Biased Estimates of Effect when a Strong but Balanced Covariate is Omitted from a Proportional Hazards Model

NOTES, 2025.06.16

It's based on a MSc Dissertation at McGill University in 2001. But, neither Jack (the then student, now Vice President and Executive Director, Center of Excellence for Statistics at Evidera, Greater Montreal Metropolitan Area [evidera.com — linkedin.com/evidera — twitter.com/evideraglobalhere], nor I (Jim Hanley, JH, his main supervisor, and now an emeritus professor) can now remember which journal(s) we submitted the manuscript to, or whether we actually ever did submit it: it looks like it's in submittable form, but JH can't find evidence that we did. The most recent reference is 2003, so that provides a left censored estimate of when we wrote the ms. !

Over the years, JH has used it in several graduate courses, including in this one, and as the basis for this assignment in this course and this course.

The fact that an ignored but balanced covariate can have different effects in different models (or with different 'comparative' parameters) was quite a surprise to many of us when, in their Biometrika paper, Gail, Tan and Piantadosi pointed it out in 1988. [Looking back on it now, it's interesting that the title of their paper focused on *tests* rather than on *estimators*]. JH suspects that the reason it didn't get as much initial 'traction' as it should have was because of the reliance on heavy algebra, and mere simulations.

We thought that our real example – albeit from the insect kingdom[1] with a covariate effect that was huge – made it much clearer when and why the phenomenon occurs.

At first, Jack and I thought it was peculiar to the Cox model, and the fact that the parameter-fitting is based on risk-sets. When one fits a 'regular' regression model to the 'subjects' in a dataset, these 'people' or 'individuals' do not move along time , bound together inside of risksets: they are not inter-connected together in any way.

But later, it became clearer that the phenomenon was a wider one, and more related to the concept of 'collapsibility', and to which type of effect measure one was modelling/pursuing.

In recent decades, its 'non-collapsibility' had been used by some authors to disparage the hazard function. In a 2024 piece, JH addressed this 'campaign', carried out primarily by 'causal inference' advocates who prefer to study 'people' (cohorts) and their 'risks', rather than rates (and hazards that are a functions of some time-scale) – that have a (possibly dynamic) population as their time-base.

If you would like to 'revitalize' this ms., we would be delighted to hear from you!

Sincerely,

James Hanley

webpage: https://jhanley.biostat.mcgill.ca | email: `james.hanley@mcgill.ca`

[1] Cf. the `FruitFlies` dataset in the `Stat2Data R` package, or JH's c622 webpage.

Biased Estimates of Effect when a Strong but Balanced Covariate is Omitted from a Proportional Hazards Model

Khajak Ishak, James A. Hanley, J. Jaime Caro

ABSTRACT

In the analysis of epidemiologic data, potential confounders that are balanced between study groups are often omitted from models estimating the effect of the determinant of interest. An extreme example is that of randomized trials where all potential confounders are omitted if groups appear balanced at baseline due to the randomization of patients. It has been shown, however, that baseline comparability of groups ensures unbiased estimates of the effect from linear regression models but not necessarily from non-linear models such as the logistic and the proportional hazards regression and other survival models. Omission of strong covariates from these models can lead to an attenuation of the effect estimate and a reduction in the power of statistical tests. The evidence presented in the literature is based on simulations and theoretical arguments. Here the bias in the proportional hazards model is illustrated using actual data from a study of the effect of sexual activity on mortality, and the mechanism that produces the bias is discussed.

Keywords: proportional hazards model, omitted covariate, bias to the null, confounding.

Abbreviations

HR Hazard Ratio

SE Standard Error

E8 High sexual activity, crowded group — male fruit flies received eight receptive female flies

C8 Inactive crowded group — male fruit flies received eight newly inseminated female flies

E1 Low sexual activity, not crowded group — male fruit flies received one receptive female fly

C1 Inactive, not crowded group — male fruit flies received one newly inseminated female fly

INTRODUCTION

Model misspecification due the omission of important covariates (i.e., potential confounders) is often a concern in the analysis of data from non-experimental studies. Researchers are less concerned, however, when analyzing data from randomized trials. With large enough samples, randomization is expected to yield groups that are balanced with respect to factors known, and possibly unknown, to be associated with the outcome. It is, therefore, common practice to estimate the treatment effect by comparing study groups directly, using simple statistical models that do not adjust for potential confounders.

Whether this approach will produce unbiased estimates of the treatment effect depends, in part, on the type of model being fitted. It is well known that in linear regression models (i.e., with Gaussian errors), inclusion of determinants of the outcome that are uncorrelated with the treatment will improve precision, even if it does not materially alter the point estimate. Non-linear models, such as logistic regression, some parametric survival and the semi-parametric proportional hazards model[1] differ in an important way in this regard. Gail and colleagues[2] were one of the first to point out (in 1984) that non-linear models are prone to underestimate the treatment effect when important determinants are omitted, even when groups are well balanced at baseline. Furthermore, the inclusion of other determinants in these non-linear models will typically not affect the standard errors of the parameter estimates, and may even reduce precision[3 - 5].

The behavior of the proportional hazards model when a strong covariate is omitted has been of particular interest in the literature due to the extensive application of this model in practice. The findings from the seminal paper by Gail and colleagues[2] have been confirmed with further

theoretical work and simulations[4, 6-12]. Gail[13] suggests that the bias to the null is likely to be important when all of the following conditions are met:

i) the omitted covariate is a strong determinant of the outcome;

ii) the determinant effect is strong;

iii) most individuals reach the endpoint by the end of the study period.

An important consequence is a loss in power of the score test to detect a statistically significant treatment effect since the point estimate is biased to the null[14, 15]. Corrections for the test have been proposed by Gail and colleagues[5] and more recently by Broët and colleagues[16].

To our knowledge, the only illustrations of the bias in the proportional hazards model have been from simulation studies. Although the theoretical aspects of the problem have been examined in detail, a detailed heuristic explanation of the mechanism that produces the bias has not appeared. In this paper, we report a striking illustration of the phenomenon in a real data set from an experiment that investigated the effect of sexual activity on the longevity of male fruit flies[17, 18]. We use this example to explore the mechanism that produces the bias, which provides useful insight about the proportional hazards model in general, particularly in the non-experimental context.

MATERIALS AND METHODS

The example

The design of the study has been described in detail elsewhere[17,18]. Briefly, 125 male fruit flies were randomly divided into five groups of 25 to determine whether increased reproduction

reduces the longevity of male flies. This effect is known to occur in female flies. Sexual activity of individual males was manipulated by providing each male in one group with eight new receptive females every two days and those in a second group with only one. These groups were denoted E8 and E1, respectively. To take into account the effect of competition for food or space due to the presence of the female flies, two other groups (denoted by C8 and C1) were created by providing males in these two groups with the same number of newly inseminated females every two days. It is known that newly inseminated females will not mate again for at least two days. This way, any differences between E8 and E1 can be attributed to the increased sexual activity. A fifth group kept males alone, but this group is not used in these analyses. All groups were treated the same way in terms of provision of fresh food. Compliance with the intended amount of sexual activity was assessed by monitoring insemination rates and was similar for all males in each group.

Each male fly was observed until death with longevity recorded in days. The length of the thorax, known to be an important determinant of fruit fly life expectancy, was measured at the beginning of the study. Since the thorax is fully grown at the time of hatching, there is no concern that it might grow during the life of the fly and, thus, become a time-dependent factor.

Statistical analysis

The effect of increased sexual activity was estimated separately for those with one partner (E1 vs. C1) and eight partners (E8 vs. C8) under two modeling paradigms. In the first, we used a linear regression model with Gaussian errors, which is possible in this case since none of the death times are censored. In the second analysis, we used a proportional hazards model to compare the mortality rate of sexually active flies with that of inactive flies. In both approaches,

we first fitted a model that only included a term for sexual activity, parameterized as a dichotomous covariate. These models provide the *crude*, or un-adjusted, estimate of the determinant effect. We also fitted an *adjusted* estimate by adding thorax length to the model. We then compared the point estimates and associated standard errors from the crude and adjusted models.

RESULTS

Table 1 summarizes the distribution of thorax length. In those provided one partner, the mean thorax length was slightly shorter (by 0.01mm) in C1 compared to E1. In those with eight partners, the difference in means was almost identical favoring C8 (0.81 vs. 0.80mm). In both groups, the quartiles of the distribution of thorax length were very similar between groups, confirming that the random allocation produced study groups that were well balanced in terms of thorax length.

Linear regression analysis

The effect of increased sexual activity in this analysis is measured in terms of a reduction in longevity of the flies. Table 2 summarizes the distribution of observed lifetimes. Among males provided with only one partner, the average lifespan of a fly was 64.8 days in C1 compared to 56.8 days in E1. Thus, a crude estimate of the cost of increased sexual activity in this group was 8 days (standard error (SE) = 4.3 days), which falls just short of statistical significance at the conventional five percent level of significance (two-sided $p = 0.07$). Among those with eight

partners, male flies in E8 lived, on average, 38.7 days compared to 63.4 days in C8, a difference of 24.6 days (SE = 3.8 days; two-sided $p < 0.01$).

This crude comparison is equivalent to fitting a linear regression model that includes only an indicator for sexual activity. The estimates obtained from this crude model and those obtained when controlling for thorax length are shown in Table 3. For flies provided with only one partner, adjusting for thorax length amplified the effect of increased sexual activity to 9.7 days, compared to 8.0 days in the crude comparison; the adjustment is effectively compensating for the slight initial advantage E1 in terms of thorax length. The standard error of the estimate of effect was reduced by 20 percent from 4.3 days in the crude comparison to 3.5 days. This, combined with the change in the point estimate increased the t-statistic sufficiently for the estimate to become statistically significant.

Among flies with eight partners, adjusting for thorax length produced only a slight change in the estimates of the determinant effect: the adjusted effect was 23.9 days in favor of C8, compared with 24.6 days in the crude analysis. The adjusted effect estimate is slightly weaker than the crude in this analysis, owing to the mild imbalance in thorax length favoring C8. The effect of the adjustment on the precision of the estimates is stronger in this analysis: the standard error of the adjusted estimate was only half of that of the crude estimate (2.2 days versus 3.8 days).

Proportional hazards analysis

The effect of the determinant is measured in terms of the relative mortality rate between the sexually active and inactive groups, quantified as a hazard ratio (HR). The results from crude and adjusted models are shown in Table 4.

Based on the crude model, increased sexuality among flies with one partner appears to have increased the mortality rate by about 60 percent (HR = 1.6, SE for the log HR = 0.3); but, as in the crude linear regression analysis, the estimate is not statistically significant at the five percent level of significance. When thorax length was taken into account, however, the effect estimate increased by 50 percent (HR = 2.4, SE for the log HR = 0.3) and reached statistical significance. Although the adjustment produced statistically significant results in both linear regression and proportional hazards analyses, we note an important difference between the two cases. In the linear model, a combined increase in the point estimate and its precision produced a statistically significant estimate in the adjusted model. In the proportional hazards model, however, the standard errors are the same to one decimal place in both the crude and adjusted models; the statistical significance of the adjusted estimate can, therefore, be attributed to the change in the point estimate alone.

Among those receiving eight partners, the crude analysis suggests that the mortality rate in the sexually active is eight times higher (HR = 8.0, SE for the log HR = 0.4). When thorax length is taken into account, however, the hazard ratio increased more than *three-fold* (HR = 25.1, SE for the log HR = 0.5), despite the slight advantage of sexually inactive at baseline which would be expected to reduce the effect size. Contrary to what we observed in the linear regression analysis, the standard error of the estimates hardly changed when thorax length was included in the model; in fact, the adjusted estimate of effect is slightly more variable than the crude.

Mechanism of the bias

We believe that the bias observed when a strong determinant (like thorax length) is omitted from the proportional hazards model is caused by the way the (partial) likelihood function of the

model is constructed. The probability at the time of each failure is quantified as the chance that, of all those who might have failed at that time (i.e., the risk set), a failure would be observed in the subject who actually failed. Due to the convenient parameterization of the model, these probabilities depend only on the covariates in the model and not on the baseline hazard function. More formally, the general form of each probability is given by $\exp(\beta' x_i) \Big/ \sum_{j \in R(T_i)} \exp(\beta' x_j)$, where

$\boldsymbol{\beta}$ is a vector of unknown parameters (the log-hazard ratios), $\boldsymbol{x}_i$ represents the vector of covariate values and T_i is the failure time of the i^{th} subject and *R(t)* denotes the risk set, defined as the set of subjects that are still at risk of failing at time *t*; that is, $R(t) = \{\text{individuals } j \mid T_j \geq t\}$.

The parameters of the model are estimated by maximizing the log-likelihood function given by the sum of the log of the probabilities defined above. Since these depend only on the covariates, the parameter estimates will depend on the discrepancy of the covariate patterns of those who fail and those at risk of failing at the time of the event. In fact, the first derivatives of the log-likelihood function, which are used to find the maximum, are of the form

$$\frac{\partial LPL(\beta)}{\partial \beta_r} = \sum_{i=1}^{k} \left[x_{ir} - \bar{x}_{w_i r} \right], \qquad [1]$$

where x_{ir} is the value of the r^{th} covariate for i^{th} subject who failed and $\bar{x}_{w_i r}$ is a weighted average of the r^{th} covariate in the risk set, and *k* is the total number of failures in the data set[19]. We note that the expression does not involve failure times. Thus, the parameters are estimated by finding values that make this weighted average equal to the covariate value of subject who failed in each risk set (since the likelihood is maximized by setting the first derivatives to 0).

This is in stark contrast with the way in which the parameters of the linear model are estimated. The kernel of the likelihood function is $\sum_{i=1}^{n}(T_i - \beta' x_i)^2$. Here, the observed failure times are being "compared" directly to their expected or "fitted" values determined by the parameters and covariates. In the simplest case, with a single dichotomous covariate, this would be equivalent to comparing the mean failure times in each of the levels of the covariate. The corresponding analogy for the proportional hazards model is to think of each subject who has the event as a "case" and the remainder of those in the risk set at that time as time-matched "controls" and imagine a 2x2 cross-tabulation of the covariate by case/control status. In that sense, one might think of each term in the partial likelihood as representing a stratum in a time-matched case-control study with one case at each time point. The overall estimate of effect can then be thought of as a weighted average of the stratum specific effect estimates, in the same spirit as the Mantel-Haenszel summary odds-ratio[20].

This highlights the "local" nature of the likelihood function of the proportional hazards model, since the probabilities depend on the composition of the risk sets. Thus, the model is sensitive to the order in which events occur, which is not the case in the likelihood of the linear model. To illustrate this, consider the simple example of a trial with four subjects evenly randomized between treated and control groups. Suppose now that the two treated subjects fail on days 35 and 45, and consider two scenarios for the control subjects: in the first, the failures occur on days 30 and 50 and in the second, they are observed on days 39 and 41. A linear regression analysis of the two scenarios would produce the same result, namely that there is no difference in the survival of patients in the treated and control groups since the average survival times are the same throughout. The proportional hazards analysis on the other hand would produce very different results in the two scenarios: in the first, the hazard ratio estimate is 1.62, while in the

second, the estimate is 0.62. Thus, the order of failures is crucial as it determines the composition of the risk sets and, hence, the estimates of effect.

Following this line of thought, we explored the composition of the risk sets in our analysis. Figure 1 shows the composition of the risk sets with respect to sexual activity for the eight and one partner groups. The top two panels (a and c) show the order in which failures occurred in the sexually active and inactive groups; each point on the graph represents a single death. The lower two panels (b and d) illustrate the composition of the risk sets at the beginning of the study and immediately following the 5^{th}, 15^{th}, 25^{th}, 35^{th} and 45^{th} failures.

The strength of the determinant can be seen clearly in the graphs of those with eight partners in terms of both the order of deaths and the composition of the risk sets. The first 12 deaths in this analysis, representing 25 percent of the entire sample, occurred in sexually active flies (panel c). As a result, the proportion of flies from this group in the risk set diminished fairly rapidly (panel d). A similar but far less pronounced pattern can also be seen for the one-partner group where the effect of sexual activity is much weaker.

We used a similar approach to examine the distribution of thorax length in the risk sets of the models (Figure 2, Table 5). For graphing purposes, we categorized thorax length at the quartiles of the distributions in each group. The strong impact of both thorax length and increased sexual activity can be for those with eight partners (panel c), where sexually active flies with short thorax were more likely to die early. More interestingly, though, the distribution of thorax length in each of the groups appears to be changing noticeably over time (panel d). The groups appear to be well-balanced at the start (risk set 0); by the 15^{th} risk set, however, the sexually active group no longer contains flies from the first quartile and very few from the second quartile, while the distribution of thorax length in the inactive group has hardly changed. The discrepancy

between the two groups becomes more striking in later risk sets; at the 25th death, the sexually active flies are solely from the third or fourth quartile of the thorax length. Table 5 shows the relative odds of being above the median (i.e., low risk). Values near unity suggest the study groups are balanced with respect to thorax length, while those above unity suggest a higher proportion of low risk flies among the sexually active. Thus, the increasing pattern reveals how the groups become increasingly unbalanced, with high risk sexually active flies dying out at a higher rate than in the sexually inactive. We observe a similar but less pronounced pattern in the one partner group.

In this light, the problem becomes one of *confounding*. Although the two groups were balanced at baseline, the distribution of thorax length changed *differentially* over time. Hence, thorax length became associated with sexual activity – surviving active flies were more likely to have a longer thorax. As a result, the estimate from the crude model is confounded due to imbalances in the distribution of thorax length between the two groups in later risk sets.

DISCUSSION

Removing confounding from the estimates of effect of a determinant is a great preoccupation in epidemiologic research. When comparison groups are balanced at baseline with respect to covariates, either by design (as in randomized studies) or by lack of association with the determinant or chance (as in observational studies), researchers typically do not control for these factors in the analysis. In a review of 50 consecutive clinical trial reports from four major medical journals published between July and September 1997, Assmann and colleagues[22] found that there was little consistency in the use of crude or adjusted measures of effect and most focus on the crude results.

In this paper, we have demonstrated the existence and potential severity of the bias resulting from omitting a balanced covariate from a proportional hazards model with a real dataset, which was particularly well suited for this exercise. These data met all three conditions that Gail[13] postulated would be necessary for the bias to occur: first, there existed a strong prognostic factor, thorax length in this case; second, the determinant of interest had a strong effect – in fact, we had a low and high "dose" of the determinant and found that the bias was far less serious for the analysis of the "low dose"; and third, the outcome of interest was common – all flies died (no censoring). We showed that, although comparison groups were close to perfectly balanced at baseline, they quickly became unbalanced with respect to thorax length in the risk sets of the partial likelihood function.

In this context, the rationale behind Gail's three conditions becomes intuitively clear. Generally speaking, subjects at the "higher risk" levels of the prognostic factor are more susceptible to a determinant with a detrimental effect (or conversely for a protective one); therefore, those exposed will likely fail (or die) and be removed from future risk sets at a much higher rate than those in the reference group. This then gradually creates an imbalance between the study groups in later risk sets. The degree of imbalance induced between the study groups will depend on both the strength of the determinant and prognostic factor. In the presence of a weak determinant and/or prognostic factor, the composition of the study groups still changes but at a much slower rate. This is where the frequency of occurrence of the outcome plays an important role. Since the disparity between the study groups becomes increasingly important with each successive risk set/failure, many more failures are required for the imbalance to have a noticeable impact on the estimates. The commonness of the outcome is also an important factor in the case where the prognostic factor and determinant are strong: with each additional failure,

an increasingly more imbalanced risk set would be included in the likelihood, leading to a more biased estimate. In fact, when we randomly censored 25 percent of the death times among those with eight partners, the magnitude of the bias weakened: the adjusted hazard ratio increased only to 17.9. In practice, the problem may be further complicated if censoring depends on the risk factor since this directly impacts the composition of the risk sets. Even if the censoring occurs similarly in the study groups, it will influence the estimates of the effect of the determinant. For instance, in our analysis, all early failures occurred in the exposed group, so that the composition of the unexposed group remained unchanged. But, in the presence of informative censoring, its composition would change and depending on the subgroup that is more likely to drop out, the effect estimate may be attenuated or amplified.

Our explanation of the bias highlights the dynamic nature of the relationships between covariates in the proportional hazards model. This has specific implications for identifying potential confounders in this type of analysis for both observational and experimental studies. Since the estimation of parameters relies entirely on the comparison of the characteristics of the subjects who fail to those still at risk of failing – a changing group of patients – confounding becomes a *time-dependent* issue. Therefore, formal testing for differences in baseline characteristics between exposed and unexposed groups, a popular practice,[22, 23] will not necessarily identify all potential confounders.

A consequence of this *time-dependent* confounding in the proportional hazards model is a violation of the proportionality assumption. In keeping with our analogy of the model and a time-matched case-control study, the overall estimate of effect is, in some sense, an average of the "stratum" (or risk set)-specific effects. But, since study groups within successive risk sets are increasingly confounded, the "stratum" specific effect would be heterogeneous. Since the risk

sets also index time in the proportional hazards model, this translates to a non-proportional (or time-dependent) hazard ratio. To verify this in our analyses, we included an interaction term between determinant and the log of time in the crude eight-partners group model; the main effect estimate was 22.15 for sexual activity and the estimate of the interaction parameter was negative (HR = 0.76), suggesting that the strength of the effect declined over time. Neither was statistically significant but this may well be due to sample size, but the pattern suggests that the effect of the sexual activity would appear to decline over time when thorax length is ignored.

An alternative view on the omission of covariates under balanced designs has been suggested by Hauck and colleagues[24] who argue that both crude and adjusted measures are valid, each reflecting a different effect. The crude hazard ratio represents a "population-averaged" effect (i.e., obtained by combining all flies, regardless of thorax length), while the adjusted hazard ratio is a "subject or covariate specific" measure (i.e. obtained by only combining flies with the same thorax length). Thus, differences between the two should not be surprising. It is important to keep in mind, however, that in this case, the "population-averaged" effect is not only averaged over subjects[25] but also over time, since the overall effect estimate is an aggregate of the risk set (and so, time) specific effect estimates. In the presence of a strong determinant and risk factor, the composition of the population will change over time; furthermore, the immediate benefit or harm may be very high initially since susceptible subjects are likely to be affected by it more quickly, but its effect will decline thereafter. Thus, a "population-averaged" estimate is not appropriate in this context.

REFERENCES

1 Cox DR. Regression models and life-tables (with discussion). J R Stat Soc B 1972; 34:187-220.

2 Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with non-linear regressions and omitted covariates, Biometrika 1984; 17(3):431-44.

3 Ford I, Norrie J, Ahmadi S. Model inconsistency illustrated by the Cox proportional hazards model. Stat Med 1995; 14:735-46.

4 Anderson GL, Fleming TR. Model misspecification in proportional hazards regression. Biometrika 1995*;* 82:527-41.

5 Gail MH, Tan WY, Piantadosi S. Tests for no treatment effect in randomized clinical trials. Biometrika 1988; 75:57-64.

6 Chastang C, Byar D. A quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival models. Stat Med 1988; 7:1243-55.

7 Struthers CA, Kabfleish JD. Misspecified proportional hazard models. Biometrika 1986; 73:363-9.

8 Schumacher M, Olschewski M, Schmoor C. The impact of heterogeneity on the comparison of survival times. Stat Med 1987; 6:773-84.

9 Bretagnolle J, Huber-Carol C. Effects of omitting covariates in Cox's model for survival data. Scan J Stat 1988; 15:125-38.

10 Ford I, Norrie J, Ahmadi S. Model inconsistency, illustrated by the Cox proportional hazards model. Stat Med 1995; 14:735-46.

11 Anderson GL, Fleming TR. Model misspecification in proportional hazards regression. Biometrika 1995; 82:527-41.

12 Schmoor C, Schumacher M. Effect of covariate omission and categorization when analyzing randomized clinical trials with the Cox model. Stat Med 1997; 16:225-37 (1997).

13 Gail MH. Dave Byar's contribution to epidemiology. Control Clin Trials 1995; 16:230-48.

14 Lagakos S, Schoenfeld DA. Properties of proportional hazards score tests under mis-specified regression models, Biometrics 1984; 40:1037-48.

15 Morgan TM. Omitting covariates from the proportional hazards model. Biometrics 1986; 42:993-5.

16 Broët P, Moreau T, Lellouch J, et al. Unobserved covariates in the two-sample comparison of survival times: a maximin efficiency robust test. Stat Med 1999; 18:1791-1800.

17 Partridge L, Farquhar M. Sexual activity and the lifespan of male fruitflies. Nature 1981; 294:580-1.

18 Hanley JA, Shapiro SH. Sexual activity and lifespan of male fruitflies: a data set that gets attention. JSE 1994; 2(1).

19 Hosmer DW Jr., Lemeshow S. Applied Survival Analysis. New York, NY: John Wiley and Sons Inc., 1999.

20 Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959; 22:719-48.

21 Altman DG. Comparability of randomized groups. Statistician 1985, 34:125-36.

22 Assmann SF, Pocock SJ, Enos LE , et al. Subgroup analysis and other (mis) uses of baseline data in clinical trials. Lancet 2000; 355:1064-9.

23 Senn S. Testing for baseline balance in clinical trials. Stat Med 1994; 13:1715-26.

24 Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? Control Clin Trials 1998; 19:249-56.

25 Hanley JA, Negassa A, Edwardes MD, et al. Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation. Am J Epidemiol 2003; 157:364-75.

TABLES

Table 1. Distribution of thorax length (in mm) in the E8, C8, E1 and C1 groups at baseline.

	One Partner		Eight Partners	
	Sexually Active	Inactive	Sexually Active	Inactive
Mean (SD)	0.84 (0.07)	0.83 (0.07)	0.80 (0.08)	0.81 (0.08)
Lower Quartile	0.80	0.80	0.76	0.76
Median	0.84	0.84	0.82	0.80
Upper Quartile	0.90	0.88	0.88	0.84

Table 2. Distribution of longevity (in days) in the E8, C8, E1 and C1 groups at baseline.

	One Partner		Eight Partners	
	Sexually active	Inactive	Sexually active	Inactive
Mean (SD)	56.8 (14.9)	64.8 (15.7)	38.7 (12.1)	63.3 (14.5)
Lower Quartile	48	50	32	56
Median	56	65	40	65
Upper Quartile	68	72	47	77

Table 3. Parameter estimates and 95 percent confidence intervals obtained from fitting linear regression models separately for those with one partner and those with eight; models included terms for the study group indicator alone, the thorax length alone and both together.

	One-Partner		Eight-partners	
	$\hat{\beta}(SE)$	*95% CI*	$\hat{\beta}(SE)$	*95% CI*
Crude Model				
Active	-8.0* (4.3)	-16.4, 0.4	-24.6 (3.8)	-32.0, -17.2
Model with Adjustment				
Active	-9.7 (3.5)	-16.6, -2.8	-23.9 (2.2)	-28.2, -19.6
Thorax	134.3 (25.0)	85.3, 183.3	136.9 (14.0)	109.5, 164.3

* Not statistically significant at five percent level of significance.

Table 4. Parameter estimates and 95 percent confidence intervals obtained from fitting proportional hazards models separately for those with one partner and those with eight; models included terms for the study group indicator alone, the thorax length alone and both together.

	One-partner		Eight-Partners	
	*HR (SE)**	*95% CI*	*HR (SE)**	*95% CI*
Crude Model				
Active	1.6 (0.3)†	0.9, 2.9	8.0 (0.4)	3.7, 17.5
Model with Adjustment				
Active	2.4 (0.3)	1.3, 4.3	25.1 (0.5)	9.4, 66.9
Thorax	0.3‡ (2.8)	0.2, 0.5	0.2‡ (3.1)	0.1, 0.4

* The standard error (SE) is that of the parameter estimate (log of the hazard ratio) from the model.

† Not statistically significant at 5 percent level of significance.

‡ Hazard ratio associated with a 0.10 mm increase in thorax length.

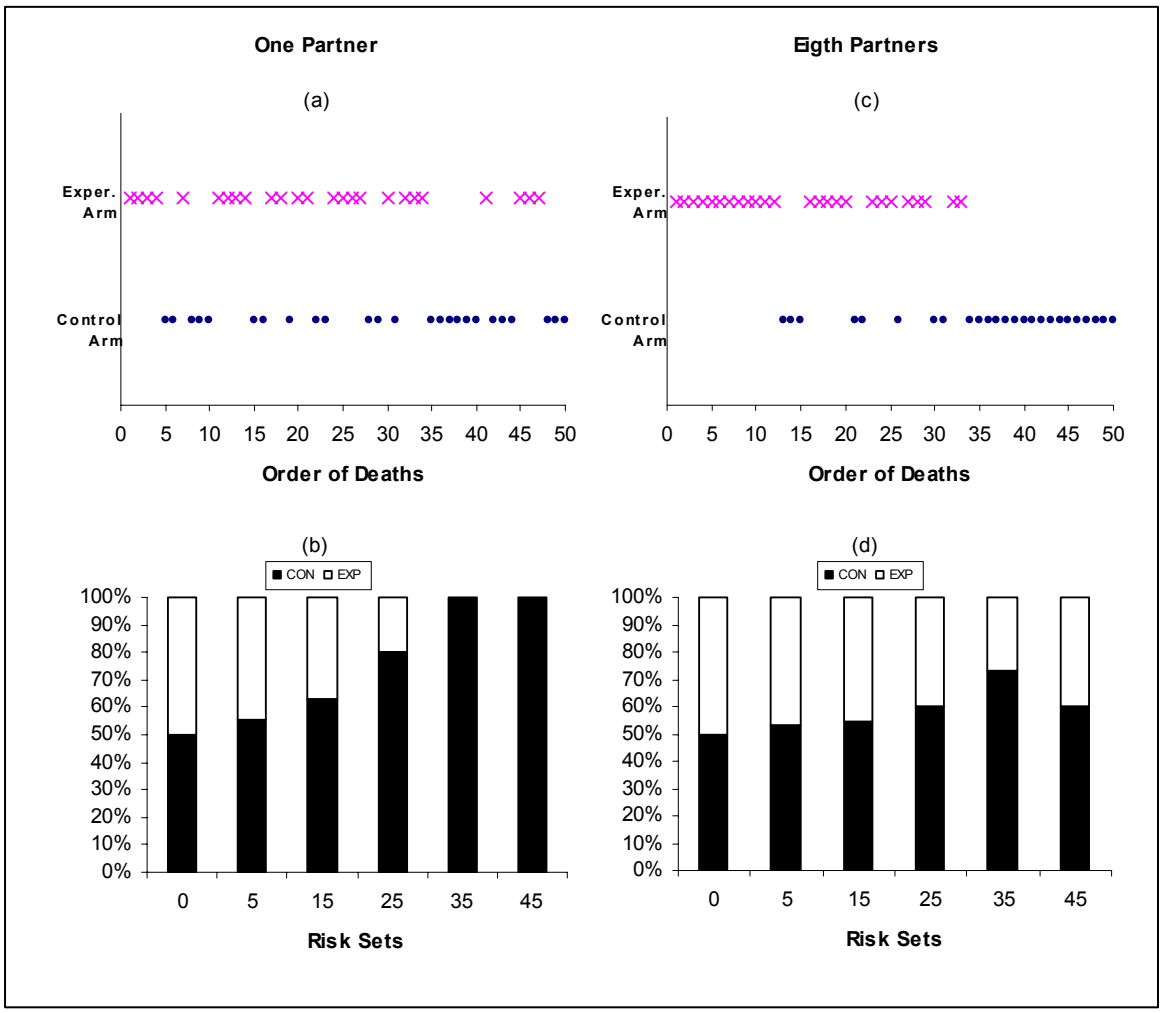
Table 5. Relative odds (and 95% confidence intervals) of being above the median thorax length in sexually active and inactive groups.

Risk Set	One-Partner	Eight-partners
0 (Baseline)	1.40 (0.45, 4.35)	1.17 (0.39, 3.56)
5	1.96 (0.60, 6.39)	1.76 (0.54, 5.72)
10	1.63 (0.47, 5.60)	4.69 (1.07, 20.6)
15	1.55 (0.41, 5.78)	5.04 (0.91, 28.0)
20	1.90 (0.45, 7.98)	∞^{*}
25	2.00 (0.41, 9.84)	∞^{*}
30	2.57 (0.37, 17.8)	∞^{*}

* No flies with thorax below the median length survived to the 20^{th} risk set, therefore, the odds ratio at and beyond this point is infinite.

FIGURES

Figure 1. Distribution of deaths over time by study group (top two panels) and composition of risk sets over time (bottom two panels).



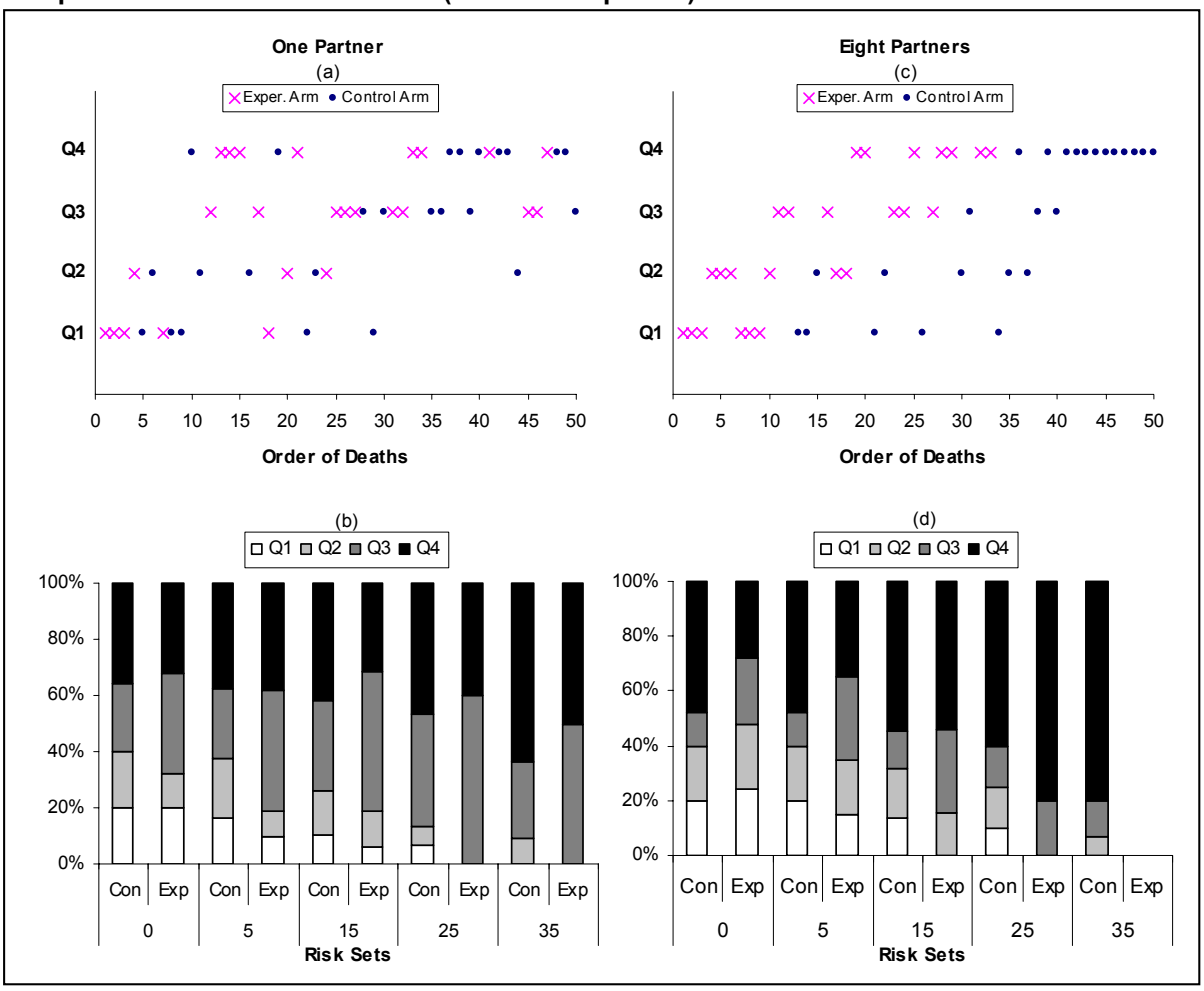
(a) Order of deaths by study group among those with one partner;

(b) Composition of risk sets at time zero and at the time of the 5th, 15th, 25th, 35th and 45th deaths among those with one partner.

(c) Order of deaths by study group among those with eight partners;

(d) Composition of risk sets at time zero and at the time of the 5th, 15th, 25th, 35th and 45th deaths among those with eight partners.

Figure 2. Distribution of deaths according to thorax length and study group (top two panels) and composition of risk sets over time (bottom two panels).



Q1, Q2, Q3, Q4 identify groups of flies in the four quartiles of the thorax length distribution

(a) Order of deaths by study group and thorax length quartiles among those with one partner;

(b) Composition of risk sets with respect to thorax size in flies remaining from the sexually active and inactive groups at time zero and at the time of the 5th, 15th, 25th, 35th and 45th deaths among those with one partner;

(c) Order of deaths by study group and thorax length quartiles among those with eight partners;

(d) Composition of risk sets with respect to thorax size in flies remaining from the sexually active and inactive groups at time zero and at the time of the 5th, 15th, 25th, 35th and 45th deaths among those with eight partners.