opportunity is irrelevant to validity, but produces a loss of precision.[8] Although low efficiency is not as serious a problem as selection bias, it is important in some instances: for rare diseases, such as mesothelioma, it is difficult to assemble a large series of cases. Furthermore, proper assessment of environmental (non-occupational) exposure to asbestos requires the exclusion of occupational exposure. In the end, an unbiased point estimate of the relative risk with a very wide confidence interval could become useless. Any substantial gain or loss of precision is then relevant.

Other conditions to ensure comparability of the study base are considered whenever hospital controls are used.[9] The validity of hospital controls relies on the assumption that the distribution of exposure among them is the same as in the study base; in other words, the exposure is unrelated to admission. The usual practice is to exclude as controls the subjects admitted for diseases related to the exposure of interest. These exclusion criteria apply only to the diagnosis that brought the person into the hospital; past history of an exposure-related disease should not be a basis for exclusion. It is easier to apply this rule when, as in the present setting, only one exposure is being studied, and it has no known or suspected effect on a wide variety of diseases. The problem of referral bias, arising when cases and controls do not have the same catchment area, applies to studies with a secondary base, defined by cases. Our study had a primary base; no subjects out of the study area were included, and, as far as we know, it is unlikely that residents in the study base were hospitalized in centres outside the base. Finally, Berkson's bias is frequently mentioned when hospital controls are under consideration. Flanders *et al.*[10] have discussed this bias, associated with differential rates of hospitalization. They concluded that when incident cases are used the bias will be negligible. Furthermore, if controls are restricted to subjects recently diagnosed with conditions unrelated to exposure and the prevalence of hospitalizations because exposure is small, the estimated odds ratio is nearly the same as the one estimated from people in the community. All these conditions seem to be fulfilled in our study.

The principles of deconfounding and efficiency have been considered together when discussing overmatching. From a general point of view, designs with matched samples have become widely used, but it does not mean that they should always be preferred, and often they conflict with efficiency. We selected controls on the basis of sex and age distribution of cases; this is usually called frequency-matching, although we prefer stratified sampling. Such designs allow for control of confounders by modelling them in the analysis, and they may serve the same purpose intended by the use of matching. The principle of comparable accuracy relates to the quality of information collected for cases and controls.[3] Recall bias, the fact that cases may have a better selective recall of health-related events, has been put as an argument in favour of selecting hospital rather than population controls, although theoretical justification for this is weak.

Some issues regarding the practical application of our method must be considered. First, we used a nominal population census that was already available, but actually it is not necessary. The 'population group' must be seen as a theoretical framework, not a real sample: it may be enough to know the population distribution regarding the variable that will determine the choice of the hospital. This will provide the probabilities to construct the theoretical framework. In our study the distribution by municipalities would be enough; in fact we included the distribution by age and sex just because we wanted to get an age–sex stratified sample, but it was not strictly necessary to the method. On the other hand, an alternative method to achieve the same objectives could be considered: simply choosing a control group randomly among the whole set of patients admitted to all hospitals in this area, independently of residence. From a theoretical point of view it would produce the same results, and it seems an easy procedure if a relatively small area and/or a few hospitals are involved. However, when the study base is large and the number of centres is high it becomes difficult to apply in practice. Finally, the procedure could be reasonably applied in a study with a secondary base, provided that cases are ascertained in a relatively large set of hospitals covering a substantial part of a well-defined population. It may be considered as an alternative when studying environmental factors or, generally, when matching cases and controls by hospital is to be avoided. Assessment of validity of the results obtained must be considered in the light of comparability principles.

## Acknowledgements

The authors are grateful to other members of the Study Group of Pleural Mesothelioma: J Ramírez, S Hernández, A Calleja, R Panadès, F López, D Turuguet, and MJ Bleda (Barcelona); and A Escolar, M Beltrán and JE González-Moya (Cádiz).

## References

[1] Cole P. The evolving case-control study. *J Chron Dis* 1979;**32:**15–27.

[2] Miettinen OS. The 'case-control' study: valid selection of subjects. *J Chron Dis* 1985;**38:**543–48.

[3] Wacholder S, McLauglin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I. Principles. *Am J Epidemiol* 1992; **135:**1019–28.

[4] Miller AB. Hospital or population controls? It depends on the question. *Prev Med* 1994;**23:**263–66.

[5] Mollo F, Magnani C. European multicentric case control study on risk for mesothelioma after non-occupational (domestic and environmental) exposure to asbestos. *Med Lav* 1995;**86:**496–500.

[6] Miettinen OS. Matching and design efficiency in retrospective studies. *Am J Epidemiol* 1970;**91:**111–17.

[7] Grupo de Estudio del Mesotelioma en Barcelona (GEMEBA). Mortality from pleural mesothelioma in the province of Barcelona. (SPA). *Med Clin (Barc)* 1993;**101:**565–69.

[8] Poole C. Exposure opportunity in case-control studies. *Am J Epidemiol* 1986;**123:**352–58.

[9] Wacholder S, Silverman DT, McLauglin JK, Mandel JS. Selection of controls in case-control studies. II. Types of controls. *Am J Epidemiol* 1992;**135:**1029–41.

[10] Flanders WD, Boyle CA, Boring JR. Bias associated with differential hospitalisation rates in incident case-control studies. *J Clin Epidemiol* 1989;**42:**395–401.