

Analysis of Time to Event Data

Kaplan-Meier and Cox Regression Analysis

**Global MECOR Course
Kenya, 2011**

Kaplan-Meier Analysis

- We will motivate the construction of Kaplan-Meier survival curves, and the logrank test for comparing them, by beginning with the analysis of life tables.
- Kaplan-Meier analysis is the limiting case of this methodology.

Kaplan-Meier Analysis

Analysis of Life Tables

- Suppose that we are measuring survival on a cohort of n individuals and that we are only able to assess their status at $k+1$ points in time:
 t_1, t_2, \dots, t_{k+1} .
- For the interval (t_i, t_{i+1}) we know only the number who started the interval alive, the number who finished alive, and the number who died.
- This might be the case, for example, with the analysis of vital statistics data.

Kaplan-Meier Analysis

Analysis of Life Tables

A hypothetical example:

Interval	# alive at start	# died in interval	# lost to follow-up
(t_1, t_2)	1000	10	5
(t_2, t_3)	985	20	10
(t_3, t_4)	955	15	30
↓	↓	↓	↓
(t_k, t_{k+1})	753	13	18

Kaplan-Meier Analysis

Analysis of Life Tables

More generally, our data may be arrayed as follows:

Interval	# alive at start	# died in interval	# lost to follow-up
(t_1, t_2)	$L_1=n$	D_1	W_1
(t_2, t_3)	L_2	D_2	W_2
(t_3, t_4)	L_3	D_3	W_3
↓	↓	↓	↓
(t_k, t_{k+1})	L_k	D_k	W_k

where

$$L_2 = L_1 - D_1 - W_1$$

and in general that

$$L_{i+1} = L_i - D_i - W_i \quad \text{for } i = 1, 2, \dots, k$$

Kaplan-Meier Analysis

Analysis of Life Tables

We wish to calculate

$S_i = S(t_i)$ = Probability of surviving to
the start of the i^{th} interval

for which we will also find it helpful to calculate

$H_i = H(t_i)$ = Probability of dying during i^{th} interval
given that you survived to the start of
the i^{th} interval

We refer to $S(t_i)$ as the **survival function** and to $H(t_i)$ as the **hazard function**. Note that $H(t_i)$ is a **conditional** probability and is thus quite distinct from the **unconditional** probability of death

Kaplan-Meier Analysis

Analysis of Life Tables

$H(t_i)$ can be readily estimated by

$$h(t_i) = \frac{D_i}{\left(L_i - \frac{W_i}{2} \right)}$$

} assumes LTF occurs uniformly over the interval

Kaplan-Meier Analysis

Analysis of Life Tables

How do we compute an estimate, $s(t_i)$, of $S(t_i)$?

$$s_1 = \mathbf{1} \quad \text{by definition}$$

$$s_2 = (\mathbf{1} - \mathbf{h}_1) \quad (\text{prob don't die in 1}^{\text{st}} \text{ interval})$$

$$s_3 = s_2 * (\mathbf{1} - \mathbf{h}_2) = (\mathbf{1} - \mathbf{h}_1) * (\mathbf{1} - \mathbf{h}_2)$$

and more generally

$$s_i = (\mathbf{1} - \mathbf{h}_1) * (\mathbf{1} - \mathbf{h}_2) * \dots * (\mathbf{1} - \mathbf{h}_{i-1})$$

Kaplan-Meier Analysis

Analysis of Life Tables

A hypothetical example:

Interval	# alive at start	# died in interval	# lost to follow-up	H(t)	(S(t))
(t_1, t_2)	1000	10	5	.0100	1
(t_2, t_3)	985	20	10	.0204	.9900
(t_3, t_4)	955	15	30	.0160	.9698
↓	↓	↓	↓		
(t_k, t_{k+1})	753	13	18	.0175	.7341

Kaplan-Meier Analysis

The logrank test

Consider the i th time interval, (t_i, t_{i+1}) , and assume that we have data for two groups, A & B. Our data might look as follows:

	Group A	Group B	
died	13	20	33
survived	167	212	379
Total at risk	180	232	412

$$E_{ia} = 180 \left(\frac{33}{412} \right)$$

$$E_{ib} = 232 \left(\frac{33}{412} \right)$$

Kaplan-Meier Analysis

The logrank test

Now let

\mathbf{O}_a = sum all deaths in group A

\mathbf{O}_b = sum all deaths in group B

\mathbf{E}_a = sum of E_{ia} s = expected # deaths in group A

\mathbf{E}_b = sum of E_{ib} s = expected # deaths in group B

It is easy to show that

$$[\mathbf{O}_a + \mathbf{O}_b = \mathbf{E}_a + \mathbf{E}_b]$$

Observed # deaths = Expected # deaths

Kaplan-Meier Analysis

The logrank test

The logrank test statistics is given by

$$X^2 = \frac{(O_a - E_a)^2}{E_a} + \frac{(O_b - E_b)^2}{E_b}$$

This is the same form as the Pearson χ^2 test for 2-way tables!

Under the null hypothesis of no difference in survival rates, X^2 will have a chi-square distribution with one degree of freedom.

We reject H_0 if X^2 gets too big.

Kaplan-Meier Analysis

The logrank test -- more than 2 groups

Now suppose instead of just 2 groups we have some arbitrary number, g , of groups.

Calculate O_a, O_b, \dots, O_g and E_a, E_b, \dots, E_g as before.

$$X^2 = \left. \frac{(O_a - E_a)^2}{E_a} + \frac{(O_b - E_b)^2}{E_b} + \dots + \frac{(O_g - E_g)^2}{E_g} \right\} \sim \chi^2_{g-1}$$

Kaplan-Meier Analysis

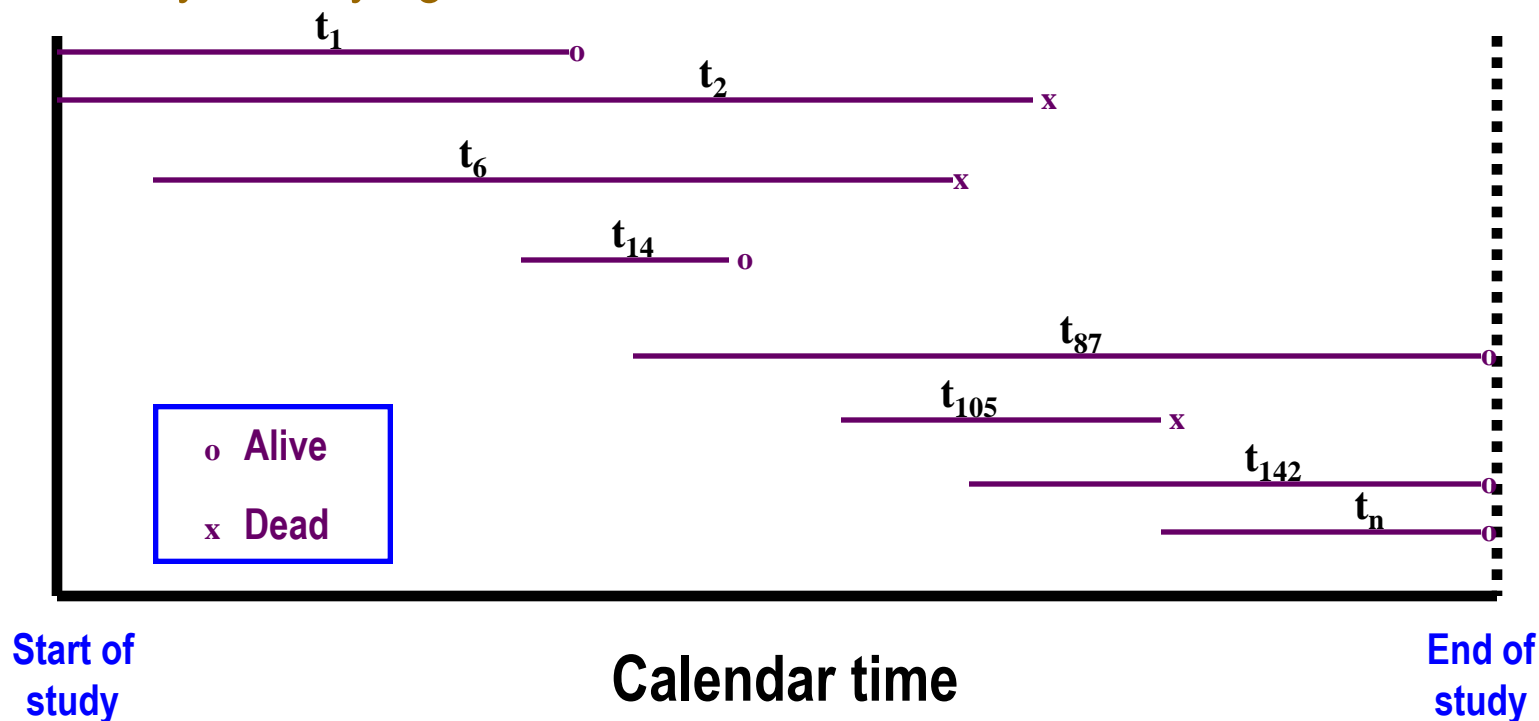
Kaplan-Meier: the limiting case

- Kaplan-Meier survival curves, and the corresponding logrank test for comparing them, are just the limiting case of the life table methodology when our time intervals get small (e.g., time measured in days rather than in years).
- Multiple deaths and/or loss to follow-up at the same timepoint become less and less common
- Otherwise, the calculations for $H(t)$, $S(t)$, and the logrank statistic are unchanged!

Kaplan-Meier/Cox Analysis

The data

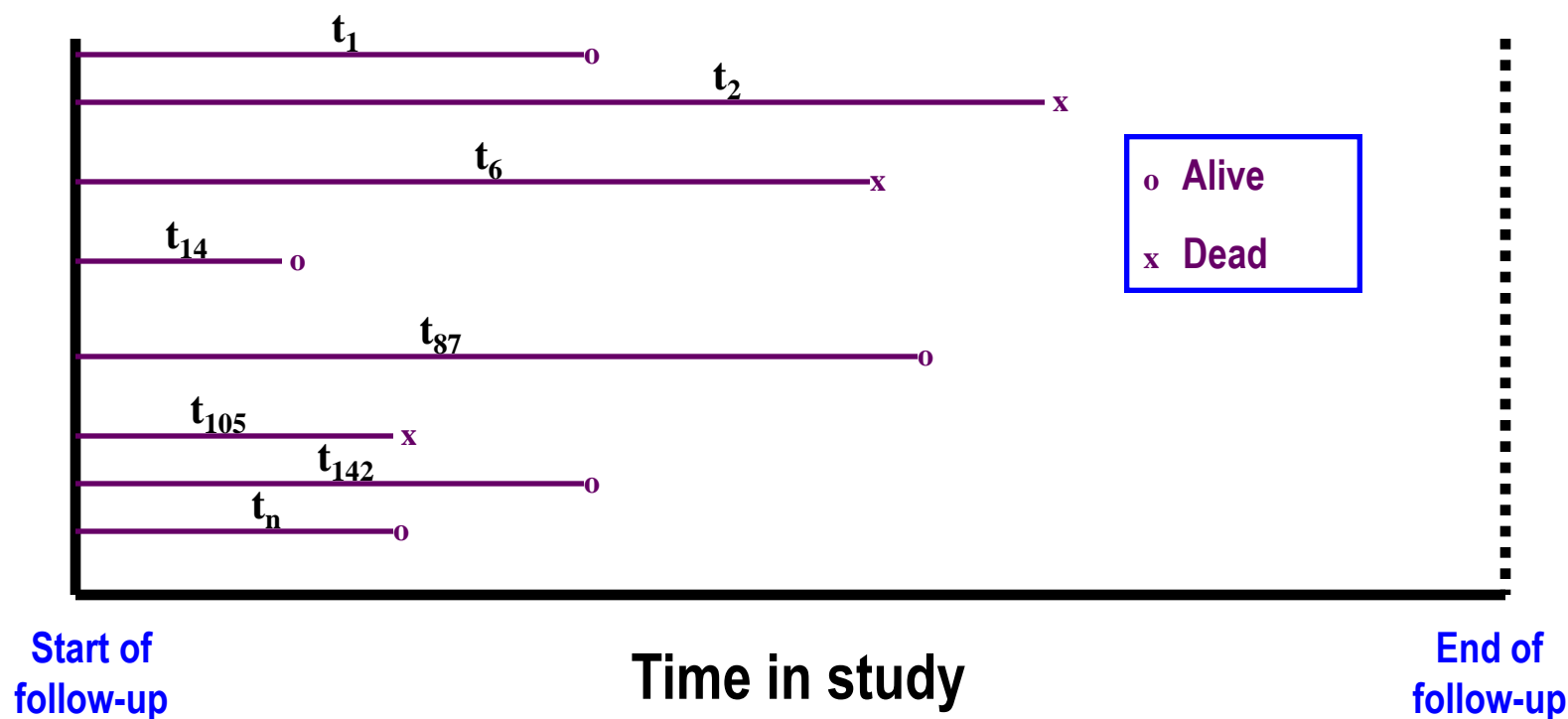
In the limiting (K-M) case, we assume we observe n individuals over time, and that they enroll in, and drop out of, the study at varying times.



Kaplan-Meier/Cox Analysis

The data

For analysis we focus on time since entry into study and so rearrange the data as follows:



Kaplan-Meier/Cox Analysis

The data

Thus our data are in the form of

- An observation time
- An indicator of whether this time ended in the event of interest (e.g., death), or whether it was “censored”
 - Censoring can occur either to early dropouts or because the participant was still “alive” at the end of the study
 - Be sure you know how to code for your stat package!
- Either a single variable indicating the groups to be compared (for K-M) or an arbitrary set of predictor variables (the Cox model)

Cox Regression Analysis

Overview

- A strength of the Kaplan-Meier analysis is that it is totally nonparametric. We have to make no assumptions about the underlying true distribution of failure times.
- On the other hand, we can only compare a finite number of groups, and we have no way to adjust our comparison of curves for potentially confounding variables.

Cox Regression Analysis

Overview

- While a number of fully parametric models for time to event data exist, perhaps the most common regression model that is in use for survival analysis is the **Cox Proportional Hazards Regression** model.
- The Cox model combines aspects of Kaplan-Meier analysis with parametric modelling, and thus provides a very flexible tool for modelling time to event data.

Cox Regression Analysis

The proportional hazards model

In the Cox PH model we construct a linear model for the “instantaneous” incidence rate, which is also called the instantaneous hazard function.

Recall that for the life table analysis we defined the hazard function as

$\mathbf{H}_i = \mathbf{H}(t_i)$ = Probability of dying during i^{th} interval given that you survived to the start of the i^{th} interval

The instantaneous hazard is just the limiting case of H_i as the interval (t_i, t_{i+1}) gets very, very small.

Cox Regression Analysis

The proportional hazards model

So let

$\lambda(t|X_1, X_2, \dots, X_k)$ = probability of “dying” on day t
given survival up to day t and
baseline covariates X_1, X_2, \dots, X_k

define the instantaneous hazard at time t .

The Cox proportional hazards model assumes that $l(t)$
can be written as

$$\lambda(t | X_1, X_2, \dots, X_k) = \lambda_0(t) e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}$$

$$\ln[\lambda(t)] = \ln[\lambda_0(t)] + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Let's break this equation down some to better understand it.

Cox Regression Analysis

The proportional hazards model

$$\lambda(t | X_1, X_2, \dots, X_k) = \lambda_0(t) e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}$$

Unspecified “baseline” hazard function estimated via Kaplan-Meier methods. Think of it as the intercept, or β_0 , term in our other regression models. It is considered to be a nuisance term that carries no information about the influence of the Xs on survival.

nonparametric portion

The regression model. The β s are the coefficients estimated by your statistics package. We use the term **proportional hazards** because the hazard functions are proportional for different values of the Xs.

parametric portion

Cox Regression Analysis

Interpretation of coefficients

$$\lambda(t | X_1, X_2, \dots, X_k) = \lambda_0(t) e^{\beta_1 \text{Age} + \beta_2 \text{Male}}$$

$$\begin{aligned} \text{RR (male vs female)} &= \frac{\cancel{\lambda_0(t)} e^{\beta_1 \text{Age} + \beta_2}}{\cancel{\lambda_0(t)} e^{\beta_1 \text{Age}}} \\ &= e^{\beta_2} \end{aligned}$$

$$\Rightarrow \beta_2 = \ln(\text{RR}_{\text{males vs females}})$$

Cox Regression Analysis

Generalizations

- The Cox model may be generalized to handle time-dependent variables.
 - the model conditions on the value of the covariate(s) at each failure time when estimating the β s.
 - your software package may not offer this option, and even if it does your options for modelling the time-dependency may be limited.
- We can get around the proportional hazard assumption to some extent by allowing the baseline hazard to vary arbitrarily for, say, smokers and nonsmokers.

Cox Regression Analysis

Assumptions

- Changes in any time-dependent covariates are not related to the outcome of interest (e.g., you don't quit smoking because your health is getting worse in a study of mortality)
- Censoring is not related to the outcome of interest (e.g., healthy people aren't more likely to leave the study early)

Cox Regression Example

Vollmer et al., NEJM, 1983

Background

- Very early days of transplantation, prior to federal funding of transplantation
- Evidence seemed to suggest huge benefits from transplantation
- Highly selected patient populations may bias results -- only healthiest patients were receiving transplants

The Question

- Does survival differ for patients on dialysis vs transplantation?

Cox Regression Example

Data features

Population

- Referral center for patients with ESRD
- Renal failure might be due to primary renal disease or secondary to diabetes or hypertension

Treatment Protocol

- Start on dialysis
- May get a transplant later
- Transplant may fail and patient go back on dialysis
- Transplants may come from either a living-related donor or a cadaveric donor

Cox Regression Example

Data features

Baseline characteristics of treatment groups				
	Dialysis Only	LRD Transplant	CAD Transplant	Sig
Age (yrs)	50	27	33	<.001
# Asso Diseases	2.1	1.4	1.4	<.001

Obviously we have the potential for serious confounding in favor of the transplant groups.

Cox Regression Example

Getting started

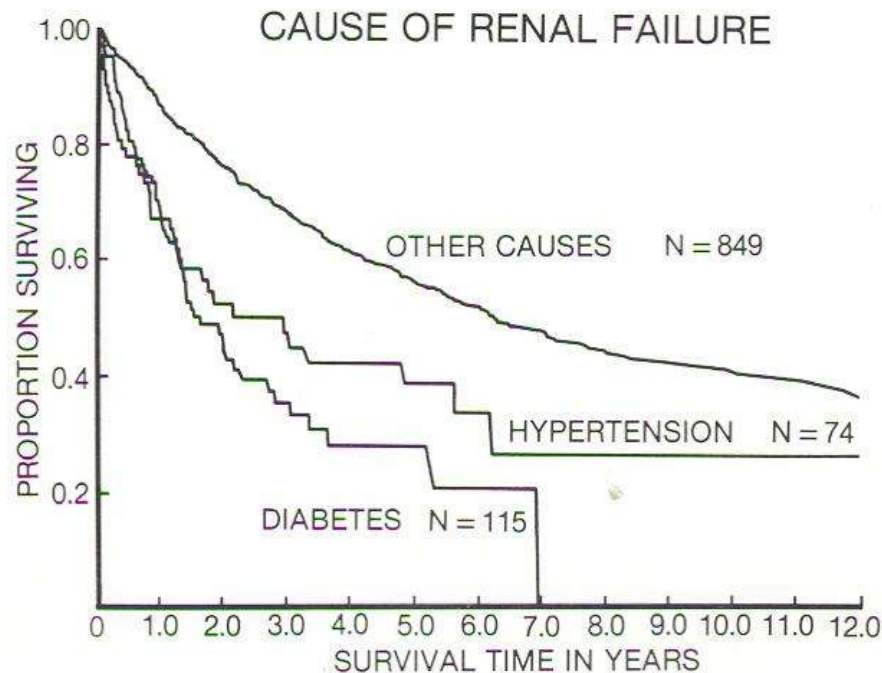


Figure 1. The Probability of Survival for 1038 Patients, According to the Cause of Renal Failure.

Estimates were obtained by using the method of Kaplan and Meier.

Checking assumptions:

- Patients with diabetes and hypertension had different disease process
- Hazards not proportional, & expected diff covariate effects than for those w/primary renal failure
- Therefore chose to conduct totally separate analyses for these individuals

Cox Regression Example

Telling a story

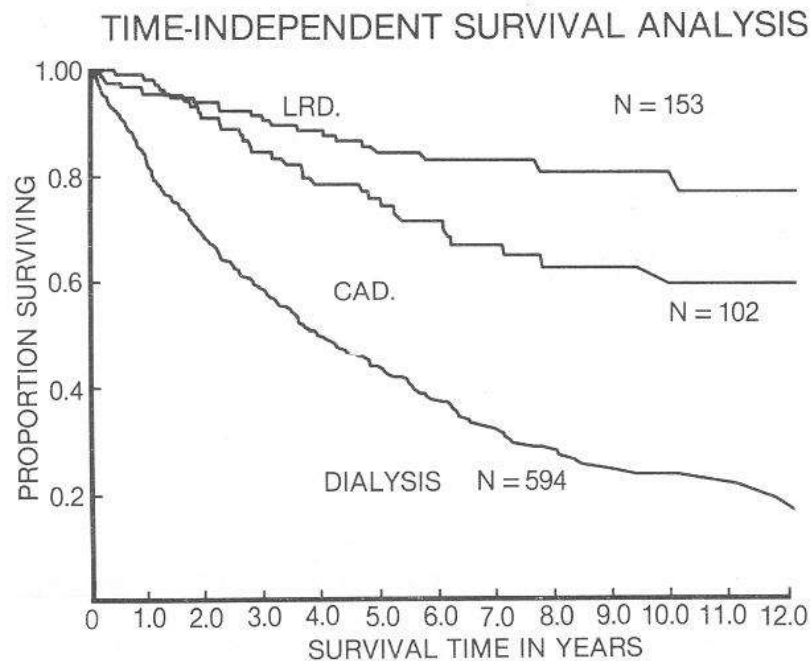


Figure 2. The Probability of Survival for 849 Patients with Primary Renal Disease Receiving Transplants from Living Related Donors (LRD) or Cadaveric Donors (CAD) or Being Treated with Dialysis Only.

Estimates were obtained by using the method of Kaplan and Meier.

Totally unadjusted analysis:

- Kaplan-Meier analysis with patients classified according to ever transplant status
- Observation time is time since enrollment into NKC, which credits transplant with pre-transplant survival
- This represents a very biased, but not atypical, analysis for the time

Cox Regression Example

Telling a story

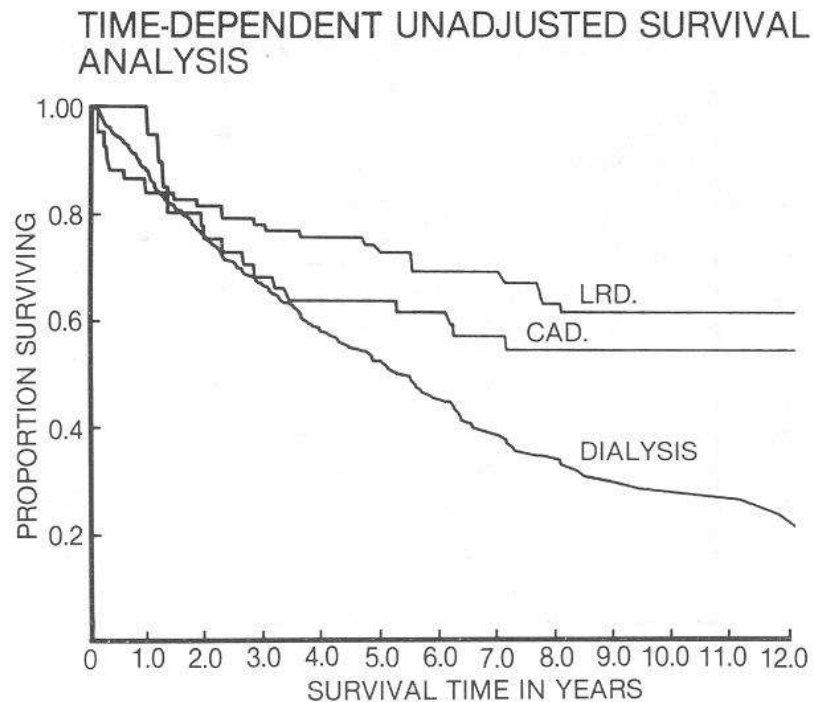


Figure 3. The Probability of Survival for 849 Patients with Primary Renal Disease, According to Transplant Classification.

Estimates were obtained by using the method of Kaplan and Meier, taking into account the time-dependent nature of the treatment status.

Time-dep, unadj. analysis:

- K-M analysis again
- For transplant pts, now use time since transplantation
- Since K-M, still no way to give dialysis credit for pre-transplant survival
- No covariate adjustment
- Starting to see curves come together

Cox Regression Example

Modeling covariates: checking assumptions

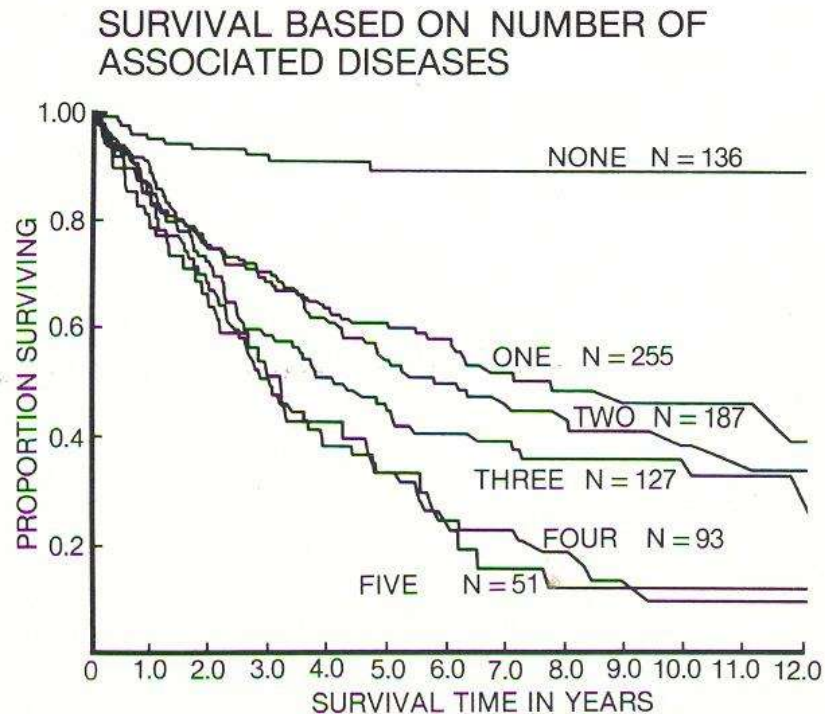


Figure 5. The Probability of Survival for 849 Patients with Primary Renal Disease, According to the Number of Associated Diseases.

Estimates were obtained by using the method of Kaplan and Meier.

How to model co-morbidities?

- PH assumption clearly not met
- Used separate model strata for “none” vs. “any” and used linear trend for latter, with 4-5 co-morbidities combined into a single group
- Again, intent was to provide best fit to this “nuisance” variable

Cox Regression Example

Telling a story

Table 2. Survival Analysis with Covariate Adjustment for Patients with Primary Renal Failure.

COVARIATES *	COEFFICIENT	STANDARD ERROR	SIGNIFICANCE †
Age at entry	0.3226	0.0467	P<0.0001
No. of associated diseases	0.1734	0.0510	P=0.0003
Year of entry	0.1297	0.1802	P=0.2358
LRD transplant	-0.5988	0.2697	P=0.0132
CAD transplant	0.0083	0.2750	P=0.5120

$RR_{LRD \text{ vs Dial}} = e^{-.60} = 0.55$

*Age at entry, number of associated diseases, and year of entry are standardized as follows: (age-44)/10, (number-2), and (year-74)/10. LRD denotes living related donor, and CAD cadaveric donor. A LRD (or CAD) transplant is assigned a value of 1 if the first transplant was from a living related (or cadaveric) donor and a value of 0 otherwise.

†All P values are one-sided.

Cox Regression Example

Telling a story

Table 2. Survival Analysis with Covariate Adjustment for Patients with Primary Renal Failure.

COVARIATES *	COEFFICIENT	STANDARD ERROR	SIGNIFICANCE †
Age at entry	0.3226	0.0467	P<0.0001
No. of associated diseases	0.1734	0.0510	P=0.0003
Year of entry	0.1297	0.1802	P=0.2358
LRD transplant	-0.5988	0.2697	P=0.0132
CAD transplant	0.0083	0.2750	P=0.5120

$$RR_{CAD \text{ vs. Dial}} = e^{0.01} = 1.01$$

*Age at entry, number of associated diseases, and year of entry are standardized as follows: (age-44)/10, (number-2), and (year-74)/10. LRD denotes living related donor, and CAD cadaveric donor. A LRD (or CAD) transplant is assigned a value of 1 if the first transplant was from a living related (or cadaveric) donor and a value of 0 otherwise.

†All P values are one-sided.

$$RR_{LRD \text{ vs. CAD}} = 0.55/1.01 = 0.55$$

Cox Regression Example

Telling a story

Table 3. Relative Risk According to Type of Analysis.

TREATMENTS COMPARED *	TYPE OF ANALYSIS		
	TIME-INDEPENDENT (UNADJUSTED)	TIME-DEPENDENT (UNADJUSTED)	TIME-DEPENDENT (ADJUSTED) †
LRD with dialysis	0.23 ‡	0.29 ‡	0.55 §
CAD with dialysis	0.53 ‡	0.58 §	1.01
LRD with CAD	0.43 ¶	0.50 §	0.54 §

*LRD denotes living related donor, and CAD cadaveric donor.

†Adjusted for age, number of associated diseases, and year of entry.

‡One-sided P value less than 0.001.

§One-sided P value between 0.01 and 0.05.

¶One-sided P value between 0.001 and 0.01.

Cox Regression Example

Summary

The Cox model is a powerful and flexible tool that can handle:

- Covariate information
- Time-dependent data
- Time-dependent RR effects
- Departures from PH assumption (e.g., strata)
- Individual and group data

Caution:

- As with any complex model, requires care in use and interpretation