The Hazard Ratio and the Cox PH model

Limitations, common misuses, and potential alternatives

TIMI stats seminar, 12/14/2021
Andrea Bellavia

1.1 Motivation

1.2 Hazard ratios and Cox model racap

1.2.1 Hazard curve and hazards ratio

- **Survival analysis is about evaluating the time T to the development of a given endpoint D**

- **Its unique features are that we are equally and jointly interested in these two objects , and that the event is often censored**

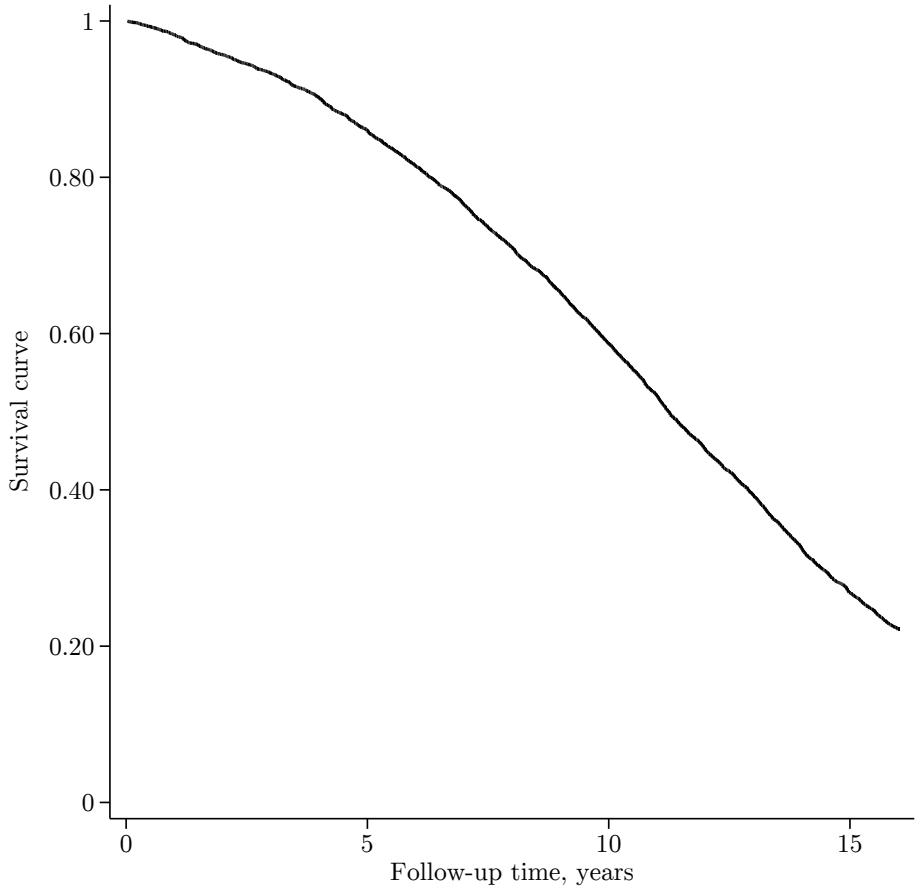
- **The main tools we use to summarize survival data are the survival function and the hazard function**

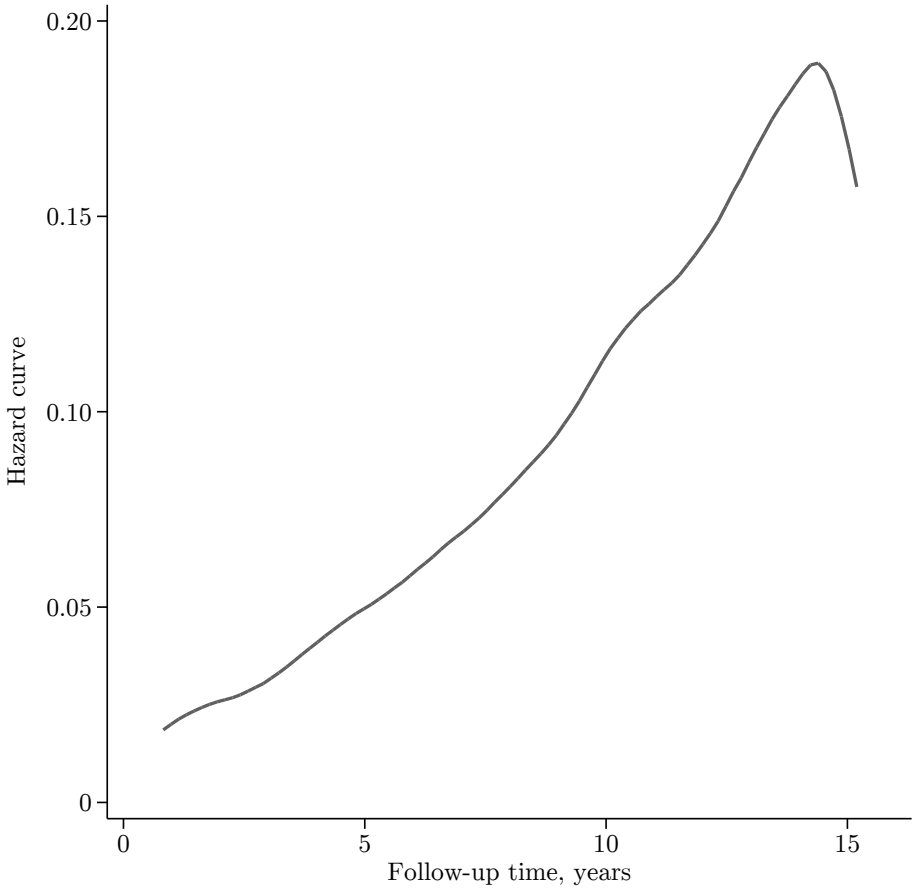
The survival curve combines information on the risk of the event of interest D and the time T by which this risk is achieved. It is bounded between 0 and 1, and risks are interpreted as probabilities of the event at a given time.

$$S(t) = P(T > t)$$

The hazard function represents the instantaneous failure rate , which is, the conditional probability of experiencing the event on a given time point given survival until that time.

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t) = \frac{f(t)}{S(t)}$$


Survival curve
1
0.80
0.60
0.40
0.20
0
0
5
10
15
Follow-up time, years


Hazard curve
0.20
0.15
0.10
0.05
0
0
5
10
15
Follow-up time, years

- **The hazard is a conditional probability and it is not bounded between 0 and 1 (only non-negative).**

- **Depending on the context, one or another parameter may be most relevant/revealing**

- **The survival probability is usually more relevant and easier to interpret**

- **On the other hand, the hazard is easier to estimate from censored data , and is used to build the major regression models and estimators**

1.2.2 *Hazard ratios and proportionality of the hazards*

In a given dataset/population, assuming proportionality of the hazards corresponds to assuming that, for every couple of individuals and at any given time point, the ratio of their hazards will be constant (and therefore, that the shape of their hazard curves is the same).

$$HR = \frac{\lambda_1(t)}{\lambda_0(t)} = \theta$$

- **Under this assumption the hazard ratio will not depend on time . This implies that a single summary measure can comprehensively summarize the overall difference in the survival experience between the 2 individuals.**

- **Note that we do not have to specify the functional form for the hazard.** $\lambda_0(t)$ **could be increasing, decreasing, constant, or anything else you can imagine. What we are assuming is that, whatever the general shape, this is the same for every individual in the study.**

- **The HR is a relative measure of association that captures how stronger/weaker the force of mortality is in one of the group , while assuming the two hazard curves are proportional over time**

1.2.3 *Proportional hazard models*

- **In a RCT where groups are balanced over covariates and the only difference is the assigned binary treatment, PH would imply that the HR summarizes the overall treatment-effect**
- **We can formalize this concept for more than one covariate and generalize this idea by writing the hazard ratio as a function of the covariates in what we call a PH model for survival data:** $\theta = f(\boldsymbol{Z})$
- **Cox (1972) was the first to discuss estimating the hazard ratio as a function of the covariates, and proposed the following (exponential) functional relationship:**

$$\theta = e^{\boldsymbol{\beta Z}}$$

which implies

$$\lambda_1(t) = \lambda_0(t) e^{\boldsymbol{\beta Z}}$$

- **Alternatives exist but are very rarely used.**

Sir David Cox: Statistics - past, present and future

9,380 views • Apr 10, 2014

109 DISLIKE SHARE SAVE ...

Regression Models and Life-Tables

By D. R. Cox

Imperial College, London

[Read before the Royal Statistical Society, at a meeting organized by the Research Section, on Wednesday, March 8th, 1972, Mr M. J. R. Healy in the Chair]

Summary

The analysis of censored failure times is considered. It is assumed that on each individual are available values of one or more explanatory variables. The hazard function (age-specific failure rate) is taken to be a function of the explanatory variables and unknown regression coefficients multiplied by an arbitrary and unknown function of time. A conditional likelihood is obtained, leading to inferences about the unknown regression coefficients. Some generalizations are outlined.

Keywords: LIFE TABLE; HAZARD FUNCTION; AGE-SPECIFIC FAILURE RATE; PRODUCT LIMIT ESTIMATE; REGRESSION; CONDITIONAL INFERENCE; ASYMPTOTIC THEORY; CENSORED DATA; TWO-SAMPLE RANK TESTS; MEDICAL APPLICATIONS; RELIABILITY THEORY; ACCELERATED LIFE TESTS.

1. Introduction

Life tables are one of the oldest statistical techniques and are extensively used by medical statisticians and by actuaries. Yet relatively little has been written about their more formal statistical theory. Kaplan and Meier (1958) gave a comprehensive review of earlier work and many new results. Chiang in a series of papers has, in particular, explored the connection with birth–death processes; see, for example, Chiang (1968). The present paper is largely concerned with the extension of the results of Kaplan and Meier to the comparison of life tables and more generally to the incorporation of regression-like arguments into life-table analysis. The arguments are asymptotic but are relevant to situations where the sampling fluctuations are large enough to be of practical importance. In other words, the applications are

In the general case, we think of the i-th individual having a set of covariates $\boldsymbol{Z}_i = (Z_{1i}, Z_{2i}, ..., Z_{pi})$, and we model their hazard rate as some multiple of the baseline hazard rate:

$$\lambda_i(t, \boldsymbol{Z}_i) = \lambda_0(t) \exp(\beta_1 Z_{1i} + \cdots + \beta_p Z_{pi})$$

This means we can write the log of the hazard ratio for the i-th individual to the reference group as:

$$\log\left(\frac{\lambda_i(t)}{\lambda_0(t)}\right) = \beta_1 Z_{1i} + \beta_2 Z_{2i} + \cdots + \beta_p Z_{pi}$$

This model is semiparametric . The baseline hazard (ie the hazard function in the baseline group) is unspecified, and the covariates are assumed to have an effect on this baseline function that is constant over time. Moreover, it is multiplicative. Parameters are interpreted as the effect of the covariates on the logarithm of the HR.

1.2.4 *Parameters estimation*

Cox (1975) extended his original results and derived an estimation procedure for censored data using the novel idea of a partial likelihood

Biometrika (1975), **62**, 2, *p.* 269
Printed in Great Britain

Partial likelihood

BY D. R. COX
Department of Mathematics, Imperial College, London

SUMMARY

A definition is given of partial likelihood generalizing the ideas of conditional and marginal likelihood. Applications include life tables and inference in stochastic processes. It is shown that the usual large-sample properties of maximum likelihood estimates and tests apply when partial likelihood is used.

Some key words: Asymptotic theory; Censoring; Conditional likelihood; Life table; Marginal likelihood; Regression; Stochastic process.

1. INTRODUCTION

Likelihood is central to much theoretical discussion of statistical inference, from whatever viewpoint. In simple cases, the likelihood is just the joint density of the observed values considered as a function of the unknown parameters. The introduction of modified definitions for complicated problems is due to Bartlett (1937), notable recent contributions being by Fraser (1968), Kalbfleisch & Sprott (1970), and Andersen (1973).

Possible aims of modified likelihood functions include:

(i) the achievement of robustness;

(ii) the study of problems, especially in stochastic processes, for which the full likelihood is difficult or impossible to compute;

(iii) the need to develop methods based on second-moment properties rather than full distributional assumptions (Wedderburn, 1974);

(iv) the reduction of dimensionality in situations with many nuisance parameters.

This paper concentrates on (iv). The need for special discussion of this situation arises in the sampling theory and pure likelihood approaches especially because of the failure of the method of maximum likelihood as a general technique when there are many nuisance parameters. In a Bayesian approach it may be a convenient approximation and simplification to bypass the prior distribution of the nuisance parameters.

Likelihood function with survival data

This is the likelihood function for a set of censored data. Parametric models deal with that by imputing a known functional form for $\lambda_i(X_i)$

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\lambda_i(X_i)}{\sum_{j \in \mathcal{R}(X_i)} \lambda_j(X_i)} \right]^{\delta_i} \left[\sum_{j \in \mathcal{R}(X_i)} \lambda_j(X_i) \right]^{\delta_i} \exp[-\int_0^{X_i} \lambda_i(u) du]$$

Cox (1975) argued that the first term in this product contained almost all of the information about $\boldsymbol{\beta}$, while the second two terms contained the information about $\lambda_0(t)$, i.e., the baseline hazard. In a Cox model $\lambda_0(t)$ is unspecified, so these can be treated as nuisance parameters.

Under the Cox PH assumption:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\lambda_i(X_i)}{\sum_{j \in \mathcal{R}(X_i)} \lambda_j(X_i)} \right]^{\delta_i}$$

This is the partial likelihood defined by Cox.

There are three aspects of this estimation procedure that are relevant to our discussion:

- **Only the rank of event times matter. There is no information on elapsed time and the likelihood is updated every time there is an event**

- **The way this update occurs is by comparing the individual who is experiencing the event with all other individuals who were still at risk at that same time point**

- **The removal of the portions related to the baseline hazards is what makes the model semi-parametric. As a consequence of this, the Cox model does not have an intercept.**

1.3 Cox model: problems and misinterpretations

Several concerns have been raised around the use (and possible misuses) of the Cox model. Cox himself (1994) said: "in the light of some of the fursther results one knows since, I thinkg I eould want to tackle problems parametrically". In general, these can be broadly summarized into:

- **Problems related to the Hazard Ratio and its interpretation**
- **Problems related to the way the partial likelihood estimation works**

Importantly, here we are not referring to problems related to the Cox model assumptions (e.g. log-linearity, PH ...) that can usually be addressed within a Cox environment

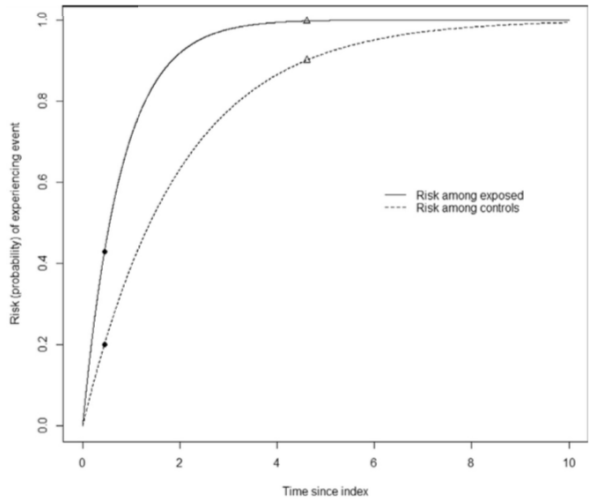
1.3.1 The Hazard Ratio is a rate ratio, not a relative risk

- **The question of interest is often about the risk of the event. This is what Kaplan-Meier tells us at the univariate level**
- **When comparing survival curves we can calculate risk ratios (or relative risks) , as the ratio between the risk in two different groups**
- **This measure is:**
  1. **difficult to estimate at the multivariable level**
  2. **time-dependent even under PH**
- **We tend to report HRs because they are easier to estimate at the multivariate level, however these are not relative risks (RR), nor approximations**
- **The HR is a rate ratio , the ratio of the instantaneous rates of the event**

A recent paper (Sutradhar and Austin, 2018) provides a nice description of how a HR can be translated into a RR.

Table 1
Magnitude of relative risk under various values of the hazard ratio and baseline risk

		Hazard ratio/relative rate obtained from Cox model					
		0.1	0.5	1	1.5	2	2.5
Risk/probability of experiencing the event by time *t* among controls	0.9	0.228	0.759	1	1.076	1.1	1.107
	0.8	0.185	0.691	1	1.138	1.2	1.227
	0.7	0.162	0.646	1	1.193	1.3	1.358
	0.6	0.145	0.612	1	1.245	1.4	1.498
	0.5	0.133	0.585	1	1.292	1.5	1.646
	0.4	0.124	0.563	1	1.338	1.6	1.802
	0.3	0.116	0.544	1	1.381	1.7	1.966
	0.2	0.110	0.527	1	1.422	1.8	2.137
	0.1	0.104	0.513	1	1.461	1.9	2.315



Ref. *Annals of Internal Medicine*, Guideline for Authors

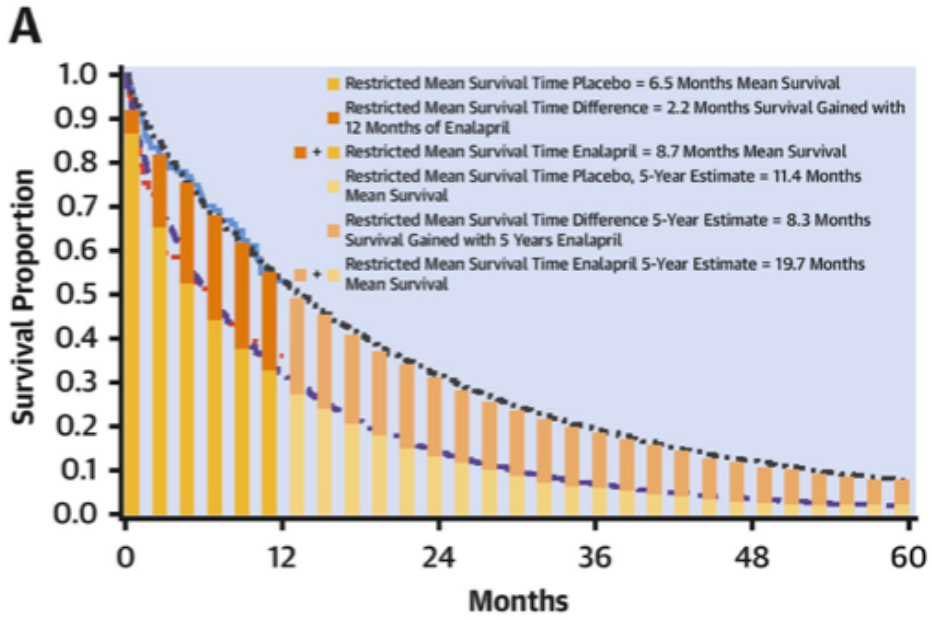
http://annals.org/public/authorsinfo.aspx

Statistical Guidelines

Presentation	
Issue	Notes
Percentages	Report percentages to one decimal place (i.e., xx.x%) when sample size is >=200. To avoid the appearance of a level of precision that is not present with small samples, do not use decimal places (i.e., xx%, not xx.xx%) when sample size is < 200.
Standard	Use "mean (SD)" rather than "mean ± SD" notation. The ± symbol is ambiguous and can represent standard deviation or standard error.
Cox models	When reporting the findings from Cox proportional hazards models: ▪ Do not describe hazard ratios as relative risks. ▪ Do report how the assumption of proportional hazards was tested, and what the test showed.
P values	For *P* values between 0.001 and 0.20, please report the value to the nearest thousandth. For *P* values greater than 0.20, please report the value to the nearest hundredth. For *P* values less than 0.001, report as "*P*<0.001."
"Trend"	Use the word *trend* when describing a test for trend or dose-response. Avoid the term *trend* when referring to *P* values near but not below 0.05. In such instances, simply report a difference and the confidence interval of the difference (if appropriate) with or without the *P* value.
Statistical software	Specify in the statistical analysis section the statistical software—version, manufacturer, manufacturer's location, and the specific functions, procedures, or programs—used for analyses.
Cox models	When reporting the findings from Cox proportional hazards models: ▪ Do not describe hazard ratios as relative risks. ▪ Do report how the assumption of proportional hazards was tested, and what the test showed.

- **The RR-HR difference is not a problem of linguistic and terminology. The critical point here is about the interpretation of the HR and whether this is really what we are after**
- **For example, if the goal is to estimate a measure that can summarize the information presented in a KM figure, or a multivariable-adjusted version of a KM, then the HR is not the best choice**
- **In this context, presenting a HR with a RR interpretation might be misleading**
- **Alternative options have been presented and are nicely summarized in this paper (Harvard/Dana Farber faculty)**

A useful and increasingly popular option is to focus on restricted mean survival time (RMST), which can also be directly estimated from parametric and flexible parametric models adjusting for covariates (Fig from Perego et al. JACC HF 2020).



Treatment	RMST (5-yr), Months	ΔRMST, Months	Bootstrap 95% CI, Months
Enalapril	19.7	8.3	1.5 to 15.8
Placebo	11.4		

Type

Observed (KM Enalapril)

Observed (KM Placebo)

Predicted from Weibull Model (Enalapril)

Predicted from Weibull Model (Placebo)

Table from Uno et al.

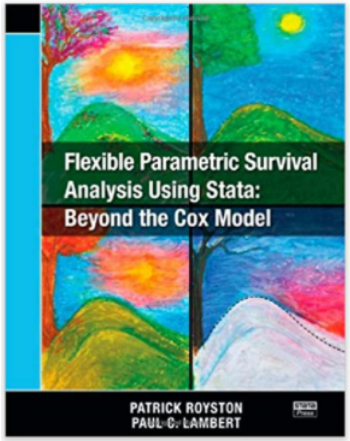
Table 1. Summary of Various Estimates for the Between-Group Difference and Corresponding 95% CIs

	Study					
	Rajkumar et al[10] (myeloma)		Zukin et al[15] (NSCLC)		Allegra et al[17] (colon cancer)	
Group Contrast Measure	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Hazard ratio (PH model)	0.87	0.60 to 1.27	0.62	0.46 to 0.83	0.95	0.79 to 1.13
t-year survival	Month 40		Month 24		Month 60	
Difference	−0.04	−0.15 to 0.06	0.11	0.02 to 0.21	0.02	−0.02 to 0.05
Ratio	0.95	0.82 to 1.10	2.74	1.09 to 6.93	1.02	0.98 to 1.06
Percentiles	10th		50th		10th	
Difference (months)	10.9	2.6 to 19.1	3.7	1.3 to 6.0	1.5	−3.9 to 7.0
Ratio	2.15	1.17 to 3.96	1.66	1.21 to 2.27	1.04	0.90 to 1.21
Restricted mean survival time	Month 40		Month 35		Month 60	
Difference (months)	2.2	0.1 to 4.2	3.9	1.5 to 6.3	0.3	−0.7 to 1.3
Ratio	1.06	1.00 to 1.13	1.49	1.17 to 1.91	1.00	0.99 to 1.02
Ratio of restricted mean time lost	0.68	0.47 to 0.98	0.86	0.77 to 0.94	0.95	0.78 to 1.16

Abbreviations: NSCLC, non–small-cell lung cancer; PH, proportional hazards.

The clinical interpretation of the HR

- The HR is not a RR but is the ratio between the instantaneous rates of the event in two groups of individuals
- When estimating HRs via Cox , the absence of an intercept makes the interpretation of this measure even less straightforward. Suppose we obtain a HR=1.3 and conclude that treated individuals have 30% higher rate of the event. We don't really know what the "30% higher" is compared to. Is is a strong or weak effect?
- Estimating the baseline hazard would definitely improve this interpretation.
- This can be accomplished in several ways, from semi-parametric models (e.g. Breslow) to flexible parametric models that estimate the baseline hazard with smooth splines functions

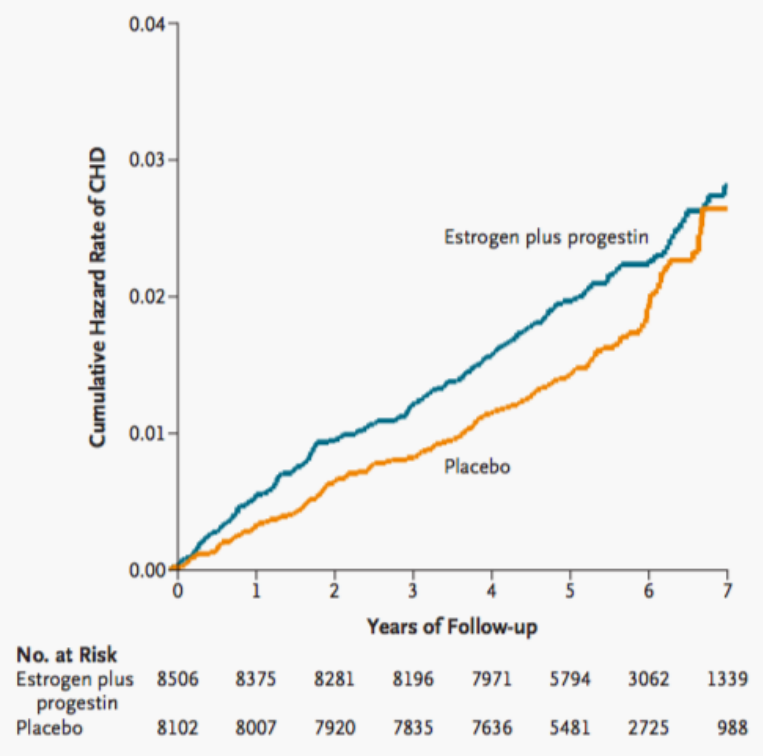


1.3.2 Hazard Ratios depend on the length of follow-up

This is a minor caveat that could however severely impact the interpretation of RCTs results. It is easily explained using a famous example of a RCT conducted on 16,000 women for an average follow-up of 5.2 years, halted due to safety concerns.

From the abstract: "Combined hormone therapy was associated with a hazard ratio for CHD of 1.24 (nominal 95 percent confidence interval, 1.00 to 1.54; 95 percent confidence interval after adjustment for sequential monitoring, 0.97 to 1.60)."

These are the cumulative hazard curves. There is no evidence of departures from PH.



In Table 2 they report the year-specific HRs:

Table 2. Estrogen plus Progestin and the Risk of CHD, According to Year of Follow-up.*

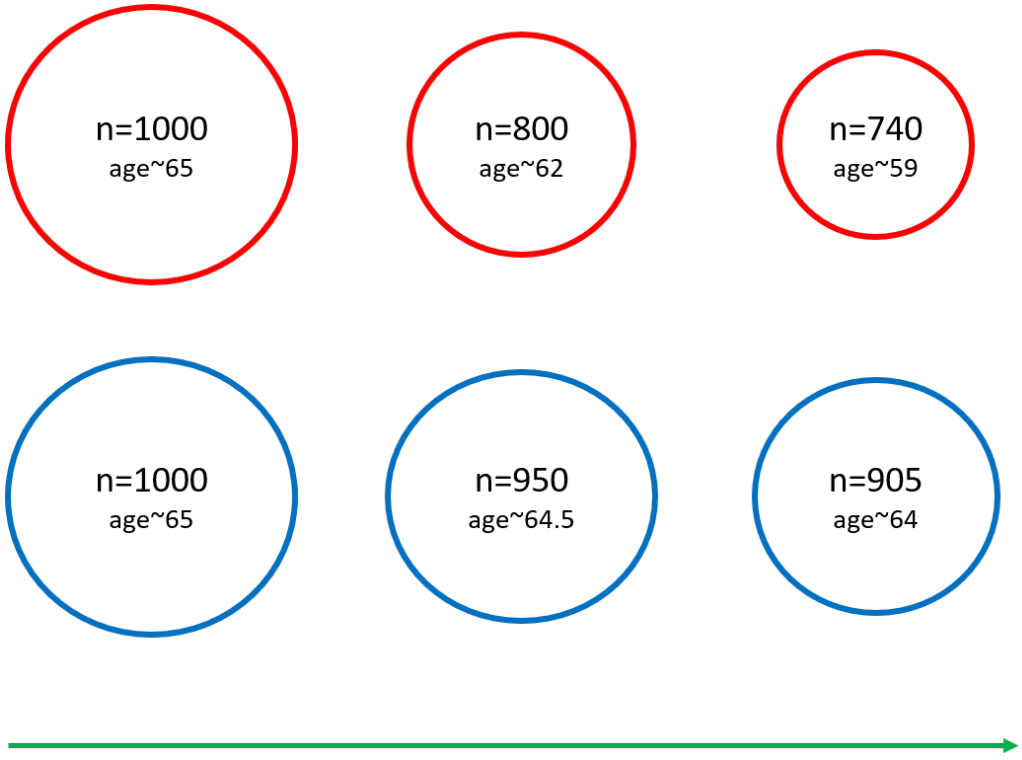
Year of Follow-up	CHD		Hazard Ratio for CHD (95% CI)
	Estrogen-plus-Progestin Group	Placebo Group	
	no. of cases (annualized percentage)		
1	42 (0.50)	23 (0.29)	1.81 (1.09–3.01)
2	38 (0.45)	28 (0.35)	1.34 (0.82–2.18)
3	19 (0.23)	15 (0.19)	1.27 (0.64–2.50)
4	32 (0.39)	25 (0.32)	1.25 (0.74–2.12)
5	29 (0.41)	19 (0.28)	1.45 (0.81–2.59)
≥6	28 (0.37)	37 (0.56)	0.70 (0.42–1.14)

The HR from the abstract is a weighted average of these year-specific ones. Had we stopped the study after 1 year, we would have observed an 80% higher rate. Had we continued the study for another 5 years, we might have observed a non-significant difference.

The HR depends on the length of follow-up because the average HR ignores the distribution of events during the follow-up (only the rank matters).

1.3.3 Built-in selection bias

Think of a RCT where participants are assigned to 2 treatments: the red treatment has a much higher hazard of events.



- **Participants experiencing the event in the second years are selected , as they had to survive the first time point**

- **Survivors, however, are not the same among groups. For example, if age is also a predictor, older individuals may experience the event faster**

- **Over time, the beneficial group will be older, and groups will be unbalanced**

- **The true effect will likely be underestimated**

- **This is due to the way the partial likelihood procedure operates. Events are ranked and covariates comparison occurs across individuals who have survived up to the given point of interest**

- **An easy fix is to validate results by adjusting for covariates that were used in the matching (age, for example)**

- **Using inverse probability weights is also a good way to address this problem**

- **If parametric assumptions can be made, AFT models or GLM models with Poisson distribution can be better choices in settings where this bias is likely to occur**

1.4 Practical implications

It is NOT time to replace the Cox model. At the same time, since several alternative techniques have been presented, it might be time to take such alternatives into account when Cox's limitations are a threat, and to revisit/expand guidelines for routine analyses

1. **What is the clinical question of interest?**
2. **Do not take for granted that the HR is the metric that better addresses such question**
3. **Include additional metrics (RMST, median survival . . .) and modeling techniques into routine analyses (e.g. revise macros)**
4. **When interpreting Cox's results do not lose the focus on the clinical interpretation**
5. **Consider accompanying HRs and COX models with other metrics and techniques already in SAPs and protocols (and possibly not as secondary/confirmatory analyses)**

I added relevant papers/material on the metrics and modeling techniques mentioned here in `Z:/_TIMI_resource/AndreaB/Survival Analysis`

Next stats seminars:

- **Jan 20 - Splines modeling (Thursday meeting)**
- **Feb 15 - Comparison of predictive models (NRI, C-stats, alternatives)**
- **Mar 24 - Bayesian adaptive trial design (Thursday meeting - invited speaker)**