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. 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
  Ch 19: Individually matched case-control studies
  Ch 29: Conditional logistic regression
  Ch 33: Nested case-control studies
Several key historical and modern articles/reviews

+ 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

1926 Lane-Claypon: breast cancer
1950 Doll and Hill
1951 Cornfield 1951 ... odds ratio
1955 Woolf : quasi-denominators
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) $c_1/(d_1 \times \text{some constant})$.

The incidence in the unexposed(0) will be $c_0/(d_0 \times \text{the same constant})$.

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 ]


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[chd | 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 [i.e., outcome]-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

1976 Miettinen (Estimability & Estimation in case-referent studies AJE, 103; 226-235.)

"The concepts that case-referent studies provide for the estimation of "relative risk" only if the illness is "rare", and that the rates and risks themselves are inestimable, are overly superficial and restrictive. The ratio of incidence densities (forces of morbidity) -- and thereby the instantaneous risk-ratio -- is estimable without any rarity-assumption. Long-term risk-ratio can be computed through the coupling of case-referent data on exposure rates for various age categories with estimates, possibly from the study itself, of the corresponding age-specific incidence-densities for the exposed and nonexposed combined -- but again, no rarity-assumption is involved. Such data also provide for the assessment of exposure-specific incidence-rates and risks. Point estimation of the various parameters can be based on simple relationships among them, and in interval estimation it is sufficient to couple the point estimate with a value of the chi square statistic used in significance testing."
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 ... ....

Null model...

no scientific meaning [ c: d ratio determined by Investigator ]

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

logit[ Prob [case | x z1 z2 z3 ] ] = β₀ + β₁ × E + γ₁ × z₁ + ...
adjusted or[ E=1 vs. E=0 ] = exp[β₁_hat]
( again, β₀ has no scientific meaning)

Confounding, interaction, collinearity: as in earlier chapters)

Factors that affect precision of β₁_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[ β₁_hat ] = (Woolf variance) × VIF_{E<->Z}

\{ VIF_{E<->Z} = \frac{1}{1 - \text{Mult. } r^2 \text{ of } E <-> \text{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
n₁n₂/(n₁+n₂) = (reciprocal of (1/n₁ + 1/n₂)).. jh] measures the relative
information in the comparison of two averages based on sample sizes of
n₁ and n₂ 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, n₂, become arbitrarily large if the size of the
experimental group, n₁, 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

\[ \log \frac{P[Y = 1 | E Z]}{P[Y = 0 | E Z]} = \log \frac{f_{\text{cases}}}{f_{\text{controls}}} + \log \frac{P[Y = 1 | E Z]}{P[Y = 0 | E Z]} \]

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 1 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... [ technically*, denominators are person-weeks ]
No. of Houses with...

<table>
<thead>
<tr>
<th>Water</th>
<th>Impure</th>
<th>Pure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40 046</td>
<td>26 107</td>
</tr>
</tbody>
</table>

* p-w = # houses x ave. # per house x 4 weeks

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 * ]

<table>
<thead>
<tr>
<th>Water</th>
<th>Impure</th>
<th>Pure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>286</td>
<td>14</td>
</tr>
</tbody>
</table>

Attack rates in houses with...

<table>
<thead>
<tr>
<th>Water</th>
<th>Impure</th>
<th>Pure</th>
<th>Ratio</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>286</td>
<td>14</td>
<td>71.4 / 10K</td>
<td>5.4 / 10K</td>
</tr>
<tr>
<td></td>
<td>40046</td>
<td>26107</td>
<td>66 / 10K</td>
<td>13.3</td>
</tr>
</tbody>
</table>

This is the so-called "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?


2. Snow, John, 1813-1858. On the mode of communication of cholera : 191p, map, 23 cm. Location WC 262 S784 1936 [Regular Loan] Osler Library
The purpose of the 100 (or 1000, or however many 'proxies for person-moments' 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 & unexposed denominators within the source population. From the relative denominators sizes, the relative size of the incidence rates (or incidence proportions, depending on 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"...

Let's 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 14 13.3 no meaning
605 395 (±)

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 (see Resources):

"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 Study of Med. students' compliance with simple administrative tasks and success in final examinations? [med2 website]
Regression models for "case-control" and matched studies

/* 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 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 chol dbp sbp mrw smok ;

Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>2848 non-cases</th>
<th>283 cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>min</td>
<td>max</td>
<td>min</td>
</tr>
<tr>
<td>I_MALE (0/1)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>AGE in 1948 y</td>
<td>40</td>
<td>59</td>
</tr>
<tr>
<td>HT inches</td>
<td>51</td>
<td>76</td>
</tr>
<tr>
<td>WT lbs</td>
<td>67</td>
<td>300</td>
</tr>
<tr>
<td>CHOL</td>
<td>96</td>
<td>517</td>
</tr>
<tr>
<td>DBP diastolic</td>
<td>50</td>
<td>160</td>
</tr>
<tr>
<td>SBP systolic</td>
<td>82</td>
<td>300</td>
</tr>
<tr>
<td>MRW Rel.Weight</td>
<td>67</td>
<td>268</td>
</tr>
<tr>
<td>SMOK cigs/day</td>
<td>0</td>
<td>60</td>
</tr>
</tbody>
</table>

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.

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 a 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**

- **a** Ignore the matching variables (break the matches) and use an unconditional logistic regression with E and other (unmatched) z's
- **b** As in (a) but include E, the other (unmatched) z's, AND the matching variables
- **c** If matching variables are not 'measurable' (e.g. if match on family, or use twin pairs or siblings, 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).
- **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)
- **e** Use matching in analysis when didn't need to.

**Consequences**

- 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
- Not as severe as under a, but or still biased towards null

If 1 case and M controls per set, and exposure E is binary, the absolute value of $\beta = \log[OR]$ 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 it 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

Avoids the over-estimation in (c) and the under-estimation in (a) and (b). But may lead to larger SE's.

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


Methods We undertook a double blind, multiple crossover trial of 771 participants. In office buildings in Montreal, Canada, Ultraviolet Germicidal Irradiation (UVDGI) was alternately off for 12 weeks, then turned on for 4 weeks. We did this three times with UVDGI 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.

```
<table>
<thead>
<tr>
<th>UVGI: On</th>
<th>Off</th>
<th>On</th>
<th>Off</th>
<th>On</th>
<th>Off</th>
<th>On</th>
<th>Off</th>
<th>On</th>
<th>Off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sx +</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sx -</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Totals</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
```

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
```

e...
Regression models for "case-control" and matched studies

Riskset, and associated Likelihood contribution from subject # 2

UVGI Sx (relative) Odds
0 - \( \exp[0 \times \beta + 25 \times \gamma + ...] \) [1]
1 - \( \exp[1 \times \beta + 24 \times \gamma + ...] \) [2]
0 + \( \exp[0 \times \beta + 25 \times \gamma + ...] \) [3]
1 - \( \exp[1 \times \beta + 26 \times \gamma + ...] \) [4]
0 - \( \exp[0 \times \beta + 24 \times \gamma + ...] \) [5]
1 - \( \exp[1 \times \beta + 24 \times \gamma + ...] \) [6]

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

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

Conditional Logistic Regression via

SAS
PROC PHREG1;
MODEL time*Sx(0) = UVGI Temp Humidity CO2;
STRATA subject;

Stata
clogit sx uvgi temp humidity co2, group(subject) or
R ("survival" package)
clogit( sx ~ uvgi+temp+humidity+co2+strata(subject) )

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; strata() in R), 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 prob. 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 lengthy computations. Peto and Breslow gave approximations for similar continuous-time situations. If the data for #3 were

subject assessment UVGI Sx Temp Humidity CO2 time(pm)
3 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

\[
\]

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 examples of matched case-control studies, see H&L, Schlesselman, or Breslow&Day Vol 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.)
Walkers 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.

WORKED ANALYSIS OF MATCHED PAIR DATA

<table>
<thead>
<tr>
<th>Vas</th>
<th>Ob</th>
<th>Sm</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coef</th>
<th>SE</th>
<th>Wald</th>
<th>Pr &gt; Wald</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vas</td>
<td>0.20</td>
<td>0.40</td>
<td>0.24</td>
</tr>
<tr>
<td>Obse</td>
<td>1.18</td>
<td>0.78</td>
<td>2.28</td>
</tr>
<tr>
<td>Smok</td>
<td>1.42</td>
<td>0.64</td>
<td>4.97</td>
</tr>
</tbody>
</table>

R (survival package)

clogit(mi ~ vas + obse + smoke + strata(pair))

SAS 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  2  0  1  1  1  50
2  2  1  1  0  1  25
36 36 2 1 1 1
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 IDF (p=0.50)

Analysis of Maximum Likelihood Estimates

| Vas  | 0.20| 0.40| 0.24| 0.62|
| Obse | 1.18| 0.78| 2.28| 0.13|
| Smok | 1.42| 0.64| 4.97| 0.03|

Stata

input pairno vas obse smoke mi
clogit mi vas obse 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
obse 1.18 .78 1.51 0.13 -.35 2.70
smoke 1.42 .64 2.23 0.03 .17 2.67

EPIB 681:

CONDITIONAL LOGISTIC REGRESSION MODELS FOR (NOT NECESSARILY) "CASE-CONTROL" STUDIES
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

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

Each subject's record split into 2-year segments

<table>
<thead>
<tr>
<th>ID</th>
<th>I_MALE</th>
<th>AGE</th>
<th>MRW</th>
<th>SMOK</th>
<th>SBP</th>
<th>DBP</th>
<th>CHOL</th>
<th>A_NEWCHD</th>
<th>I_NEWCHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1</td>
<td>50</td>
<td>126</td>
<td>20</td>
<td>144</td>
<td>80</td>
<td>286.0</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>50</td>
<td>126</td>
<td>20</td>
<td>144</td>
<td>80</td>
<td>286.0</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>50</td>
<td>126</td>
<td>20</td>
<td>144</td>
<td>80</td>
<td>286.0</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>50</td>
<td>126</td>
<td>20</td>
<td>144</td>
<td>80</td>
<td>286.0</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>50</td>
<td>126</td>
<td>20</td>
<td>144</td>
<td>80</td>
<td>286.0</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>5050</td>
<td>1</td>
<td>59</td>
<td>97</td>
<td>40</td>
<td>148</td>
<td>86</td>
<td>213.0</td>
<td>64</td>
<td>1</td>
</tr>
</tbody>
</table>
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)

<table>
<thead>
<tr>
<th>Listing of a few of the 283 matched sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>M</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>I</td>
</tr>
<tr>
<td>L</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

On the left, shown in blue are 4 subjects chosen at random, without replacement, from the 50 subjects. (4 controls per case)
### Conditional logistic regression models for "Nested" case-control studies

**EPIB 681: Conditional logistic regression models for "Nested" case-control studies**

**proc sort data=cc ; by case_id;**

**proc phreg data = cc;**

**Variable Risk Ratio   Lower  Upper  Label**

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

**proc sort data=cc ; by i_male case_id;**

**proc phreg data = cc; by i_male ;**

**Variable Risk Ratio   Lower  Upper  Label**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Risk Ratio</th>
<th>Lower</th>
<th>Upper</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRW</td>
<td>1.005</td>
<td>0.995</td>
<td>1.013</td>
<td></td>
</tr>
<tr>
<td>SMOK</td>
<td>1.002</td>
<td>0.972</td>
<td>1.031</td>
<td></td>
</tr>
<tr>
<td>SBP</td>
<td>1.014</td>
<td>1.002</td>
<td>1.044</td>
<td></td>
</tr>
<tr>
<td>DBP</td>
<td>0.989</td>
<td>0.965</td>
<td>1.024</td>
<td></td>
</tr>
<tr>
<td>CHOL</td>
<td>1.005</td>
<td>1.000</td>
<td>1.009</td>
<td></td>
</tr>
</tbody>
</table>

**proc sort data=products ; set cc;**

**proc phreg data = products;**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Standard</th>
<th>Wald</th>
<th>Pr &gt;</th>
<th>Chi-Square</th>
<th>Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRW</td>
<td>0.008849</td>
<td>0.00365</td>
<td>5.88612</td>
<td>0.0153</td>
<td></td>
</tr>
<tr>
<td>SMOK</td>
<td>0.023597</td>
<td>0.00594</td>
<td>15.79099</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>SBP</td>
<td>0.007210</td>
<td>0.00425</td>
<td>2.87955</td>
<td>0.0897</td>
<td></td>
</tr>
<tr>
<td>DBP</td>
<td>0.009001</td>
<td>0.00163</td>
<td>1.19995</td>
<td>0.2733</td>
<td></td>
</tr>
<tr>
<td>CHOL</td>
<td>0.007091</td>
<td>0.00101</td>
<td>18.99726</td>
<td>0.0001</td>
<td></td>
</tr>
</tbody>
</table>

* 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Conditional: likelihood</th>
<th>Unconditional: single α</th>
<th>Unconditional: multiple α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnic group⁵</td>
<td>1.27 ± .33</td>
<td>1.42 ± .36</td>
<td>1.62 ± .25</td>
</tr>
<tr>
<td>Chinese wine⁶</td>
<td>.51 ± .29</td>
<td>.54 ± .28</td>
<td>.68 ± .27</td>
</tr>
<tr>
<td>Cigarettes⁷</td>
<td>.11 ± .10</td>
<td>.11 ± .09</td>
<td>.16 ± .09</td>
</tr>
<tr>
<td>Temperature⁸</td>
<td>.79 ± .16</td>
<td>.76 ± .15</td>
<td>1.12 ± .15</td>
</tr>
</tbody>
</table>

⁵1 = Teochew and Hokkien, 0 = Cantonese and other.
⁶1 = Consumer, 0 = Nonconsumer.
⁷Per pack of 10.
⁸Number of beverages (0–3) drunk "burning hot."

Source: Breslow 1982.

one tries to explicitly estimate the stratum parameters $\alpha_j$ 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 $x$, the disease incidence rate at time $t$ is specified as $\lambda(t|x) = \lambda_0(t) \exp(x\beta)$ (Cox 1972). At the time $t_j$ of occurrence of the $j$th 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 random, 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

$$\hat{\Lambda}(t; \hat{\beta}) = \frac{1}{\sum_{t_j \leq t} \sum_{u \in R_j} \exp(x_i \hat{\beta})},$$

where $R_j$ denotes the full risk set at the time $t_j$ of occurrence of the $j$th case and $\hat{\beta}$ is the partial likelihood estimate. Suppose that $R_j$ contains $N_j$ subjects including the case and that $M$ controls are sampled for the reduced risk set $\bar{R}_j$. 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).