Agenda

History, case-control methods .. up to "modern" times

Unmatched c-c study

'Synthetic Case-Control Studies' (unmatched)

Historical example, with a few "twists"

Matched Case-Control Studies

Conditional logistic regression e.g. double-blind multiple crossover trial

The case-crossover study

Matched retrospective cohort study to ascertain the long term health consequences of vasectomy

Worked e.g. of Nested case-control study (& link to Cox model)

References

- Breslow & Day
  Vol I Ch 6 (Unconditional logistic regression for large strata) & Ch 7 (conditional logistic regression for large strata)
  Vol II Ch 5 (Fitting Models to Continuous Data (nested cc))
- Hosmer & Lemeshow ALR: Ch 6.3 (Logistic Regression for Case-Control Studies) and Ch 7 (Logistic Regression for Matched Case-Control Studies)
- Clayton & Hills

Several key historical and modern articles/reviews

`[cf. http://www.epi.mcgill.ca/hanley/c681/case_control]`

+ Pair of expository articles by JH

Quoted in Breslow 1996.

The sophisticated use and understanding of case-control studies is the most important methodologic development of modern epidemiology ***(Rothman textbook 1986, p. 62)***

Epidemiologists who have done case-control studies during the past 20 years... have stood on the shoulders of giants. And, lest we epidemiologists lose sight of one major root of our discipline, we should remember that all of these men are, or were, statisticians ***(Cole 1979, p 15 in "The evolving case-control study" J Chron Dis 32, 15-27)***

Some historical landmarks

1951 Cornfield 1951 ... odds ratio

1959 Mantel & Haenszel ... summary odds ratio

1961 Cornfield ... logistic regression (inside cohort)

1970 Cox's textbook on logistic regression

1972 Cox 1972 ... p h model .. estimated from risksets

1973 Mantel 1973 ... 'synthetic' case-control study

1976 Miettinen ... incidence density sampling .. ~~rare disease~~

1977 Liddell, McDonald & Thomas .. sampling from risksets

1978 Breslow et al. conditional LR for matched c-c studies

1986 Case-cohort studies

1982/88 Two-stage sampling

1990's Case-only designs / case-crossover / case-time

2000 Daniel McFadden (Economist): Nobel Prize for his development of theory and methods for analyzing discrete choice: the economist he shared it with does work on causal models (similar to work of Jamie Robins in epidemiology)

Case-Control Studies (developments in analytic methods**)**

(see Breslow's 1996 paper)

1951 Cornfield developed odds ratio in c-c study as estimator of relative risk
2 x 2 tables: Inferences about "relative risk" made by applying to case-control data the same calculations as would be applied to cohort data from same population {B&D Vol I. p 202].

1955 [overlooked] Woolf uses 'quasi-denominators' derived from what he called the 'control series' Here (with some more user friendly notation, c for sizes of case series and d for denominator series and some rewording) is what he said: Even in case-control studies, one should think in terms of and "work with incidence rates in exposed and unexposed.. Case-control data do not permit calculation of absolute rates, nor are they needed. What is wanted and readily obtained is an estimate of the ratio of one rate to another.

The incidence in the exposed (1) will be (Hanley notation)

$c_1/(d_1 \times$ some constant$)]$.

The incidence in the unexposed(0) will be

$c_0/(d_0 \times$ the same constant$)$.

Thus the estimate of the rate ratio will be

$(c_1/d_1) \ / (c_0/d_0)$ " [Woolf p 251]

Notice that the focus (very enlightened, even in 1955!!) is on comparison of exposed with unexposed, **not** of cases with controls, i.e., he does not compare 'exposure odds in the cases' with the 'exposure odds in the controls'. Woolf was, as we should be, (in OSM's words) "a student of rates".

1961 Cornfield, using cases of chd that occurred in Framingham cohort study, developed (prospective) logistic regression (LR) equation to model risk[chd | determinants]

1966-1970: estimation of LR coefficients by Maximum Likelihood rather than discriminant analysis

1973-1979: even in 'retrospective' (c-c) studies, where overall probability that a subject is a case is fixed by the design, one can use the prospective logistic regression risk[case | determinants] to estimate odds ratios (estimators of rel. risk). "implications: ***analysis of data from case-control studies via logistic regression may proceed in same way and using the same computer programs as cohort studies"*** (H&L p 208)

1976-1978 conditional logistic regression for matched cc-studies (no explicit cohort required ...)
- likelihood similar to that used in 'choice-based' sampling in consumer research [why do some (and which) customers buy a particular brand of merchandise?] ..
- has same Likelihood as Cox's partial likelihood for survival analysis and risk set samples.
- If use matching in design, best to use this matching in analysis

Case-Control Studies

Unmatched c-c study

(c cases, d controls i.e. case&denominator series sized c&d)

datafile records

Case	"Exposure"	Confounder(s)	etc.
0/1 (Y)	"E"	z1 z2 z3 ...	....
etc....			

Null model...

logit[Prob [case]] = log [c / d] = log[case/control ratio] has no scientific meaning

Analysis model (using "E" and z's generically)

logit[Prob [case | x z1 z2 z3]] = $\beta_0 + \beta_1 \times E + \gamma_1 \times z_1 + ...$

adjusted or[E=1 vs. E=0] = exp[β_1_hat]

(again, β_0 has no scientific meaning)

Confounding, interaction, collinearity: as in earlier chapters)

Factors that affect precision of β_1_hat, and thus of OR_hat

- no. of cases (c);
- case-control ratio (d / c);
- distrn. of exposure among d [cf. notes m_m_ch_9_epi]
- OR[E=1 vs. E=0]
- collinearity of E with {z1, z2, .. }
- see Breslow and Day Vol II Ch. 7, or Schlesselman

In case of binary E, (to a first approximation)

var[ln or] = var[β_1_hat] = (Woolf variance) × $VIF_{E<-->Z}$

$$\{ VIF_{E<-->Z} = \frac{1}{1 - \text{Mult. } r^2 \text{ of E <-> remaining terms in model}} \}$$

'Synthetic Case-Control Studies (unmatched)

"In a large prospective study in which comparatively few cases of disease have occurred, computational problems* can be so burdensome as to preclude a comprehensive and imaginative analysis of the data. The prospective study can be converted into a synthetic retrospective study by selecting a random sample of the cases and a random sample of the noncases, the sampling fraction being small for noncases, but essentially unity for cases. It is demonstrated that such sampling will tend to leave the dependence of the log odds on the variables unaffected except for an additive constant."

(abstract) Mantel 1973

* or cost of analyzing stored sera, or entering questionnaire data

"A particular prospective-study situation which I encountered gave rise to only 165 cases of a particular condition in a cohort of about 4,000 individuals". computations were arduous given the computing facilities at that time

"Suppose we included in the analysis a random proportion (sampling faction), f_{cases} , of our cases and another random proportion, $f_{controls}$, of the negatives. If we chose f_{cases} as 1 and $f_{controls}$ as 0.15, we would have all the cases and 3.5 negatives per case. By the reasoning that $n_1 n_2/(n_1+n_2)$ {= reciprocal of $(1/n_1 + 1/n_2)$.. jh} measures the relative information in the comparison of two averages based on sample sizes of n_1 and n_2 respectively, we might expect by analogy, which would of course not be exact in the present cases, that this approach would result in only a moderate loss of information. (The practising statistician is generally aware of this kind of thing. There is little to be gained by letting the size of the control group, n_2, become arbitrarily large if the size of the experimental group, n_1 , must remain fixed.)

If we refer to P' as the probability in the synthetic study, and P as the probability in the full cohort study, then we have

$$\log \frac{P'[Y = 1 \mid E\ Z\ ]}{P'[Y = 0 \mid E\ Z\ ]} = \log \frac{f_{cases}}{f_{controls}} + \log \frac{P[Y = 1 \mid E\ Z\ ]}{P[Y = 0 \mid E\ Z\ ]}$$

i.e., **the expected relationship in the synthetic dataset is the same as in the full dataset, with the exception that the intercept is now shifted by the (known) log of the ratio of the sampling fractions.**

(see H&L pp 205 - 208, or B&D Vol I p 202-203 for fuller and more modern versions of this important insight)

Historical example, with a few "twists"

(excerpts from JH's notes used for medical students in Fall 2002)

Recall [excerpt from Rothman & Greenland] .. there are two primary types of non-experimental studies in epidemiology.

The first, the ***cohort study*** (also called ***the follow-up study*** or ***incidence study), is*** a direct analogue of the experiment; different exposure groups are compared, but (as in Snow's study) the investigator does not assign the exposure.

The other, the incident case- control study, or simply the ***case-control study,*** employs an extra step of sampling according to the outcome of individuals in the population. This extra sampling step can make a case-control study much more efficient than a cohort study of the entire population, but it introduces a number of subtleties and avenues for bias that are absent in typical cohort studies.**{Case-control studies are best understood by defining a source population, which represents a hypothetical study population in which a cohort study might have been conducted.** If a cohort study were undertaken, the primary tasks would be to identify the exposed and unexposed denominator experience, measured in person-time units of experience or as the number of people in each study cohort, and then to identify the number of cases occurring in each person- time category or study cohort. In a case-control study, the cases are identified and their exposure status is determined just as in a cohort study, but denominators from which rates could be calculated are not measured. Instead, a control group of study subjects is sampled from the entire source population that gives rise to the cases.

The purpose of the control group is to determine the relative (as opposed to absolute) size of the exposed and unexposed denominators within the source population. From the relative size of the denominators, the relative size of the incidence rates (or incidence proportions, depending on the nature of the data) can be estimated. **Thus, case-control studies yield estimates of relative effect measures.** Because the control group is used to estimate the distribution of exposure in the source population,

In sum, case-control studies of incident cases differ from cohort studies according to how subjects are initially selected. A cohort study identifies and follows a population or populations to observe disease experience; a case- control study involves an additional step of selecting cases and controls from this population. [end of excerpt]

NOTE[JH] The statistical precision of the ratio measure of risk is largely a function of the number of cases. The same amount of person time is needed to generate a given no. of cases in a cohort study as in a case-control study. The latter's efficiency derives from the reduced amount of data-gathering, and the investigator's time-scale -- IF the exposure of past cases and "non-cases" can be accurately established after the fact.

The essential difference can be illustrated using the data from John Snow's investigation

"According to a return which was made to Parliament, the Southwark and Vauxhall Company supplied 40,046 houses from January I to December 31, 1853, and the Lambeth Company supplied 26,107 houses during the same period; " [but no list available to Snow!]

So, the **denominators** were...

No. of Houses with... ***("impure" and "pure" is overstating it)***

Water	
Impure	Pure
40 046	26 107

286 fatal attacks of cholera took place, in the first four weeks of the epidemic, in houses supplied by the former company, and only 14 in houses supplied by the latter

No. of CASES (**numerators**) in houses with... ["shoe-leather +" method *]

Water	
Impure	Pure
286	14

Attack **rates** in houses with...

Water			
Impure	Pure	**Ratio**	Difference
$\frac{286}{40046}$	$\frac{14}{26\ 107}$		
71.4 / 10K	5.4 / 10K	13.3	66 / 10K

This is the cohort approach -- start with denominators of known sizes and then determine the numerators.

But what if sizes of the two denominators not readily available (but the numerators were) ???. it would be a lot of leg work to determine the water source of each of 40046 + 26107 = 66153 houses!

*** And a non-statistical (numerator) Q : how did John Snow determine which of the 300 houses had which source of water?**

Cf. 1. Shephard, D. John Snow : anaesthetist to a queen and epidemiologist to a nation : a biography 1995 WZ 100 S674S 1995 [Regular Loan] Osler Library;

2. Snow, John, 1813-1858. On the mode of communication of cholera : 191p, map, 23 cm. Location WC 262 S764 1936 [Regular Loan] Osler Library

No. of CASES (**numerators**) in houses with... **[Snow e.g. continued..]**

Water	
Impure	Pure
286	14

If is a huge amount of work to determine the sizes of the two denominators, how about we take a sample and estimate their estimate their relative sizes ?

Say we survey **100** houses selected at random; we might find that the **sources** were...

No. (± sampling variation) of 100 **sampled Houses** with...

Water		
Impure	Pure	
61 (±10)	39 (±10)	100

We can take the 61 and 39 as "quasi-denominators" and make two "quasi-rates"

Quasi-attack **rates** in houses with...

Water			
Impure	Pure	**Ratio ***	Difference
286/61	14/39	13.1 (±)	no meaning

Lets say that instead we survey **1000** houses selected at random and that the sources were...

No. (± sampling variation) of 1000 **sampled** Houses with...

Water	
Impure	Pure
605 (±32)	395 (±32)

Quasi-attack **rates** in houses with...

Water			
Impure	Pure	**Ratio ***	Difference
286/605	14/395	**13.3 (±)**	no meaning

***** Inappropriate to use Woolf's formula for var(log or) as there many have been multiple cases in (numerator contributions from) the same house, but broad principle re efficiency of denominator estimation still holds

Thus the purpose of the 100 (or 1000, or however many are selected, depending on the budget, and the statistical precision required) houses selected at random is to determine the relative (as opposed to absolute) size of the exposed and unexposed denominators within the source population. From the relative size of the denominators, the relative size of the incidence rates (or incidence proportions, depending on the nature of the data) can be estimated.

A good descriptor of these houses selected at random is "the **denominator series"**. The cases, already in hand, constitute the "numerator series". [terminology of McGill Prof Miettinen]

To make the calculation of the statistical errors associated with the estimated ratio less complicated, most epidemiologists would exclude the "case houses" from the sampling frame of 66153 houses and would instead sample the "source to be determined" houses from the remainder - i.e. from the "non-case houses". See for example Fletcher et al.'s Figure 10.3, where they write of "non-cases".

Unfortunately, the more common (and older) name for these "non-case" houses is the "control" houses. This creates considerable **confusion** among non-epidemiologists, since we **now have 2 meanings for "control" ..**

1 in an experiment (e.g. clinical trial), those who do not receive the experimental (new) treatment are sometimes referred to as the "controls" ("comparison group" or --if it is the situation -- "unexposed group" is a more informative label) The same applies in a (non-experimental) cohort study (e.g. what should one call the wives of the male resident physicians when their pregnancy outcomes are compared with those of the female resident physicians?)

Notice that Fletcher et al. themselves use confusing terminology -- in describing the characteristics of a cohort study (Table 10.2 3rd row, 1st column) they say "Controls, the comparison group (i.e. noncases), not selected -- evolve naturally.

2 in a "study that relies on quasi-denominators", (commonly known as a "case-control" study), the "controls" are the denominator series. Their exposure status (or exposure history) is the focus of the inquiry. Even though it is not entirely accurate, it is less confusing to call them "non-cases" than to call them "controls".

"Being epidemiologically correct"... Most epidemiology textbooks still describe case-control studies as "comparing cases with controls". In fact, as the above example [that views the "controls (or non-cases) as a denominator series] shows, e**ven in a case-control study one compares** (quasi-rates) for the **exposed** with quasi-rates for the **non-exposed** (in the ratio of these quasi-rates, the hidden sampling fraction cancels out in the arithmetic)

This last point about the sampling fraction is very important: the "controls" [i.e., the "non-case" or "the denominator series"] must be selected without regard to their exposure.. see page 1 re "this **cardinal requirement"**

Other simple e.g.'s of denominator issue:

" Pour battre Patrick Roy, mieux vaut lancer bas" (JH course 626)

"WOMEN ARE SAFER PILOTS": newspaper article (JH course 626)

Could we use a case-control approach to the Study of Medical students' compliance with simple administrative tasks and success in final examinations?

```
/* 'Synthetic Case-Control Studies (rough* example)
Framingham: cases:       new chd within 10 years
            noncases: no new chd within 10 years
           (similar to Cornfield's analysis in 1961)
*/

data cc_prev;
keep i_male age ht wt chol dbp sbp mrw smok case ran_no;

set sasuser.fram;
case = .;
if i_newchd = 1 and t_newchd <  10 then case = 1;
if i_newchd = 0 or t_newchd  >= 10 then case = 0;

ran_no = ranuni(12345677); /* for sampling */

if (40 <= age <= 59);
```

```
proc means data=cc_prev maxdec=2 mean;
 class case ;
 var i_male age ht wt chol dbp sbp mrw smok ;
```

	2848 non-cases			283 cases		
Variable	min	max	mean	mean	min	max
I_MALE (0/1)	0	1	0.42	0.64	0	1
AGE in 1948 y	40	59	48.45	51.28	40	59
HT inches	51	76	64.53	64.96	55	73
WT lbs	67	300	153.47	164.17	91	256
CHOL	96	517	232.79	248.28	155	493
DBP diastolic	50	160	87.07	93.06	60	150
SBP systolic	82	300	140.31	152.84	100	270
MRW Rel.Weight	67	268	121.51	127.54	86	191
SMOK cigs/day	0	60	7.96	11.64	0	60

```
Total Number of Observations: 3036
Some observation(s) were deleted due to missing values for the
response or explanatory variables.
```

*** WARNING**: The following analyses are for demonstration purposes only; a more refined analysis would including possibly separate analyses for men and women, the proper representation of each determinant, etc.., and one would not define 'non-cases' at the ***end*** of the 10 year follow-up. See nested c-c study later for 'modern' way.

```
title all/40%/10% of the <<non-cases>> ;
proc logistic descending data=cc_prev;
model case = i_male age ht wt chol dbp sbp mrw smok
                                          /risklimits *;
```

```
                    % of <<non-cases>> ;
                  where( (case = 1)
                      or (case = 0 and  ....
                   ran_no < 0.4) );   ran_no < 0.1) );
```

	CASE	Count	Count	Count
	1	275	275	275
	0	2761	1095	284
% of non-cases		100%	40%	10%
missing values		(162)	(47)	(18)

Maximum Likelihood Estimates

	Param.	Stand.	OR	Param.	Stand.	OR	Param.	Stand.	OR
Var	Est	Error	*	Est	Error	*	Est	Error	*
B0	-2.669	6.28	.	-1.880	7.21	.	1.295	9.10	.
MALE	1.212	0.23	3.36	1.184	0.28	3.27	1.079	0.32	2.94
AGE	0.084	0.03	1.08	0.071	0.01	1.08	0.086	0.02	1.09
HT	-0.156	0.10	0.85	-0.148	0.11	0.86	-0.161	0.14	0.85
WT**	0.021	0.020	1.02	0.019	0.022	1.02	0.020	0.029	1.02
CHOL	0.007	0.001	1.01	0.007	0.002	1.01	0.005	0.002	1.01
DBP	0.006	0.008	1.01	0.001	0.009	1.00	0.002	0.011	1.00
SBP	0.010	0.004	1.01	0.013	0.005	1.01	0.009	0.006	1.01
MRW	-0.013	0.024	0.99	-0.010	0.028	0.99	-0.014	0.035	0.99
SMOK	0.022	0.005	1.02	0.026	0.006	1.03	0.023	0.008	1.02

```
** Should have used continuous variables with bigger units e.g. 10
years 10 mm mercury, chol in units of 10 etc.. otherwise the
coefficients are small and off the scale.

[* 95% CIs (from RISKLIMITS option) were printed but not shown here]
```

As 'intuited' by Mantel, The SE's with the different numbers of non-cases are roughly proportional to

sqrt[1/275 + ***1/2761 or 1/1095 or 1/ 284***]

Critical factor is not sampling fraction, but control/case ratio

Matched Case-Control Studies (B&D I; H&L Ch 7; Schlesselman)

Preamble: Matching & stratification are the same concept: what most call "matched data" are just "finely stratified" data. Frequency matching is coarser form of stratification. The "fineness" of the stratification/matching (how many cases and controls in same stratum or "matched set") affects the amount of distortion of the OR estimate if, in the analysis, the analysis does not fully take account of the matching.

	Options	Consequences *
a	Ignore the matching variables (break the matches) and use an unconditional logistic regression with E and other (unmatched) z's	or tends to be biased towards null. See Rothman & Greenland (e.g. they like to use extreme examples to scare readers) or B&D I section 7.6, or H&L p 243. But smaller SE's
b	As in (a) but include E, the other (unmatched) z's, AND the matching variables	Not as severe as under a, but or still biased towards null
c	If matching variables are not 'measurable' (e.g. if match on family, or use twin pairs or siblings, to control for genetic or familial factors), in an unconditional logistic regression include this matching variable as a categorical variable with as many levels as there are matched sets (effectively adds a separate intercept for each matched set.).	If 1 case and M controls per set, and exposure E is binary, the absolute value of β = log[OR] is overestimated (i.e., away from null) by a factor of (M+1)/M. For example, if matched pairs, so with matched pairs, i.e., M=1, log[or] from unconditional LR is 2/1 = double what its should be, i.e. the or is the square of what it should be. (see B&D 7.1). This is consequence of fitting too many parameters to too little data
d	Eliminate 'separate intercepts' in (c) by conditioning on total number of exposed individuals in the matched set. (same as conditioning on total # of ***cases*** in a prospective study)	Avoids the over-estimation in (c) and the under-estimation in (a) and (b). But may lead to larger SE's.
e	Use matching in analysis when didn't need to.	Can lose efficiency (SE's larger than they should be) See Fig 7.1, B&D I pp. 271-272.

Conditional Logistic Regression Analysis: matched data

Not just for matched case-control studies !!

Effect of ultraviolet germicidal lights installed in office ventilation systems on workers' health and well being: double-blind multiple crossover trial

D Menzies, J Popa, J Hanley, et al. Lancet 2003; 362: 1785–91 (Nov 23, 2003)

Methods We undertook a double blind, multiple crossover trial of 771 participants. In office buildings in Montreal, Canada, Ultraviolet germicidal irradiation (UVGI) was alternately off for 12 weeks, then turned on for 4 weeks. We did this three times with UVGI on and three times with it off, for 48 consecutive weeks. Primary outcomes of self-reported work-related symptoms, and secondary outcomes of endotoxin and viable microbial concentrations in air and on surfaces, and other environmental covariates were measured six times.

Response patterns in 5 different participants (subjects), 2 present at all 6 assessment occasions, 1 at 5, 1 at 3, 1 at 1.

UVGI:	On	Off	On	Off	On	Off	On	Off	On	Off	On	Off
Sx +	0	0(0)	0	1(1)	1	1(2)	2	1(3)	1	0(1)		
Sx -	3	3	3	2	2	1	0	0	0	0		
Totals	3	3	3	3	3	2	2	1	1	0		

To provide within-person comparisons of symptoms with UVGI on and off, we used conditional logistic regression adjusted for changing environmental covariates (the PHREG procedure in SAS, version 8). This method analysed every person as a stratum if they completed at least one questionnaire with UVGI on, and one with UVGI off, and had some variation in response. Individuals' characteristics, such as age or sex, were not included, since they could not alter the within-person estimate of effect. Potential building effects, that could cause variations in the adjusted odds ratios, were assessed by adding three interaction terms of condition and building to the regression models. To assess potential effect modification by personal or medical characteristics, conditional logistic regression was repeated within subgroups, and by trial.

datafile record for subject # 2

subject	assessment	UVGI	Sx	Temp	Humidity	CO2	time(pm)
2	1	0	0	25	42	500	3
2	2	1	0	24	38	650	3
2	3	0	1	25	36	480	3
2	4	1	0	26	41	510	3
2	5	0	0	24	43	710	3
2	6	1	0	24	40	450	3

etc...

Riskset, and associated Likelihood contribution from subject # 2

UVGI	Sx	(relative) Odds	
0	-	exp[0 × β + 25 × y + ...]	[1]
1	-	exp[1 × β + 24 × y + ...]	[2]
0	+	exp[0 × β + 25 × y + ...]	[3]
1	-	exp[1 × β + 26 × y + ...]	[4]
0	-	exp[0 × β + 24 × y + ...]	[5]
1	-	exp[1 × β + 24 × y + ...]	[6]

Likelihood* contribution: $\dfrac{[3]}{[1] + [2] + [3] + [4] + [5] + [6]}$

***CONDITIONAL on 1 occasion of Sx=1 & 5 of Sx=0 [cf. Fisher's exact test]**

Conditional Logistic Regression via

SAS

```
PROC PHREG ;
MODEL time*Sx(0)  =  UVGI  Temp  Humidity  CO2;
STRATA subject;
```

Stata

```
clogit sx uvgi temp humidity co2, group(subject) or
```

Notes

1 We could have used a 'fake' time here. The sole purpose is to force the data into the mode expected by the survival program PHREG, any set of times will work as long as the times associated with the Sx=0 occasions are greater than or equal to the times associated with the Sx=1 occasions. For example, one could even use the subject number as the 'time'. Notice that the specialized clogit program in Stata does not require this trick. (If use Stata's stcox , do have to 'fake' the time)

2 The key to keeping the different subjects in different strata is the use of the STRATA statement (group statement in Stata) , so that the within-person (log) likelihoods from the different subjects are multiplied (added). The likelihoods from subjects with no variation in response (Sx) and those with no variation in UVGI (e.g. subjects # 1, 4 and 5 in the example) do not contribute to the estimation of the parameters of the conditional logistic model. i.e. subjects with a zero in a margin of their table cannot contribute.

3 The likelihood contribution from subject # 3 is more complex, and is akin to the situation of 'tied' failure times in the Cox model. We now have to calculate the probability of the 2 Sx=1 occasions being recorded on the 2 occasions they were (1&4, the observed situation), rather than on any other of the 20 pairs of occasions. If matched set (riskset) is large, the substantial number of combinations of candidate occasions can lead to considerable computations. Peto and Breslow gave approximations for such situations.. If the data for #3 were

subject	assessment	UVGI	Sx	Temp	Humidity	CO2	time(pm)
3	1	0	1	24	40	530	3
3	2	1	0	26	39	560	3
3	3	0	0	25	44	520	3
3	4	1	1	23	38	490	3
3	5	1	0	25	31	610	3

the Likelihood contribution from this subject ('stratum') would be

$$\frac{[1]\times[4]}{[1]\times[2] + .. + [1]\times[5] + [2]\times[3] + .. + [2]\times[5] + .. + [4]\times[5]}$$

4 The OR estimates (calculated as exp[beta_hat]) are interpreted in the same way as those from an unconditional logistic model.

5 Since persons are (self-)matched, cannot assess the impact of personal characteristics (e.g. sex, history of atopy) that remain constant across the occasions. But we can assess whether odds ratios for UVGI are different in males and females, or those with/without a history of atopy. It is also possible to do this by including a UVGI*atopy or UVGI*male product term in the model (more economical than separating them)

6 ***Not possible to distinguish*** conditional likelihood for this "prospective" matched study (or vasectomy/MI study analyzed next page) from conditional likelihood from a matched case-control study ('cases' = occasions with Sx=1, 'controls' = occasions with Sx=0). For e.g.s of matched case-control studies, see H&L, Schlesselman, or Breslow&Day Volume 1.

7 The 'case-crossover' study (e.g. the D Redelmeier & R Tibshirani NEJM Vol336 Feb 13, 1997 study of "association between cellular-telephone calls and motor vehicle collisions") is nothing more than a self-matched case control study.

> "Methods We studied 699 drivers who had cellular telephones and who were involved in motor vehicle collisions resulting in substantial property damage but no personal injury. Each person's cellular-telephone calls on the day of the collision and during the previous week were analyzed through the use of detailed billing records."

The separate records for the collision occasion [numerator] and the non-collision occasions [***person-moments, denominator series***] of the previous week could be laid out just as in the UVGI example (replace Sx by collision, UVGI on/off by on/off cell phone, and temperature, humidity etc. by relevant driving conditions that affect the risk of a collision, and might not be the same on the ***compared occasions***.]

Worked analysis of matched pair data [more details in 626 website, and part I of JH draft article]

Walker et al. (1981) undertook a matched retrospective cohort study to ascertain **the long term health consequences of vasectomy**. The data shown 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, ...

The listing indicates which of the pair members suffered an MI, and records for each pair member presence or absence of obesity predating vasectomy and a history of smoking. Analysis of the 36 matched sets with a matched proportional hazards model, as described above, yields 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.

The number of clinical records which needed to be abstracted 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.

```
DATA a; INPUT PairNo Vas Obese Smoke MI;
time = 10; /* a 'fake' time for PHREG */
```

Pair	Vas	Ob	Sm	MI		Vas	Ob	Sm	MI
1	1	0	0	0	19	1	1	1	1
1	0	1	0	1	19	0	0	0	0
2	1	1	0	1	20	1	0	0	0
2	0	0	1	0	20	0	0	1	1
3	1	0	1	1	21	1	0	0	0
3	0	0	0	0	21	0	0	1	1
4	1	0	1	0	22	1	0	1	0
4	0	0	1	1	22	0	0	0	1
5	1	1	0	0	23	1	0	1	1
5	0	1	1	1	23	0	0	1	0
6	1	0	0	1	24	1	0	1	1
6	0	0	1	0	24	0	0	0	0
7	1	1	0	1	25	1	1	1	1
7	0	0	1	0	25	0	0	0	0
8	1	0	0	1	26	1	0	0	0
8	0	0	0	0	26	0	0	0	1
9	1	0	0	0	27	1	0	1	1
9	0	0	1	1	27	0	0	1	0
10	1	0	1	1	28	1	1	0	1
10	0	0	1	0	28	0	0	0	0
11	1	0	1	0	29	1	0	0	0
11	0	0	1	1	29	0	0	1	1
12	1	0	0	0	30	1	0	0	1
12	0	0	0	1	30	0	0	0	0
13	1	1	1	1	31	1	0	1	0
13	0	0	1	0	31	0	1	1	1
14	1	0	0	0	32	1	0	1	1
14	0	0	1	1	32	0	0	0	0
15	1	0	1	1	33	1	1	0	0
15	0	0	0	0	33	0	0	0	1
16	1	1	1	1	34	1	1	0	0
16	0	1	1	0	34	0	0	1	1
17	1	0	0	1	35	1	0	0	1
17	0	0	0	0	35	0	0	0	0
18	1	1	1	0	36	1	0	0	1
18	0	0	1	1	36	0	0	0	0

data 'wrap around' to save space

```
PROC PHREG;
  MODEL time*MI(0) = Vas;
  STRATA PairNo;
```

```
Dep Variable: TIME Cens. Variable: MI
Cens. Value(s): 0 Ties Handling: BRESLOW

Stratum  PAIRNO   Total  Event Cens %Cens
      1       1       2      1     1   50
     ..      ..      ..     ..    ..   ..
     36      36       2      1     1   50
  Total              72     36    36   50

Testing Global Null Hypothesis: BETA=0
-2 LOG L 49.90 Without  Diff Chi_sq
         49.46 With     0.44 1DF (p=0.50)

Analysis of Maximum Likelihood Estimates

       Par.  SE  Wald    Pr > Risk
Var DF Est      Chi-Sq Chi-Sq Ratio

VAS  1 0.22 0.34 0.44  0.50  1.25 (20/16)
```

```
PROC PHREG;
MODEL time*MI(0) = Vas Obese Smoke;
  STRATA PairNo;
```

```
Testing Global Null Hypothesis: BETA=0

-2 LOG L 49.90 Without  Diff Chi_sq
         41.28 With     8.62  3DF (p=0.03)
 Score                  7.71  3DF (p=0.05)
 Wald                   6.23  3DF (p=0.10)

      Par.  SE  Wald Pr >    Risk   95%
Var   Est     Chi-Sq Chi-Sq  Ratio Limits

VAS   0.20 0.40 0.24 0.62   1.2 [0.6, 2.7]
OBESE 1.18 0.78 2.28 0.13   3.3 [0.7,15.0]
SMOKE 1.42 0.64 4.97 0.03   4.1 [1.2,14.4]
```

Stata

```
input pairno vas obese smoke mi
clogit mi vas obese smoke, group(pairno)
```

```
Conditional (fixed-effects) logistic regrn.

LR chi2(3) = 8.62 Prob > chi2  =  0.03
Log likelihood = -20.64  Pseudo R2 = 0.17

mi      Coef.  SE.  z  P>|z| [95% Conf. Int]
vas     0.20  .40  0.49 0.62   -.59   .99
obese   1.18  .78  1.51 0.13   -.35  2.70
smoke   1.42  .64  2.23 0.03    .17  2.67
```

Worked e.g. of Nested case-control study [cf next page]

Framingham Study, the first 10 years of follow-up on each subject

The 30 year story (in retrospect) for selected subjects

ID	I_MALE	AGE	MRW	SMOK	SBP	DBP	CHOL	A_NEWCHD	I_NEWCHD
100	1	50	126	20	144	80	286.0	63	1
850	1	50	116	.	126	86	.	53	1
1100	1	53	128	0	138	78	148.0	56	0
1200	0	47	113	0	160	110	315.0	77	0
1650	1	57	121	15	102	72	162.0	62	1
1700	1	53	111	60	120	86	209.0	56	1
2300	1	57	142	0	220	118	205.5	62	1
5050	1	59	97	40	148	86	213.0	64	1

Each subject's record split into 2-year segments

ID	I_MALE	AGE	MRW	SMOK	SBP	DBP	CHOL	AGE_THEN	AGE_NEXT	AGE_DX	CASE
100	1	50	126	20	144	80	286.0	50	52	51	0
100	1	50	126	20	144	80	286.0	52	54	53	0
100	1	50	126	20	144	80	286.0	54	56	55	0
100	1	50	126	20	144	80	286.0	56	58	57	0
100	1	50	126	20	144	80	286.0	58	60	59	0
850	1	50	116	.	126	86	.	50	52	51	0
850	1	50	116	.	126	86	.	52	54	53	1
1100	1	53	128	0	138	78	148.0	53	55	54	0
1200	0	47	113	0	160	110	315.0	47	49	48	0
1200	0	47	113	0	160	110	315.0	49	51	50	0
1200	0	47	113	0	160	110	315.0	51	53	52	0
1200	0	47	113	0	160	110	315.0	53	55	54	0
1650	1	57	121	15	102	72	162.0	57	59	58	0
1650	1	57	121	15	102	72	162.0	59	61	60	0
1650	1	57	121	15	102	72	162.0	61	63	62	1
1700	1	53	111	60	120	86	209.0	53	55	54	0
1700	1	53	111	60	120	86	209.0	55	57	56	1
2300	1	57	142	0	220	118	205.5	57	59	58	0
2300	1	57	142	0	220	118	205.5	59	61	60	0
2300	1	57	142	0	220	118	205.5	61	63	62	1
5050	1	59	97	40	148	86	213.0	59	61	60	0
5050	1	59	97	40	148	86	213.0	61	63	62	0
5050	1	59	97	40	148	86	213.0	63	65	64	1

Those males, born the same year, who were at risk in the second time-segment, ie when subject # 1650 developed CHD at age 62.

51 in riskset (50 in addition to # 1650, the case)

	ID	I_MALE	AGE	MRW	SMOK	SBP	DBP	CHOL	AGE_THEN	AGE_NEXT	AGE_DX	CASE
	14	1	57	118	15	128	76	.	61	63	62	0
	275	1	57	89	35	112	68	276.0	61	63	62	0
	358	1	57	92	0	158	96	215.0	61	63	62	0
	412	1	57	118	0	220	124	192.0	61	63	62	0
	651	1	57	117	0	140	85	200.0	61	63	62	0
->	681	1	57	131	0	148	84	209.0	61	63	62	0<-
	718	1	57	113	5	110	78	250.0	61	63	62	0
	909	1	57	118	0	120	80	259.0	61	63	62	0
	965	1	57	121	0	142	92	159.0	61	63	62	0
->	1000	1	57	105	.	96	70	.	61	63	62	0<-
	1217	1	57	111	0	106	70	179.0	61	63	62	0
	1305	1	57	132	0	114	70	292.0	61	63	62	0
	1339	1	57	124	20	126	70	252.0	61	63	62	0
	1499	1	57	114	20	140	90	318.0	61	63	62	0
	1650	1	57	121	15	102	72	162.0	57	59	58	0
	1650	1	57	121	15	102	72	162.0	59	61	60	0
	1650	1	57	121	15	102	72	162.0	61	63	62	1
	1741	1	57	119	0	160	80	292.0	61	63	62	0
	1801	1	57	124	20	120	76	255.0	61	63	62	0
	1810	1	57	121	5	110	80	150.0	61	63	62	0
	1877	1	57	129	0	184	110	219.0	61	63	62	0
	1976	1	57	95	0	130	90	269.0	61	63	62	0
	1988	1	57	110	0	118	78	292.0	61	63	62	0
	2008	1	57	102	0	128	68	263.0	61	63	62	0
	2056	1	57	105	5	130	78	252.0	61	63	62	0
	2281	1	57	127	0	184	100	162.0	61	63	62	0
	2333	1	57	117	0	110	82	196.0	61	63	62	0
	2854	1	57	114	50	120	70	232.0	61	63	62	0
->	3044	1	57	118	0	160	94	213.0	61	63	62	0<-
	3115	1	57	130	0	234	134	255.0	61	63	62	0
	3286	1	57	115	5	150	90	174.0	61	63	62	0
	3300	1	57	119	0	140	84	275.0	61	63	62	0
	3378	1	57	120	0	120	66	198.5	61	63	62	0
	3379	1	57	115	0	120	78	239.0	61	63	62	0
	3398	1	57	100	0	110	75	233.0	61	63	62	0
	3625	1	57	142	0	134	84	242.5	61	63	62	0
	3710	1	57	103	5	156	78	275.0	61	63	62	0
	3756	1	57	83	5	138	85	199.0	61	63	62	0
	4052	1	57	116	40	160	80	237.5	61	63	62	0
	4306	1	57	108	0	125	85	209.5	61	63	62	0
	4346	1	57	116	0	118	76	200.5	61	63	62	0
	4473	1	57	113	15	164	94	253.0	61	63	62	0
	4501	1	57	135	15	154	100	226.0	61	63	62	0
	4592	1	57	114	0	124	74	238.5	61	63	62	0
	4653	1	57	131	5	126	78	161.0	61	63	62	0
	4806	1	57	163	20	168	86	221.0	61	63	62	0
	4841	1	57	114	20	136	88	237.0	61	63	62	0
	5001	1	57	150	20	158	100	261.5	61	63	62	0
->	5054	1	57	96	20	154	66	203.5	61	63	62	0<-
	5055	1	57	123	20	147	68	186.5	61	63	62	0
	5129	1	57	107	50	138	78	182.0	61	63	62	0

On the left, shown in blue are **4 subjects chosen at random**, without replacement, from the 50 subjects . **(4 controls per case)**

Listing of a few of the 283 matched sets

SET_NO	I_MALE	ID	MRW	SMOK	SBP	DBP	CHOL	AGE_DX	CASE
41M_40_01	1	197	101	20	136	94	205.0	41	0
41M_40_01	1	1070	106	20	138	82	.	41	0
41M_40_01	1	1420	136	25	152	110	258.0	41	1
41M_40_01	1	2076	116	20	126	90	206.0	41	0
41M_40_01	1	2564	136	0	164	106	227.0	41	0
62M_57_03	1	681	131	0	148	84	209.0	62	0
62M_57_03	1	1000	105	.	96	70	.	62	0
62M_57_03	1	1650	121	15	102	72	162.0	62	1
62M_57_03	1	3044	118	0	160	94	213.0	62	0
62M_57_03	1	5054	96	20	154	66	203.5	62	0
65M_56_01	1	571	126	0	148	80	288.0	65	0
65M_56_01	1	1783	100	10	143	80	219.0	65	0
65M_56_01	1	2132	145	0	180	100	178.0	65	0
65M_56_01	1	3548	131	0	158	100	265.5	65	0
65M_56_01	1	4070	103	20	190	110	283.5	65	1
68M_59_04	1	1651	127	10	128	78	167.0	68	0
68M_59_04	1	2120	192	0	124	80	191.0	68	0
68M_59_04	1	3656	101	20	180	104	199.5	68	0
68M_59_04	1	3718	128	0	200	130	239.0	68	1
68M_59_04	1	4801	117	0	184	120	211.0	68	0

```
proc sort data=cc ; by case_id;
proc phreg data = cc;
 model age_dx*case(0)= mrw smok sbp dbp chol / risklimits;
 strata case_id;
```

Dependent Variable: AGE_DX Censoring Variable: CASE
Ties Handling: BRESLOW Censoring Value(s): 0

Summary of the Number of Event and Censored Values

Stratum	CASE_ID	Total	Event	Censored	Percent Censored
1	61	4	1	3	75.00
2	70	5	1	4	80.00
80	1387	5	1	4	80.00
-> 81	1420	4	1	3	75.00
82	1427	5	1	4	80.00
-> 95	1650	4	1	3	75.00
96	1688	5	1	4	80.00
196	3714	5	1	4	80.00
-> 197	3718	5	1	4	80.00
-> 214	4070	5	1	4	80.00
283	5202	5	1	4	80.00
Total		1370	275	1095	79.93

Notes..
1. I have identified each riskset by the id number of the case
2. Some risksets are smaller than 5, because information on one or more of the covariates is missing

Criterion	Without Covariates	With	Model Chi-Square
-2 LOG L	868.737	794.039	74.698 with 5 DF (p=0.0001)
Score	.	.	77.282 with 5 DF (p=0.0001)
Wald	.	.	67.812 with 5 DF (p=0.0001)

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
MRW	0.008849	0.00365	5.88612	0.0153
SMOK	0.023597	0.00594	15.79099	0.0001
SBP	0.007210	0.00425	2.87955	0.0897
DBP	0.009001	0.00822	1.19995	0.2733
CHOL	0.007091	0.00163	18.99726	0.0001

Conditional Risk Ratio and 95% Confidence Limits

Variable	Risk Ratio	Lower	Upper	Label
MRW	1.009	1.002	1.016	Metropol Rel. Weight
SMOK	1.024	1.012	1.036	cigarettes/day
SBP	1.007	0.999	1.016	systolic BP
DBP	1.009	0.993	1.025	diastolic BP
CHOL	1.007	1.004	1.010	

```
proc sort data=cc ; by i_male case_id;
proc phreg data = cc; by i_male ;
 model age_dx*case(0)= mrw smok sbp dbp chol /
risklimits;  strata case_id;
```

Female (I_MALE=0) 99 Risksets **Male (I_MALE=1) 176 Risksets**

Variable	RiskRatio	Lower	Upper	RiskRatio	Lower	Upper
MRW	**1.005**	0.995	1.015	**1.013**	1.002	1.025
SMOK	**1.002**	0.972	1.032	**1.031**	1.017	1.044
SBP	**1.014**	1.002	1.026	**1.004**	0.991	1.016
DBP	**0.989**	0.965	1.014	**1.024**	1.002	1.047
CHOL	**1.005**	1.000	1.009	**1.009**	1.004	1.013

```
data products; set cc;
m_mrw  = i_male*mrw; m_smok = i_male*smok;
m_sbp  = i_male*sbp; m_dbp  = i_male*dbp; m_chol = i_male*chol;

proc sort data=products ; by case_id; proc phreg data = products;
 model age_dx*case(0)=  mrw   smok   sbp   dbp   chol
                    m_mrw m_smok m_sbp m_dbp  m_chol/ risklimits;
 strata case_id;
```

	b	SE	RR	Lower	Upper
RW	0.005200	0.00502	**1.005**	0.995	1.015
MOK	0.001934	0.01529	**1.002**	0.972	1.032
BP	0.013855	0.00610	**1.014**	1.002	1.026
BP	-0.011035	0.01256	**0.989**	0.965	1.014
HOL	0.004500	0.00249	**1.005**	1.000	1.009
m_MRW	0.008131	0.00753	1.008 *	0.993	1.023
m_SMOK	0.028160	0.01667	1.029 *	0.995	1.063
m_SBP	-0.010358	0.00868	0.990 *	0.973	1.007
m_DBP	0.035059	0.01683	1.036 *	1.002	1.070
m_CHOL	0.003988	0.00333	1.004 *	0.997	1.011

* these are the amounts by which the HR's in women are to be multiplied to obtain the HR's in men (check the ratio of the HRs for men & women)|

Q: what would happen if we added i_male to the last model above?

Table 4. Regression Coefficients ± Standard Errors

Variable	Conditional: likelihood	Unconditional: single α	Unconditional: multiple α_j
Ethnic group[a]	1.27 ± .33	1.42 ± .36	1.62 ± .25
Chinese wine[b]	.51 ± .29	.54 ± .28	.68 ± .27
Cigarettes[c]	.11 ± .10	.11 ± .09	.16 ± .09
Temperature[d]	.79 ± .16	.76 ± .15	1.12 ± .15

[a] 1 = Teochew and Hokkien, 0 = Cantonese and other.
[b] 1 = Consumer, 0 = Nonconsumer.
[c] Per pack of 10.
[d] Number of beverages (0–3) drunk "burning hot."
Source: Breslow 1982.

one tries to explicitly estimate the stratum parameters α_j is somewhat greater than the factor $M/(M+1) = 1.25$ predicted by results for a single binary exposure (Breslow 1981). It well illustrates the problems of likelihood inference with large numbers of parameters. The fact that the original analysis that ignored the matching agreed with the new, correct analysis was, of course, fortuitous and suggested that the matching variables were not strongly associated with the exposures. Unmatched analyses of matched data generally yield conservative estimates of relative risk (Armitage 1975; Breslow and Day 1980, table 7.12).

Prentice and Breslow (1978), in a paper that further clarified the conceptual foundations of the case-control study, derived the conditional likelihood (10) from failure time considerations. One starts with a large (voire infinite) population that is followed forward in time. For an individual with exposures $\mathbf{x}$, the disease incidence rate at time t is specified as $\lambda(t|\mathbf{x}) = \lambda_0(t)\exp(\mathbf{x}\boldsymbol{\beta})$ (Cox 1972). At the time t_j of occurrence of the jth disease case, M controls are sampled at random from the population. Conditioning on the unordered set of exposures for the case and controls then leads to (10). This derivation helps to explain why, with "incidence density sampling" (Miettinen 1976) where controls are sampled at the times of occurrence of the cases, the exposure odds ratio approximates the ratio of instantaneous disease rates and thus why the odds ratio is useful even for the study of common diseases (Greenland and Thomas 1982).

Nested Case-Control Studies

Although these conditional likelihood arguments were developed in the context of sampling from an infinite population, there is no reason why they cannot be applied also to sampling from an actual finite cohort. As noted earlier, this idea was already implicit in the 1959 Mantel–Haenszel paper. Mantel (1973) explicitly proposed sampling from a defined cohort, using an independent toss of a biased coin to decide whether or not each control would be included in the final sample. Motivated by a desire to reduce the computational burden, he termed the result a "synthetic" case-control study. Thomas was the first to propose sampling from the risk sets formed during a Cox regression analysis (Liddell, McDonald, and Thomas 1977). Figure 2 is a schematic of the risk sets in a cohort study. The basic idea is to replace each of them by a reduced risk set consisting of the case and a random sample (without replacement) of the remaining risk set members. Thomas proposed using the conditional likelihood (10) for inference, which of course has exactly the same form as Cox's (1975) partial likelihood for the original risk set. Here too the initial motivation was primarily computational. But it quickly became clear that the real value of such nested case-control sampling, as it came to be called, was for selection of individuals on whom additional data could be collected. The technique is particularly valuable when stored sera or other biological materials are available for a large cohort, but expensive laboratory assays are needed for quantitative exposure assessment.

Although the intuition underlying the nested case-control study is strong, and the use of the likelihood (10) is rendered plausible by the results for matched studies, more formal justification has taken time to develop. Oakes (1981) led the way with his derivation of (10) as a partial likelihood, but these arguments were still regarded as incomplete. Only recently have rigorous proofs appeared of the asymptotic consistency and normality of relative risks estimated by partial likelihood under nested case-control sampling (Goldstein and Langholz 1992). The most interesting of these proofs develop the theory in terms of marked point processes (Borgan, Goldstein, and Langholz in press). Besides confirming the asymptotic properties of the relative risk estimates, this approach also neatly solves the problem of how to use the nested case-control sample for estimation of the baseline cumulative incidence function.

Estimation of absolute risk functions as well as relative risk functions is in principle possible from a nested case-control sample, because one knows the sampling probabilities. If data from the full cohort are available, then the standard estimator of $\Lambda_0(t) = \int_0^t \lambda_0(u)\,du$ in the Cox model is

$$\hat{\Lambda}(t;\hat{\beta}) = \sum_{t_j \le t} \frac{1}{\sum_{l\varepsilon\mathcal{R}_j} \exp(\mathbf{x}_l\hat{\boldsymbol{\beta}})}, \qquad (11)$$

where $\mathcal{R}_j$ denotes the full risk set at the time t_j of occurrence of the jth case and $\hat{\boldsymbol{\beta}}$ is the partial likelihood estimate. Suppose that $\mathcal{R}_j$ contains N_j subjects including the case and that M controls are sampled for the reduced risk set $\tilde{\mathcal{R}}_j$. Borgan and Langholz (1993) and Borgan et al. (in

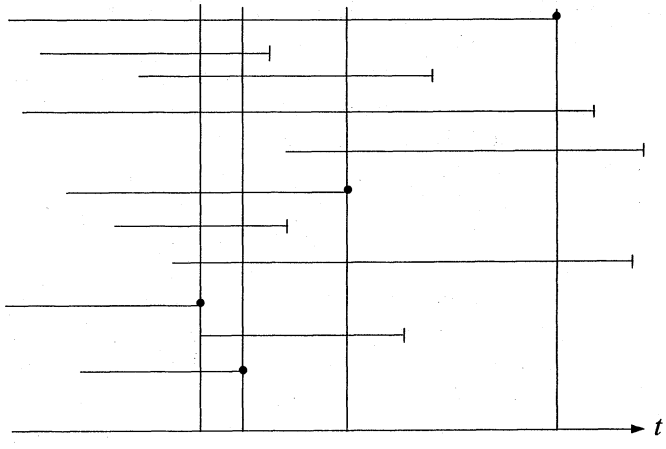

Figure 2. Definition of Risk Sets. Each horizontal line (——) denotes the observation period for a single subject as a function of time or age. Lines that terminate in a bullet (•) correspond to cases diagnosed at that time, whereas those that terminate with a bar (|) are noncases. The risk sets defined at each time of diagnosis contain those subjects whose observation period intersects the corresponding vertical line. (Adapted from Langholz and Clayton 1994, Fig. 1).