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 ***othe**r* 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.

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.

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 × IDR + 215 × 1), and so on to the last probability of 1/(19 × IDR + 65 × 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.

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.

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 $h_0[\ ]$.

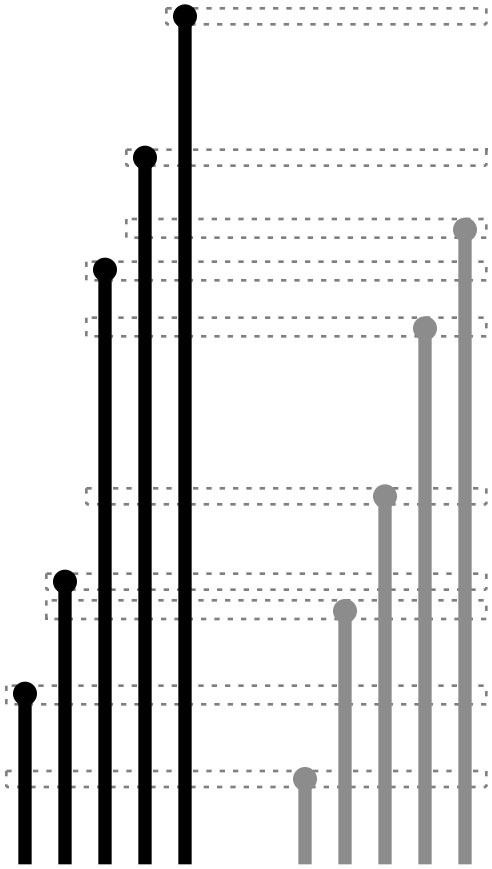
* 2 + 2 × S + 2 × HR + HR × S abbreviated to 'Sum'

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

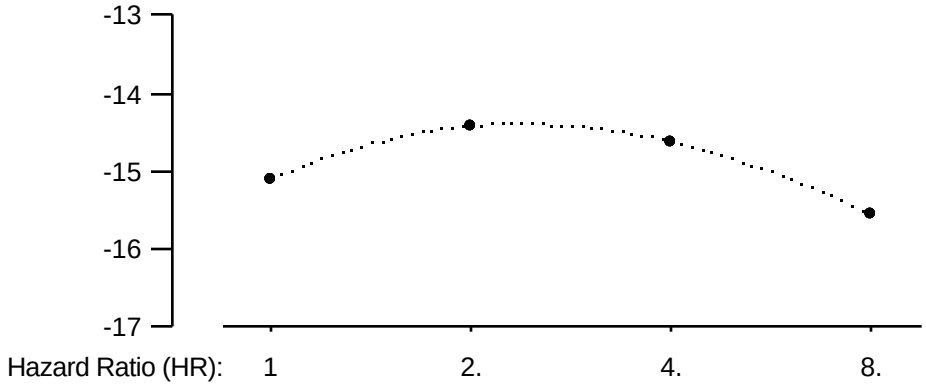
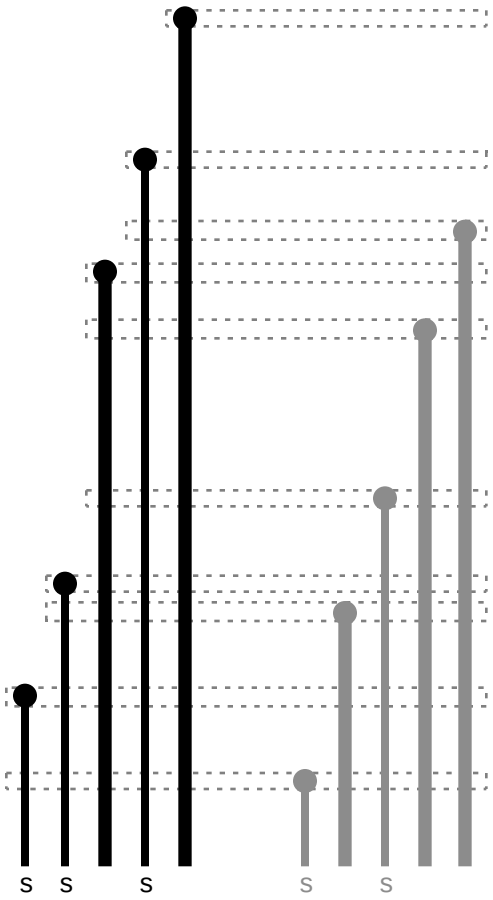


Figure 2 Maximum Likelihood estimation of 2-parameter proportional hazards model. Vertical lines represent the longevity of $n = 5$ sexually active fruitflies (gray) and $n = 5$ sexually inactive male fruitflies ((black, reference group). Three of the latter, and two of the former have shorter than average 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

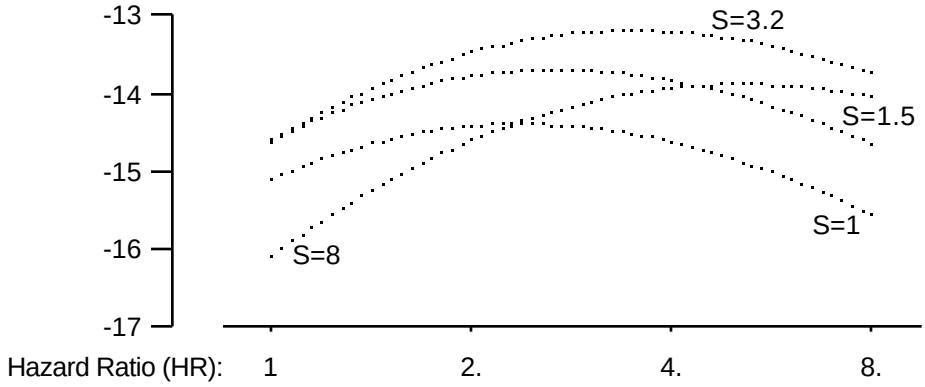
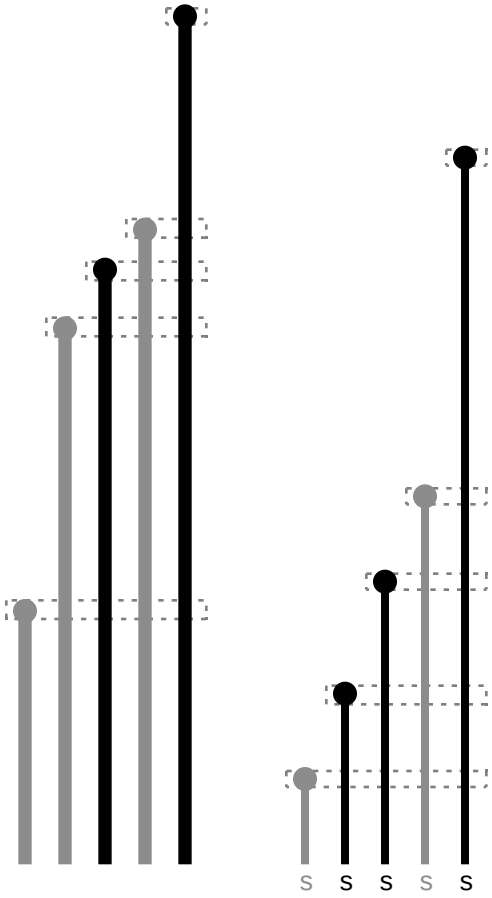


Figure 3 Maximum Likelihood estimation of a 1-parameter proportional hazards model using stratification/matching to eliminate confounding/variation produced by an extraneous variable. Vertical lines represent the longevity of $n = 5$ sexually active fruitflies (shaded line) and $n = 5$ sexually inactive 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

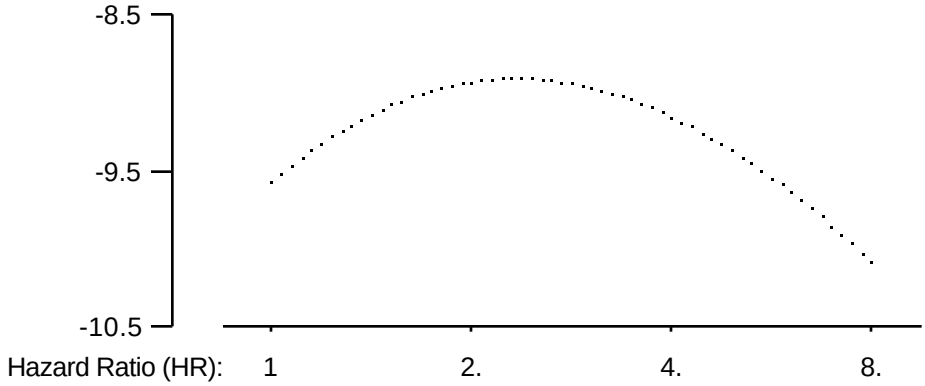


Figure 4 *Top*: Zone-and-year-specific exposure data used in the analysis, adapted from table 1 in original article. Shown are estimates, for each year indicated, of the fraction of each household's water supply that arose from the two contaminated wells. For estimates for the years 1960-1969, the town 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).

Figure 4

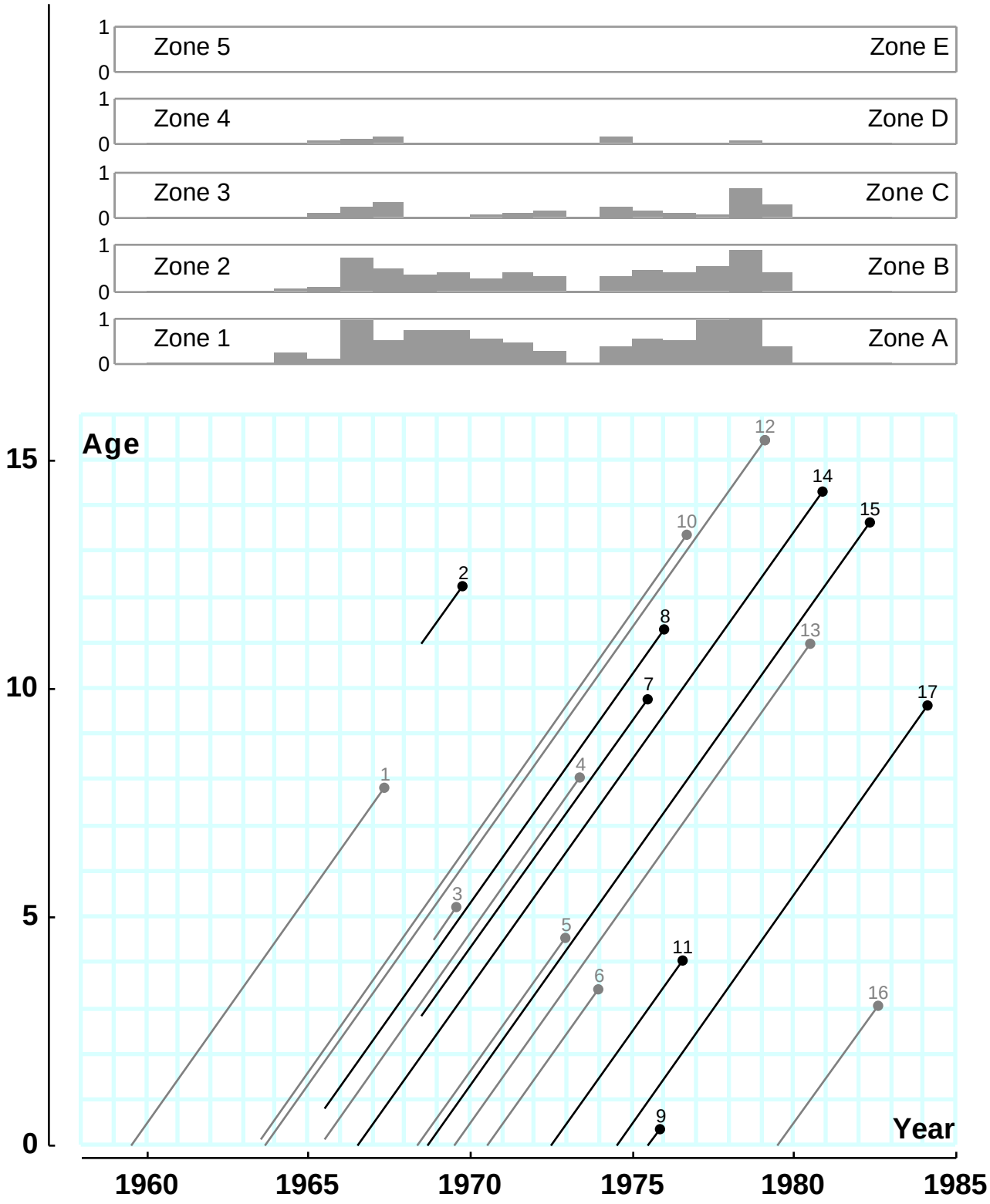

Zone 5
Zone E
Zone 4
Zone D
Zone 3
Zone C
Zone 2
Zone B
Zone 1
Zone A
1
0
Age
15
10
5
0
12
14
15
10
2
8
13
7
17
1
4
3
5
11
6
16
9
Year
1960
1965
1970
1975
1980
1985

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

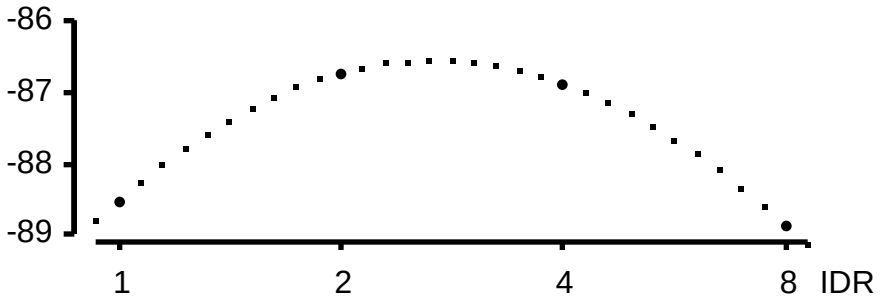


Figure 6 Log Likelihood calculations of the IDR comparing those with (x+1) vs. those with x "well-years" (W-Y) of cumulative exposure, illustrated using 4 selected leukemia cases [15, 13,12,7), and for just 2 values of the regression coefficient b. Cumulative exposure is depicted on the x axis. The cumulative 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

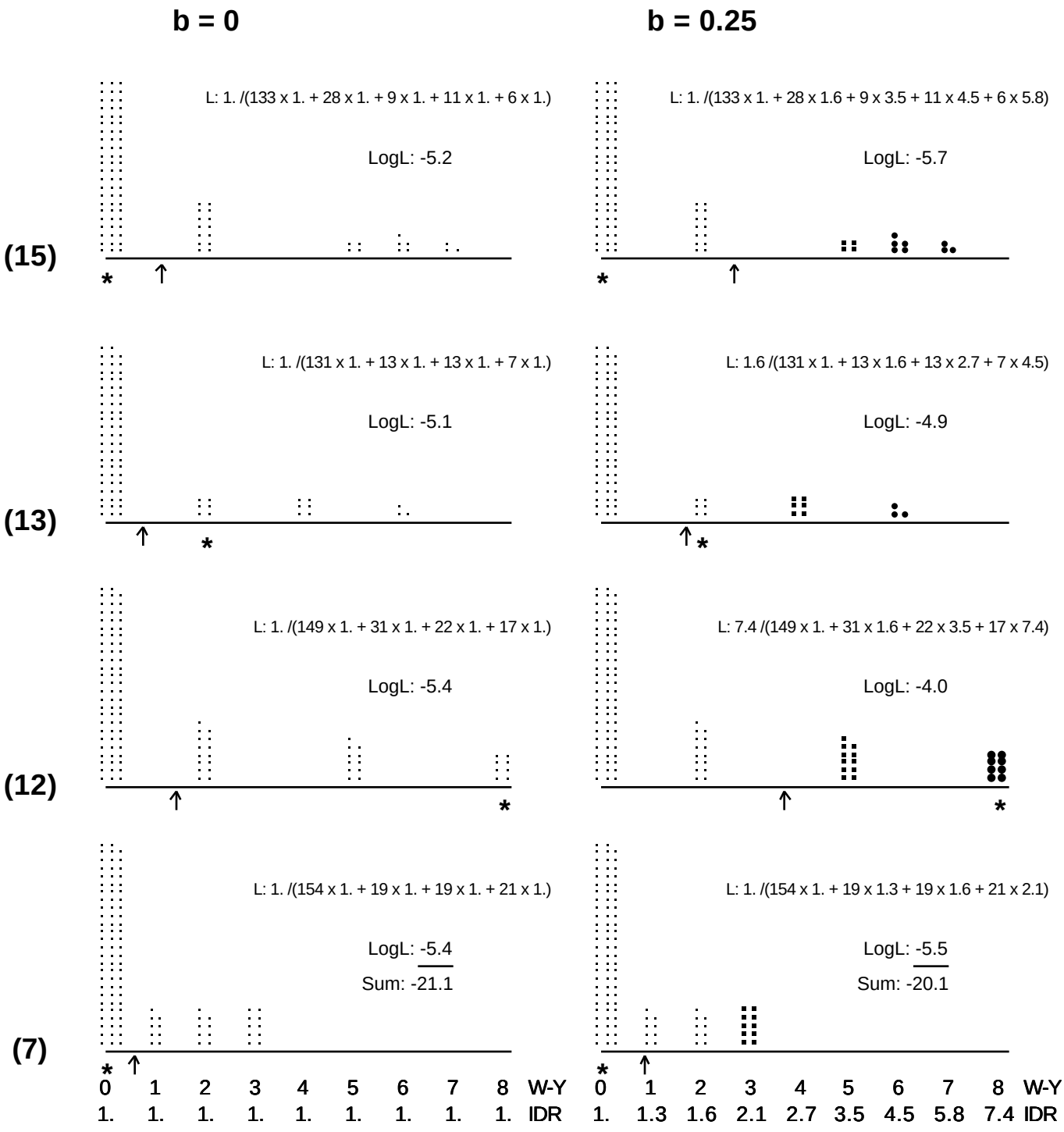
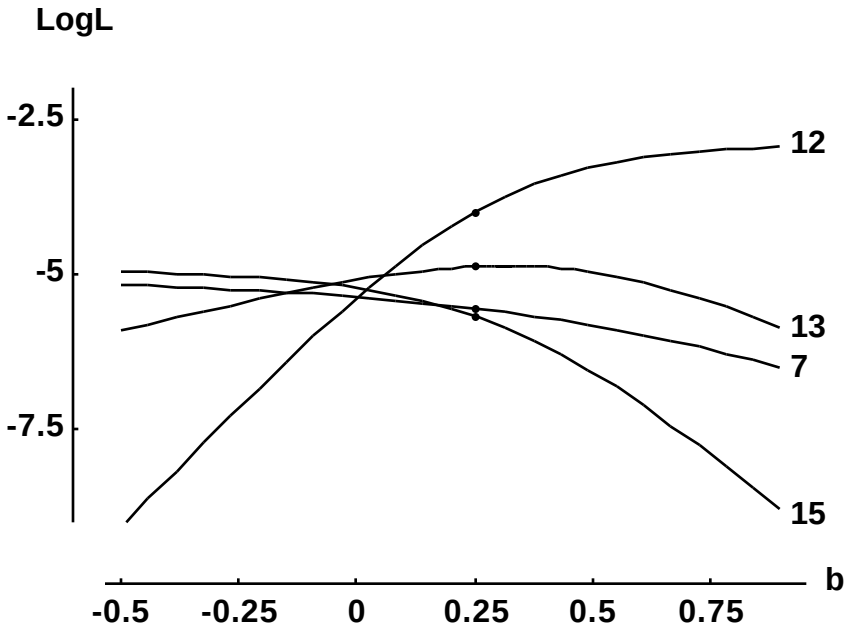
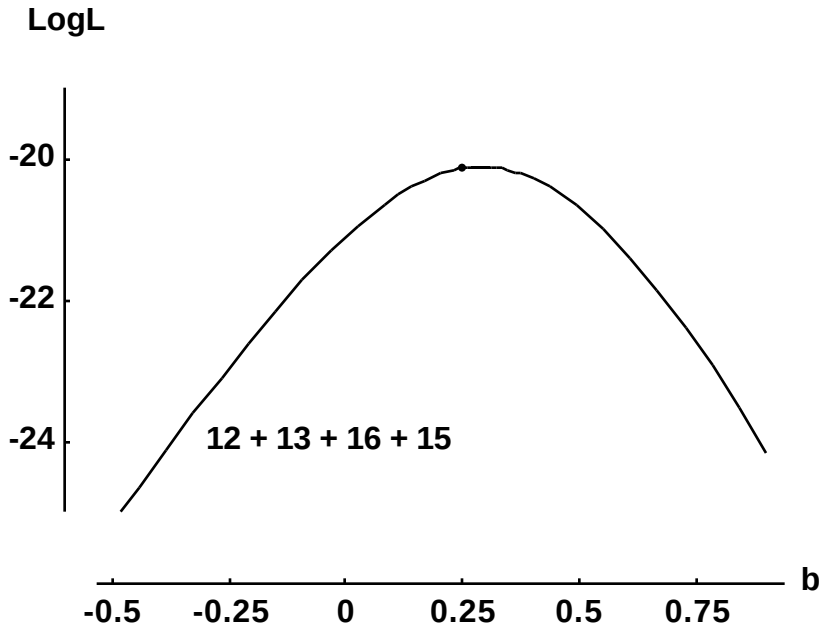

b = 0
b = 0.25
(15)
L: 1. /(133 x 1. + 28 x 1. + 9 x 1. + 11 x 1. + 6 x 1.)
LogL: -5.2
L: 1. /(133 x 1. + 28 x 1.6 + 9 x 3.5 + 11 x 4.5 + 6 x 5.8)
LogL: -5.7
(13)
L: 1. /(131 x 1. + 13 x 1. + 13 x 1. + 7 x 1.)
LogL: -5.1
L: 1.6 /(131 x 1. + 13 x 1.6 + 13 x 2.7 + 7 x 4.5)
LogL: -4.9
(12)
L: 1. /(149 x 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