Oct 5, 2004

CONFIDENTIAL DRAFT

Survival analysis; risk sets; matched case control studies:
a unified view of some epidemiologic data-analyses.

Part I

James A. Hanley
Department of Epidemiology, Biostatistics and Occupational Health
McGill University, Montreal, Canada

ABSTRACT

Over the past decades, case-control studies gave gained wide acceptance and respectability. Two developments in particular have contributed most to this change. One is the concept of incidence-density sampling, put forward by Miettinen in 1976. The second is the concept of risk sets, used by Cox in 1972 primarily to develop likelihoods that allow one estimate hazard ratios without having to estimate the hazard functions themselves. One might also add a third, the arrival in textbooks and statistical packages of logistic regression in 1970, and the broader generalized linear model in the mid 1970's.

In this article, which is divided into two parts, I describe how these developments, and the extensions and insights that flowed from them, have (a) helped resolve confusion about the appropriate choice of controls in case-control studies, (b) extended the analytic tools available for epidemiologic data, and (c) even if the unity has not always been evident, allowed for different epidemiologic designs and analyses to be seen in a unified way. As a by-product, I discuss the proportional hazards model, its flexibility and the price one must pay for this, and how its parameters are fitted by the method of Maximum Likelihood. Throughout, the aim is to de-mystify concepts, using pictograms rather than equations whenever possible, and using upper case for an unknown parameter value, and lower case for an estimate of it. The concepts and principles will be illustrated by three examples drawn from investigations of the risk of myocardial infarction following vasectomy and, in part II, the effect of sexual activity on male longevity and the possible leukemic effects of contaminated drinking water.

INTRODUCTION

Compared with so-called cohort studies [Doll2001a.b], case-control studies [Paneth2002a,b] have a shorter and much more turbulent history. Views on their value and credibility have changed dramatically over the last several decades, from harsh criticism and controversy[Feinstein73,81l Mayes et al.88], to wide acceptance and respectability today [Breslow96]. Part of the reason for this dramatic improvement has to do with the convergence of two, initially parallel methodologic developments, one in epidemiology and one in biostatistics, in the 1970's. In this article, which is divided into two parts, I describe how these developments, and the extensions and insights that flowed from them, have (a) helped resolve confusion about the appropriate choice of controls in case-control studies, (b) extended the analytic tools available for epidemiologic data, and (c) even if the unity has not always been, or its not yet fully evident [Miettinen2004], allowed for different epidemiologic designs and analyses to be seen in a unified way. As a by-product, I discuss the proportional hazards model, its flexibility and the price one must pay for this, and how its parameters are fitted by the method of Maximum Likelihood. Throughout, the aim is to de-mystify concepts, using pictograms rather than equations whenever possible, and using upper case for an unknown parameter value, and lower case for an estimate of it. After a short discussion of incidence density and hazard functions, and their ratios, the concepts and principles will be illustrated by three examples, one in this article and two in the next.

Measures

Much of epidemiologic data analysis focuses on the "incidence density" measure, and on comparisons using incidence density ratios, while "hazards" and their ratios, are used in survival analyses. Incidence density is different from, and yet similar to, the statistical concept of 'hazard'. An incidence density has as a denominator a 'volume of experience' measured in person-time, for example the number of "driver-use-minutes" motor vehicle drivers drove while using cellular telephones. It is shown in two formats in Figure 1a, first in detail, driver by driver, with the turnovers in the pool of drivers using a cellular telephone shown explicitly, and then collectively, i.e., de-personalized. The number of motor vehicle collisions, divided by the driver-use-time in which they occurred, yields an incidence density. In another example, from the paper which introduced the term)Miettinen1976), the denominator consisted of a constant number of men

(77.4K), in the age-category 50-54, observed for 1.5 years (77.4 x 1.5 = 116.1K man-years). By the end of this period, there would have been a turnover of approximately 1.5/5 or 30% in the initial membership of the age-group. The 35 new cases of bladder cancer in this amount of experience yielded an incidence density of 30 cases / 100,000 man-years.

In contrast, hazards (plural) usually involve a *closed* population, and are indexed by (i.e., a function of) the time elapsed since a specific 'time zero'. This time zero defines the time of entry into a 'state'; the *exits from this state* are the events (transitions) of interest. The hazard function is best understood in the context of life tables or survival functions (where the focus is on remaining in, or exiting from, the state of 'being alive'). Whereas hazard functions can be defined mathematically as "conditional exit probabilities per unit time", they can also (informally at least) be thought of as follow-up-time-specific incidence densities or rates, but within a narrow follow-up time-windows. The denominator associated with the hazard at a specific follow-up time t consists of the thin slice of person time during the short interval (t,t+ δt) contributed by those who have not yet suffered (or -- if it the event is pleasant -- reached) the event of interest. The numerator is the number of events (exits/transitions) in this interval. Or, in terms of a survival curve that descends from 1 at time zero, the denominator is the area under the 'survival' curve between t and t+ δt. and the numerator is the absolute amount by which the survival curve decreases in that interval[1]. Strictly speaking, the hazard at

[1] In an internet site [http://planetmath.org/encyclopedia/HazardFunction.html, motto: Math for the people, by the people] the hazard function is defined in words as

"the *probability* of death (non survival) at time t, given survival up to time t",

and in a formula as

h[t] = limit, as δt -> 0, of Prob[event in (t,t+ δt) given that event had not occurred by time t] / δt

The *formula*, which is correct, shows that the hazard is not a probability, but a *probability per unit time*, and that, depending on the time scale, (and unlike a probability) it can be bigger than 1. For example, if the time scale is expressed in centuries since takeoff, the failure rate after take-off is of the orfer of

The (conditional) *probability* in the numerator can itself be approximated by

$$\frac{S[t] - S[t+ \delta t]}{S[t]} .$$

Divided by δt , so that it has dimensions probability per unit time, it becomes

$$\frac{S[t] - S[t+ \delta t]}{S[t] \times \delta t} .$$

follow-up time *t* is the *limit* of the incidence density as the δt defining the time-window is made narrower and narrower; and it could equally be arrived at by considering windows that are *centered on*, rather than that have their *left boundary*, at *t.* The degree to which the *theoretical* incidence density centered at a specific time *t* varies as the window around *t* is made narrower depends on the specific context: e.g., it would be sensitive to the δ*t* if we were studying the rate of failures in the first few several minutes after takeoff of an aircraft, or adverse events after commencing a new medication, or onset of fever in the days after receiving a certain vaccination) but maybe quite insensitive at other times. And even at a *t* where the theoretical hazard is expected to be quite stable, the sensitivity of the empirical hazard to the choice of δ*t* is also influenced by the numbers of events from which it is derived.

Examples of hazard functions -- over somewhat longer horizons -- are rates of first motor vehicle accidents in the years after obtaining a driver's licence, or the occurrence of an MI in the years following a vasectomy. If one ignores the competing risks that accompany longer and longer follow-up, then rates of childhood leukemia, first experience with tobacco or alcohol, or first marriage, can all be though of as hazard functions if measured as a function of time since birth (i.e. age). Since only certain persons enter graduate school, or marriage, and do so at differing ages, rates of exit from these institutions ('states') are expressed as functions of the person-specific times from entry to these states. In some situations, there may have a choice of "meters" of elapsed time or of cumulated experience. For example, follow-up of new drivers might be indexed by the *distance* driven or the *length of time* they have been licensed. Just like incidence densities, hazard functions can be further divided or made specific by using particulars of person, place, and calendar time.

I: STUDY OF EFFECT OF VASECTOMY ON LONG-TERM RISK OF A MYOCARDIAL INFARCTION (MI)

Walker et al. (1981, 1982) studied myocardial infarctions in 4,830 pairs of vasectomized and non-vasectomized men from the membership files of a large group medical plan. The design is shown schematically in Figure 2(a). Each vasectomized man was matched on year of birth and year of surgery

The numerator of this quotient is (proportional to) the number of new events in the window. The denominator is (the same proportional to) the number of *persons* at risk, S[t], multiplied by the amount of *time*, δ*t*. i.e., to the area under the S curve, between *t* and (*t*+ δ*t*). Thus, the quotient has the same dimensions as an incidence density.

with a man having another minor surgery. For each pair, follow-up began when the pair members underwent surgery and (effectively) ended when the first of the pair suffered an MI or was lost to follow-up, or the data were analyzed.

Classical estimator of rate ratio

The "crude" results are shown in the top left of Figure 2(b): in 20 pairs, the vasectomized man had an MI during the follow-up period, in 16 pairs the non-vasectomized man had an MI, and in the remaining 4794 pairs neither man had the event of concern.

The Mantel-Haenszel estimate of the rate (i.e. incidence density) ratio, this *ratio* presumed constant over age/follow-up time and calendar years, is [Rothman2002]

$$\text{Sum}[(\text{MI}_{\text{V}} \times \text{FT}) / (2 \times \text{FT})] / \text{Sum}[(\text{MI}_{\text{not-v}} \times \text{FT}) / (2 \times \text{FT})]$$

where the sum is over all 4830 pairs, MI_{v} and $\text{MI}_{\text{not-v}}$ are a (0/1) indicators of an MI in the vasectomized , ('v') and non-vasectomized ('not') men in the pair, FT is the common length of time the pair is followed until the first event or the end of the common follow-up, and 2 × FT is the sum of these. Some 4810 of the pairs contribute zero to the numerator of this estimator, while the other 20 pairs contribute 1/2 each, for a total numerator of 10. Some 4814 pairs contribute zero to the denominator, while the other 16 contribute 1/2 each, for a total denominator of 8. Thus the estimate is rr_{MH} = 10/8, or 20/16, or 1.25. If one subdivides each follow-up interval into several smaller ones, e.g., using (horizontal) age-categories, or (vertical) calendar-time slices, or (diagonal) follow-up-time-slices, the estimate is still the same. No matter how finely one subdivides the follow-up time, only the 36 'last' slices containing the 36 MI's contribute to the numerator and denominator of the estimate. This key feature of the traditional estimator will come to the fore again below, where the 36 event-containing time slices, and the men they involve, will be called 'risksets'.

There are a number of ways to construct a confidence interval to accompany the point estimate of 1.25. One particularly instructive way, one that is similar to the conditioning used by Cox(1972) , is by 'conditioning' on the total of 36 events. To do so, one determines the range of *theoretical* rate ratio (RR) values that are compatible with how the observed events were distributed between 'v' and 'not-v' men. If this *theoretical* rate ratio were IDR or RR (I use Rate Ratio and Incidence Density Ratio interchangeably, and denote theoretical, i.e., unknowable, parameter values by upper case), then, *on average*, for ever pair in which the MI would occur to the 'non-v' member, the *expected* number of pairs where it would occur to the 'v' member of the pair is RR. Thus, *given* a pair with one MI, the *'after the fact'* probability P that it occurred in a 'v', rather than a 'non-v' member of the pair, is

$$\text{P = Probability[MI occurred in v, rather than non-v, member of pair]} = \frac{\text{RR}}{1+\text{RR}}$$

Inverting this leads to the theoretical relationship

$$\text{RR} = \text{P}/(1-\text{P})$$

The *observed* proportion p = 20/36 in our data leads to the following point estimate, rr, (lower case, for empirical, 'estimate of' RR):

$$\text{rr} = \text{p}/(1-\text{p}) = (20/36) / (1 - 20/36) = 20/16 = 1.25.$$

However, p = 0.56 is only a point estimate of the parameter P. Since it is based on an '*n*' of 36, it can be accompanied by a (say 95%) binomial-based confidence interval $\{P_{lower}, P_{upper}\}$ of {0.40, 0.72}, leading to the following 95% lower and upper limits for RR, the comparative parameter of interest,

$$\{ RR_{lower} , RR_{upper} \} = \{ P_{lower}/(1-P_{lower}), P_{upper}/(1-P_{upper}) \} = 0.7 \text{ to } 2.6.$$

The need for adjusted estimates of rate ratio

The data used to match pair members on birth date, age at surgery and follow-up were available in the computerized membership file; however, other relevant risk factors for MI, such as smoking and obesity had to be abstracted from the clinical records and thus were not used in the matching. (For simplicity, only the baseline values of these will be considered here)

As was the case in the matched-pair Mantel-Haenszel calculations for the 4830 pairs, we can again, without any loss of precision, restrict our adjusted analyses to the 36 "MI-containing" pairs -- these contain virtually all of the information on the true rate ratio. And again, we can focus only on the 36 last, narrow time-slices containing these events. Table 1 shows the composition of these 36 *risksets* with respect to smoking and obesity. Only 19 of the 36 pairs were concordant in their smoking history; in 7 of the remaining pairs, the vasectomized member smoked, but his counterpart did not; the pattern was reversed in the other 10 pairs. Some 25 pairs were similar with respect to obesity, but only 13 of the 36 pairs were matched on *both* factors. The CI associated with the $rr_{\text{fully matched}}$ = 9/4 = 2.25 based only on these 13 is very wide: 95%CI 0.6 to 10.0)

To add to these the information on these 13, the information from the 23 mismatched pairs one requires a *set of assumptions i.e., an explicit 'statistical model'*. For cohort studies such as this, one might posit a Poisson regression model for the observed numbers of events in the different sub-divisions of person time indexed by age, calendar year, follow-up time, smoking, obesity, and vasectomy; the variation in *absolute* incidence rates or hazards for *non-vasectomized* men across these "cells" would be taken as a parametric function, typically multiplicative, of these variables. The model would overlay this 'grid' of rates with the corresponding rates in *vasectomized* men; in the absence of effect modification, these latter rates in the vasectomized are assumed to be in the *same* proportion to their non-vasectomized counterparts -- over all possible covariate patterns. Such assumptions/models were used long before Cox, sometimes explicitly, and sometimes only implicitly, as, for example, when one calculates a rate ratio using a Mantel-Haenszel summary estimator.

The substantial number of parameters in such regression models can be a serious constraint, if, as here, the amount of data available -- 36 events in all -- is small. Many of the model parameters go towards constructing the absolute age-specific or follow-up-time-specific *incidence rates*, even though these parameters are merely a 'nuisance', i.e., of no direct interest. Even if one wished to estimate the absolute incidence rates as a function of age or calendar-time, one cannot reliably do so from so few events, even if one made strong -- and unverifiable -- assumptions.

The proportional hazards model

The first innovation introduced by Cox(1972) was to avoid specifying any explicit form for the *absolute* rates in each of the two compared groups, and to instead concentrate directly on the *relative* rates, i.e., on rate *ratios*. We already do this in classical case-control situations, and indeed, as we proceed, Cox's analysis will more and more resemble a so called case-control approach. Instead Cox used a form of the 'proportional hazards' or 'common rate ratio' assumption. His second, and to this author, fundamental innovation had to do not with the assumed *pattern* of rates, but with the *way the analysis was set up* -- i.e., with the special use of *conditioning* to eliminate what were in any case just nuisance parameters.

The proportional hazards model is illustrated schematically in Figure 2(c), using fully matched pairs, unspecified incidence rates (hazards), and a fictitious HR of 1.5. Technically speaking, because of its 'instantaneous' referent, it is difficult to conceptualize the 'hazard' at a specific time t, for example, that for non-smoking non-vasectomized men, operated on during their 39th year, for whom the 'follow-up clock' now shows exactly t = 3 years and 197 days post entry. For practical purposes however, the hazard can be adequately represented by taking the time window (the δt in the definition) to be the next 24 hours. In this window, the incidence density of MI's might be of the order of x.x per thousand man-years, or y.y per million man-days (numerical values for the densities at the different times t are deliberately left unspecified in our exposition, to emphasize that the hazard function h[t] is not estimated directly (it *can* be estimated, if even quite unreliably, but seldom is). Theoretically, the average density during a 1 week, or 1 month, or even a 1 year window centered on, or even immediately following, t = 3 years and 197 days would be very similar. The proportional hazards model, with a hazard ratio of 1.5, posits that if the hazard for non-smoking non-vasectomized man at time t is h.h events per unit of person-time, that for the vasectomized counterparts is 1.5 × h.h. At t = 7 years and 233 days, the absolute hazards would both be higher, but still in the ratio of 1.5:1. This constant hazard ratio over the span of follow-up t is also posited to hold for men entering the cohort, i.e., beginning follow-up, at different ages and different calendar years.

Risk sets

In order to estimate the postulated constant hazard ratio HR, the analysis, just as above, uses a *conditional* approach and considers *only* those pairs in which there was an 'event' (MI). For these, it considers only the small time slice that contains the event. For each of these 'risk sets', the data analyst pretends to be situated immediately after the event, and using the profile of the candidates just before the event, asks "why would/did the MI happen to the one its happened to (the 'case'), rather than to the other candidate(s) in the risk set? Of course, since it is not possible to answer this "why did it happen to the 'case'?" question case by case, the epidemiologic question is posed for the *collective*: why did the even*ts tend* to happen to *those* they happened to? Risk sets need not be restricted to just 2 candidates; in traditional survival analyses, there will be many candidates, while in nested, or other incidence density, case control studies there may be several. Also, even though the risk set is finalized just *before* the event, the logic is more in the after-the event, 'case-control', "why did it happen to the case?" spirit.

Estimating the HR parameter by the method of Maximum Likelihood

The method of Least Squares (LS), developed in the early 1800's for quantitative responses, seeks as the 'best' parameter value(s) that(those) which minimize(s) the discrepancies between the observed values and those 'predicted' by the model, whereas the method of Minimum Chi-Square, developed in the early 1900's for data in the form of, or converted into, frequencies, seeks the parameter value(s) which minimize(s) the discrepancies between the observed and 'predicted' frequencies. The method of Maximum Likelihood, introduced in 1912, determines the 'best' parameter value(s) by maximizing the 'Likelihood', defined as the probability of obtaining the observed data, calculated as a function of the parameter(s) in the model. If the outcomes in the *n* different units of observation (here pairs) are independent, this probability for the entire dataset is the product of the probabilities for each separate unit.

In order to avoid having to specify probability models for other -- and as Cox argued, irrelevant -- aspects of the data-generation process, such as how long men could *potentially* be followed, and hazard ratios at follow up times t at which there were no events, Cox instead focused only on the product of the *conditional*, *after-the-event*, probabilities associated with the data for each of the 36 pairs. Each probability is couched as the answer to the question "Given an MI-containing pair, what is the (a-posteriori) probability that the MI *occurred to the man it happened to rather than to the other man*?". Figure 4 illustrates the estimation procedure for a reduced dataset, containing the data from 4 'event-containing' pairs.

Three features are of note: (i) as a product of individual probabilities, the likelihood becomes quite small, and the 'support' for different HR values is more easily tracked using the natural log of the likelihood (ii) here, as with other comparative epidemiologic parameters measured in the ratio scale, the likelihood and log likelihood are more symmetric when the horizontal scale is linear in the log[HR] scale rather than the HR scale itself (iii) the Maximum Likelihood Estimate (MLE) of HR is found by locating where the tangent [derivatives] of the log likelihood function with respect to its parameter is zero; in this simple example, the MLE can be found from a closed-form equation derived by calculus. When this is not possible, iterative methods are used; these are similar to the use of a walking stick by a blind person who, when climbing a mountain, has to repeatedly decide 'which way is up'.

MLE of HR parameter in a more complex situation

Whereas in this example the ML estimate $\mathrm{hr}_{\mathrm{ML}}$=3 is no surprise, and other estimate would be, the MLE is not quite so obvious, and not even closed form, when we are forced to incorporate data from unmatched pairs. The additional assumptions required to do this can best be understood by examining the assumptions behind a Mantel-Haenszel summary estimate derived from all of the fully-matched pairs. The pursuit of a single ('overall') estimate makes the -- at least implicit -- assumption that the true hazard ratio HR in the V:NonV contrast is the same, not only at each age, but also in smokers (S) and non-smokers (NonS), i.e.,

$$\frac{\text{Hazard [V ; S]}}{\text{Hazard [NonV ; S]}} = \frac{\text{Hazard [V ; NonS]}}{\text{Hazard [NonV ; NonS]}}$$

It is instructive to rewrite this so as to *contrast smokers and non-smokers* who are concordant with respect to vasectomy:

$$\frac{\text{Hazard [S ; V]}}{\text{Hazard [NonS ; V]}} = \frac{\text{Hazard [S ; NonV]}}{\text{Hazard [NonS ; NonV]}}$$

which states that the S:NonS hazard ratio is homogeneous across vasectomized and non-vasectomized men. This equivalent formulation will be useful below, as will yet another:

$$\text{Hazard [V ; S]} = \text{Hazard [NonV ; NonS]} \times \frac{\text{Hazard [V ; NonS]}}{\text{Hazard [NonV ; NonS]}} \times \frac{\text{Hazard [S ; NonV]}}{\text{Hazard [NonS ; NonV]}}$$

which is equivalent to assuming that, relative to a comparison population with *neither* factor, the hazard ratio associated with *two* factors is the (multiplicative) *product of the two separate hazard ratios*. Clayton and Hills(1993) call this model the 'corner' model, since it begins with the hazard in the (0,0) or ('unexposed, no other riskfactor') corner, and works outwards from there.

This assumption, coupled with the conditional approach used earlier, now allows us to use *all 36* pairs, *whether fully matched or not*. We first illustrate how we adjust just for smoking, and later deal with smoking and obesity simultaneously. The approach is illustrated in Figure 4. In order to focus on the HR

parameter associated with vasectomy, and on how we 'adjust' for smoking, we begin by *assuming* that the two HR's associated with *smoking* i.e. Hazard [S ; V] / Hazard [NonS ; V]) and
Hazard [S ; NonV] / Hazard [NonS ; NonV]) are *both known to be 2* (one could for example, imagine that these 'external' values are based on *other* data, e.g., from the literature). Later, we will let this parameter value free to vary and estimate it from the dataset. The proportional hazards model now allows us to calculate, for different values of the HR for vasectomy, the conditional *probabilities of observing what we did observe* in the 36 pairs. The critical use of the assumption is in cases where the pairs are unmatched with respect to smoking. The calculations are illustrated in Fig 4. As above, in pairs where both the vasectomized and non-vasectomized man are non-smokers, the probability that the MI occurs in the vasectomized man is HR/(1+HR) and that it occurs in the non-vasectomized man is 1/(1+HR). In pairs where both smoke, the probability that the MI occurs in the vasectomized man is (2 × HR)/(2 + 2 × HR), which again simplifies to HR/(1+HR), and that it occurs in the non-vasectomized man is, after simplification, 1/(1+HR). In pairs where the vasectomized man smokes, but the non-vasectomized man does not, the hazards are now in the ratio of (2 × HR) : 1, so that the probabilities that the MI occurs in the vasectomized man is (2 × HR)/(1 + 2 × HR), and that it occurs in the non-vasectomized man is 1/(1 + 2 × HR). In the opposite configuration, where the vasectomized man does not smoke but the non-vasectomized man does, so that the hazards are in the ratio HR : 2, the probability that the MI occurs in the vasectomized man is HR/(2 + HR), and that it occurs in the non-vasectomized man is 2/(2 + HR). In the figure, we examine several scenarios for the HR and show the Likelihood (the product of the 36 probabilities) for each one. The maximum log likelihood of **-2x.x** is achieved at HR value of hr_{ML} = 1.34.

The hazard ratio associated with smoking was fixed at 2 simply for illustration, in order to make the search for the HR one-dimensional, and thus easier to visualize; However, the HR associated with vasectomy, and the one associated with smoking can be estimated simultaneously from the 36 pairs themselves: The 50:50 distribution of vasectomy:nonvasectomy in each riskset was designed for efficiency in estimating the HR for v:non-v, but using this dataset one can just as easily, if not as efficiently, concentrate on estimating the effect of smoking, while thinking of *vasectomy as the nuisance or confounding* variable. To estimate both parameters simultaneously, one again searches, but now simultaneously over two dimensions, for the values which mazimize the likelihood. As is seen in Figure 5, the maximum log likelihood of **-2xx** is achieved when the HR for vasectomy is **1.xx** and that for smoking is **3.xx**. And, as one can see from the greater 'sharpness' of the likelihood function with respect to the HR value for smoking, it turns out that the data set provides more information about it than about vasectomy. **[Check coeff. of variation]**

Regression formulation, 2-person risksets

Each probability in this "*2-person* risksets" example happens to have a traditional "unconditional logistic regression" form, but with the *pair* as the unit of analysis ('observation'). To see this, consider the binary 'outcome', Y, which is set to *1 for each pair where the vasectomized man was the one to suffer the MI,* and to 0 for each pair where his counterpart was. The probabilities of Y=1 for the four possible pair configurations are given in Table 2. By expressing HR_V and HR_S as $\exp(\beta_v)$ and $\exp(\beta_s)$ respectively, the logit of the probability that Y=1 can be written in a single 'master' equation that covers the four possible pair configurations

$$\text{logit [Prob[Y=1]]} = \beta_v + \beta_s \, d$$

where d is the difference between the indicators of smoking in the vasectomized and non-vasectomized men, i.e., d = 0 if both smoke, or both do not; 1 if the vasectomized smokes but the non-vasectomized does not; and -1 if the converse. This regression formulation can be extended to differences in other confounding variables such as obesity, to obtain, as Walker(1982) did, HR estimates of 1.2, 4.1, and 3.2 for vasectomy, smoking and obesity, respectively.

Some statistical asides

This clever use of unconditional logistic regression would not work if, unlike V, the 'X' variable of interest were a continuous variable, or if a riskset contained more than two men. To accommodate these more general situations, one would need to *reverse* the question, from *"were vasectomized men more likely to have a MI?"*, to the opposite, *conditional*, and closer to 'case-control', one: *"were men who suffered MI's more likely to have had a vasectomy?"* i.e., *"why did the MI happen to the man it occurred to?"* By using this *reverse* way of asking the question, the way in which the resulting probabilities are linked to the explanatory variables forms "what biostatisticians and epidemiologists now call the conditional logistic regression model for matched case-control groups; (...) economists and other social scientists call (it) a fixed-effects logit model for panel data"[ref: introduction to the clogit procedure in Stata]. The model is identical to McFadden's choice model (Breslow1996). Before software for conditional logistic regression and Cox's proportional hazards model became widely available, and when risksets were limited to pairs, some authors, notably Holford(1978), and Breslow and Day[ref], estimated the model parameters of this conditional logistic regression model by applying the standard unconditional logistic regression software, with each pair as a single observation. With the data now defined in terms of focus on the *case*, i.e., the man who suffered the MI, the "Y" was set to 1 for each observation; the "X's" were I_v: whether the case was the one who had had the vasectomy, and the differences (0,1, or -1) between the 'case' and the 'non-case' with respect to smoking [and obesity]. The HR for vasectomy is estimated by exponentiating the coefficient of I_V obtained from a 'no-intercept' unconditional logistic regression, fit with software such as GLIM. Software for unconditional regression available in SAS and Stata check for variation in Y, and finding none in this special case, would refuse to fit the parameters, If concern is with a binary V, one can still use the "V-based" (rather than case-based") approach described in the previous paragraph.

Larger risksets (but still just 1 case per riskset)

For larger risksets, such as arise in case-control studies with each set containing 1 case and several matched controls, the data cannot generally be accommodated within the family of the generalized linear models, of which unconditional logistic regression is a special case. The one exception is when 'exposure' is binary, and there are no other covariates. A closed form, non-ML solution is possible using the classical Mantel-Haenszel odds ratio estimator(1959). Miettinen (19xx) obtained a closed-form ML estimator in the special case where all sets are of size 3 (1 case and 2 controls). Breslow and Day (198x, pages xx-yy) treat the 'variable number of controls per case' situation, using special-purpose ML estimation algorithms. Nowadays, and for risksets in which members are unmatched with respect to certain measured covariates, packages such as Stata include a standalone conditional logistic regression module, similar to that described by Breslow and Day. Others, such as SAS, do not; instead they recommend the module for stratified proportional hazards model used initially in survival analysis. In doing so, they take advantage of the equivalence between the likelihood (and thus ML parameter estimates) under the conditional logistic regression model and the 'partial likelihood' used by Cox to fit the parameters of the proportional hazards model (see part II for a more detailed exposition on the use of proportional hazards model for classical 'survival' data, and on the link with case-control analyses).

Stratified Cox model

In a larger vasectomy study [Massey et al.,1984], identified men in four U.S. cities who had undergone vasectomies and paired each one with a neighbour of the same age and circumstances at the time the surgery was done. In a follow-up ranging from 1 to 41 years, some 200 of the vasectomized, and 250 non-vasectomized men suffered MI's. Some of their analyses used the matching, but did not include any unmatched variables. Other analyses broke the individual matching and either (a) compared incidence rates (events per person years) or (b) stratified the men by age and city and performed what were termed "stratified Cox-covariate analyses". These last two types of analyses are not as different as they might appear. First, in the limit, if one "slices" the Lexis diagram into very small rectangles and ignores rectangles in which there were no events, the resulting person-years analysis can be seen as a variant on the approach used here. Second. the stratified Cox-covariate analysis first divided the men into separate 'city'-'age at surgery' cells or *strata*. For the men within any one such cell, their time of surgery, or the time when their neighbour had surgery, becomes their 'time zero '. The successive risk sets in the same stratum are defined by the *order* in which MI's occurred along this time-scale; each riskset contains those men who are still being followed up, and thus "at risk", at the time of the riskset-defining MI. In this approach, a man who had an MI a certain number of years after vasectomy is in the risk set for each 'earlier' MI'. The likelihood for each risk set is calculated much as in the paired case, except that the risk set may now have several 'candidates'; as before, for each riskset, one calculates the probability of the MI happening to the member its happened to (the "case") rather than to the others. In this approach, one never contrasts those in one stratum with those in another stratum. The final step calculates the overall likelihood of the alignment of events within the each risk set within each stratum as the product of the individual likelihoods over risksets and strata. The benefit of *stratifying* on age at surgery, rather than using it as a variable in the regression model, is that it makes no assumptions about the hazard of an MI as a function of age. One is already making a large enough assumption that the relevant HR is constant over all times from surgery within each age-at-surgery stratum without postulating further structure in the data. A computational advantage is the smaller size of each riskset.

Risksets with several 'cases' ('ties' in survival analyses)

If the time scale is coarse, two or more (k) persons may suffer an MI at the 'same' time point. A number of approaches to this situation are available in most software. One way is to simplify the likelihood is to randomly break the tie, so that the risksets are ordered in time: in this approach, the later 'cases' become controls in the earlier risksets. Another, the basis for conditional logistic regression, is to calculate the probability of the events happening to the k candidates they happened to, rather than to any other combination of k candidates. Breslow and Day (1980, volume 1, and section 52. of Volume II) illustrate these calculations. If k and the size of the riskset are large, the substantial number of combinations of candidates can lead to considerable computations. Peto and Breslow (in Discussion of Cox1972) gave approximations for such situations. Subsequently, Gail offered a faster and more elegant, recursive solution(Gail, 1981; Storer1983).

DISCUSSION

'Time to event' versus its reciprocal (event rate): the statistical divide

Examples of 'matched' cohorts of the type illustrated here, are few, although the one by Walker was itself followed up by a much larger cohort study of a similar pair-matched design (Massey, 1984). So rare are they that a 2003 follow-up study which, for every 'exposed' person, used 10 randomly selected unexposed people, matched on 3 variables, a elicited a special commentary (Evans, 2003, commenting on Helms at al. 2003). Sadly, this commentary is a striking example of the wide 'divide' that the present article tries to bridge. The chasm is more related to the background of the data *analysts* than to the differences, which are fewer, in the statistical methods themselves. On one side are statisticians who grew up in a clinical trials culture, and who see Cox's model as a *survival analysis* technique, and who focus on, in Evans' words, "the *time taken to an event* that is the outcome under study, a *survival analysis*". The present author himself began as this type of biostatistician in 1973, and only saw the other (epidemiologic) side when he joined his present department in 1980, and found that his new colleagues Liddell and Thomas were, along with Breslow, the first to see how the same Cox model software could be used to analyze data from a 'nested' case-control study (Liddell 1976, Breslow 197x). (He will comment further on the epidemiologic analyses carried out by his former 'survival analysis' colleagues Lagakos and Zelen when, in Part II, he discusses their study of leukemia in relation to contaminated well water).

Using SAS procedure PHREG, the authors of the 2003 matched cohort study (which included 48,857 persons with foodborne infections, each one matched with 10 non-infected persons, matched for age, sex, and county of residence) compared the mortality of the infected and non-infected data "using conditional proportional hazard regression". They included a comorbidity index as an important, but unmatched, confounding variable. They reported their comparisons using relative mortality (effectively hazard ratios), over the entire 12 month follow-up window, and -- because of the sharp decline in this ratio over the follow-up period -- in several sub-windows. These analyses are similar to those we have illustrated for MI and vasectomy: "Elevens" (i.e. matched sets) in which there was no event (death) during the time window do not contribute to the (partial) likelihood, and so can be ignored. If each of the deaths (4707 in all) occurred in a different matched set (and most probably did!), then 48,857 - 4707 = 44,150 (i.e., more than 90%) of the matched sets were uninformative with respect to mortality ratios. It is not obvious from their report whether the authors took advantage of this). Given that they were largely obtained from administrative databases, the marginal cost of obtaining and processing these uninformative 441,500 records may have been minimal; however, as Walker(1982) emphasizes, if obtaining important exposure or covariate information involves substantial unit costs, these should be expended on the *informative*, i.e., *event-containing,* matched sets. The commentary does point out that "cohort studies usually have to be very large to obtain a sufficient number of outcome events". To this, one might add "*Once* the large number of events has been generated, we should use the data in the most cost-efficient and statistically-efficient way".

The other interesting lesson from both these studies is the artificial distinction between 'case-control and "cohort" studies, one that is unfortunately maintained by the BMJ commentary (Evans 2003)

> "Most *BMJ* readers are familiar with matched case-control studies but fewer will be familiar with matched cohort studies. Case-control studies are based on selecting cases of a disease and then finding people who are as similar as possible to the cases. The study by Helms *et al* is not a case-control study; people were selected not on the basis of having, or not having, the outcome of interest (in this instance mortality) but on the basis of being exposed or not to something that may affect mortality."

Helped by those who helped break this 'trinity' (which used to teach that "there are 3 kinds of study -- cross-sectional, cohort case-control") (Miettinen99), the case-control study is increasingly being seen as 'nested' in a cohort, either a virtual or --as here -- a real one. Moreover, as our examples show, even though the *authors* may have begun with a *cohort*, the *analyses begin* with the *event*, which in turns *defines* the *riskset*, which in turn makes this a matched case-control study. Suppose one did not know the community prevalence of vasectomy (or of persons having had food poisoning in the last year), or their ages, but was presented with several 2 × 2 tables, each consisting of 11 (2) persons, with 1 and 10(1) in the 'outcome' margin, and also 1 and 10(1) in the 'exposure' margin. From these, one would not know whether the data arose directly from a case-control study, or indirectly from a case-control type *analysis* based on cases arising from an actual cohort. The likelihood, and the resulting parameter estimates, are the same whether one sets up the likelihood based on answers to the question "Were the vasectomized(infected) more likely to have a MI (die)?" or "Were those who had an MI (died) more likely to have had a vasectomy (infection)? As we hinted at earlier, the choice of design is guided by *efficiency*: one fixes the frequencies in the margin that achieves, with a fixed overall total, the lowest values for the variance of the log HR.

[aside: As modern teachers emphasize, we are students of rates, not of exposure, and so -- *even* in 'case control' studies --conceptually we compare event rates in the exposed vs. rates in the unexposed, and *not* 'exposure in the cases vs. exposure in the controls' (Miettinen2004). The controls simply serve as *denominators*: quasi-denominators in case-control studies, and real denominators in cohort studies, (Miettinen2004). Second, the immediate effects of *transient* exposures, such as use of cellular telephones on motor vehicle accident rates, can *only be studied with a case-control approach*. A database may be helpful in *identifying* accidents or cell phone use, but a traditional cohort analysis would be *very* wasteful. Imagine, for the sake of illustration, that in Figure 1 there were 3 accidents, the first of which happened while the driver was, and the second and third while the driver was not, using the phone. Suppose that one carried out 'second by second' Mantel-Haenszel cumulations of the same type used in our first analysis of the vasectomy dataset to arrive at the rr_{MH} of 20/16 . One might measure, at each instant, the exposed and unexposed person moments ($PM_{exposed}$ and $PM_{unexposed}$, total PM). Most of the contributions to the Mantel-Haenszel numerator and denominator would be $0 \times PM_{unexposed} / PM = 0$ and $0 \times PM_{exposed} / PM = 0$! The *entire information* is contained in the PM distributions *just at the times of* *the 3 accidents*, [1], [2] and [3], . Thus, replacing person moments (PM) by numbers of persons (P) on or not on the telephone at these 3 instants, rr_{MH} is simply

$$\frac{1 \times P_{unexposed[1]} / P_{[1]} + 0 \times P_{unexposed[2]} / P_{[2]} + 0 \times P_{unexposed[3]} / P_{[3]}}{0 \times P_{exposed[1]} / P_{[1]} + 1 \times P_{exposed[2]} / P_{[2]} + 1 \times P_{exposed[3]} / P_{[3]}}$$

Little statistical efficiency is lost if, instead of the *exact* numbers of P's, one uses *estimates* of them, based on the exposed/unexposed distribution in random *samples* of the persons at risk (risksets) at the time of each event.

Case-crossover studies: new name for a very old way of evaluating causal associations

This last example raises the obvious question: why compare accident rates in people who are on the phone while driving with those in others who are not? why not compare accident rates in the *same* people when they are and are not on the phone while driving? In terms of the orientation of the people in Figure 1, this becomes a 'horizontal' rather than a 'vertical' comparison, and corresponds to the design used by Redelmeier and Tibshirani in one of the most important investigations of this question(Redelmeier1997). This design has become known as the case-crossover design, and is generally ascribed to MacClure, although variations on this design have been used in epidemiology for quite some time (e.g., New York pedestrian deaths(Friedman1974), and by individuals since time immemorial to investigate the origins of rashes, headaches, computer crashes, and other untoward personal events. The statistical analyses are just as in the matched case-control examples discussed above. We wonder why the design was not given the more informative "self-paired case control" label.

Acknowledgment

This work was supported by a grant from The Natural Sciences and Engineering Research Council of Canada.

References

Doll, R. Cohort studies: history of the method. I. Prospective cohort studies. Soz Praventivmed. 2001;46(2):75-86.

Doll R. Cohort studies: history of the method. II. Retrospective cohort studies. Soz Praventivmed. 2001;46(3):152-60.

Liddell FD. The development of cohort studies in epidemiology: a review. J Clin Epidemiol. 1988;41(12):1217-37.

Paneth N, Susser E, Susser M. Origins and early development of the case-control study: Part 1, Early evolution. Soz Praventivmed. 2002;47(5):282-8.

Paneth N, Susser E, Susser M. Origins and early development of the case-control study: Part 2, The case-control study from Lane-Claypon to 1950. Soz Praventivmed. 2002;47(6):359-65.

Feinstein AR. Clinical biostatistics. XX. The epidemiologic trohoc, the ablative risk ratio, and "retrospective" research. Clin Pharmacol Ther. 1973 Mar-Apr;14(2):291-307.

Feinstein AR, Horwitz RI, Spitzer WO, Battista RN. Coffee and pancreatic cancer. The problems of etiologic science and epidemiologic case-control research. JAMA. 1981 Aug 28;246(9):957-61.

Mayes LC, Horwitz RI, Feinstein AR. A collection of 56 topics with contradictory results in case-control research. Int J Epidemiol. 1988 Sep;17(3):680-5.

Breslow NE. Statistics in epidemiology: the case-control study. J Am Stat Assoc. 1996 Mar;91(433):14-28.

Miettinen OS. Lack of evolution of epidemiologic "methods and concepts". Soz Praventivmed. 2004;49(2):108-9.

Miettinen OS. Estimability and estimation in case-referreny studies.. Mmer J of Epidemiology, 103,226-235.

Walker AM et al. Vasectomy and non-fatal myocardial infarction. Lancet 1, 13-15, 1981.

Walker AM. Efficient assessment of confounder effects in matched follow-up studies. Applied Statistics, 31(3), 293-297, 1982.

Rothman KJ. (2002) *Epidemiology: An Introduction* . New York, Oxford university Press.

Cox DR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society B, 34, 269-276, 1972.

Clayton D and Hills M. Statistical models in epidemiology. Oxford ; New York: Oxford University Press, 1993.

Holford TR.The analysis of pair-matched case-control studies, a multivariate approach. Biometrics. 1978 Dec;34(4):665-72.

Mantel N & Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J. National Cancer Institute 11, 719-748, 1959

Massey FJ et al. Vasectomy and Health: results from a large cohort study. Journal of American Medical Association 252(8), 1023-1029, 1984.

Breslow NE & Day NE. Statistical methods in cancer research I. the analysis of case-control studies. Lyon: Intnl. Agency for Research on Cancer 1980.

Helms M, Vastrup P, Gerner-Smidt P, Mølbak K, Short and long term mortality associated with foodborne bacterial gastrointestinal infections: registry based study BMJ 2003; 326: 357.

Evans, S. Matched cohorts can be useful. Commentary. BMJ 2003; 326: 357

Liddell FDK et al. Methods of cohort analysis: appraisal by application to asbestos mining (with discussion). Journal of the Royal Statistical Society A, 140, 469-491, 1977.

Miettinen OS (1999). Etiologic research: needed revisions of concepts and principles. Scand J Work Envir Health 25 (6, special issue): 484–90.

Redelmeier DA and Tibshirani R. Association between cellular-telephone calls and motor vehicle accidents. N Engl J Med 1997;336:453-8.

Friedman GF. Primer of Epidemiology, 4th Edition 1994. Chapter 7 (**Case Control Studies**). McGraw-Hill; (1994)

Additional Reading

Cornfield J. A method of estimating comparative rates from clinical data: applications to cancer of the lung, breast and cervix. J. National Cancer Institute 11, 1269-1275, 1951.

Breslow NE et al. Multiplicative models cohort analysis. Journal of the American Statistical Association, 78, 1-12, 1983.

Breslow NE. Design and analysis of case-control studies. Annual Review of Public Health, 3: 29-54, 1982.

Breslow NE & Day NE. Statistical methods in cancer research I. the analysis of case-control studies. Lyon: Intnl. Agency for Research on Cancer 1980.

Breslow NE. Elementary methods of cohort analysis. International Journal of Epidemiology, 13(1) 112-115, 1984.

Table 1

Degree of concordance with respect to smoking (S) and obesity (O) in the 36 informative (MI-containing) pairs. Entries are the numbers of pairs in which vasectomized man had the covariate pattern shown in a given row and non-vasectomized man had the pattern shown in a given column. Shown in parentheses are the numbers of these pairs in which the MI occurred in the vasectomized/unvasectomized man. Numbers of pairs which are matched on the variable(s) are shown in bold. Data from Walker[ref].

Non-Vasectomized man

Vasectomized man

	S–	S+
S–	**10** **[6/4]**	*10* *[3/7]*
S+	*7* *[6/1]*	**9** **[5/4]**

	O–	O+
O–	**23** **[13/10]**	*2* *[0/2]*
O+	*9* *[6/3]*	**2** *[1/1]*

	S– O–	S– O+	S+ O–	S+ O+
S– O–	**7** **[5/2]**	*1* *[0/1]*	*6* *[1/5]*	
S– O+	*2* *[1/1]*		*3* *[2/1]*	*1* *[0/1]*
S+ O–	*5* *[4/1]*		**5** ***[3/2]***	*1* *[0/1]*
S+ O+	*2* *[2/0]*		*2* *[1/1]*	**1** **[1/0]**

FIGURE LEGENDS

Figure 1. Schematic of driver_use_time and driver_non-use_time that form the denominators of incidence densities. Shown are when, and for how long 100 different motor vehicle drivers drove while using or not using cellular telephones, during a specific time-window on a particular morning. The raw data are depicted in two formats (1) in detail, driver by driver, with the driving time shown as thin horizontal lines, and the time driving while using a cellular telephone in darker and thicker horizontal lines and (2) collectively, i.e., de-personalized. The height of the upper curve indicates how many were driving at the indicated instant, and the lower curve how many of them were at that instant using the phone while driving. The area under the lower curve represents the total number of driver-moments 'on-the-phone', and that between the two curves the total driver-moments 'off-the-phone'.

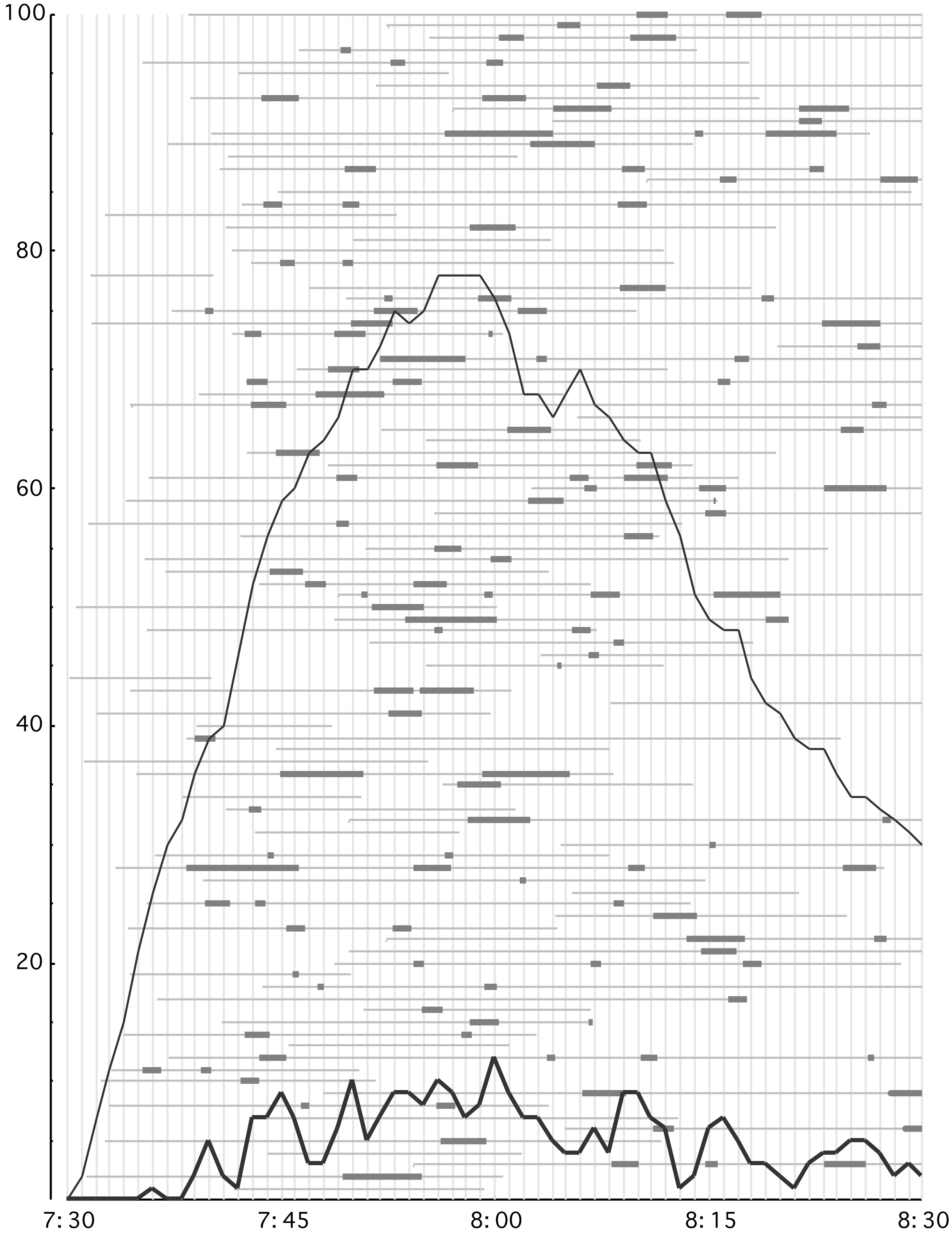

100
80
60
40
20
7: 30
7: 45
8: 00
8: 15
8: 30

Figure 2. Schematic representation of matched follow-up study of myocardial infarction in pairs of vasectomized and non-vasectomized men, and of proportional hazards assumption.

(a) each vasectomized man (shaded line) was matched on the basis of year of birth, age at surgery and calendar time of follow-up with a man who underwent another minor surgery (solid line). Follow-up began when the pair members underwent surgery and ended when the first of the members suffered an MI, denoted by a circle, (pairs 3, 4) or was lost to follow-up (pair 2), or the analysis was performed (pair 3).

(b) overall results in the 4830 pairs.

(c) the proportional hazards model, shown for two of the pairs:

c(3): excerpts of the hazard function h[t], with follow-up time denoted in this legend by 't', over the 13 years of follow-up, starting from "t-zero" = (1969, age 35). h[t] is shown in black for non-vasectomized men, and in gray for the vasectomized counterparts. At each follow-up instant t, the t-specific hazards -- the h[t, v] for vasectomized and the h[t, non] for non-vasectomized men of this age and era -- are assumed to be in constant proportion to each other; for illustrative purposes a constant hazard ratio (HR) of 1.5:1 is shown (this ratio is to be estimated). Although it may appear to be regular and parametric, the h[t] in the non-vasectomized has an *unspecified* form. A numerical scale for h has been deliberately omitted to emphasize that the two h[t] *functions are not estimated*. Rather, the object of the estimation is the *ratio* of the two functions, i.e. *the* (singular) hazard ratio.

c(4): excerpts of the $h[t]_{v}$ for vasectomized and $h[t]_{non}$ for the follow up of pairs from t-zero = (1970, age 44). Relative to pairs of type (3), the absolute value of the $h[t]_{non}$ function is higher, reflecting the older ages at t-zero, but the HR remains at 1.5:1.

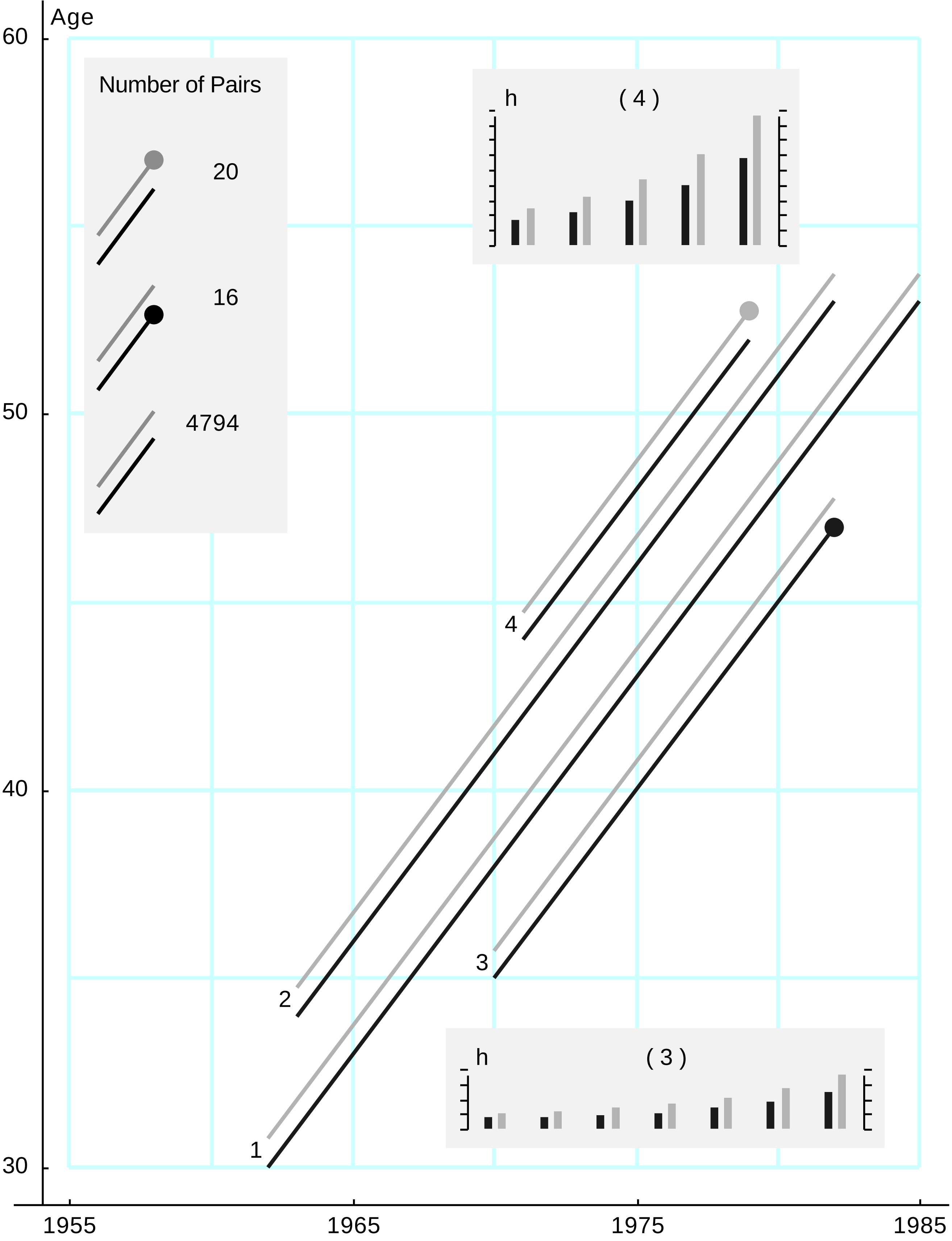
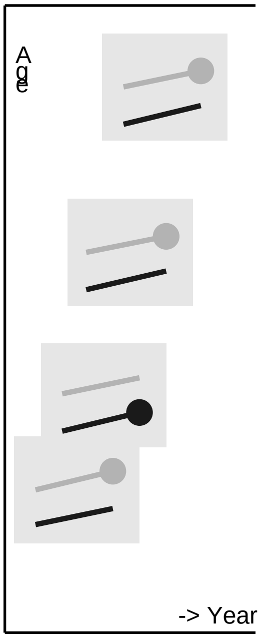

Age
60
50
40
30
1955
1965
1975
1985
Number of Pairs
20
16
4794
h
(4)
h
(3)
1
2
3
4

Figure 3. Illustration of Maximum Likelihood estimation of HR, based on data from 4 'event-containing' pairs shown at top left. In 3 of these 4 "risksets", the MI (the 'event'), indicated by a disk, occurred in the vasectomized man (shaded), while in 1 pair, it occurred in the non-vasectomized man. *The Likelihood function is the probability of observing this configuration of outcomes*, calculated as a function of the parameter(s) of interest, here the theoretical hazard ratio, HR. Since the results in the 4 risksets are independent of each other, the Likelihood is the product of the probabilities of the observed results in each the 4 risksets. In two men with the same risk profile, one of whom had suffered the MI, the probability that it would occur in the vasectomized rather than the non-vasectomized man is HR/(1+HR), and that it would be in the opposite configuration is 1/(1+HR). The probabilities of the 4 observed configurations are shown under the heading "Prob.". The 4 numerical probabilities in a particular column are the probabilities of the 4 observed configurations, calculated with the value of HR shown at the top of that column. The product, or 'Likelihood', is the probability of the observed alignment of MI's. Of the 7 HR values shown, the value of HR=4 produces the highest likelihood. A more refined numerical search, shown in smaller dots, illustrates that the Maximum of the Likelihood occurs at HR=3.

Fig 3

Observed data



	Hazard Ratio (HR):						
Prob.	0.5	1	2.	4.	8.	16.	32.
$\frac{HR}{1 + HR}$	1/3	1/2	2/3	4/5	8/9	16/17	32/33
$\frac{HR}{1 + HR}$	1/3	1/2	2/3	4/5	8/9	16/17	32/33
$\frac{1}{1 + HR}$	2/3	1/2	1/3	1/5	1/9	1/17	1/33
$\frac{HR}{1 + HR}$	1/3	1/2	2/3	4/5	8/9	16/17	32/33
Product [i.e., Likelihood]:	0.025	0.062	0.099	0.102	0.078	0.049	0.028
Log Likelihood:	-3.7	-2.77	-2.32	-2.28	-2.55	-3.02	-3.59

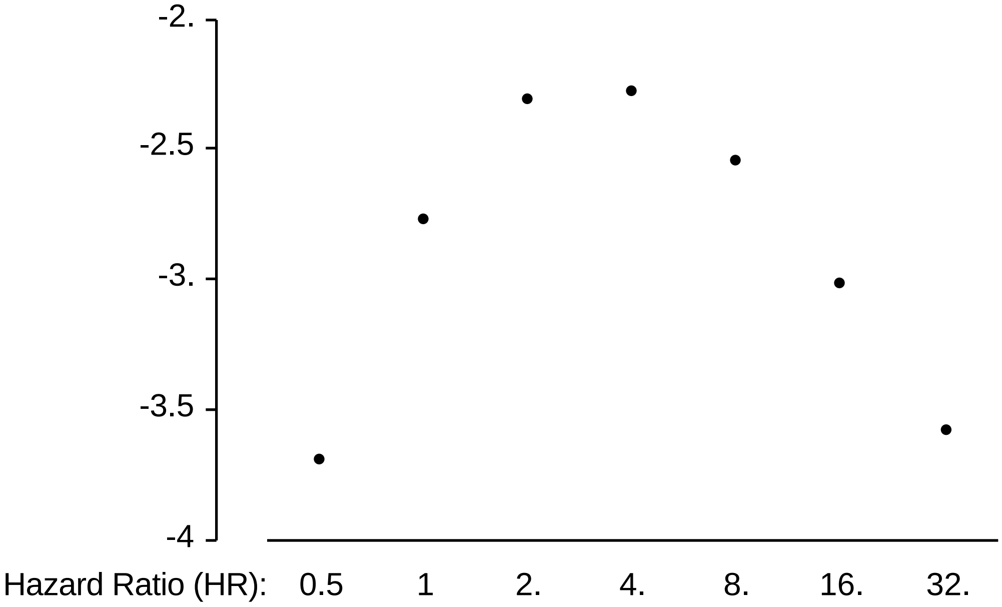


Figure 4. Illustration of Maximum Likelihood estimation of HR, based on data from all 36 'event-containing' pairs, 17 of which were not matched with respect to smoking (S). The hazard ratio associated with smoking is (for illustrative purposes) fixed at 2. Of the 10 "risksets" involving non-smokers, the MI (the 'event'), indicated by a disk, occurred in the vasectomized man (shaded) in 6, each with an associated probability of 1/(1+HR) while in 4 such pairs, it occurred in the non-vasectomized man(dark) with associated probabilities of 1/(1+HR) each. The 9 MI's in the non-smoking pairs split 5:4, with the same associated probabilities. Of the 7 pairs where only the vasectomized man smoked, the MI's split 6:1, with associated probabilities of 2HR/(1+2HR) and 1/(1+2HR) each. Conversely, of the 10 pairs where only the non-vasectomized man smoked, the MI's split 7:3, with associated probabilities of HR/(2+HR) and 2/(2+HR). The Likelihood function, namely the probability of observing the *entire* configuration of 36 outcomes, is the *product* of the 36 probabilities, calculated using the HR parameter shown at the top of each column. Shown at the head the 4 columns are different 'trial' values of HR, while shown in each row are the individual probabilities calculated using the trial HR value. To emphasize the *multiplicative* nature of the hazard, the components of each conditional probability are also shown – as small rectangles with bases of 1 and HR, and heights of 1 or 2(the assumed hazard ratio associated with smoking). The relative probabilities that the MI occurred in the vasectomized/non-vasectomized man are therefore proportional to the areas of these little rectangles: for example, if the vasectomized man did not smoke, and the non-vasectomized man did (last row), and if the HR for vasectomy is 4, then their relative probabilities of an MI are 1 × 4 = **4** and 2 × 1 = **2**. Thus if one was told one of these had an MI, one would say that the probability that it happened to the vasectomized/nonvasectomized man is 4/(4 + 2) = 2/3 and that it happened to the non-vasectomized man is 2/(4 + 2) = 1/3. Of the 4 trial HR values shown, the value of HR=1 produces the highest likelihood. A more refined numerical search, shown in smaller dots, illustrates that the Maximum of the Likelihood occurs at HR=1.34.

Fig 4

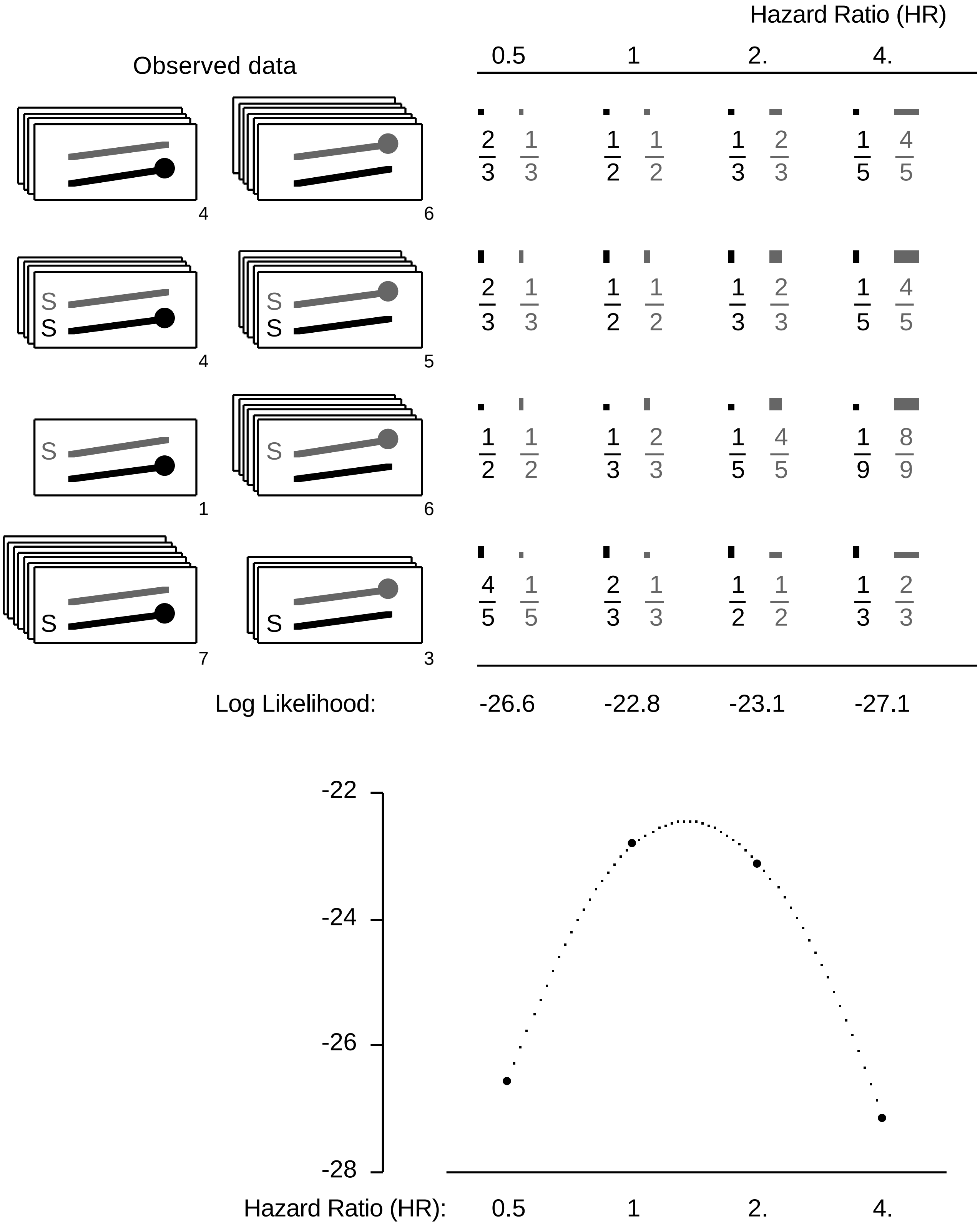

Hazard Ratio (HR)
0.5
1
2.
4.
Observed data
4
6
2/3 1/3 1/2 1/2 1/3 2/3 1/5 4/5
S
S
S
S
4
5
2/3 1/3 1/2 1/2 1/3 2/3 1/5 4/5
S
S
1
6
1/2 1/2 1/3 2/3 1/5 4/5 1/9 8/9
S
S
7
3
4/5 1/5 2/3 1/3 1/2 1/2 1/3 2/3
Log Likelihood:
-26.6
-22.8
-23.1
-27.1
-22
-24
-26
-28
Hazard Ratio (HR):
0.5
1
2.
4.

Figure 5. (3-D) MLE of HR's for **both** vasectomy and smoking

Fig to come...

Contour Plot of the Log Likelihood, as a function of the (logs of) the hazard ratios for Vasectomy (HR_V) and Smoking (HR_S)

For simplicity, logs of hazard ratios are to base 2, so that
LOG2HR_V = 1 means HR = 2^1 =2, LOG2HR_V = 3 means HR = 2^3 =8, etc.

For plotting purposes, Log Likelihoods coded so that they range '0' = -28, to '9' = -22

```
LOG2HR_S |
         |
       6 +
         |
         |                           00011110000
         |                        00111111111111100
         |                      011112222222222211110
       5 +                    0111222223333332222211100
         |                  0111222333333333333332221110
         |                 011222333444444444443333221110
         |                01122333444445555544444333222110
         |               012233344455555555555544443322110
       4 +              0122334445555666666665555444332211
         |             012233445556666666666666555444332110
         |            0112334455666677777777666655544332210
         |            1123344556667777777777776665554432210
         |           01223445566777778888887777666554432210
       3 +           11234455667778888888888877766554432210
         |          012334556677788888888888877766554432210
         |          01234455677788889999998888776655443211
         |          12234556677888899999999888776655433210
         |          12334556777888999999999888776655432210
       2 +          1233456677888999999999988877665443210
         |          1233456677888999999999888776655433210
         |          123345667788899999999888877665443210
         |          12334556777888999999888877665443211
         |          01234556677888888888888776655432210
       1 +          0123445667778888888887776655433210
         |           12234556667777888777766655433210
         |           0123345566677777777666554432210
         |            01233445566666666665544332210
         |            0112334445555555555544322110
       0 +             0012233444444444443322110
         |               0112233333333333221100
         |                0011122222222211100
         |                  00001111110010
         |                        000
      -1 +
         |
         ---+---------+---------+---------+---------+---------+--
           -2        -1          0         1         2         3

                                LOG2HR_V
```

```
options ls=80 ps=55;
data a;

do log2HR_V = -4 to 4 by 0.1;
  HR_V = 2**log2HR_V;
  do  log2HR_S = -3 to 6 by 0.1;
   HR_S = 2**log2HR_S;
   logLik = 4*log(1    *  1   / (1    *  1   +   1  * HR_V)) +
            6*log(1    * HR_V / (1    *  1   +   1  * HR_V)) +

            4*log(HR_S *  1   / (HR_S *  1   + HR_S * HR_V )) +
            5*log(HR_S * HR_V / (HR_S *  1   + HR_S * HR_V )) +

            1*log(1    *  1   /( 1    *  1   + HR_S * HR_V )) +
            6*log(HR_S * HR_V /( 1    *  1   + HR_S * HR_V )) +

            7*log(HR_S *  1   /( HR_S *  1   +   1  * HR_V )) +
            3*log(1    * HR_V /( HR_S *  1   +   1  * HR_V )) ;

   ll = round( 1.5*(logLik +28), 1.0);
   if (logLik  > -28) then output;
end;
end;
run;

proc plot;
 plot log2HR_S * log2HR_V = ll;

run;
```

Table 2

Probability that the vasectomized man was the one to suffer the MI (or equivalently, that the MI occurred in the vasectomized man), for each possible configuration of the riskset.

		Non-Vasectomized man ($I_v = 0$)	
		S– ($I_s = 0$)	S+ ($I_s = 1$)
Vasectomized man ($I_v = 1$)	S– ($I_s = 0$)	$\frac{HR_V}{HR_V + 1}$	$\frac{HR_V}{HR_V + HR_S}$
	S+ ($I_s = 1$)	$\frac{HR_V \times HR_S}{HR_V \times HR_S + 1}$	$\frac{HR_V \times HR_S}{HR_V \times HR_S + HR_S}$

S– , S+:	non-smoker, smoker
HR_V :	Hazard ratio, vasectomized : non-vasectomized
HR_S :	Hazard ratio, smokers : non-smokers