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[Feinstein73,811 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 $(t, t+\delta t)$ 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 $t+\delta t$. 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 $\delta t \to 0$, of Prob[event in $(t, t+\delta t)$ 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.

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.

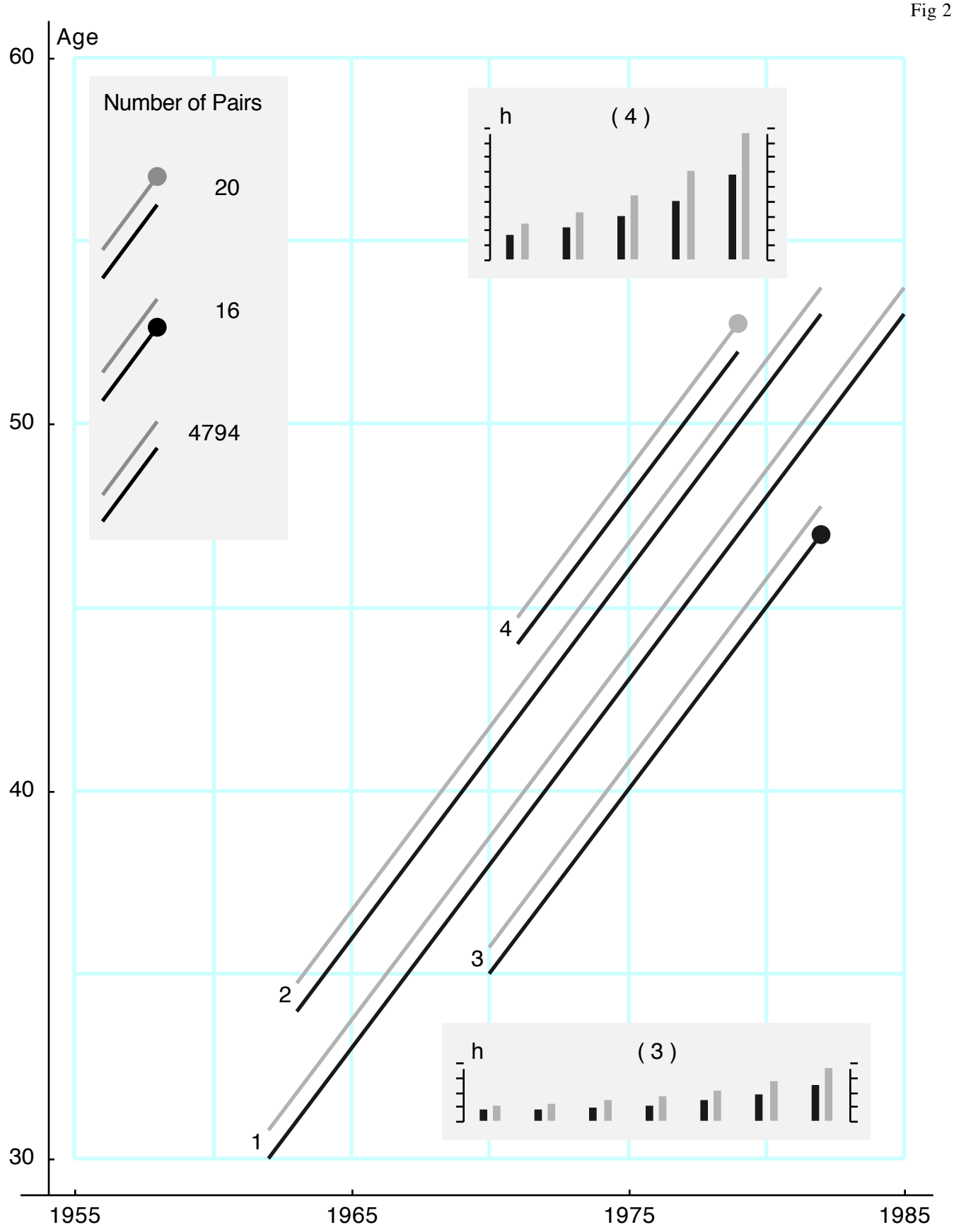
(a) each vasectomized man (shaded line) was matched on the basis of year of birth, age at surgery and calendar time of follow-up with a man who underwent another minor surgery (solid line). Follow-up began when the pair members underwent surgery and ended when the first of the members suffered an MI, 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

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]

$$\text{Sum}[(MI_V \times FT) / (2 \times FT)] / \text{Sum}[(MI_{not\text{-}v} \times FT) / (2 \times FT)]$$

where the sum is over all 4830 pairs, MI_v and $MI_{not\text{-}v}$ 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

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

Inverting this leads to the theoretical relationship

$$RR = P/(1-P)$$

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

$$rr = p/(1-p) = (20/36) / (1 - 20/36) = 20/16 = 1.25.$$

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

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]**

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 event*s 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.

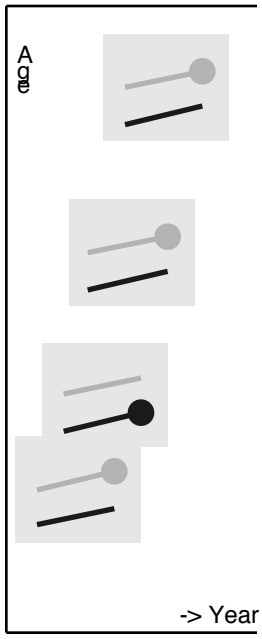
Estimating the HR parameter by the method of Maximum Likelihood

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

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.



Observed data	Prob.	Hazard Ratio (HR): 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

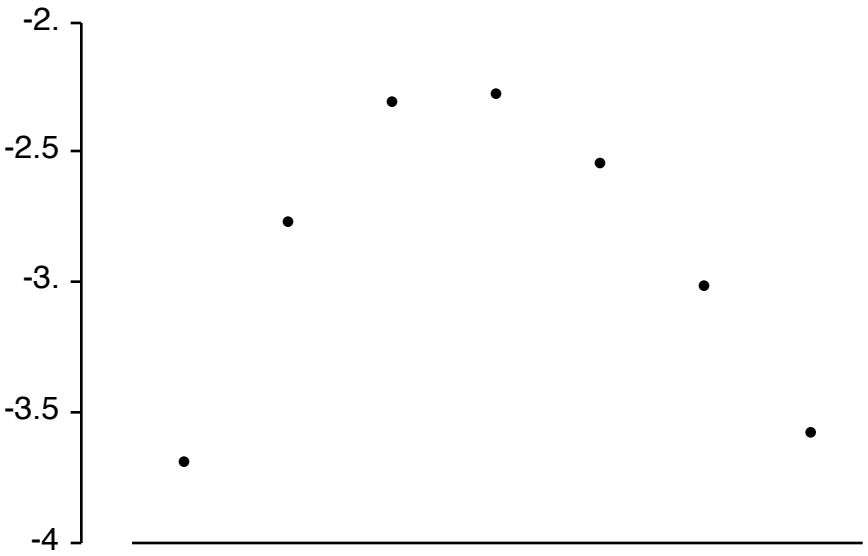
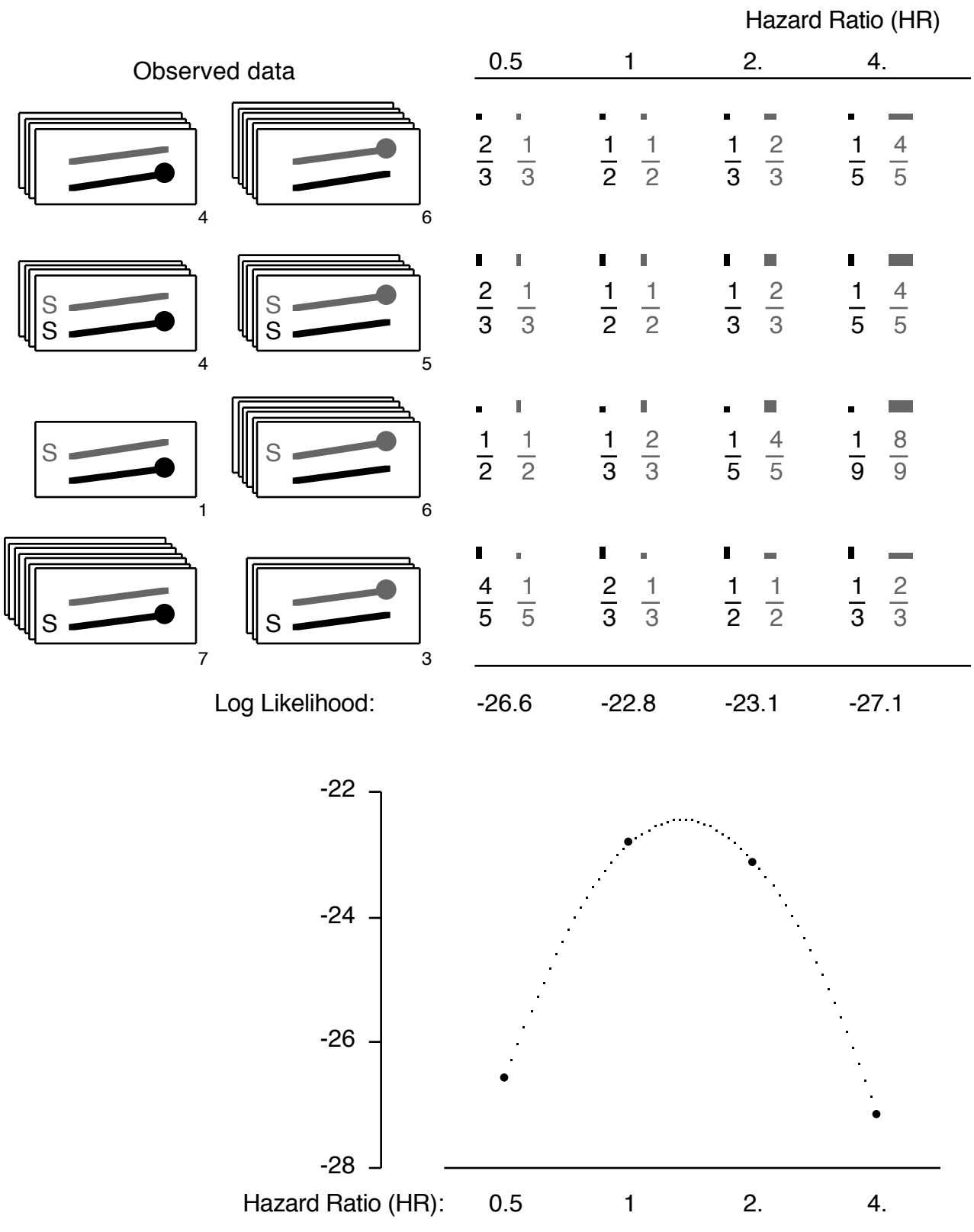


Figure 4. Illustration of Maximum Likelihood estimation of HR, based on data from all 36 'event-containing' pairs, 17 of which were not matched with respect to smoking (S). The hazard ratio associated with smoking is (for illustrative purposes) fixed at 2. Of the 10 "risksets" involving non-smokers, 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 $1 \times 4 =$ **4** and $2 \times 1 =$ **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.


Hazard Ratio (HR)
Observed data
0.5
1
2.
4.
4
6
S
S
4
5
S
1
6
S
7
S
3
2/3 1/3
1/2 1/2
1/3 2/3
1/5 4/5
1/2 1/2
1/3 2/3
1/5 4/5
1/9 8/9
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.

[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 $hr_{ML}=3$ 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:

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

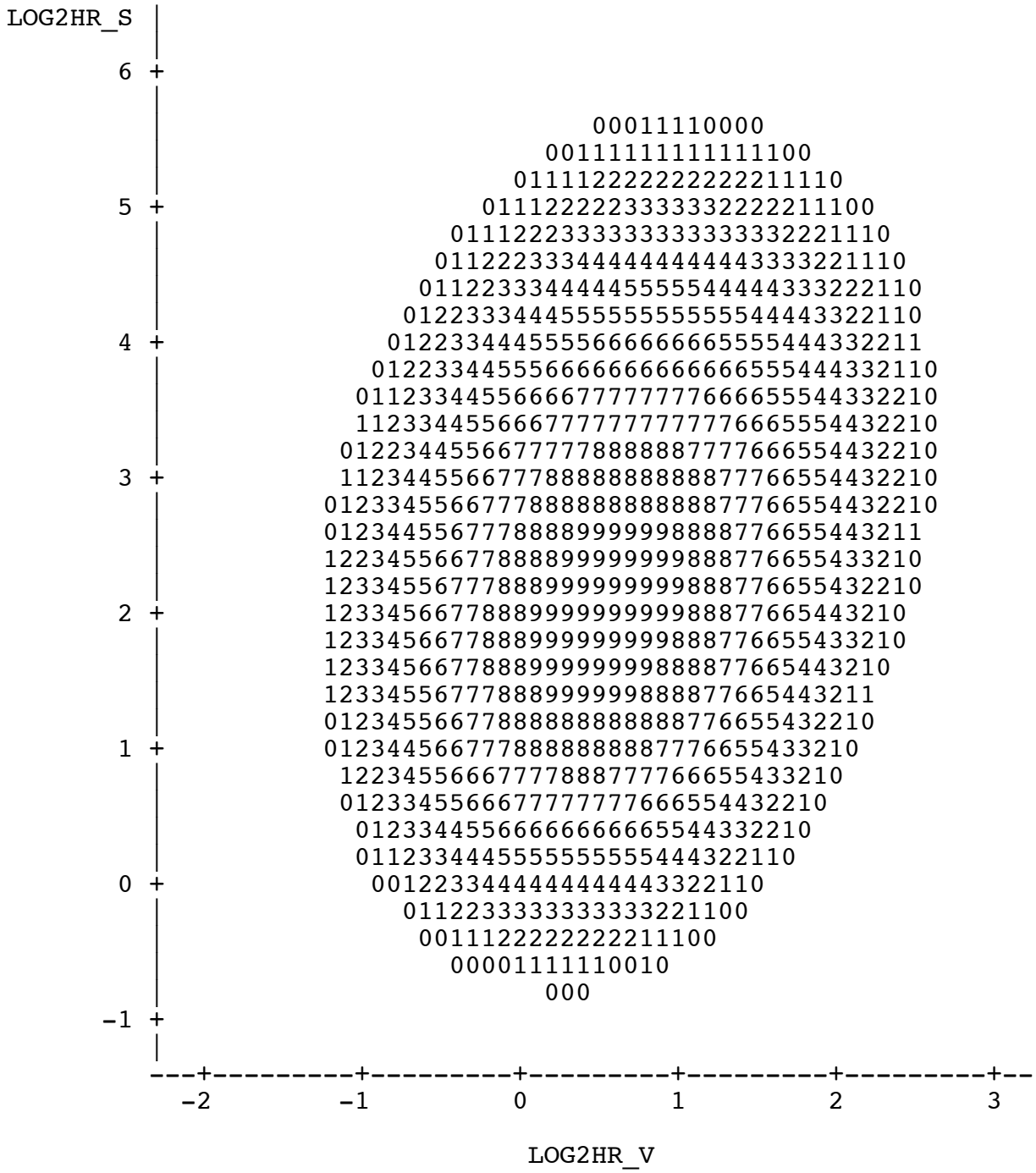
which is equivalent to assuming that, relative to a comparison population with *neither* factor, the hazard ratio associated with *two* factors is the (multiplicative) *product of the two separate hazard ratios*. Clayton and Hills(1993) call this model the 'corner' model, since it begins with the hazard 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

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



```
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;
```

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

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

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

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

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.

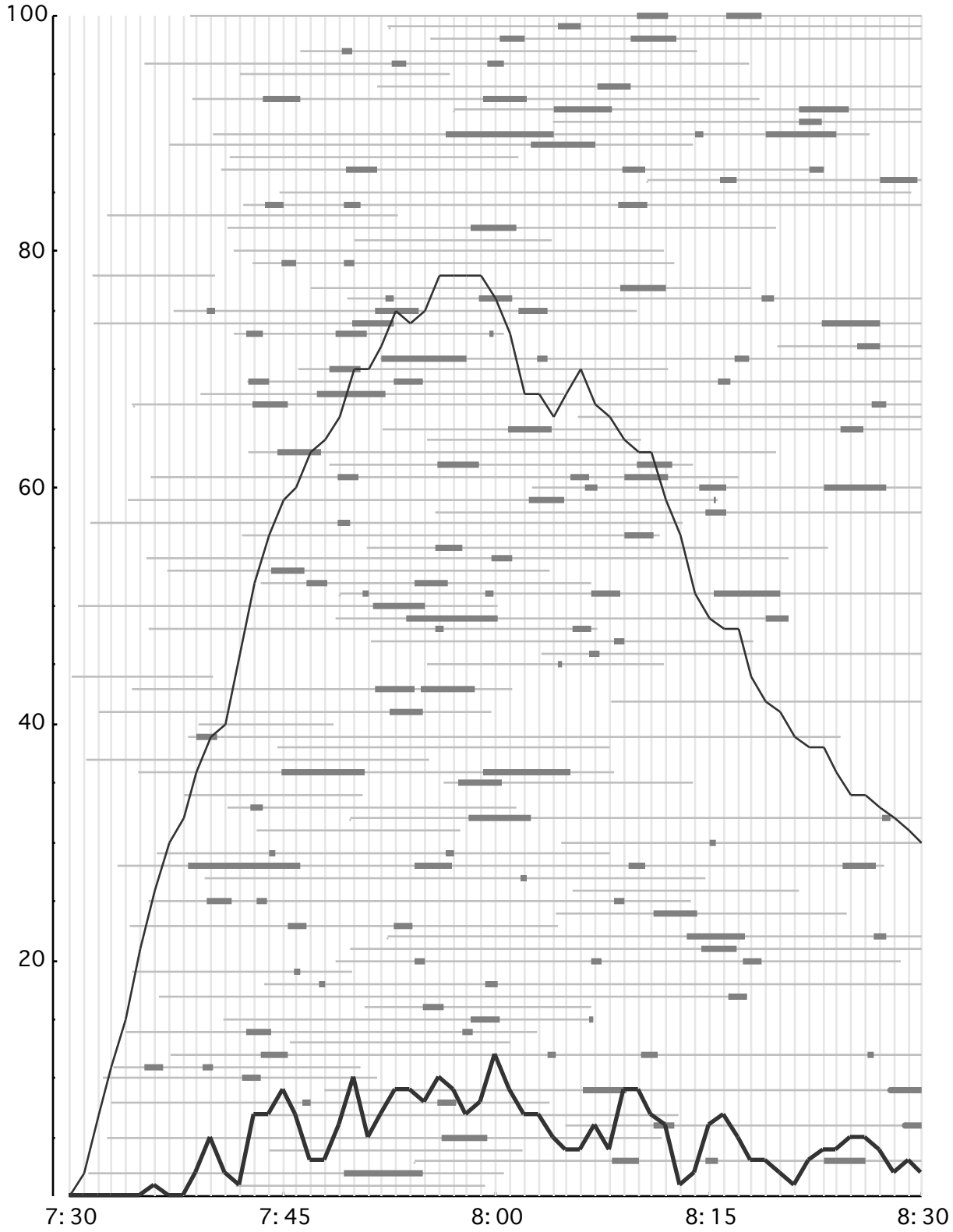
Breslow NE & Day NE. Statistical methods in cancer research I. the analysis of case-control studies. Lyon: Intnl. Agency for Research on Cancer 1980.

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

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'.

Fig 1



Oct 5, 2004

CONFIDENTIAL DRAFT

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

Part II

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

ABSTRACT

The first of the two articles in this series presented the proportional hazards model to analyze data arising from matched pairs followed until one of the pair-members had the event of concern. Using worked calculations and diagrams, I attempted to show what the model is, its flexibility and its assumptions, how its parameters are fitted, and how it can help us to see different epidemiologic designs and analyses in a more unified light. These are further illustrated in this second paper via two additional examples. In one, the longevity of two groups in an experimental study is compared via lifetable regression; in the other -- non-experimental -- the focus is on the degree of exposure to a contaminated source and its possible role in the etiology of cancer

To illustrate data analysis in an easier-to-compute situation, the first of the two articles in this series presented the proportional hazards model to analyze data arising from a very uncommon design, matched pairs followed until one of the pair-members had the event of concern. This second article focuses on the more traditional and more common: a survival analysis, via what Cox calls "lifetable regression" and a case-control study.

ILLUSTRATION II: DOES SEXUAL ACTIVITY DECREASE THE LONGEVITY OF MALES?

This experimental study investigated whether sexual activity reduces the lifespan of male fruitflies. The study, and the teaching dataset derived from it, are described in detail elsewhere (Partridge and Farquhar, 1981; Hanley and Shapiro, 1994). In all, in order to control for the possibility that reductions might be the result of competition for food etc., rather than sexual activity, five groups of 25 males each were formed by random allocation. In the two "experimental" groups, sexual activity was manipulated by supplying individually housed males with ("e") a smaller, or ("E") a larger number of, receptive virgin females each day. In two other "control" groups, individually housed males were supplied with ("c") a smaller or ("C") a larger number of, sexually *in*active (i.e., newly inseminated) females; the third control group consisted of 25 individually housed males who lived alone. Very large longevity differences were noted between the males in groups E and C, so here we restrict our analyses to the longevity comparison of "e" vs. "c". In order to show detailed hand-calculations, we analyze the data for just 10 males, 5 of whom we selected at random from the 25 in "e" and 5 from the 25 in "c". And, as the original authors did, we consider one important covariate, thorax size, which is a strong determinant of longevity.

There were no losses to follow-up, and all subjects had died by the time the data were analyzed. Thus the biologists analyzed the difference in the mean longevity using classical analysis of covariance. Today, we might accomplish this using thorax size as a covariate in a multiple regression model: if thorax length is included as a centered variable, then the fitted intercept denotes the mean in the reference group, the coefficient associated with the indicator variable for the experimental group is the adjusted difference in mean longevity, and the coefficient associated with the covariate denotes the independent relationship between it and longevity, This analysis makes the comparison "fairer" by adjusting for (what was a very

slight) imbalance in the groups with respect to thorax size. More importantly, even if the comparison had turned out, by a lucky randomization, to be perfectly fair, the inclusion of the covariate would make the comparison "sharper" (Hanley, 1983; Anderson et al. 1980). It does this by removing the (extraneous) variation in longevity caused by variation in thorax size: the smaller standard error of the difference in means leads to a more precise comparison.

Few epidemiologists work in research contexts where a long lifetime is 100 days rather than 100 years, where informed consent, missing data, confounding factors and multiple time scales are not an issue, and all subjects reach the endpoint of interest, leaving no censored observations. Nevertheless, we will analyze the data using survival analysis to see how this analysis is intimately linked with other data analysis approaches in epidemiology. We will take advantage of the simple structure -- no censored observations, and just two explanatory variables, both binary -- to see more clearly the *essential* elements of the Cox model, to understand how the likelihood is set up and maximized, and to illustrate how the Cox model. with stratification on nuisance covariates, can be an alternative to modeling them.

Figure 1 shows the longevity data, together with the associated risksets, and Maximum Likelihood estimation of the hazard ratio (HR) parameter of the proportional hazards model (vertical timelines, rather than the more common horizontal left to right ones, were used to allow the primary focus, the Likelihood function, to be drawn in the standard orientation). For now, the one covariate is ignored.

Time zero , time scales, and risksets

A time scale, starting at a defined "time-zero", is the point of departure for the 'survival' or other 'time-to-event' analysis version of the Cox life-table regression analysis(Cox1972). In non-experimental follow-up studies of humans, there may be several possible time scales. In analyses of the Framingham study, the most commonly employed time scale has been the time elapsed since "t_0" = 1948, even though this is simply an administrative scale, starting from the *funding* year zero. Age is often a more relevant time scale, Farewell and Cox(1979) have given some guidance on this, suggesting that the primary scale should be the one over which hazard rates vary the most, and are the most difficult to model accurately with a parametric

function and that other relevant time scales be included as regressor variables. In their example, dealing with the occurrence of breast cancer in parous women, the possible times scales were chronological age, and age since the birth of the woman's first child. In the fruitfly study, longevity was measured from when fruitflies emerged as adults (the fourth stage of their life cycle) and were allocated to one of the two conditions being investigated.

Using the selected time scale, each distinct event-time (death) defines an unambiguous riskset, namely those subjects alive just before the event in question. Subjects appear in successive risksets until they themselves suffer a riskset-defining event, or are lost to follow-up or otherwise censored. Thus, in this mortality study, each riskset contains those fruitflies who were alive just before the death which defines the riskset. And, unusually, since all subjects were followed until death, and each death occurred at a different age, there are as many risksets as there are subjects.

The elements of the (partial) Likelihood function

Using a specific HR value, one can, for each riskset, calculate the (conditional) probability that the death would occur to the subject who *did* die then, rather than to one of the other candidates in the riskset, who were also alive just before. The likelihood is a product of the probabilities associated with the different risksets. By calculating the likelihood for various values of HR, one is effectively asking why the deaths occurred in the *order* they did. That the first subject to die did so on the xx-th day of life rather than on some other day, that the second died on the yy-th specifically, and that there was or was not a large amount of *potential* follow-up time over and above what was needed, are not considered in the analysis. The probabilities used in the likelihood are *conditional on the individuals dying when they did*. and this conditioning leads to what is now referred to as a *partial* likelihood. The analysis does not ask why *then*? but rather "*who* then?", i.e., *given* that there was a death then, was it likely to happen to the person to *whom it happened* ? Thus, the Likelihood is not affected by the actual times, or by the time spaces between events. [This is one of the reasons why survival analysis software can be used to analyze case-control studies with matched sets, even if there is no, or no natural, time dimension: for the 'case' in the set, one simply

designates an *arbitrary* event-time t; for each control in the same set, one creates an event-time that is censored at or beyond t]

This *conditional approach to analysis*, i.e.., posing each probability as the answer to an after-the-fact "why the event in *this* person?" question, is *one* of the two reasons why the probabilities shown in Figure 1 have the simply form they do. The other has to do with the *form* of the proportional hazards model itself. The PH form was not new, even in 1972: constant (homogeneous) odds and incidence density ratios are used implicitly in the Mantel-Haenszel summary ratio measures, and explicitly in Poisson regression models that use multiplicative rates. In this example, with "t" (= adult age), the model posits that if $h_{inactive}[age]$ is the age-specific mortality rate (hazard) for sexually inactive subjects (reference category) of that age, then the corresponding mortality rate at the same age for their sexually active counterparts (the index category) is a constant times this, viz.

$$h_{inactive}[age] = HR \times h_{inactive}[age]$$

where HR is shorthand for the hazard ratio, *presumed constant over age.*

Note: we use the square brackets [ ] in $h_{inactive}[age]$ to denote that $h_{inactive}$ is a *function of* age, rather than a single number obtained by multiplying two other single numbers $h_{inactive}$ and age. Also, it is common, but potentially misleading, to use the term "baseline" hazard here: many authors use a subscript "0" where I have used the subscript "inactive"; by doing so, they may give the mistaken impression that the zero refers to *time 0*, when in fact they are referring to the *reference category of persons who have none of the risk factors or interest*, i.e., to persons with zero levels of all covariates. For a single categorical 'determinant', the notation

$$h_{index\text{-}category}[t] = HR \times h_{reference\text{-}pattern}[t]$$

males it clearer that $h_{reference\text{-}pattern}[t]$ is a *series* of h's, indexed by t, i.e. a time-function. Moreover, if one takes logs of both sides, then on the right-hand side the log of the time-function $h_{reference\text{-}pattern}[t]$ forms the (nuisance) "intercept' of a regression model, and log[HR] becomes the regression parameter of primary interest).

The implications of this form are best seen by example. Consider the fourth subject in figure 1 to die. Some 3 individuals from the active group and 4 from the inactive group were alive at the end of the previous day. The subject who died was in the sexually inactive group. *Imagine for now that we don't know that and that* we are simply given a list of the 7 members of the riskset at the end of the previous day, showing which group each one belonged to, and asked to try to identify the individual who died. What is the chance that we could identify the 'correct' individual? [1] For concreteness, let us say that the first 4 individuals in the list were the sexually inactive ones, and the last 3 the active ones, and that [although we don't know this] the event occurred in the *1st* subject on the list.

Because the 7 individuals in the riskset are all of the same age, and because the HR is constant at all ages, the $h_{inactive}$[age] factor drops out of the calculations: the 7 individuals' *relative* chances of being the 'case' are simply1:1:1:1:HR:HR:HR. If we *know* that there is *one* event, but not to whom, then the probability that it happened to a *specific* sexually inactive subject is $1/(4 \times 1 + 3 \times HR)$, and that it happened to a sexually active subject is $HR/(4 \times 1 + 3 \times HR)$.

But in our data analysis, we know that the event befell the *1st* individual, i.e., the one indicated in bold in the list **1**:1:1:1:HR:HR:HR. The *probability that it happened to this specific individual* is therefore $1/(4 \times 1 + 3 \times HR)$. Note that this probability no longer depends on $h_{inactive}$[age], but only on HR -- the *combination* of the *assumption* that the hazards are proportional , and the specific conditional probability formulated in reference to the riskset, leave us with a probability which only involves HR. If we consider a trial value of HR=1, then the probability that *we* could pick out the 'correct' individual, or that nature would 'finger' this specific (1st) individual from among the 7, is $1/(4 \times 1 + 3 \times 1) = 1/7$. If we use a trial value of

---

[1] OSM "***a person is not a case***" [OSM]. The person *represents* an '*instance*' of the phenomenon under study (this is one of the several OED definitions of case; the first definition the OED gives for 'case' is 'a thing that be*falls* or happens; an event, occurrence, ...' ; the word *case* comes from the Latin *casus* f. *cas-*, *cadere* , words all having to do with *fall*. The Latin dictionary at http://www.nd.edu/~archives/latgramm.htm gives **casus** -us m. [a falling , fall]. Transf.: (1) [what befalls, an accident, event, occurrence]. (2) [occasion, opportunity]. (3) [destruction, downfall, collapse]; and, in gen., [end]. (4) in grammar, [a case]]. and **cado** cadere cecidi [to fall , sink, drop]; 'vela cadunt', [are furled]; 'iuxta solem cadentem', [in the west]; of living beings, often [to fall in death, die]; hence [to be destroyed, to subside, sink, flag, fail]; 'cadere animis', [to lose heart]; with in or sub, [to come under, be subject to]; with in, [to agree with, be consistent with]; of events, [to fall out, happen]; of payments, [to fall due]. Of note is the fact that the 'case' focuses on the *event*, rather than on the *person* in whom it occurred.

HR=2, *the probability of the event happening to the individual it happened to* is $1/(4 \times 1 + 3 \times 2) = 1/10$, and so on.

The (partial) Likelihood function based on all 10 risksets

We have worked out the probability of observing the data we did observe for *one particular riskset*. We now calculate the corresponding probabilities for the other 9 risksets, and multiply together the 10 probabilities derived from these 10 different risksets. Each riskset-specific probability represents the *probability of the event happening to the individual it happened to*, and is a function of the parameter of interest, HR. The Likelihood, the product of these, is the probability based on the *observed time ordering* of the collection of 10 events. It concentrates on *who* in each riskset died, but not specifically *when*.

It now remains to work out this product (or, to avoid small products, the log of the product) for the *continuum* of candidate HR values, and to plot the Likelihood (or its log) as a function of HR, to determine which HR values make the observed data pattern more 'likely' than other values. The use of the *log* likelihood, which is a *sum* of individual log-likelihoods, also emphasizes the independent *additive* nature of the information from each riskset, just like the Mantel-Haenszel adds the information from separate strata. [In his very first paper on Likelihood, Fisher(1912) did not begin with the Likelihood (i.e., the product) *per se* but rather went directly to the log Likelihood, as a sum of log-Likelihood contributions from each observation or 'atom'.]

Incorporating confounding variables/ other covariates

We now move beyond a crude comparison to one that takes account of thorax size, an important determinant of survival. To simplify matters for didactic purposes, we dichotomize thorax size into smaller (s, the index category) or larger (reference category). There are two ways of incorporating a covariate into the proportional hazards analysis. One of these is to include it as a term in a regression model, the other is to stratify/match on thorax size.

*The **model-based*** approach is similar in spirit to a classical analysis of covariance which allows comparisons of *means* to be adjusted for imbalances in the distribution of important variables. In the crude model, the hazard function $h_{active}[age]$ for those in the active group was the simple product of the $h_{inactive}[age]$ function describing the hazard in the reference group, and the parameter HR, with the same HR value for all ages. In the simplest *multivariable* model, the hazard function for a group of individuals is the product of the $h_{inactive,large}[age]$ function for those larger, inactive individuals (the reference group, shown in the upper left cell in the table below), the parameter of interest HR (if applicable), *and* (again if applicable) a factor S (also a hazard ratio). The HR value is assumed constant over all ages and both thorax sizes, and the S factor is assumed constant over all ages and both activity levels (a model which allows the combined factor in the lower right cell to be sub- or super-multiplicative, but to remain constant for all ages, is still considered a proportional hazards model, since *the primary proportionality is across the 'time' axis* (age in this example).

Hazard function for groups of individuals, in relation to thorax size, and sexual activity (reference category: upper left "corner"; models of this type are referred to by Clayton and Hills(1993) as the "corner model")

	Thorax Size	
Sexual activity	larger	smaller
inactive	$h_{inactive,larger}[age]$	$h_{inactive,larger}[age] \times S$
active	$h_{inactive,larger}[age] \times HR$	$h_{inactive,larger}[age] \times HR \times S$

Consider again the probability of observing what we did in the previously examined riskset, where 7 individuals were alive at the close of the previous day. As shown in figure 2, of the 4 individuals in the list who were not sexually active, two were smaller and two were larger; of the 3 sexually active ones, 1 was smaller and 2 were larger. Given that one of them died on day 't', the easiest way to obtain their relative probabilities of 'being the one to die' is to list the 7 absolute hazards as in Table 1, then cancel out the common factor $h_{inactive,larger}[age]$ and arrive at the expression in the last column.

The probabilities for this and the other 9 risksets, as a function of HR and the nuisance parameter S, are shown in Figure 2. Also shown are sections of the log-likelihood surface for selected values of S, allowing us to see that, with simultaneous consideration of thorax size, the MLE of the HR is found at HR=2.4. Note the *symmetry* in the estimation process: the fitting of the 2-parameter model also provides a ML estimate of 3.2 for S.

The ***other*** approach to estimating HR from these data is to use a ***combination* of *matching* and regression**. This is illustrated in Figure 3, where subjects are first *segregated* by thorax size, so that *each riskset is matched with respect to this variable*. The likelihood, for any HR value, is again the product of the probabilities associated with the different risksets, each one now smaller and more homogeneous. Because we have not included thorax size in the regression, the likelihood function involves just the 1 parameter (HR) of direct interest. Implicitly however, by the act of pooling the log-likelihoods from the smaller and larger sub-populations, the analysis makes the further assumption that the HR's in the two sub-populations are *'poolable'*, i.e., that the two series are estimating a *common* HR. See the textbook by Kalbfleisch and Prentice(2002), who were the first to suggest this less restrictive 'model' -- this 'stratified' proportional hazards model allows the hazard functions in the reference categories (i.e., the $h_{inactive}$[age] function in the *smaller* individuals and the corresponding $h_{inactive}$[age] function in the *larger* individuals) to follow different *non-proportional* shapes over time (age).

Just as with matching in other contexts, one difficulty with this approach is that if strata are narrow, some of them may only contain individuals of one kind (e.g. all those of thorax size 0.72 mm are in the active group) and so -- just as in a Mantel-Haenszel summary ratio, do not contribute to the comparison. This problem would be worse in an observational study (the present study formed groups by randomization) and where there are important uncontrolled variables. The full multivariate model circumvents this by *mathematical*, rather than *actual*, matching: the price of this convenience is the uncertainty about the assumptions made, and the consequences of miss-specifying how the covariate affects the hazards.

Table 1. Calculation of the probability that the 4th event occurred (on day 't') to the individual (fruitfly) it occurred to, as a function of the hazard ratio HR associated with sexual activity (relative to the reference category, 'inactivity') and the hazard ratio (S) associated with being short (relative to the reference category, 'larger')

order in list#	Active?	Smaller?	hazard (absolute)			hazard (relative)	probability (conditional)
1		√	$h_0[\ ]$	× 1	× S	S	$\frac{S}{Sum}$
2			$h_0[\ ]$	× 1	× 1	1	$\frac{1}{Sum}$
3		√	$h_0[\ ]$	× 1	× S	S	$\frac{S}{Sum}$
4			$h_0[\ ]$	× 1	× 1	1	$\frac{1}{Sum}$
5	√	√	$h_0[\ ]$	× HR	× S	HR × S	$\frac{S}{Sum}$
6	√		$h_0[\ ]$	× HR	× 1	HR	$\frac{HR \times S}{Sum}$
7	√		$h_0[\ ]$	× HR	× 1	HR	$\frac{HR}{Sum}$
					Total:	Sum*	1

from left to right [4th 'earliest' riskset in Figure 1]

$h_{inactive,larger}[\ t\ ]$, the hazard function for the reference category, is abbreviated to $h_0[\ ]$.

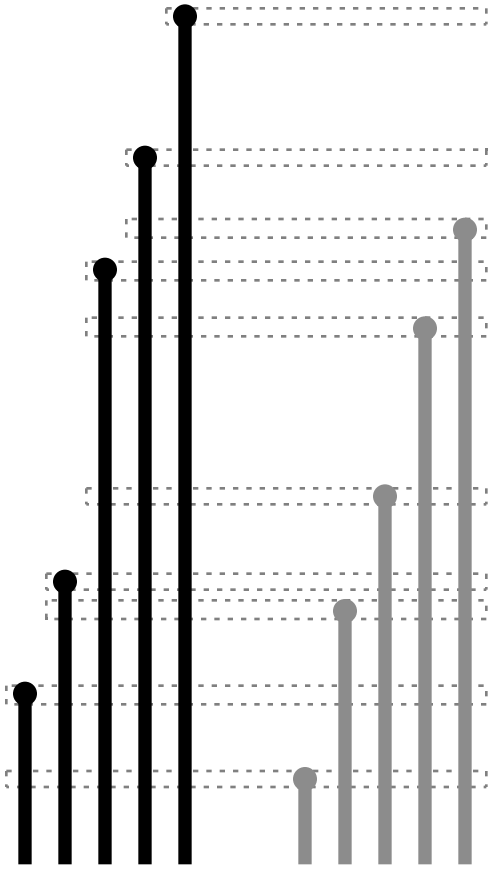
* $2 + 2 \times S + 2 \times HR + HR \times S$ abbreviated to 'Sum'

In this example, the event occurred to the 1st (leftmost) member on list.

FIGURES AND LEGENDS

Figure 1 Longevity of $n = 5$ sexually active male fruitflies (gray vertical lines) and $n = 5$ sexually inactive male fruitflies ((black vertical lines, reference group), together with the associated risksets, and Maximum Likelihood estimation of hazard ratio (HR) parameter in the (1-parameter) proportional hazards model which ignores thorax size. Circles denote age at death (longevity, survival time). In order to show all calculations clearly, the survival time axis is not perfectly to scale; the distortion is of no consequence, since the likelihood depends only on the ordering of the deaths. Risksets, one for each distinct event-time, are enclosed by dashed lines. The entries in the corresponding rows are the probabilities, calculated using the HR value in the column, that the death would occur to the subject who did die then, rather than in one of the other candidates in the riskset. As an example, consider the fourth subject to die, when the riskset consisted of 4 individuals from the inactive group and 4 from the active group. The subject who died, the leftmost of the 7, was in the sexually inactive group. If only told showing which group each of the 7 members of the riskset belonged to, and an HR value of say 2, the probably of replicating the results of this 'lottery', is 1/(1+1+1+1+2+2+2) = 1/10. The entire likelihood, for this HR value, is the product of the full (column of) probabilities associated with the different risksets. The Maximum (log-)Likelihood occurs at HR = 2.4.

Fig 1



	Hazard Ratio (HR) 1	2.	4.	8.
	1/1	1/1	1/1	1/1
	1/2	1/2	1/2	1/2
	1/3	1/2	2/3	4/5
	1/4	1/5	1/7	1/11
	1/5	2/7	4/11	8/19
	1/6	2/9	4/15	8/27
	1/7	1/10	1/16	1/28
	1/8	1/6	1/5	2/9
	1/9	1/13	1/21	1/37
	1/10	2/15	4/25	8/45
Log[Product] , i.e., log Likelihood:	-15.1	-14.43	-14.64	-15.57

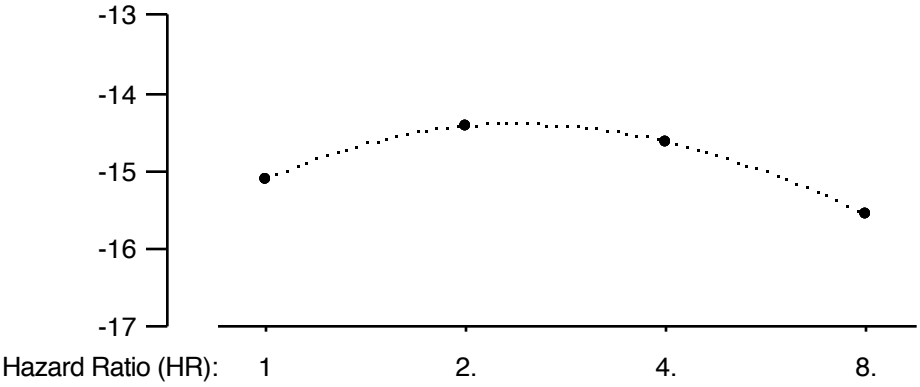
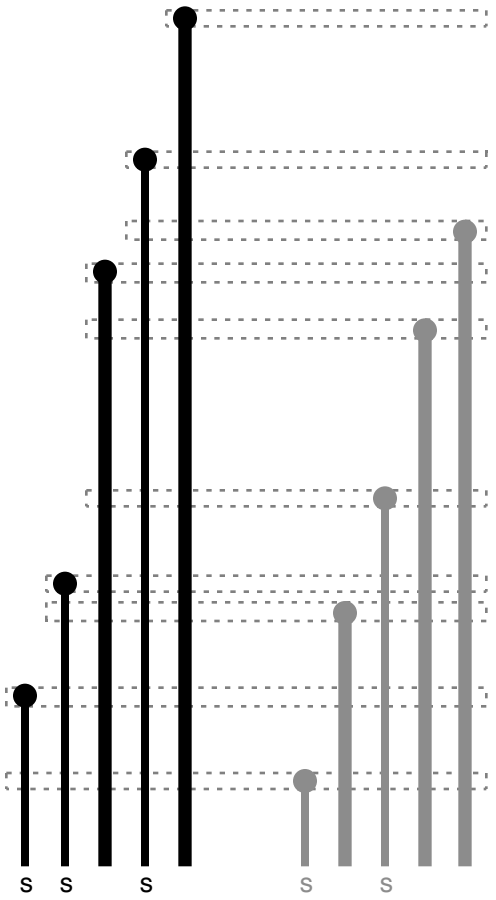


Figure 2 Maximum Likelihood estimation of 2-parameter proportional hazards model. Vertical lines represent the longevity of $n = 5$ sexually active fruitflies (gray) and $n = 5$ sexually inactive male fruitflies ((black, reference group). Three of the latter, and two of the former have shorter than average thorax lengths and are identified by the lowercase letter s and represented by thinner lines, while the remainder , with above average thorax lengths, are represented by thicker lines. Circles denote age at death and dashed lines enclose the risksets. The entries in the corresponding rows are the probabilities, calculated using the HR value in the column, and the hazard ratio S associated with a short thorax, that the death would occur to the subject who did die, rather than in one of the other candidates in the riskset. The likelihood, for a fixed value of S, and a specific HR value, is the product of the (column of) probabilities associated with the different risksets. Sections of the 2-D log-likelihood surface are shown for selected values of S: S=1 (same function as in Figure 1), 1.5, 3.2 and 8. The Maximum (log-)Likelihood occurs at HR = 3.5, S= 3.2.

Fig 2



	Hazard Ratio (HR) 1	2.	4.	8.
	1/1	1/1	1/1	1/1
	S/(1+S)	S/(1+S)	S/(1+S)	S/(1+S)
	1/(2+S)	2/(3+S)	4/(5+S)	8/(9+S)
	1/(3+S)	1/(4+S)	1/(6+S)	1/(10+S)
	1/(4+S)	2/(6+S)	4/(10+S)	8/(18+S)
	S/(4+2S)	2S/(6+3S)	4S/(10+5S)	8S/(18+9S)
	S/(4+3S) 1/(5+3S)	S/(6+4S) 2/(8+4S)	S/(10+6S) 4/(14+6S)	S/(18+10S) 8/(26+10S)
	S/(5+4S)	S/(8+5S)	S/(14+7S)	S/(26+11S)
	S/(5+5S)	2S/(8+7S)	4S/(14+11S)	8S/(26+19S)
Log[Product] , i.e., log Likelihood:	-14.62	-13.48	-13.22	-13.75

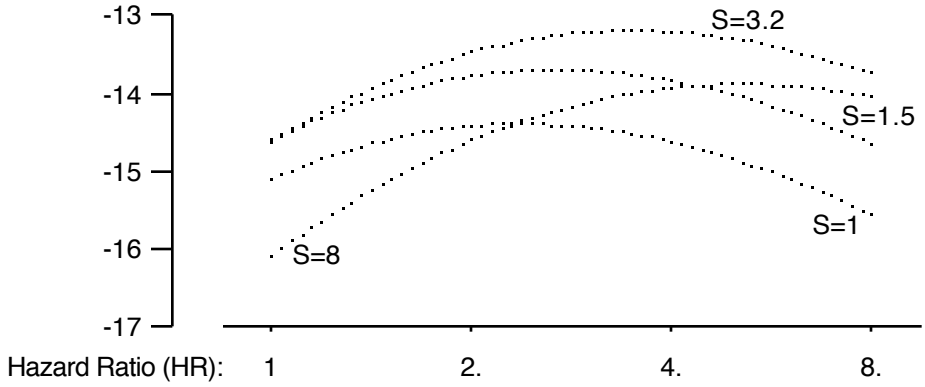
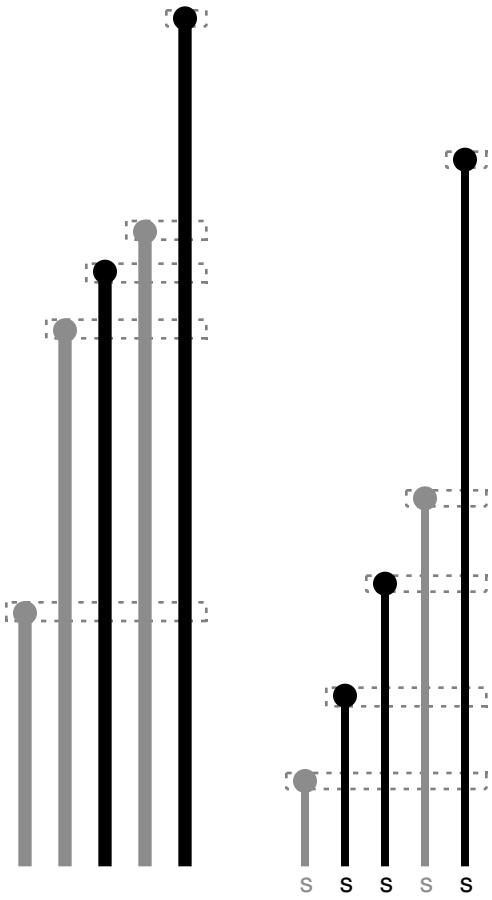


Figure 3 Maximum Likelihood estimation of a 1-parameter proportional hazards model using stratification/matching to eliminate confounding/variation produced by an extraneous variable. Vertical lines represent the longevity of $n = 5$ sexually active fruitflies (shaded line) and $n = 5$ sexually inactive male fruitflies (black, reference group). Three of the latter, and two of the former have shorter than average thorax lengths and are identified by the lowercase letter s and represented by thinner lines, while the remainder, with above average thorax lengths, are represented by thicker lines. Circles denote age at death. Subjects are first segregated (stratified) by thorax size, so that each riskset (enclosed by dashed lines) is homogeneous with respect to this variable. The entries in the corresponding rows are the probabilities, calculated using the HR value in the column, that the death would occur to the subject who did die, rather than in one of the other candidates in the riskset. The likelihood, for any HR value, is the product of the (column of) probabilities associated with the different risksets. The Maximum Likelihood occurs at HR = 2.3. The different log-likelihood scale, compared with Figure 2, stems from the fact that each riskset is smaller, so that the associated probability is larger, and the log-probability is less negative. For this reason, the log-likelihood based on these stratified series cannot be compared with the log-likelihood from the 2-parameter model.

Figure 3



	Hazard Ratio (HR) 1	2.	4.	8.
	1/1	1/1	1/1	1/1
	1/1	1/1	1/1	1/1
	1/2	2/3	4/5	8/9
	1/3	1/4	1/6	1/10
	1/4	1/3	2/5	4/9
	1/2	2/3	4/5	8/9
	1/3	1/4	1/6	1/10
	1/5	1/4	2/7	4/13
	1/4	1/5	1/7	1/11
	1/5	2/7	4/11	8/19
Log[Product] , i.e., log Likelihood:	-9.57	-8.93	-9.16	-10.09

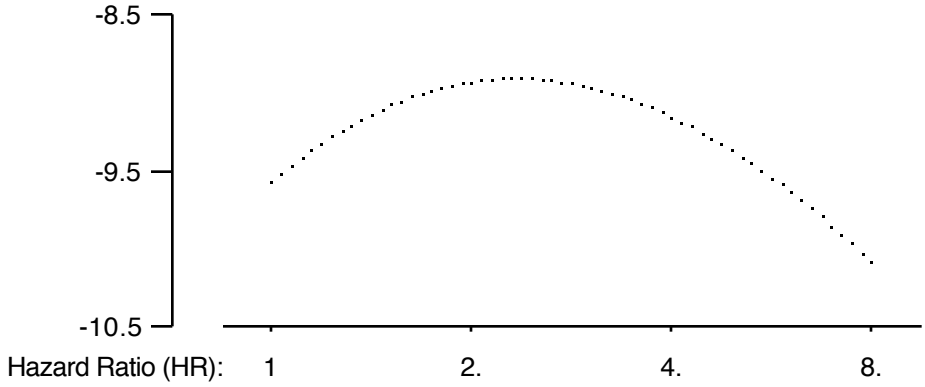


ILLUSTRATION III:
ACCESS TO CONTAMINATED DRINKING WATER: LINK WITH INCIDENCE OF CHILDHOOD LEUKEMIA?

The study by Lagakos et al (1986) compared rates of miscarriages, birth-defects and childhood leukemia in Woburn, Massachusetts residents whose households received different amounts of their drinking water from two municipal wells found to have been heavily contaminated by several chlorinated organics. The investigations were planned and supervised by university investigators (both biostatisticians, well known for their contributions to the statistical analysis of survival data from clinical trials of cancer, and now HIV, therapies). Personal data were collected by telephone interviews conducted by community volunteers. The pumping records for each of the town's 8 wells, combined with a detailed model of the water distribution system, provided estimates, some of which are shown in the top of Figure 4, of the fraction of each household's annual water supply that originated from the contaminated wells. The results of the study, which used were widely reported in the lay press and the portion on leukemia was the subject of the book and subsequent film *A Civil Action*. The scientific report used a modern approach to statistical analysis, and is an early example of the *singular* [Miettinen2004] basis for "cohort" and "case-control" studies -- entities that, even today, are widely perceived as two conceptually distinct entities. The report appeared in a technical statistical journal, and so it has taken longer for its holistic approach to epidemiologic data-analysis to be appreciated by epidemiologists.

Data

Even though childhood leukemia was the most "statistically fragile" of the outcomes studied, the data on this outcome are used here because they were reported in some detail, and were compact enough to allow the arithmetic of the parameter estimation to be carried out with a calculator or simple software. In all, some 20 cases of childhood leukemia were documented during the period studied. The bottom of Figure 4 shows the residential histories of the 17 informative ones, born before the wells were closed, and identifies, by a lighter color, the 9 cases in which the child resided for some years in zones where *some* of the water supply in those years originated in the contaminated wells. The estimated *amount*s of exposure (obtained by cumulating the yearly fractions into a "well-years of exposure") for each of the 17 are shown in Table 2.

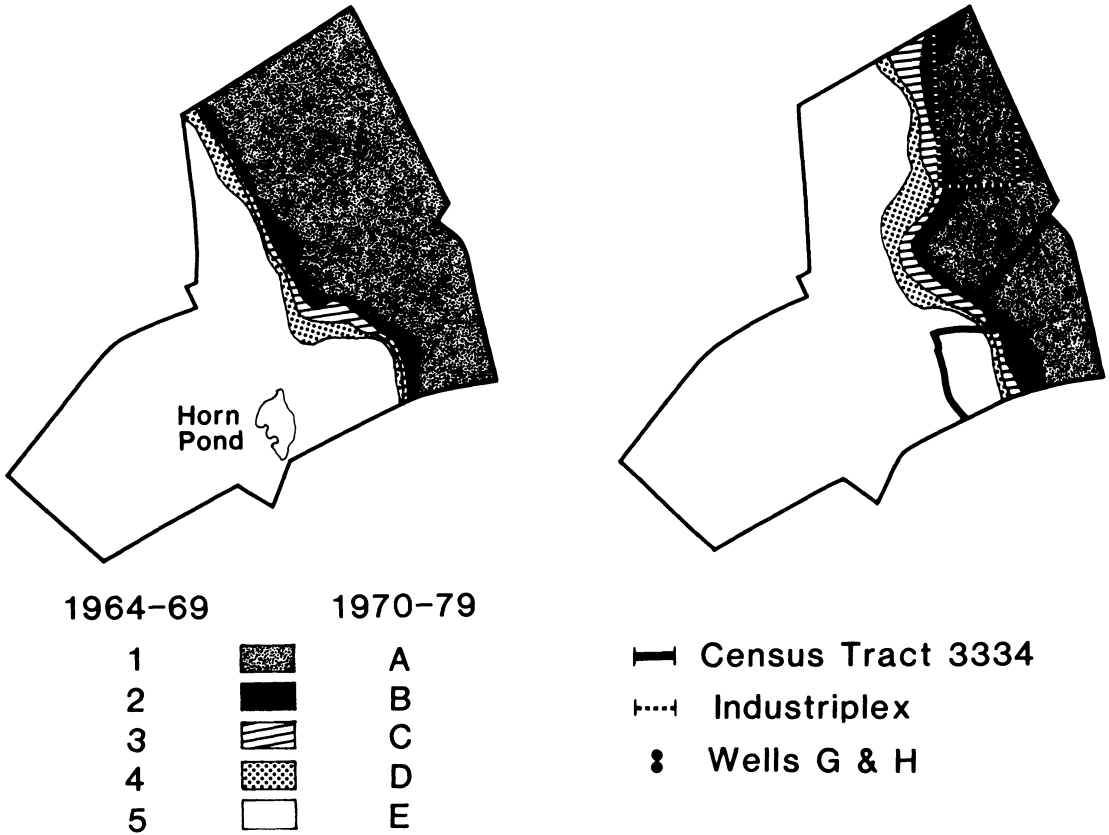


Figure 1. Outline of Woburn, Massachusetts, Showing 1960–1969 and 1970–1982 Zones of Graduated Exposure to Wells G and H. Zones 1 and A represent the greatest G and H exposure, and Zones 5 and E represent the least (no) G and H exposure.

Table 1. Annual G and H Exposure Scores by Zone

Year	1960–1969 Zones			
	1	*2*	*3*	*4*
1960–1963	0	0	0	0
1964	.23	.05	0	0
1965	.08	.08	.08	.08
1966	.95	.70	.25	.12
1967	.51	.47	.32	.17
1968	.72	.34	0	0
1969	.75	.40	0	0

Year	1970–1982 Zones			
	A	*B*	*C*	*D*
1970	.54	.27	.03	0
1971	.46	.38	.08	0
1972	.29	.29	.14	0
1973	0	0	0	0
1974	.37	.32	.25	.14
1975	.55	.44	.13	.01
1976	.49	.38	.09	0
1977	.94	.52	.04	.01
1978	1.00	.88	.61	.05
1979	.39	.39	.29	0
1980–1982	0	0	0	0

NOTE: Table entries give estimated fraction of residential water supply derived from wells G and H by year and residential zone. Refer to Figure 1 for zonal definitions. Exposure scores are zero for Zones 5 and E in all years.

Table 2. Observed and Expected Exposures to Wells G and H for 20 Childhood Leukemia Cases

Case	Year of diagnosis	Year of birth	Period of residency	Observed cumulative exposure	Size of risk set sample	Expected cumulative exposure (var)	Proportion of risk set exposed
1	1966	1959	1959–1966	1.26	218	.31 (.26)	.33
2	1969	1957	1968–1969	0	290	.34 (.36)	.26
3	1969	1964	1969	.75	265	.17 (.10)	.25
4	1972	1965	1965–1972	4.30	182	.90 (2.23)	.36
5	1972	1968	1968–1972	2.76	183	.58 (.88)	.32
6	1973	1970	1970–1973	.94	170	.20 (.20)	.19
7	1974	1965	1968–1974	0	213	.56 (1.04)	.29
8	1975	1964	1965–1975	0	239	.99 (2.78)	.38
9	1975	1975	1975	0	115	.09 (.03)	.25
10	1976	1963	1963–1976	.37	219	1.18 (3.87)	.40
11	1976	1972	1972–1976	0	132	.24 (.32)	.18
12	1978	1963	1963–1978	7.88	219	1.41 (6.23)	.40
13	1979	1969	1969–1979	2.41	164	.73 (2.56)	.31
14	1980	1966	1966–1980	0	199	1.38 (6.00)	.39
15	1981	1968	1968–1981	0	187	1.14 (4.20)	.35
16	1982	1979	1979–1982	.39	154	.08 (.02)	.23
17	1983	1974	1974–77, 1980–83	0	84	.25 (.45)	.23
18	1982	1981	1981–1983	0	—	0 (0)	0
19	1983	1980	1980–1982	0	—	0 (0)	0
20	1983	1980	1981–1983	0	—	0 (0)	0
Totals				21.06		10.55 (31.52)	5.12
Score test statistic:						1.87	2.08
Significance level:						$P = .03$	$P = .02$

NOTE: Risk set for a case consists of children born in the same year as the case and who were residents of Woburn when the case was. Variance of proportion, say p, of risk set exposed equals $p(1 - p)$. Cases 18–20 do not contribute to the test statistic because birth occurred after closure of wells G and H.

Shown in Figure 5 are the corresponding "ever/never exposed" data for the children in the 17 risksets. The riskset for a particular instance (case) of leukemia consists of the child diagnosed with leukemia, together with that child's cohort/peers. Contrary to some mis-apprehensions, the risk set *in*cludes the person who represents the 'case'. In several 'birth-cohorts'. there were 2 cases of leukemia. For children born in 1964 for example, the investigators were able to obtain 1964-1969 residential histories for 265 children who were 'at risk' -- when case number 3 occurred in 1969. The 1964-1975 residential histories were available for 239 'candidate' children from this same 'cohort' when another case (number 9) occurred in 1975 (technically, the in- and out-migration made this a dynamic population, rather than a fully-followed closed birth-cohort, so the 239 are not a pure subset of the 264). However, the child representing the 'case' in 1975 was also in the 1969 riskset, but with a shorter history at that earlier time.

Ever exposed vs. never exposed : Simple and Maximum Likelihood estimators of IDR

The data in the first 3 columns of Table 2 allow us to use a Mantel-Haenszel type estimator of the incidence density ratio, based on the "ever/never" exposure scale. In each of the 9 instances where the child diagnosed with leukemia had been exposed, the data on the n children in the case-associated 2 × 2 table contribute zero to the denominator, and $1 \times [n \times (1-p)] / n = (1-p)$ to the numerator. Conversely, in each of the other 8 cases, where there was no history of exposure, the 2 × 2 table contributes zero to the numerator, and $1 \times [n \times p] / n = p$ to the denominator. Thus, the IDR estimate is simply

$$IDR_{M\text{-}H} = \frac{0.67 + 0.75 + 0.64 + 0.68 + 0.81 + 0.60 + 0.60 + 0.69 + 0.77}{0.26 + 0.29 + 0.38 + 0.25 + 0.18 + 0.39 + 0.35 + 0.23} = \frac{6.21}{2.33} = 2.66$$

As shown in Table 5, the Likelihood can be constructed using the same scheme shown in the two previous applications. Consider the first riskset, comprising 218 children born in 1959, one of whom was diagnosed with leukemia at age seven. By that age, some 72 of the 218 had lived for some time in a part of Woburn supplied by contaminated water, and the remaining 146 had not. The probability that the leukemia would occur in the specific (exposed) child in whom it did is thus $IDR/(72 \times IDR + 146 \times 1)$. For the second

riskset, the probability that the leukemia would occur in the specific (unexposed) child in whom it did is $1/(75 \times IDR + 215 \times 1)$, and so on to the last probability of $1/(19 \times IDR + 65 \times 1)$. The product of these 17 conditional i.e. evaluated-after-the-fact, probabilities is the Likelihood. Because it is a function of just one parameter, it -- or more readily, its log, a sum-- is easily maximized with nothing more than a spreadsheet: for any 'what-if' value of IDR, the logs of the 17 probabilities can be calculated using a spreadsheet formula, then summed to form the logLikelihood. The Maximum Likelihood Estimate of IDR can be found by trial and error, i.e., by varying the 'what-if' value of the IDR, until the largest sum is found. The maximum occurs at IDR_{MLE}=2.68. Once the formula is set up, it is a simple matter to obtain enough values to sketch the logLikelihood function, the curvature of which is used to measure the precision of the MLE.

Taking advantage of calculations used in the log-rank test of IDR=1, which yields an expected number of exposed 'cases' of 5.12, Lagakos et al. used the *approximation* to the MLE

$$MLE_{approx.} = \exp[(9- 5.12)/\{0.33 \times 0.67 + ... + 0.23 \times 0.77\}] = 3.03.$$

They acknowledge -- and the exact calculation in this example shows -- that the approximation can inaccurate when the IDR is far from the null. Interestingly, in this example, the Mantel-Haenszel estimator gives an estimate very close to the MLE.

Cumulative exposure : Simple and Maximum Likelihood estimators

For each child, the *amount* of exposure was obtained by cumulating the yearly exposure fractions into a "well-years" (W-Y) of exposure". Just as the authors did, we will use this in a statistical model in which children, born in a certain year, and now age t, who had accumulated x well-years of exposure by this age, were $IDR_X = \exp[x \times b]$ times more likely to be diagnosed with leukemia in the next little while than

children born the same year, who had accumulated x=0 units exposure by this same age t. The proportional hazards model does not *force* one to assume this exposure 'metric' x, or this particular exponential function; for example, one might ignore the exposure in the previous year, or in the first year of life, etc., or use the logarithm of the exposure, or use IDR_x as some other function of x.

The cumulative exposures for the children diagnosed with leukemia were reported, but we did not have access to the separate x's for each child in each riskset,. Therefore, for illustrative purposes, for four selected leukemia cases (numbers 15, 13, 12 and 7) we used the reported mean and variance of each of the four distributions to construct four rough histograms that matched the reported mean and variance for the risksets. These four histograms are shown twice each in Figure 6,: on the left when calculating the LogLikelihood under the null value b=0, and on the right under the value b=0.25. The non-null value 0.25 was deliberately chosen to make the arithmetic easier, so that the exponents, exp[x × b], would be of integers, for example, $IDR_{4:0}$ = exp[4 × b]=exp[1] =2.7, and $IDR_{8:0}$ = exp[8 × b]=exp[2] =7.4, in relation to the reference value ID_0 = exp[0 × b]=exp[0] =1.

The calculations in each riskset can be likened to after-the fact calculations for a lottery with an undesirable prize. For example, consider the children in riskset 13, consisting of 131, 13, 13 and 7 children with 0, 2, 4 and 6 W-Y units of exposure respectively. Under the non-null value b=0.25, these children hold 1, 1.6, 2.7 and 4.5 'shares' each (total: 164 children, holding a total of 218.4 shares). If, as did happen, the undesirable prize was drawn by a child with 2 W-Y units of exposure, worth exp[2 × b]=1.6 shares, one could calculate that *this specific child*, rather than any of the others, had a 1.6/218.4 or 1 in 137 chance (Log: -4.9) of being the unlucky one. This contrasts with the 1 in 164 chance (Log: -5.1) of it happening to him if the cumulated exposures did not confer additional risk.

In the upper panel of Figure 7, the LogLikelihood is evaluated for each riskset, for a range of b values. In the lower panel, these LogLikelihoods is combined over all four risksets. The 'convenient for computation' value b=0.25 happens to be close to the b_{ML} = 0.29 found by the full search. [Lagakos et al., using *all 17* risksets, but the same approximate Ml method referred to previously, obtained an estimate of b=0.33].

The ML estimation process has often been explained "algebraically" using estimating equations. The data display in Figure 6 allows one to "see" the ML estimation process more graphically. In the usual expositions of the MLE process, including the 1972 one by Cox himself, the LogLikelihood is first written as a sum of the riskset-specific LogLikelihoods; the derivatives, with respect to b, of these summands are then obtained. Setting the sum of these derivatives to zero results in the "estimating equation", with the sum taken over risksets,

Sum[exposure of the 'case'] = Sum[weighted average of exposures of all persons in riskset].

In our example, the sum on the left is of the four x's denoted by asterisks in Figure 6. Cox(972) noted that the four weighted averages on the right were constructed using an 'exponential weighting' of the exposures in the riskset. In fact, the weights are the IDR's, -- the exp[x × b]'s themselves: a person with an exposure of zero receives a weight of 1, a person with an exposure of 3 a weight of exp[3 × b], etc., i.e., persons with larger exposures count for more. We have tried to illustrate this in the Figure using dots of increasing magnitudes. The weighted averages are shown in Figure 6 as vertical arrows. In effect then, the search for b_{ML} involves turning the 'b knob' up (so that the dots get larger, and averages move to the right) or down (so they move to the left) until the sum of the four resulting "fitted" weighted averages matches (i.e., is counter-balanced by) the sum of the four "observed" exposures. This idea of translating persons in the riskset into 'IDR-equivalents' is also a helpful way to understand how, in survival analysis applications, one can estimate the survival (and hazard) curve for persons in the reference (unexposed) category: one uses the fitted b, to convert each other ('exposed') person into a number of unexposed-equivalents, and then applying the standard Kaplan-Meier estimator to these.

Sampling from Risksets

A "primitive' form of this design was used in a study reported in 1972 (Doll, 2001). This technique of carrying out a 'case-control-within-a-cohort-study', was formally proposed by Mantel in 1973, and extended to time-based sampling by Liddell et al. in 1977. Breslow and Day(Chapter 5, 1987) use two worked examples, one involving an ever-never and one a measured exposure, to illustrate the computational

savings that can be achieved by restricting analyses to subsamples of the risksets. Extensive computations are far less of an obstacle nowadays, but the costs of obtaining the exposure and/or confounder data continue to be important considerations. Naturally, these savings come at a cost of poorer precision. Sometimes, e.g., when all of the data come from cohorts or administrative databases, with all of the data already in electronic form, cost considerations are less of an issue. In such situations, the worked examples in their textbook show that the common belief that '4 controls per case' is sufficient is not generally justified, particularly if the exposure distribution is considerable skewed, and the associated IDR's are large. In such instances, there can be considerable reductions in standard errors by taking 10 or 20 'controls' from each riskset. Even in the analysis of the 17 leukemia cases in the Woburn study, where the proportion of the riskset 'exposed' ranged from 0.18 to 0.39, simulations carried out by this author confirm that IDR estimates based on say 16 or 32 'controls' per case were often quite far from the estimate of 2.7 obtained using the *full* risksets.

DISCUSSION

The data in the three examples arise from seemingly very different 'study designs', yet the analyses follow a common approach. The unifying factor is the riskset, and the partial likelihood -- which focuses on the parameter(s) of interest, and eliminates -- by conditional arguments -- those felt to be of no direct interest.

The three examples emphasize that whereas modern-day epidemiologists *continue* to separate study designs into '*cohort*' and '*case-control*' studies, there is only *one* modern approach to their analysis. No matter whether "case-control" or "cohort 'study, the risksets used to construct the likelihoods in all three of our examples use as their point of departure the '*case* series'. Epidemiologists who analyze '*case-control*' studies are comfortable starting with the *numerators* of the to-be-compared rates. Then, rather than establishing the *total* sizes of the denominators for these -- denominators that would allow them to calculate ID's -- they resort to *samples* of the denominators, and from the computed quasi-rates, they can estimate the ID *ratios*. But the risksets, and the associated likelihoods, in the fruitfly study -- a classic '*cohort*' study, traditionally defined by its *denominators* -- *also* begin with the numerators. Risksets in a traditional matched case control study have no overlap one with the next, whereas each riskset in a survival analysis is included within the one to its 'left' along the time axis. Indeed, Miettinen (2004) argues that

Even before Cox proposed the concept of risksets to more readily *estimate* the hazard ratio, the *log rank* test (Mantel 1966, Peto and Peto 1972*)* had been used to *test* the equivalence of two survival curves, by constructing a 2×2 table (effectively a riskset) at each distinct event-time. However, this is the same test proposed by Mantel on 1959 -- for *case-control* studies that use stratification to control for confounding.

In the first part, I argued that "case-crossover' studies(McLure 1991, exemplified by Redelmeier 1997) did not need this special name; they are self-matched case-control studies. Again, the purpose of the 'control' (more appropriately called the 'denominator') series is to obtain estimates of the person-specific denominators (amounts of person-time, exposed and unexposed) underlying each 'exposed' and 'unexposed' numerator in the 'case' series. Even though at first the design appears to be very different, the likelihood used in the analysis of "self-controlled case-series studies" (Farrington1995, Andrews2002) has

the same form as that used throughout examples I-II above. And, although "case-cohort" studies have some statistical complexities, they too are examples of the fact that all epidemiologic contrasts, whether in "case-control" or "cohort" studies, involves contrasts between the "exposed" and unexposed"; the comparison never is between a "case group" and a "control group"(Miettinen2004). Rather, what distinguished the case-control from the cohort study is the completeness of the denominators -- complete in the latter, estimated in the former. This modern way of thinking of the two designs as one was nicely illustrated in a report (Hernan2002) of a meta-analysis of 48 studies (44 case-control and 4 cohort) examining the link between cigarette smoking and the risk of Parkinson's disease. The table listed the sizes of the 48 investigations using the *numbers of cases* (numerators) and the "*number of controls or the cohort size*" (i.e., the sizes of the *partial*, or *entire denominators*).

In the first article, I make limited use of the diagram on cell phone use. I use it again here to emphasize that when etiologic research involves transient -- or accumulating -- exposures, and necessarily dynamic denominators, -- the 'case-control' approach is only one viable, and even conceptually valid, option. Suppose one sought denominators by which to compare the rate of accidents in on-the-phone driver time with that in off-the -phone driver-time (whether in the same or different drivers). Imagine that person-specific records were readily available for say an entire year for each of the 1 million persons who drove at some time in a city that year. Imagine further that a person's record was divided up into 60x60x24x265 = 31 million time units, of 1 second each, each one indicating whether the person was driving at that instant, and if so whether (s)he was using the cell-phone. Even with this utopian database, few investigators would go through the laborious exercise of creating two dynamic registers with which to record the levels of on-the-phone and off-the-phone driving at each of these instants (they might, for a crude comparison -- one that ignores driver, time of day, weather, season etc. -- calculate the total numbers of on-the-phone and off-the-phone driver-moments). Instead, most would use the accidents within this base as a more efficient and informative point of departure. And, in the real world, they would estimate -- via sampling -- the levels of on-the-phone and off-the-phone driving in the relevant time windows preceding these events.

Epidemiologists have been slow to change terminology or to appreciate that, conceptually, there is only one approach to -- what Miettinen(2004) aptly calls -- *the* etiologic study. Even in a 'cohort' study, and even in one with full documentation already available, the point of departure is the case series, and that for both the 'case' and the 'base' series, ['denominator' series in a 'case-control' study] "*the etiologic histories are defined as of the time of the outcome (case occurring or not occurring*)" (Miettinen,2004). Indeed, by the very way we pursue causes, we dare forced to pursue in the *historical* direction. The way in which Cox set up his partial Likelihood reinforces the direction of this pursuit.

Despite the slow evolution in (study design) methods and concepts, the data analyses presented here do show that over the period of time covered in the review by Zhang et al (2004), considerably considerable convergence in the statistical analyses of data from *the* etiologic study. And, it is hoped that -- as a byproduct of this exposition -- the extensive arithmetic used throughout these two articles will make the inner-workings of Maximum Likelihood- estimation a little more understandable.

ACKNOWLEDGMENT

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

Figure 4 *Top*: Zone-and-year-specific exposure data used in the analysis, adapted from table 1 in original article. Shown are estimates, for each year indicated, of the fraction of each household's water supply that arose from the two contaminated wells. For estimates for the years 1960-1969, the town was partitioned into 5 zones (1-5) of graduated exposure to wells G&H. Because of a substantial change in industrial demand in 1970, different residential zones (A-E) were used for the estimates for the period from 1970 until the 2 wells were closed in 1979. The study estimated, on a monthly basis, which zones received none, some or all of their water from wells G&H. These data were used to estimate, for each year and each residential zone, the fraction of each household's water supply that arose from the two wells
Bottom: Residential histories, shown by Lexis diagram, in 17 informative cases of leukemia. Duration of the child's residence in Woburn, up until the date of diagnosis, is indicated by a line. Lighter color lines indicate children who resided for some years in zones where some of the household water supply in those years was estimated to have originated in the 2 contaminated wells. Darker lines indicate a child whose residential history suggested that during that time none of the household water supply originated in these 2 wells. Circles denote when leukemia was diagnosed. Adapted, with some simplifications, from Lagakos et al (1986).

REFERENCES

Partridge L & Farquhar M. Sexual activity reduces the lifespan of male fruitflies. Nature 294:480-581, 1981.

Hanley JA and Shapiro SH. : "Sexual Activity and the Lifespan of Male Fruitflies: A Dataset That Gets Attention" Journal of Statistics Education v.2, n.1 (1994). Both the data and the article are available on-line at http://www.amstat.org/publications/jse/jse_data_archive.html.

Hanley JA. Appropriate uses of multivariate analysis. Annual Rev Public Health. 1983;4:155-80.

Anderson S et al.. Statistical methods for comparative studies: techniques for bias reduction. New York Wiley, 1980.

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

Farewell;V. T and Cox D. R. A note on multiple time sales in life testing Applied Statistics , Vol. 28, No. 1. (1979), pp. 73-75.

Fisher RA. On an Absolute Criterion for Fitting Frequency Curves. Messenger of Mathematics, 41: 155-160 (1912)

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

Kalbfleisch ,J.D. Prentice RL. The statistical analysis of failure time data. 2nd Edition . Hoboken, N. J. : J. Wiley, 2002.

Lagakos, S.W., Wesson, B.J., and Zelen, M. (1986) An analysis of contaminated well water and health effects in Woburn, Massachusetts (with discussion). J. Amer. Statist. Assoc., 81, 583-614.

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

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

Mantel N. Synthetic retrospective studies and related topics. Biometrics. 1973 Sep;29(3):479-86.

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.

Breslow NE & Day NE. Statistical methods in cancer research. II. The design and analysis of case-control studies. Lyon: Intnl. Agency for Research on Cancer 1987.

Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep. 1966 Mar;50(3):163-70.

Peto R and Peto J. Asymptotically Efficient Rank Invariant Test Procedures Journal of the Royal Statistical Society. Series A (General) , Vol. 135, No. 2. (1972), pp. 185-207.

Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol 1991;133:144-53.

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

Farrington CP. Relative incidence estimation from case series for vaccine safety evaluation. Biometrics. 1995 Mar;51(1):228-35.

Andrews NJ. Statistical assessment of the association between vaccination and rare adverse events post-licensure. Vaccine. 2001 Oct 15;20 Suppl 1:S49-53.

Hernan MA, Takkouche B, Caamano-Isorna F, Gestal-Otero JJ A meta-analysis of coffee drinking, cigarette smoking, and the risk of Parkinson's disease. Ann Neurol. 2002 Sep;52(3):276-84.

Zhang FF, Michaels DC, Mathema B, et al. (2004). Evolution of some epidemiologic methods and concepts in selected text books of the 20th century. Soz Praventiv Med 49: 97–104.

ADDITIONAL READING

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.

Figure 4

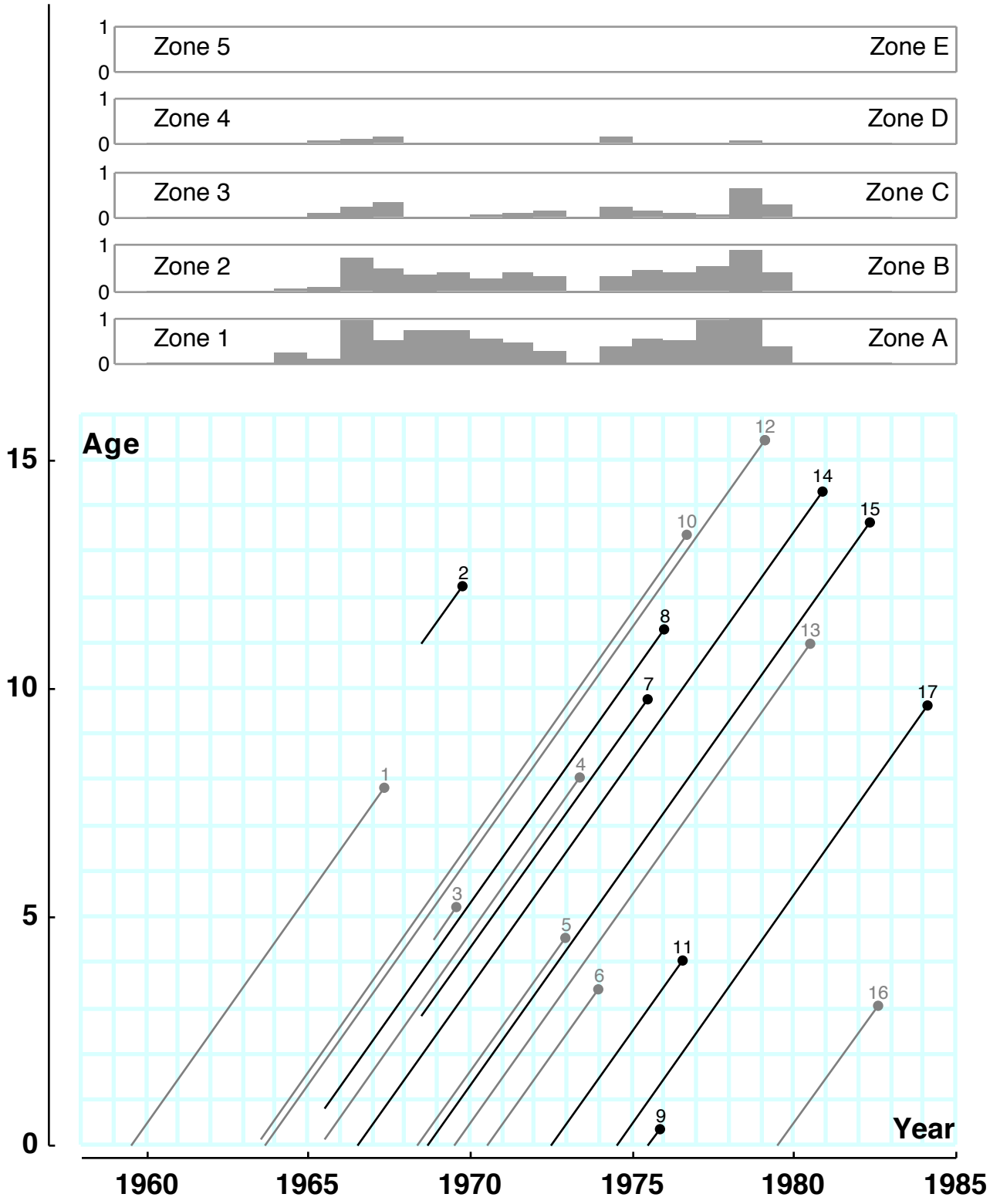

Zone 5
Zone E
Zone 4
Zone D
Zone 3
Zone C
Zone 2
Zone B
Zone 1
Zone A
1
0
Age
15
10
5
0
Year
1960
1965
1970
1975
1980
1985
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

Figure 5 Obtaining the Maximum Likelihood estimate of the IDR comparing those ever-exposed with those. never-exposed. Cases are numbered as in Figure 4, with "E" denoting that the child had lived in a zone exposed to water from the contaminated wells. The numbers of exposed and unexposed in the riskset are shown in plain and bold text respectively. The 17 likelihood contributions, one per riskset. calculated under the assumption that the IDR is 1. are shown in the first column, and the log of the product of these, i.e. the log Likelihood of -88.5, is shown at the foot of the column. The entries in the three remaining columns are calculated under the assumption that the IDR is 2, 4 and 8 respectively. The log Likelihoods are also shown in the graph for intermediate values of the IDR, allowing us to see that the Maximum Likelihood Estimate of the IDR is approximately 2.7.

Figure 5

case		Riskset	1	2	4	8 IDR
1	E	72 **146**	1/(72+146)	2/(144+146)	4/(288+146)	8/(576+146)
2		75 **215**	1/(75+215)	1/(150+215)	1/(300+215)	1/(600+215)
3	E	66 **199**	1/(66+199)	2/(132+199)	4/(264+199)	8/(528+199)
4	E	66 **116**	1/(66+116)	2/(132+116)	4/(264+116)	8/(528+116)
5	E	59 **124**	1/(59+124)	2/(118+124)	4/(236+124)	8/(472+124)
6	E	32 **138**	1/(32+138)	2/(64+138)	4/(128+138)	8/(256+138)
7		62 **151**	1/(62+151)	1/(124+151)	1/(248+151)	1/(496+151)
8		91 **148**	1/(91+148)	1/(182+148)	1/(364+148)	1/(728+148)
9		29 **86**	1/(29+ 86)	1/(58+ 86)	1/(116+ 86)	1/(232+ 86)
10	E	88 **131**	1/(88+131)	2/(176+131)	4/(352+131)	8/(704+131)
11		24 **108**	1/(24+108)	1/(48+108)	1/(96+108)	1/(192+108)
12	E	88 **131**	1/(88+131)	2/(176+131)	4/(352+131)	8/(704+131)
13	E	51 **113**	1/(51+113)	2/(102+113)	4/(204+113)	8/(408+113)
14		78 **121**	1/(78+121)	1/(156+121)	1/(312+121)	1/(624+121)
15		65 **122**	1/(65+122)	1/(130+122)	1/(260+122)	1/(520+122)
16	E	35 **119**	1/(35+119)	2/(70+119)	4/(140+119)	8/(280+119)
17		19 **65**	1/(19+ 65)	1/(38+ 65)	1/(76+ 65)	1/(152+ 65)
		log Likelihood:	−88.5	−86.7	−86.9	−88.9

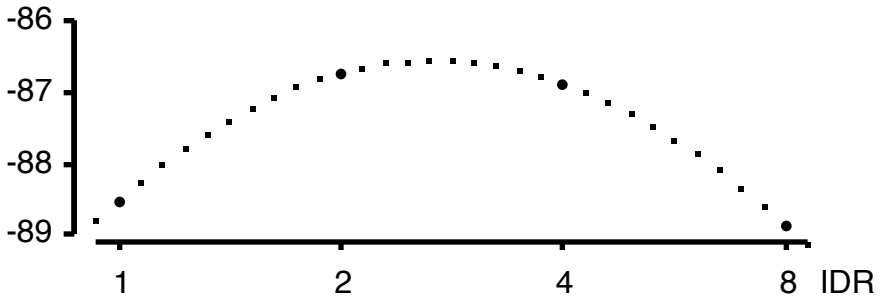


Figure 6 Log Likelihood calculations of the IDR comparing those with (x+1) vs. those with x "well-years" (W-Y) of cumulative exposure, illustrated using 4 selected leukemia cases [15, 13,12,7), and for just 2 values of the regression coefficient b. Cumulative exposure is depicted on the x axis. The cumulative exposure for the actual child diagnosed with leukemia is indicated with an asterisk. The distribution of cumulative exposure in all the children in the riskset is shown as a histogram, with 1 dot representing 2 children [For each riskset, only the mean and variance of the distribution were reported. For didactic and presentation purposes, the possible values are limited to a few values, all integers -- but the distributions shown were constructed to match the reported means and variances].

Left: Log Likelihood calculation under the null value, IDR=1, regardless of W-Y. Thus, the probability that the leukemia would be diagnosed in the child in whom it was actually diagnosed is simply 1/(number of children in risk set). The log of this "likelihood" is shown for each riskset (risksets differ slightly in size, and so the LogLikelihoods do too). The LogLikelihood based on *all 4* risksets is the sum of the 4 individual LogLikelihoods. The vertical arrows denote the average of the exposures in the riskset.

Right: Log Likelihood calculation under the assumption that, relative to children with x well-years of cumulative exposure, those with (x+1) well-years is exp[b] = exp[0.25]. Thus, relative to the reference category where W-Y=0, the IDR's for those with 1,2, ,4, , , ,8 W-Y's are exp[0.25]=1.3, exp[2 × 0.25] =1.6, , exp[4 × 0.25] =2.7, , , ,exp[8 × 0.25] =7.4. The IDR's for the children with different amounts of exposure are shown using dots whose diameters are proportional to the IDR's. The probability that the leukemia would be diagnosed in the child who was actually diagnosed -- who had W-Y units of exposure -- is $IDR_{[W-Y]}$/(Sum of IDR's for each child in riskset). Again, the LogLikelihood based on all 4 risksets is the sum of the 4 individual LogLikelihoods. The Maximum Likelihood estimate is found by varying b until the LogLikelihood, based on all 4, is the largest (i.e., least negative) possible. The vertical arrow denotes the weighted average of the exposures in the riskset, with weights given by the corresponding IDRs.

Figure 6

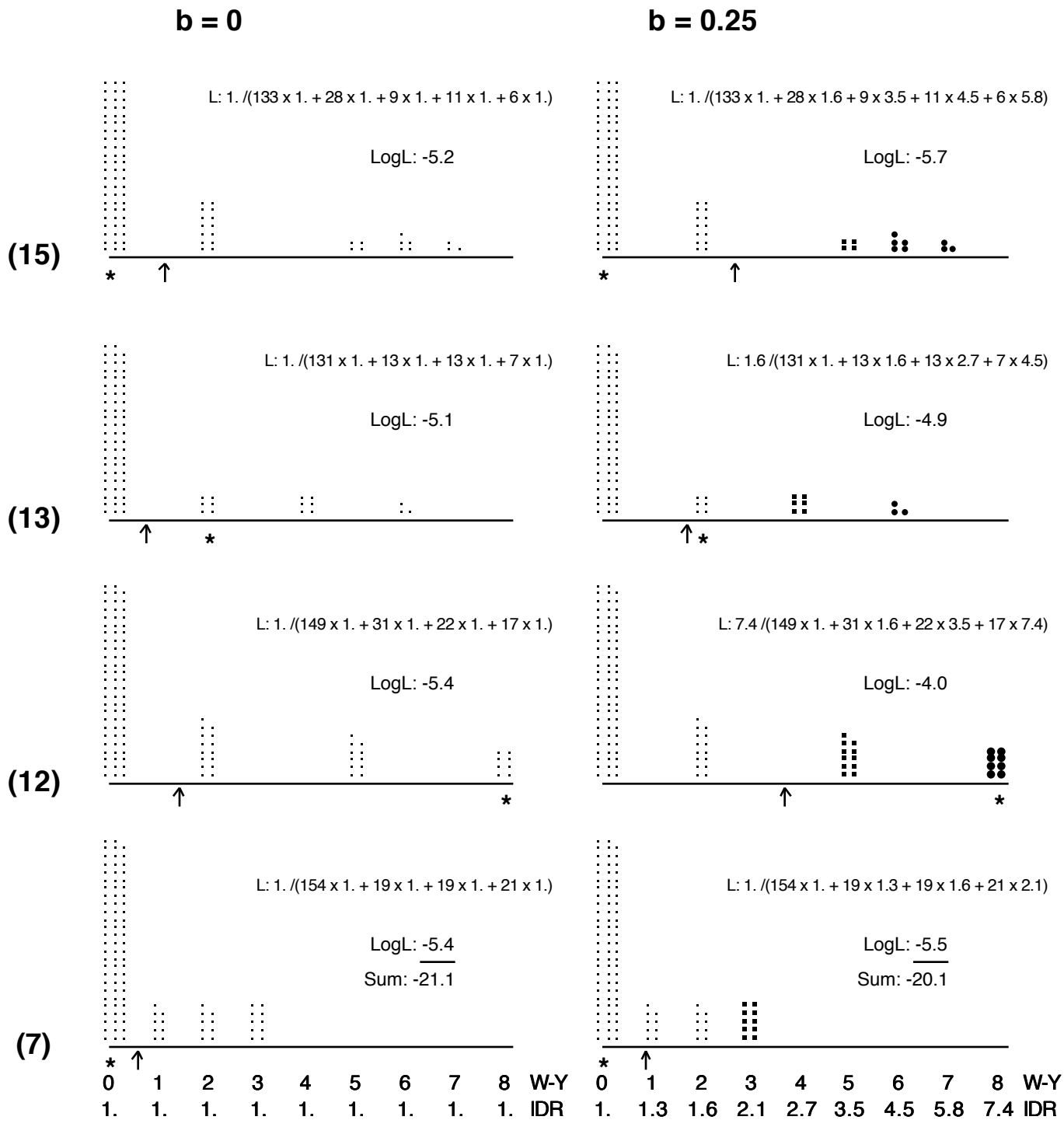
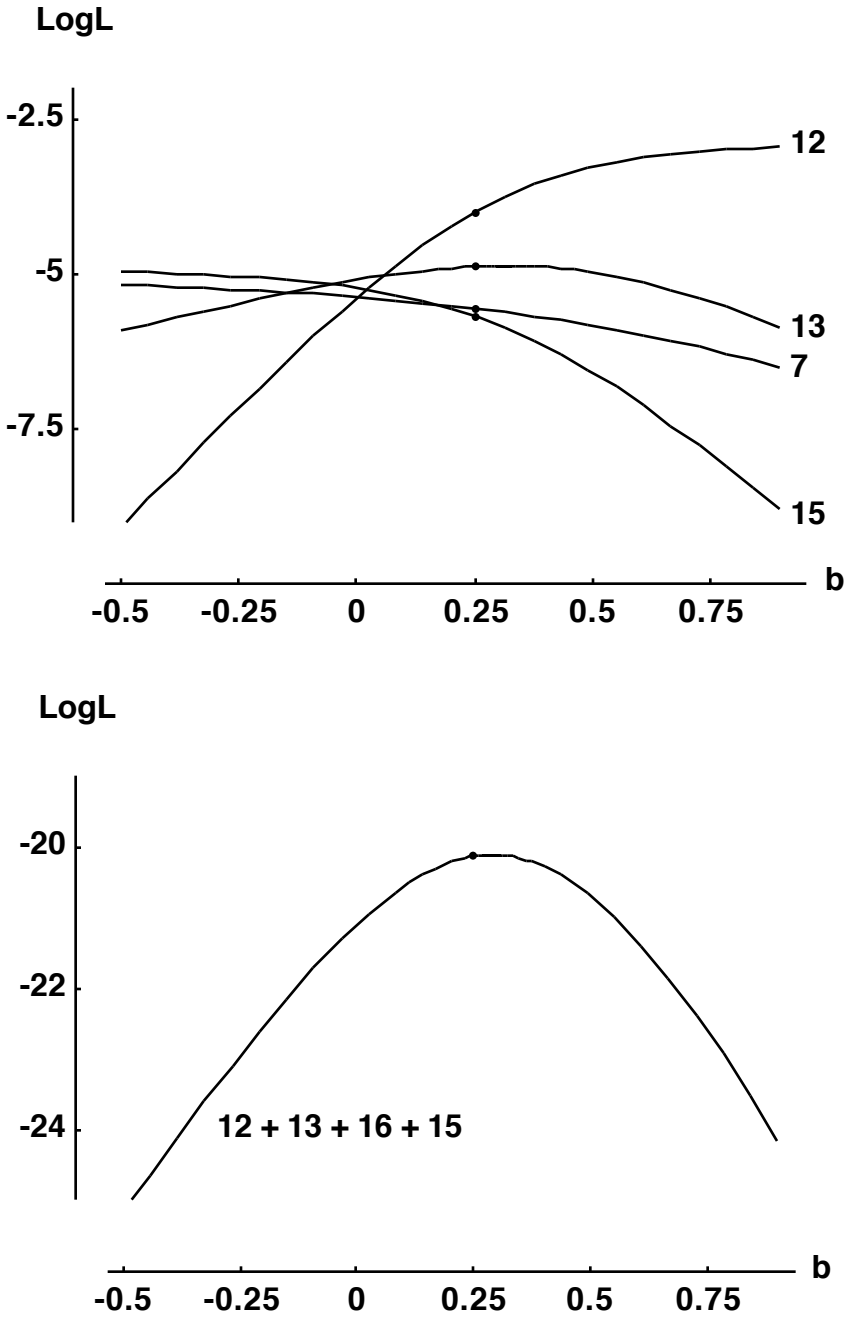

b = 0
b = 0.25
(15)
L: 1. /(133 x 1. + 28 x 1. + 9 x 1. + 11 x 1. + 6 x 1.)
LogL: -5.2
L: 1. /(133 x 1. + 28 x 1.6 + 9 x 3.5 + 11 x 4.5 + 6 x 5.8)
LogL: -5.7
(13)
L: 1. /(131 x 1. + 13 x 1. + 13 x 1. + 7 x 1.)
LogL: -5.1
L: 1.6 /(131 x 1. + 13 x 1.6 + 13 x 2.7 + 7 x 4.5)
LogL: -4.9
(12)
L: 1. /(149 x 1. + 31 x 1. + 22 x 1. + 17 x 1.)
LogL: -5.4
L: 7.4 /(149 x 1. + 31 x 1.6 + 22 x 3.5 + 17 x 7.4)
LogL: -4.0
(7)
L: 1. /(154 x 1. + 19 x 1. + 19 x 1. + 21 x 1.)
LogL: -5.4
Sum: -21.1
L: 1. /(154 x 1. + 19 x 1.3 + 19 x 1.6 + 21 x 2.1)
LogL: -5.5
Sum: -20.1
0 1 2 3 4 5 6 7 8 W-Y
1. 1. 1. 1. 1. 1. 1. 1. 1. IDR
0 1 2 3 4 5 6 7 8 W-Y
1. 1.3 1.6 2.1 2.7 3.5 4.5 5.8 7.4 IDR

Figure 7 Individual and collective LogLikelihood contributions of 4 risksets shown in Figure 6.
Top: LogLikelihood functions for the parameter b = log[$IDR_{(x+1):x}$], evaluated from b = -0.5 to b = 0.9, for each of the 4 risksets shown in Figure 6. In case 12, the calculated exposure for the child in question was 8 W-Y, well beyond the mean of 1.4 W-Y in the entire riskset, and so this case is better explained by positive values of b. In contrast, in case 15, the calculated exposure was 0, whereas the mean in the riskset was 1.1, and so the data are better explained by negative values of b. In cases 14 and 6, the observed W-Y values are just about as probable under a wide range of positive and negative values. (The slopes of the LogLikelihoods at b = 0 are called 'scores', and their sum is called the *score statistic*.)
Bottom: The summation of the 4 separate LogLikelihoods: the observed W-Y pattern in the 4 cases is 'most likely' for b values closer to 0.25, but -- with just 4 cases in this example-- the data could have been produced with any of a broad range of values of b. In practice, parameter values that produce LogLikelihoods that are within 2 units of the Maximum LogLikelihood (so that 2 times the difference is approximately 4 i.e. chi-squared critical value 3.84) are considered as 'plausible', i.e. the observed data-pattern is only exp[2] or approximately 7 times more likely under the MLE value than under the values at the edge of this range.

Figure 7


LogL
-2.5
-5
-7.5
12
13
7
15
b
-0.5
-0.25
0
0.25
0.5
0.75
LogL
-20
-22
-24
12 + 13 + 16 + 15
b
-0.5
-0.25
0
0.25
0.5
0.75