Appl. Statist. (1982),
31, *No.* 3, *pp.* 293–297

Efficient Assessment of Confounder Effects in Matched Follow-up Studies

By Alexander M. Walker

Harvard School of Public Health, Boston, MA 02115, USA

[Received July 1981. Revised May 1982]

Summary

When matched sets of individuals are entered into follow-up studies, data collection on many covariates can be eliminated without loss of relevant information. In a proportional hazards model with risk sets restricted by a matching factor, no individual's data enter into the likelihood function unless some member of his matched set experiences an event. In many epidemiological studies, the probability of an event in any given matched set is small; as a result most sets, having no event, make no contribution to effect estimates, and data collection in these sets can be avoided.

Keywords: MATCHING; PROPORTIONAL HAZARDS

Prentice and Breslow (1978) have noted that Cox's proportional hazards model (1972) can be readily adapted to stratified analysis. Those authors developed the observation in terms of its implications for case-referent studies. The purpose of this note is to point out a practical consequence of stratification for matched follow-up studies.

Matching in follow-up studies is of use when there are variables which are distributed disproportionately in the groups to be compared and which are likely to be predictive of an outcome, but whose effects are of no intrinsic interest. Particularly when such variables have many possible realizations, the efficiency of a study is improved by assuring that at every level of the uninteresting predictive factors there is some heterogeneity in terms of other factors which are of interest. For example, one may wish to control for genetic effects by studying a series of twins who differ in regard to some non-genetic characteristics, or one may wish to control simultaneously for sex, chronologic age and secular trends by matching individuals by sex, year of birth and year of observation.

Assuming that the object of study is the relationship between an outcome event and a particular exposure of interest, I will use the terms "covariate" and "confounder" as follows: "covariate" refers to any putative predictor of outcome included in the analysis including the exposure of interest; "confounder" refers to any covariate other than the exposure of interest.

The essence of the proportional hazards model is to consider the incidence density, or hazard, associated with a set of covariates to be the product of an underlying hazard rate, which may vary with time, and a multiplier which is commonly a log linear function of the covariates:

$$\lambda(t) = \lambda_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}), \tag{1}$$

where T is the vector product operator, $\lambda_0(t)$ is the underlying hazard at time t, $\mathbf{Z}$ is a covariate vector of the individual, independent of time, and $\boldsymbol{\beta}$ is a vector of coefficients common to all individuals. Here, $\boldsymbol{\beta}$ is obtained by maximizing a partial likelihood function (Cox, 1975) in which each individual in whom an event occurs at time t is compared to the set of individuals in whom the event might have occurred. The likelihood contribution of the ith event is

$$\begin{aligned} l_i &= \lambda_0(t_i)\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i)/\sum_{j\in R_i}\lambda_0(t_i)\exp(\boldsymbol{\beta}^T\mathbf{Z}_j) \\ &= \exp(\beta^T\mathbf{Z}_i)/\sum_{j\in R_i}\exp(\boldsymbol{\beta}^T\mathbf{Z}_j), \end{aligned} \tag{2}$$

where R_i is the "risk set" for event i, the set of all individuals still under observation at t_i in whom no event has yet occurred before t_i. The overall partial likelihood to be maximized is $L = \Pi_i\, l_i$.

Stratification within this model is accomplished by further restricting R_i to those individuals who share some stratification factor with the individual suffering the ith event. Matching involves selecting individuals for follow-up according to a matching criterion, and using the matching criterion as the stratification factor in the analysis. Prior specification of the distribution of covariates within the matched sets does not affect subsequent estimation of $\boldsymbol{\beta}$.

This last fact is of some importance because it permits one to specify, as part of the study design, the proportion of individuals exposed in each matched set. The sets, therefore, might be termed "exposure-balanced" matched sets.

In many follow-up studies the outcome of interest is a rare event. If the matched sets are sufficiently small that the expectation of an event in any matched set is itself small, then it follows that a great many sets, having no event, will drop out of the estimation procedure altogether because they make no contribution to the likelihood function. If estimates of $\boldsymbol{\beta}$ depend only on the distribution of covariates in those matched sets in which an event occurs, then collection of covariate data is superfluous for all eventless sets. This implies that in some circumstances, follow-up studies may be effectively carried out with comparatively reduced data gathering, while still controlling for relevant confounders.

Efficient follow-up study designs suggest themselves when exposure, matching and outcome data are available through resources which do not provide information on potential confounding factors other than those used in matching. The procedure would be to form matched sets for follow-up, balancing exposed and non-exposed individuals in the matched sets according to their availability and cost of follow-up, observing their experience over time, and subsequently obtaining confounder data only on the members of those exposure-balanced sets in which an event occurs. Those sets are then entered into a proportional hazards analysis which is algebraically equivalent to the analysis which would have been carried out, had covariate data been available for the entire initial cohort.

The matched follow-up design permits considerable flexibility in the handling of time-dependent covariates. As noted before, the covariates in the proportional hazards model have values which are fixed over time (for $t>0$). Thus, the covariates entered into a customary proportional hazards analysis ought to refer to characteristics of individuals which are similarly invariant with time. However, in the matched model, in which there will frequently be only a single event per matched set, it is possible to replace the $\mathbf{Z}_j$ of expression (2) with $\mathbf{Z}_{ji}$, the vector of covariates characteristic of the jth individual at the time that the ith individual suffers an event, that is at t_i. After the occurrence of an event in individual i, covariate data can be collected for all members of the matched set, with reference to t_i. The coefficients, $\boldsymbol{\beta}$, of time-dependent factors retain their former interpretation as (1) the log relative hazards associated with the factors, when present (in the case of dichotomous factors) or (2) the changes in the log hazard associated with unit increases in factor levels (in the case of factors measured on an interval scale).

While any time-dependent factor can be controlled using this analysis, the choice of relevant times needs to be made with some care. For factors which may themselves be influenced by exposure, levels at or prior to the time of exposure may be the most meaningful independent predictors of outcome. For factors exerting a short-term effect on risk, status at the time of the event may be most important. Finally there may exist factors whose effect may be best determined for multiple times up to the event, or even accumulated (integrated) over the entire period of an individual's pre-event experience.

Selective collection of confounder data, using this or any outcome-dependent sampling procedure, requires that the quality of confounder information be unaffected by the outcome event. The proposed procedure then shares with case-control studies limitations imposed by data availability. For example, confounder information obtainable only by interview, by analysis of blood specimens, or by examination of perishable records cannot be studied when the outcome of interest is death. Nor, for example, might it be possible to control for early history in studying the occurrence of multiple sclerosis, because the recollection of crucial events may be impaired by the presence of disease.

If there is at most one failure per matched set, restricting the risk set in (2) to members of the matched set produces a likelihood function which has been proposed previously, by Breslow *et al.* (1978), for the analysis of matched case-control studies. The identity follows immediately if the matching is construed in the same way in the prospective and retrospective models; in effect the present proposal is for a case-control study in which the controls are a 100 per cent sample from population at risk (the risk set). The identity of the likelihood functions has the useful consequence that any computer program for carrying out a stratified proportional hazards analysis can be used to carry out the conditional logistic case-control analysis of Breslow *et al.* with variable matching ratios. Multiple discrete failures in a matched set pose no special problem for the matched prospective analysis, but have no precise counterpart in the case-control formulation. Multiple concurrent failure times are similarly tractable, but special care is warranted because commonly used approximations to the partial likelihood function can lead to biased estimates of regression coefficients. Farewell and Prentice (1980) discuss this issue and offer less biased, computable alternatives.

The procedure proposed here differs in an important respect from one proposed by Mantel (1973), who suggested that confounding in cohort studies could be controlled by sampling diseased and non-diseased individuals in the cohort, obtaining covariate data on those persons alone, and analysing the data as if they derived from a case-control study. The Mantel suggestion suffers from a weakness often encountered in case-control studies: if the exposure of interest is relatively rare, the power is low relative to the number of individuals for whom covariate data are obtained. Sampling in the context of exposure-matched sets, by contrast, guarantees that the number of exposed diseased and non-diseased persons will be more substantial, with an attendant increase in power. Liddell *et al.* (1977) actually implemented Mantel's suggestion with the further refinement of matched control selection. In this case, however, the matching was necessarily on a confounding factor (year of birth), and could not serve directly to increase the proportion exposed among those sampled.

Example

Walker *et al.* (1981) undertook a matched retrospective cohort study to ascertain the long-term health consequences of vasectomy. The data in Table 1 pertain to pairs of vasectomized and non-vasectomized men. These 36 pairs arose out of a cohort of 4830 vasectomized/non-vasectomized pairs of men matched from the membership files of a large group medical plan, on the basis of year of birth and calendar time of follow-up. For each pair, follow-up began when one of the pair members underwent vasectomy. There were no pairs of which both the vasectomized and non-vasectomized man suffered a myocardial infarction (MI). Clinical records abstracted for each of the 72 MI-discordant pair members yielded information on smoking and obesity, as well as on a variety of other characteristics. Men were classed not smoking if there was (a) no mention in the medical record of smoking, (b) a mention of never smoking or (c) a definite mention before vasectomy (or the corresponding date in the non-vasectomized pair member), of having stopped in the past, with no further mention of

TABLE 1

Occurrence of non-fatal myocardial infarction and the prevalence of obesity, smoking and vasectomy in 36 pairs of men

		Pair member			
	MI occurrence	*Vasectomized*		*Not vasectomized*	
Pair number	*V = Vasectomized* *N = Not vasectomized*	*Obese*	*Smoker*	*Obese*	*Smoker*
1	N	−	−	+	−
2	V	+	−	−	+
3	V	−	+	−	−
4	N	−	+	−	+
5	N	+	−	+	+
6	V	−	−	−	+
7	V	+	−	−	+
8	V	−	−	−	−
9	N	−	−	−	+
10	V	−	+	−	+
11	N	−	+	−	+
12	N	−	−	−	−
13	V	+	+	−	+
14	N	−	−	−	+
15	V	−	+	−	−
16	V	+	+	+	+
17	V	−	−	−	−
18	N	+	+	−	+
19	V	+	+	−	−
20	N	−	−	−	+
21	N	−	−	−	+
22	N	−	+	−	−
23	V	−	+	−	+
24	V	−	+	−	−
25	V	+	+	−	−
26	N	−	−	−	−
27	V	−	+	−	+
28	V	+	−	−	−
29	N	−	−	−	+
30	V	−	−	−	−
31	N	−	+	+	+
32	V	−	+	−	−
33	N	+	−	−	−
34	N	+	−	−	+
35	V	−	−	−	−
36	V	−	−	−	−

smoking. Otherwise they were classed as smoking. Obesity was classed as (a) not present, (b) present before MI (or corresponding date in the non-MI pair member) but first noted after vasectomy (or corresponding date) or (c) present before vasectomy (or corresponding date). Table 1 indicates which of the pair members suffered an MI, and records for each pair member the presence or absence of obesity predating vasectomy and a history of smoking. Analysis of these 36 matched sets with a matched proportional hazards model, as described above, yields the incidence ratio estimates given in Table 2. After adjustment for the confounding effects of smoking and obesity, vasectomy appears not to have any strong relation to MI. This result held when further factors, interactive effects, and time trends were examined (see Walker *et al.*, 1981, for more detail). The number of clinical records which needed to be abstracted

TABLE 2

Risk factors for non-fatal myocardial infarction: analysis of vasectomy- and infarction-discordant pairs

Factor	*Relative incidence (factor present versus absent)*	*95% confidence bounds*
Vasectomy	1·2	0·6, 2·7
Obesity	3·2	0·7, 15·0
Smoking	4·1	1·2, 14·5

constituted 0·7 per cent of the total number of records in the study. Since there is a maximum of one MI per exposure-balanced set in these data, the ordering of MI's within each of the sets is not at issue, and the analysis is essentially identical to that proposed for matched pair studies by Rosner and Hennekens (1978). Had there been multiple MI's within any set, a scheme which accounts for the timing of events, such as the one described here, would have been essential.

ACKNOWLEDGEMENTS

I am indebted to George B. Hutchison and Kenneth J. Rothman for their useful suggestions and criticisms. This investigation was supported by National Research Service Award 2T32CA 0900-04 and by Grant Number CA06516, awarded by the National Cancer Institute, DHEW.

REFERENCES

BRESLOW, N. E., DAY, N. E., HALVORSEN, K. T., PRENTICE, R. L. and SABAI, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *Am. J. Epidemiol.*, **108**, 299–307.

COX, D. R. (1972). Regression models and life-tables. *J. R. Statist. Soc.* B, **34**, 187–220.

—— (1975). Partial likelihood. *Biometrika*, **62**, 269–276.

FAREWELL, V. T. and PRENTICE, R. L. (1980). The approximation of partial likelihood with emphasis on case-control studies. *Biometrika*, **67**, 273–278.

LIDDELL, F. D. K., MCDONALD, J. C. and THOMAS, D. C. (1977). Methods of cohort analysis: appraisal by application to asbestos mining. *J. R. Statist. Soc.* A, **140**, 479–486.

MANTEL, N. (1973). Synthetic retrospective studies and related issues. *Biometrics*, **29**, 470–486.

PRENTICE, R. L. and BRESLOW, N. E. (1978). Retrospective studies and failure time models. *Biometrika*, **65**, 153–158.

ROSNER, B. and HENNEKENS, C. H. (1978). Analytic methods in matched pair epidemiological studies. *Int. J. Epidemiol.*, **7**, 367–372.

WALKER, A. M., JICK, H., HUNTER, J. R., DANFORD, A., WATKINS, R. N., ALHADEFF, L. and ROTHMAN, K. J. (1981). Vasectomy and non-fatal myocardial infarction. *Lancet*, **1**, 13–15.