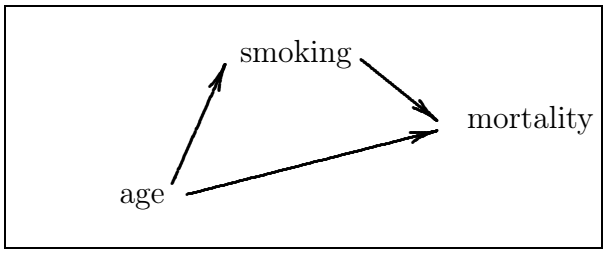
Biostats VHM 811 Course
Fall 2009, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to Home assignment 2

The study of health effects of smoking among British doctors is well documented through many published papers on the study design and the results for successive follow-up periods from study initiation in 1951 (10 years, 20 years etc.). The initial questionnaire was followed up by several subsequent questionnaires that allowed the tracking of smoking habits and health issues over time. It is instructive to follow the systematic reporting of the study over several decades. The assignment uses data for the first 10-year period because detailed raw data were published in the paper Doll, R. & Hill, A. B. (1966), Mortality in British doctors in relation to smoking: Observations on coronary thrombosis, In *Epidemiological Study of Cancer and other Chronic Diseases*, National Cancer Institute Monograph **19**, 205–68. Another part of this dataset was used as an example (Table 15.2) in Rothman *et al.* (2008) and is included in the Stata documentation (of the `ir` command).

1. Study design and causal diagram

The description of the study makes it clear that cohorts of different categories of smokers (including non-smokers) were followed over time to obtain records of mortality due to the specific heart disease. The study is therefore a prospective cohort study. The causal relation between the factors is essentially determined by the fact that age can affect (the probability of) smoking, whereas the reverse relation would be meaningless. Both age and smoking are potential risk factors for mortality, leading to the following causal diagram.



2. Crude measure of association between smoking and mortality

For a cohort study, the outcome can be measured as a rate or as a count, depending on the information available. If the follow-up time differs between subjects (because they do not all remain at risk during the entire follow-up period), a rate-based outcome more accurately reflects the risk. In this study, the subjects may exit the study cohort due to death, of either heart disease or other reasons, or due to loss of contact with the researchers carrying out the study. The preferred measure of disease is an incidence rate, and for comparison of exposure groups the natural choice of measure of association is the incidence rate ratio (IR), although an incidence rate difference (ID) might also be used. The crude incidences in the smoker and non-smoker groups are 4.1 and 2.0 deaths per 1000 person-years, giving a incidence rate difference of ID = 2.06 deaths per 1000-years and an incidence rate ratio of IR = 2.01. The statistical significance of the association may be assessed by looking at its associated confidence interval, or by referring to the test statistic for the hypothesis of no difference between groups. The 95% (exact) CI for IR is given in Stata as $(1.65, 2.46)$, and as the value 1 is well outside the interval, we have evidence at the 5% significance level (and lower) of a difference between the two groups. Stata reports a P-value <0.00005 for a two-sided exact test. Thus, the crude measure of association indicates a much increased mortality among smokers of the heart disease.

3. Crude measure of association between age and mortality

For the variable age grouped into 10-year intervals, the approach used above for comparison of IR between two groups is not directly applicable. A simultaneous comparison of multiple groups is not feasible with methods covered in the course (a Poisson regression would be the obvious choice), but we can create multiple 2-group comparisons across age groups. Two approaches are possible: comparisons with a fixed ("baseline") age group, and dichotimization at different cut-points (and comparison of risk between age groups above and below the cut-point). The age group 25–34 has no deaths, and is therefore a poor choice for a baseline group; we use instead a middle age group as the baseline group: 55–64 years, which has both a substantial number of deaths and a fairly large follow-up time. The table below gives IR and associated confidence intervals for both baseline and cut-point comparisons of age groups.

	Age group						
Parameter	25-34	35-44	45-54	55-64	65-74	75-84	85+
IR vs. baseline	0	0.072	0.36	1	2.17	3.44	4.19
95% CI	(0,0.013)	(0.050,0.10)	(0.29,0.44)	n/a	(1.83,2.57)	(2.87,4.13)	(3.02,5.73)
IR at cutpoint	n/a	∞	24.7	11.9	8.86	7.70	7.22
95% CI	n/a	(57.8,∞)	(17.7,35.5)	(10.1,14.2)	(7.80,10.1)	(6.67,8.86)	(5.29,9.66)

All IR comparisons, both with a baseline category and across different cut-points, give evidence of a strong relation between age and mortality (of the heart disease). Biologically this is no surprise: mortality of many diseases (including heart disease) increases with age. We see here that the IR increases when a higher age group is compared to the baseline, and we also see that no matter the cut-point, the older category is always associated with larger risk. Statistical significance is less relevant here where the association is so strong, but we could refer to the fact that all confidence intervals in the table exclude the value 1 (which would correspond to the same risk in the two groups).

4. Confounding

The causal diagram shows that age could be a confounder for the relation between smoking and mortality. On the other hand, smoking is on the pathway between age and the outcome — an intervening (or intermediate) variable — and can therefore not be a confounder for the relation between age and mortality. There are 3 further conditions to check for whether age is indeed a confounder for the former relation (excluding here the issue of an interaction effect which will be discussed in the next question):

1. *Substantial effect change.* The Mantel-Haenszel estimate of the effect of smoking stratified on age groups is 1.29 with an associated 95% CI of (1.06,1.57). The effect change is quite large: $(2.01-1.29)/2.01 = 36\%$. (*Note:* It could be argued to omit the lowest age category, 25–34 years, because it contains no deaths due to heart disease. If this reflects that it is biologically implausible for people in this age category to die of heart disease, the age category should be excluded. If on the other hand, it reflects a very low death risk and an insufficient sample size to actually observe any deaths in this age group, these data could be included. In any case, it is probably of interest to investigate the impact of this age group on the results. The M-H estimate is not affected by the 25–34 year age group, but without these data the crude estimate equals IR = 1.71, corresponding to a change of $(1.71-1.29)1.71 = 25\%$, still a substantial effect change.) Age does affect the IR estimate enough to qualify as a confounder.

2. *Association with outcome.* The crude associations were computed in Question 3, but the assessment should really be done among the exposure-negative subjects, i.e., the non-smokers. It is intuitively obvious and evident from an inspection of the data that the association between age and mortality exists also among the non-smokers. For reference, we repeat the calculations from Question 2 for IR relative to a baseline category (again, the 55–64 age group) although such detail is hardly necessary.

	Age group						
Parameter	25-34	35-44	45-54	55-64	65-74	75-84	85+
IR vs. baseline	0	0.022	0.23	1	2.21	4.32	6.60
95% CI	(0,0.053)	(0.003,0.086)	(0.11,0.47)	n/a	(1.26,3.87)	(2.51,7.48)	(2.96,13.7)

   The table shows that the age effect may actually be stronger among non-smokers than overall; the estimates are mostly further away from 1, although the confidence intervals are also wider. The impact of age on mortality from heart disease is evident.

3. *Association with exposure.* In a cohort study, we should consider the crude association between potential confounder and exposure. The data provided on mortality and person-years for the different age and smoking groups does not directly allow to assess this association. However, the percentages of non-smokers in the different age categories show that smoking does vary with age. Statistical significance can be assessed by converting the percentages of non-smokers to approximate counts of smokers and non-smokers in each age category, and performing a chi-square test in the resulting 2×7 table (*note:* this table is shown in Question 6). This analysis gives $X^2 = 929$ with 6 df, and the strong significance should be convincing evidence of the association despite the doubts on the exactness of the numbers (as noted in the assignment, it was not clear from the information given whether the total and the proportion of non-smokers included ex-smokers; ex-smokers are however likely to be a fairly small group (in 1951)).

We conclude that age meets all conditions to be a confounder for the effect of smoking on the mortality by heart disease.

5. Interaction

The Mantel-Haenszel analysis for the effect of smoking after stratification on age groups included a test for homogeneity of the incidence rate ratios: $X^2_{\text{hom}} = 10.74$ corresponding to $P = 0.057$ in a χ^2-distribution with 5 degrees of freedom (the 25–34 age group does not contribute to this statistic). There is some indication of an interaction, although the P-value does not meet the formal significance level of 0.05. The stratum-specific values of IR are listed in the table below.

	Age group						
Parameter	25-34	35-44	45-54	55-64	65-74	75-84	85+
IR (smoking)	n/a	5.07	2.12	1.28	1.25	0.99	0.74
95% CI	n/a	(1.30,43.6)	(1.17,4.22)	(0.86,1.97)	(0.84,1.91)	(0.67,1.49)	(0.37,1.62)

The table shows the IR to be decreasing as a function of age. Furthermore, only for men below 54 years is there a statistically significant association between smoking and mortality. The nearly significant X^2 statistic and the clear pattern in the IRs across age groups lead to the conclusion that there is indeed an interaction between age and smoking in their effect on mortality by heart disease.

Smoking is a risk factor up till mid-age (54 years in this dataset), and after any added risk of heart disease associated with smoking is questionable.

In summary, age primarily acts as a moderator variable (i.e. interacting factor; Section 13.11.9) for the relation between smoking and mortality by heart disease. The relevant measures of association are the age-specific IRs in the table above. In addition, the measures of association for the relation between age and mortality should be stratified by smoking. The crude and non-smoker associations indicated clearly increasing risks with age, and the same tendency is seen in the raw data for the smokers as well. We could compute IRs relative to a baseline category as in the previous questions, but as these estimates are not the primary focus of the analysis, we won't go into details.

6. Control of confounding by (future) study design

For this question we will consider a confounding by age in the relation between smoking and mortality by heart disease in a cohort study. The two procedures for control of confounding by design are restriction and matching.

Restriction here means restricting the study to subjects within a fairly narrow age range. The division of age into 10-year intervals could even leave some residual confounding after adjustment for age groups, so the age range would have to be fairly narrow to avoid any confounding. The limited scope would probably make it of little interest to carry out a study with only a fairly narrow age range. As the effects of smoking were strongest among the younger to mid-age subjects, these would probably be the age ranges of primary interest.

Frequency matching on age means that the non-exposed group (non-smokers) should be selected to have roughly the same distribution of age as the exposed group (smokers). Ideally this equal distribution should be of time at risk, but as person-years are difficult to control one would instead try to match the number of subjects in the different age groups. The approximate numbers of smokers and non-smokers in the different age groups are shown in the table below (these are the counts involved in the chi-square statistic for the association between age and smoking in Question 4).

	Age group						
Group	25-34	35-44	45-54	55-64	65-74	75-84	85+
smokers	6895	7457	6252	3683	2460	1224	149
(%)	24.5	26.5	22.2	13.1	8.8	4.4	0.5
non-smokers	2335	1431	853	387	228	162	28
(%)	43.1	26.4	15.7	7.1	4.2	3.0	0.5

It is seen that the distribution of smokers has a larger proportion of older men than the distribution of non-smokers. Therefore, a frequency-matched sample of non-smokers would have to include less younger men and more older men in order to achieve the same age distribution as among the smokers. It would also be possible to revert the role of exposure and non-exposure for the matching, and thereby match the distribution of smokers to the distribution of non-smokers. As there are more smokers than non-smokers in the population, this would be easier to manage in practice.