33
Nested case-control studies

Any cohort study can be used to generate a case-control study by sampling the cohort for controls to use in place of the full cohort. The case-control study is then said to be *nested* in the cohort study. For each case the controls are chosen from those members of the cohort who are at risk at that moment, in other words from the risk set defined by the case. Although the idea of nested case-control studies predates Cox's method for the analysis of cohort studies, the design and analysis of such studies has been greatly clarified by the ideas of partial likelihood and risk sets.

33.1 Reasons for using a nested case-control study

The main reason for using a nested study is to reduce the labour and cost of data collection by collecting complete data only for those subjects who are chosen for the nested study. For example, in cardiovascular epidemiology the habitual energy expenditure of subjects has been measured using detailed diary records in which subjects record their physical activities in 15-minute blocks. Coding these diary records into energy expenditure is time consuming and expensive, but with a nested case-control design this conversion is only needed for the cases and their controls. Similar considerations apply to coding diary records in cohort studies in nutritional epidemiology, and to expensive laboratory analyses on biological specimens — these can be collected for all subjects in the cohort but "banked" and analyzed only for cases and their controls.

Another use of nested case-control studies is when an on-going cohort study is to be used to address a question about an exposure or confounder not measured in the original design. Data collection can be restricted to those subjects in a nested study. For example, suppose that routine health service monitoring data shows differences in mortality between groups of patients but, because information is not available on important confounders, it is not possible to exclude confounding as an explanation. A more detailed abstraction of medical records in a nested case-control study could make it possible to measure the confounders in the nested study and hence to control for them.

The final reason for using a nested case-control study is to avoid the computational burden associated with time-dependent explanatory vari-

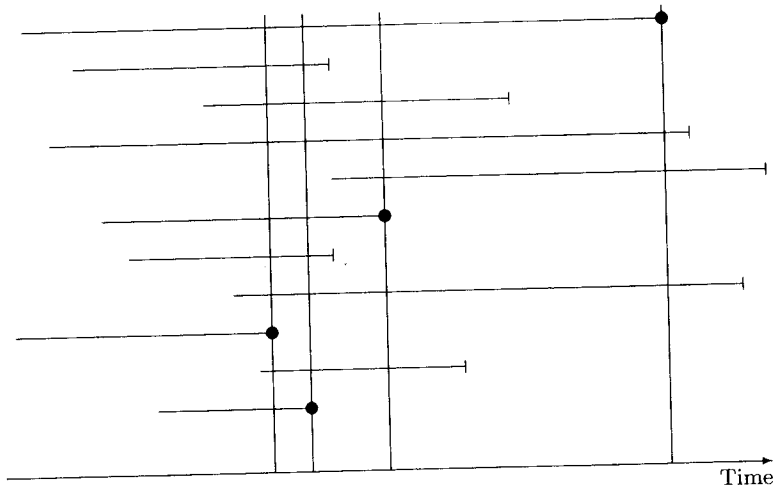

Fig. 33.1. Definition of risk sets.

ables. This problem was discussed briefly in Chapter 31, where we indicated that a natural design for such studies is to randomly sample the *risk sets* on which a full analysis by Cox's method would be based. In this chapter we discuss this suggestion in more detail.

33.2 Sampling risk sets

In nested case-control studies, controls are drawn for each case from the corresponding risk set. Fig. 33.1 shows the risk sets for a follow-up study of eleven subjects, four of whom fail. Corresponding to each of these four events is a risk set containing all those subjects under study at the moment of event occurrence — that is, all subjects whose observation lines cross the relevant vertical. To select controls we ignore the case and choose a random sample of the remaining subjects in the risk set. Sampling of a risk set must be carried out independently both of the sampling of other risk sets and of any later failure or censoring of its members.

Exercise 33.1. What are the sizes of the four risk sets? Indicate how you would select a single control for each case.

In the analysis of the full cohort study using Cox's method, each of the events contributes a term of the form

$$\log\left(\theta_{(\text{for case})} \Big/ \sum_{\text{Risk set}} \theta\right)$$

to the log partial likelihood. When the risk sets are sampled this becomes

$$\log\left(\theta_{(\text{for case})} \Big/ \sum_{\text{Case-control set}} \theta\right),$$

which is identical to the log likelihood contribution of a matched case-control set in a conditional logistic regression analysis.

CAN THE SAME SUBJECT BE INCLUDED MORE THAN ONCE?

In the procedure for sampling risk sets described above the same subject can be selected as a control more than once and may eventually become a case. This will not happen very often for rare events but when it does it should be permitted. Any intervention in the sampling procedure to prevent its happening violates the requirement for independent sampling of risk sets.

A second aspect of this question is illustrated by the fourth subject shown in Fig. 33.1 who belongs to all four risk sets. If this subject is drawn as a control in one of these risk sets it is tempting to use him or her as an extra control in the other sets. Including a subject in all samples for which he/she is eligible represents an extremely *interdependent* method of sampling risk sets. The result is that the successive terms which contribute to the partial likelihood are no longer independent — each term does not contribute quite as much *new* information as it appears. When this dependence is taken into account the expected gain in precision as a result of multiple use of controls largely evaporates. However, there may be other advantages. One is that, because controls are no longer tied to a particular risk set, they can be randomly selected at the time of recruitment into the cohort study. This design has been called a *case-cohort* study, and some logistic advantages have been claimed. One situation in which it might be considered is for studies in which several different types of event are of interest — for example, occurrence of several different cancers. Independent sampling of risk sets leads to a different set of controls for each type of event while the case-cohort design allows a single control sample to be used for all outcomes. Against this must be weighed the fact that a more complex analysis is required to take account of the interdependency in the sampling of controls.

HOW MANY CONTROLS?

If there are m times as many controls as cases, the precision of the case-control study compared to the cohort study is given by

$$\frac{\text{SD of estimate from case-control data}}{\text{SD of estimate from entire study base}} = \sqrt{1+\frac{1}{m}}.$$

This formula applies to the simple situation where the exposure effect is small and there is no control for confounding, but it can also be used as a rough guide more generally. Since $\sqrt{1+1/m}$ is only slightly greater than 1 for $m > 5$ little accuracy is lost by taking five or at most ten controls for each case, rather than the whole risk set.

33.3 Matching

In an occupational study of lung cancer, smoking will be a strong confounder, and the comparisón of occupational groups should therefore be controlled for smoking. An overall sample of (say) five controls per case could lead to a very different ratio within smokers and non- smokers. Since there will be many more cases among the smokers than among the non-smokers it is likely that there will fewer than five controls per case among smokers and many more than five per case among non-smokers. In such cases it would be better to match controls to cases with respect to smoking habits. Of course, this requires that smoking data are available for the entire cohort. The contribution to the log likelihood now becomes

$$\log\left(\theta_{(\text{for case})} \Big/ \sum \theta\right)$$

where the $\sum\theta$ denominator refers to summation over the case and the matched controls. Matching the controls to the cases on smoking does not allow estimation of the smoking effect, but when smoking is a confounder this need not concern us.

⋆ 33.4 Counter-matching

In the previous section we discussed the situation where the values of the confounding variables are known for all subjects in the cohort and a nested case-control study is used to reduce the cost of measuring the exposure. Matching controls to cases on the confounding variables can improve the precision of the comparison of exposure groups although, as a side-effect, the effects of the confounding variables cannot be estimated. What about the opposite situation in which the exposure variable is measured for all subjects in the cohort and a nested case-control study is used to reduce the cost of measuring the confounding variables? In this case it would be disastrous to match the controls to the cases on exposure since we would then be unable to estimate the effect of exposure. However, the information available for the full cohort can still be used to sample controls more efficiently.

To illustrate this we consider first the case in which all subjects are classified as exposed or unexposed. For any particular risk set let the numbers of exposed and unexposed subjects be N_1 and N_0 respectively, and suppose we are to draw m controls. The nested case-control set will

contain $n = m + 1$ subjects (the case plus m controls). Let the split of these n subjects between exposed and unexposed be n_1 and n_0. When controls are drawn by simple random sampling of the risk sets this can produce a very uneven split of exposed and unexposed subjects and lead to inefficiency. The efficiency of the study can be improved by fixing the split in advance — usually to be 50:50.

For example, suppose that there are 10 exposed and 100 unexposed subjects in the risk set and we wish to select a sample of 5 exposed and 5 unexposed, including the case which defines the risk set. If the case is exposed this means we need 4 exposed controls and 5 unexposed controls. If the case is unexposed we need 5 exposed controls and 4 unexposed controls. For a sample of one exposed and one unexposed an exposed case will always be paired with an unexposed control and an unexposed case with an exposed control. It is from this that the term *counter-matching* is derived.

When sampling in this way the contribution of each risk set to the partial log likelihood must be adjusted to reflect the fact that the exposure distribution in the sample is different from the exposure distribution in the risk set. The modified log partial likelihood contributions take the form

$$\log\left((W\theta)_{(\text{for case})} \Big/ \sum_{\text{Case-control set}} (W\theta)\right),$$

where W are *risk weights* for each subject which compensate for the sampling. These weights take the values

$$W = \begin{cases} N_1/n_1 & \text{for an exposed subject} \\ N_0/n_0 & \text{for an unexposed subject.} \end{cases}$$

Note that the choice of weight depends only on exposure status and not upon whether the subject is a case or a control.

Exercise 33.2. What are the weights for exposed and unexposed subjects in a risk set with $N_1 = 10$ exposed subjects and $N_0 = 100$ unexposed subjects, in a 1:1 counter-matched study?

Exercise 33.3. For the special case where there are no confounders θ takes the value 1 for an unexposed subject and the value ϕ for an exposed subject, where ϕ is the (multiplicative) exposure effect. Show that, using the correct weights, the partial log likelihood contribution for the 1:1 sampled set is identical to the contribution of this risk set to the full cohort analysis.

The design and analysis extends readily to the case where there are more than two exposure categories. If the risk set contains N_i subjects in exposure category i and the case-control set is to contain n_i, then we draw either $n_i - 1$ or n_i controls at random according to whether or not the

case falls into this category. The risk weight for subjects in this category is N_i/n_i.

The same design and analysis may be used when exposure data is difficult or expensive to collect, but in which we have a surrogate measure available for all subjects. If exposure is rare, it makes sense to use the surrogate exposure measurements to construct a more efficient nested study in which there is a more even split between exposed and unexposed subjects. In a 1:1 study, for example, a case classified as exposed by the surrogate measure would be paired with a control classified as unexposed, and a case classified as unexposed paired with a control classified as exposed. Remembering that in the 1:1 study only exposure discordant pairs are informative for the estimation of the exposure effect, this design is more efficient since it should increase the number of such pairs.

An area in which counter-matching by surrogate exposure measurement could prove particularly useful is pharmacoepidemiology. Exposure to any one drug is rare and can usually only be ascertained after detailed checking of medical records. However, a simple questionnaire might be very successful at identifying a subgroup particularly likely to have taken the drug of interest. The nested case-control study should contain all subjects in the group likely to have taken the drug, and a random sample of the remainder. With this design, the introduction of the correct risk weights into the partial likelihood analysis provides a valid estimate of the drug effect.

★ 33.5 Two-stage sampling of controls

Both matching and counter-matching require that some information is available for all subjects in the cohort. The general rule is that, when this concerns a confounder we should consider using it for matching controls to cases while, if it concerns an exposure of interest, we should consider counter-matching.

Similar ideas may be useful even when we have no such data for the full cohort or, indeed, in a conventional case-control study. The information to be used in the final matching or counter-matching is collected in an initial study but complete data collection is only followed through in a subsample. This is known as a *two-stage* case-control study.

Solutions to the exercises

33.1 The risk set for the first event contains 10 subjects, the others contain 9, 7, and 4 subjects respectively. A control for the first case is selected at random from the remaining 9 subjects in the risk set. Similarly the remaining controls are sampled at random from the 8, 6, and 3 eligible subjects in the remaining risk sets.

33.2 In the 1:1 counter-matched study each set contains $n = 2$ subjects,

1 exposed and 1 unexposed so that $n_1 = n_0 = 1$. The risk weights used in the analysis are therefore,

$$W = \begin{cases} 10 & \text{for an exposed subject} \\ 100 & \text{for an unexposed subject.} \end{cases}$$

33.3 Suppose the case is exposed. Using the whole risk set the contribution to the log partial likelihood is

$$\log\left(\frac{\phi}{10 \times \phi + 100 \times 1}\right).$$

Using the 1:1 counter-matched design, the contribution to the partial log likelihood is

$$\log\left(\frac{(10\phi)}{(10\phi) + (100)}\right) = \log(10) + \log\left(\frac{\phi}{10 \times \phi + 100 \times 1}\right).$$

These two expressions are the same except for a constant term, $\log(10)$, which does not depend on ϕ and can be ignored. The same is true when the case is unexposed.