3. THE CASES—A MULTIVARIATE APPROACH

3.1 Introduction

Since a worker is seldom subject to a constant exposure, there is a need to account for the fact that the workers have been at several work sites and for various lengths of time during the follow-up period. We have applied multivariate analysis as a screening device to select for further studies those work sites or job combinations that exhibited a deviating cause-of-death pattern. Thus, the proportional distribution of causes has been related to whether or not a worker had been in a certain job and also to the length of exposure (Table 3.1).

The AID (Automatic Interaction Detection) method and its supplement, the THAID version, are exploratory methods for the classification of individuals into disjoint, homogeneous groups with respect to a dependent variable. These groups result from separate as well as interaction effects in the 'explanatory' variables, which should either be categorical or categorized. The AID version requires a metric dependent variable, ie it should be at least on an interval scale, while the THAID method was designed to handle the nominal case. Despite these methods usually require large data sets they seem to be applicable to the data furnished by this study where the aim, at this screening stage, is to explore variations in the cause of death pattern among workers exposed at various sites within the smelter. In this way we will try to identify work places or combinations of work history where unusual mortality patterns prevail, e.g. where an extraordinary low or high proportion of workers died from a specific cause, e.g. cancer.

3.2 The techniques

The sub-groups in the AID analysis are constructed through a stepwise application of one-way analysis of variance techniques where the original sample is subdivided through a series of dichotomous splits (Fig. 3.1). The criterion for these successive splits is to select, at each stage, the explanatory variable which results in the greatest between-groups sum of square in the dependent variable, y,

$$BSS = \sum_{i=1}^{2} N_i(\bar{y}_i - \bar{y})^2 = \frac{N_1 \cdot N_2}{N_1 + N_2}(\bar{y}_1 - \bar{y}_2)^2$$

Table 3.1

Dependent and independent variables used in multivariate analysis

Dependent variable	Classification	Independent variable	Classification
Overall distribution	1 = all cancer 2 = circulatory dis. 3 = violent deaths 4 = other causes	9 main work sites	0 = unexposed 1 = exposed for at least some time
Lung cancer (See Fig. 3.3)	1 = lung cancer 0 = other cancers	work sites 1-28 as shown in Table 1.1	0 = unexposed 1 = 0-11 months 2 = 12-59 months 3 = 5 years or more
Cancer site distribution (See Fig. 3.4)	1 = lung cancer 2 = digestive cancer 3 = other cancers	9 main work sites	-"-

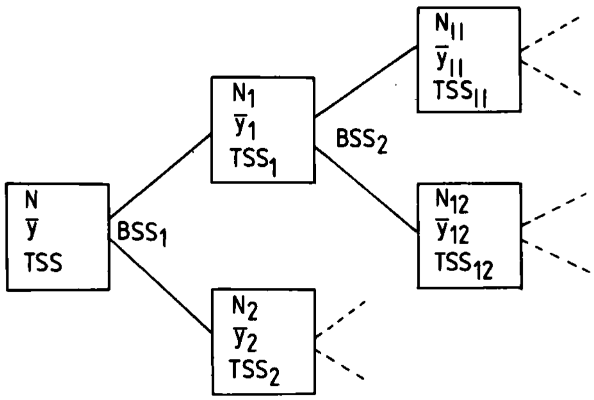

Fig. 3.1. A schematic description of the AID-method. N = group size, TSS = total sum of square, BSS = between groups sum of square.

The chosen predictor (in this case work site) is the one which gives the maximum reduction in total variation, i.e. which results in the largest BSS for each split.

For a split to be considered as important, ie warranting further sub-division, it should reduce the 'unexplained' variation with a certain predetermined fraction of the original variation and not construct groups of small sizes. When these restrictions are no longer met, there is no explanatory variable powerful enough and the hierarchic 'tree' process terminates.

While the AID method uses a group mean to 'predict' the outcome, the THAID version focuses on the distribution shape of a criterion. The sequential partition is analogous to that in AID but, in view of the metrics lacking in the dependent variable, variance reduction would be an improper split criterion. We have used the Delta-statistic,

$$\delta = N_1 \sum_{j=1}^{G} |p_j - p_{1j}| + N_2 \sum_{j=1}^{G} |p_j - p_{2j}|$$

on the basis of which the dichotomy is identified for which the weighted probability distribution over G classes differs maximally from the original group and hence from each other. Thus in a group where 20, 50 and 30% of all deaths are due to cancer, circulatory diseases and other causes respectively, we have $p_1 = 0.20$, $p_2 = 0.50$ and $p_3 = 0.30$. Suppose that this group of say 500 individuals is split into two groups of 200 and 300 individuals, with an over-representation of cancer deaths (40%) in the first group, we may have

$p_{11} = 0.40$, $p_{12} = 0.40$, $p_{13} = 0.20$, which will give $p_{21} = 0.067$, $p_{22} = 0.567$ and $p_{23} = 0.367$. Thus,

$$\delta = 200(|0.20 - 0.40| + 0.50 - 0.40| + |0.30 - 0.20|) + 300(|0.20 - 0.067| + |0.50 - 0.567| + |0.30 - 0.367|) = 80 + 80 = 160.$$

It should be pointed out that large samples are required in order to avoid a tendency to capitalize too much on chance variations. To our knowledge, the AID method has rarely been used in occupational epidemiology and we think it may serve as a didactic screening method in large retrospective cohorts such as the one under consideration.

3.3 The overall mortality pattern

Fig. 3.2 displays the results of an initial THAID-analysis where the dependent variable is the proportional distribution among the 953 deaths between tumours, circulatory diseases, violent deaths and other causes. Thus, of the total group, tumours account for 26%, circulatory diseases for 51%, violent deaths for 9% and other causes for 14%. Out of the 9 main work sites within the smelter a split between those who did and did not work at the converter hall gives maximally different proportional distributions. Thus among those who did work there, 25% died from cancer, 58% from circulatory diseases and only 5% from violent causes. Among those who did not work at the converter hall, the proportion of circulatory diseases is decreased (48%), and consequently other causes increased.

Following the upper branch of the tree in Fig. 3.2, we see that among those who have not worked at the converter hall *but* who have worked at the roasters, the increase in the proportion of cancer deaths is more evident (and still more so among those who have *also* worked at the arsenic departments). Of those 55 workers, 37 died from cancer. Among those who did not work at the roaster, work at the copper smelting furnace also seems to be associated with relatively many cancer deaths. Turning to the lower branch of the tree, we see that work at the shipping department *and* at the roasters displays a similar pattern while work at the anode furnace seems to produce the opposite results, with a high proportion of deaths due to circulatory diseases.

3.4 Lung cancer mortality

We have chosen to demonstrate the use of the AID method as a means of finding which work sites are associated with a specific cancer form, lung cancer. It is obvious, that the roasters again play a significant role. Thus, among those 80 workers who worked at the roasters and died from cancer, 43% of the deaths were due to lung cancer while the corresponding figure among those who did not work at the roasters was 25%. Work at the mechanical department or at the arsenic departments also seems to be associated with an increased proportion of lung cancer deaths, while work at the converter hall again seems to be negatively associated with lung cancer. Fig. 3.3 also gives an indication of a dose-response relation in

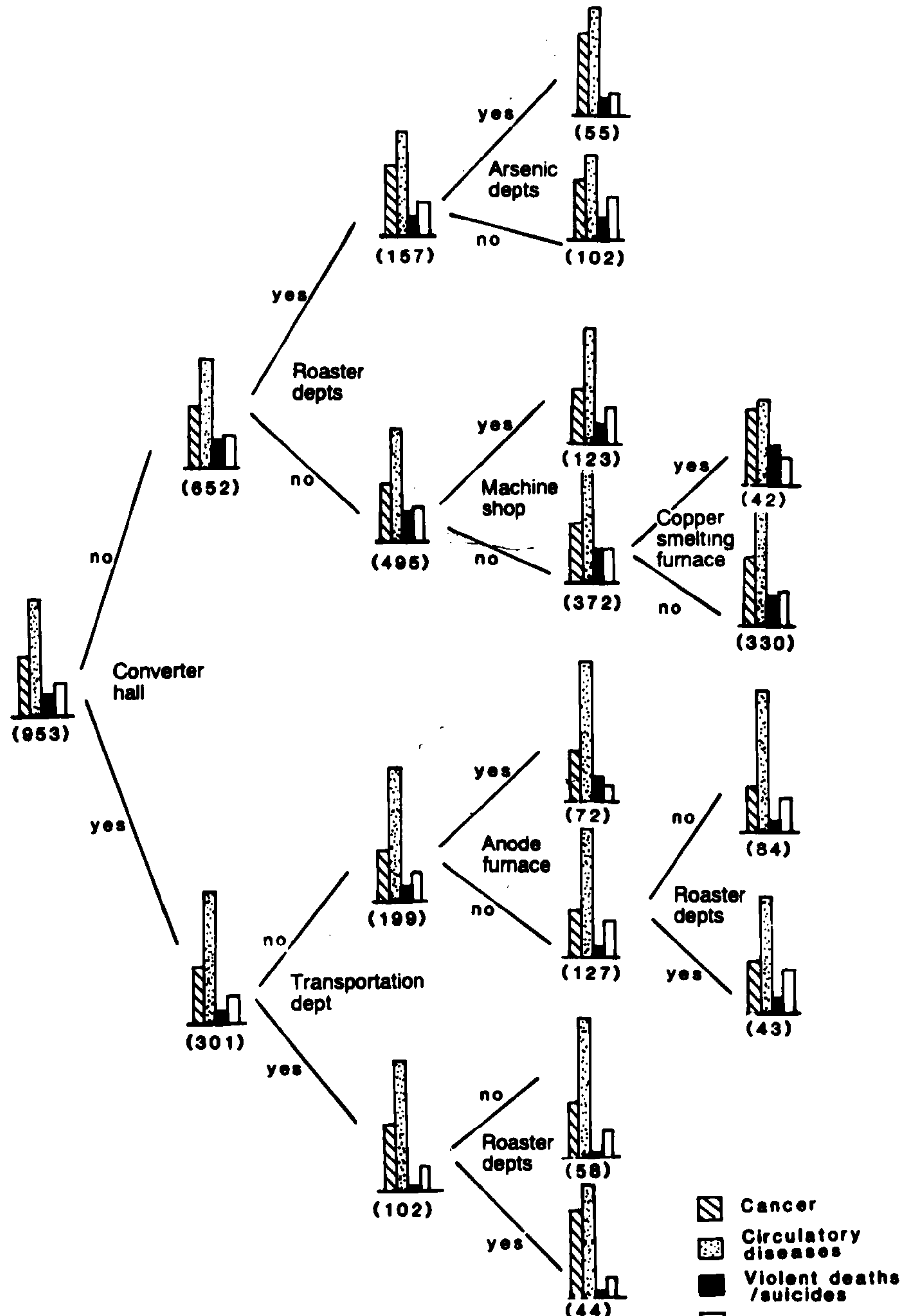

Fig. 3.2. THAID-analysis of the cause-of-death-pattern among 953 deaths by work site predictors.

that an increased length of exposure at the roasters further increases the proportion of lung cancer deaths (of the 32 workers who had been there for 5 years or more and who died from cancer, 18 had lung cancer).

3.5 Other cancer sites

Analogous with the previous THAID-analysis of the overall mortality patterns, we have analysed the proportional distribution among cancer deaths. We have used only three categories for this purpose: cancer in respiratory organs, cancer in digestive organs and cancer of other sites (Fig. 3.4).

The roasters again dominate the picture. Work at the machine shop (lower branch in Fig. 3.4) is associated with a relatively high proportion of cancer in digestive organs while the opposite pattern prevails among roasters. The two extreme groups are further illustrated in Fig. 3.5.

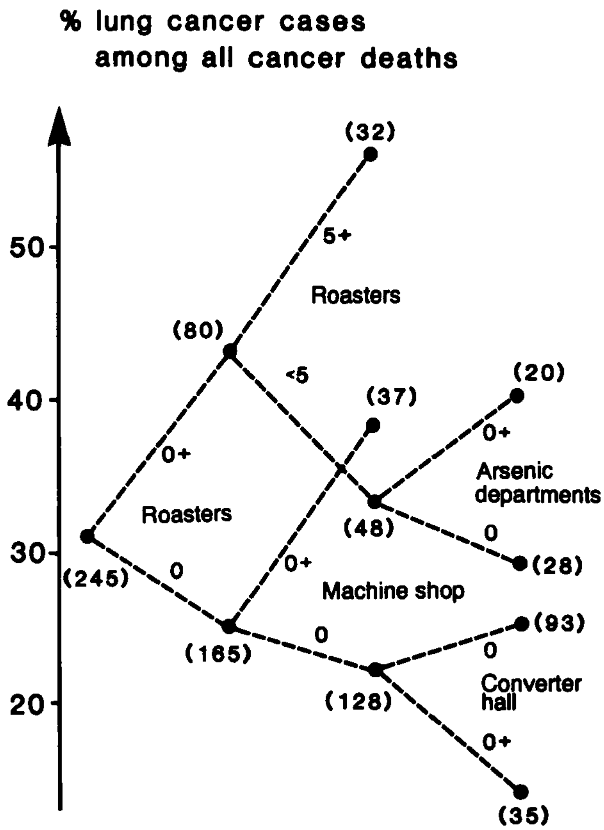

Fig. 3.3. AID-analysis of the proportion of lung cancer cases among all cancer deaths using work site and length of exposure as predictors (figures within brackets denote number of cases).

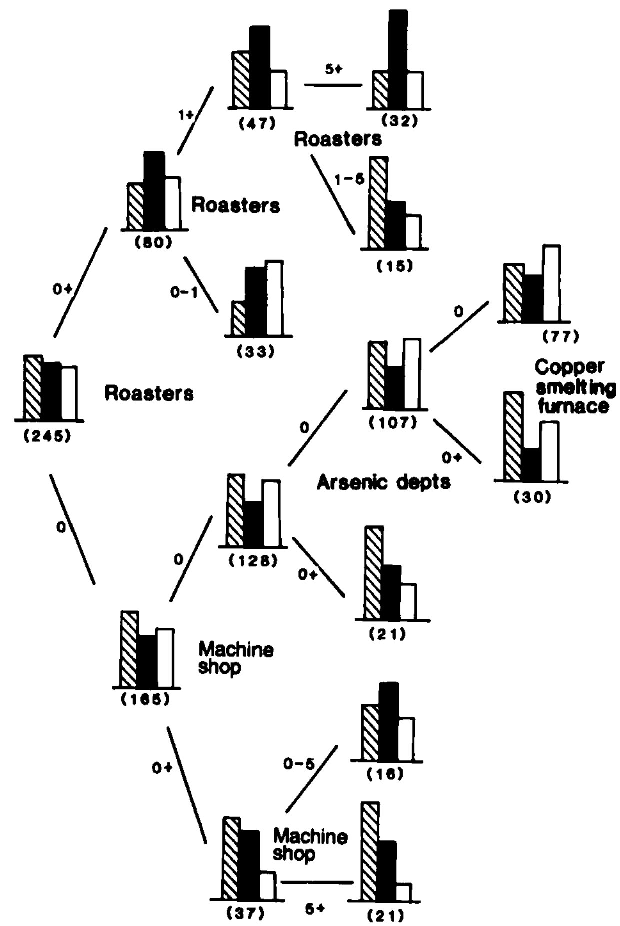

Fig. 3.4. THAID-analysis of the distribution of three cancer sites using work sites as predictors. ▧ Digestive organs; ■ Respiratory organs; □ Other sites.

3.6 Conclusions thus far

A multivariate analysis of the mortality pattern relative to work site combinations again revealed that work at the roaster departments dominates the picture. Among those who did *not* work at the converter hall *but* at the roasters *and* at the arsenic departments, 37% died from cancer. There is also an apparent dose-response relationship between lung cancer mortality and work at the roasters. It is obvious that proportionate mortality analysis does not take into account the absolute risk of dying in the population studied. Thus, it is possible that the mortality rate among Rönnskär workers is less than that of the general population even though there was a disproportionate number of cancer deaths among them. A high proportionate ratio may thus be due either to an excess of cancer incidence or to a deficit of other causes.

The fact that we have been able to locate proportionate cause-specific mortality to certain jobs, however, may serve as a first guide in the identification of occupational hazards and also for the purpose of revealing which sites ought to be studied more in detail. It should be mentioned, however, that this is a tool for priority decisions, not for acquitting sites from blame.

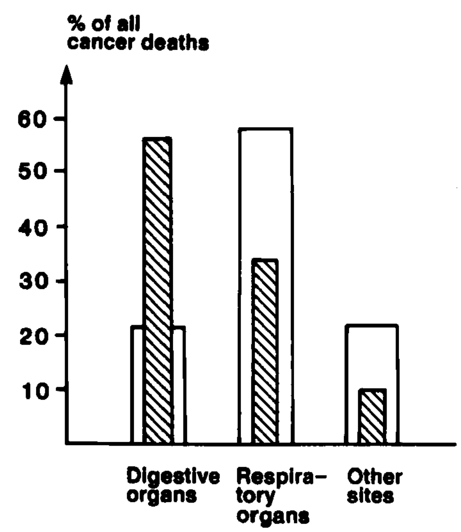

Fig. 3.5. Extreme groups split off in the THAID-analysis of the distribution of cancer sites by work-site predictors. □ 32 workers exposed to roaster work ⩾ 5 years; ▨ 21 workers who never worked at the roasters but at the machine shop ⩾ 5 years.