Epidemiology for biostatisticians

January 2007.

Survival analysis, competing risks, regression models.

Per Kragh Andersen

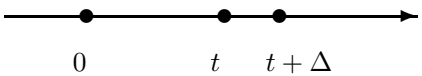
Survival analysis

Crucial part of epidemiology:

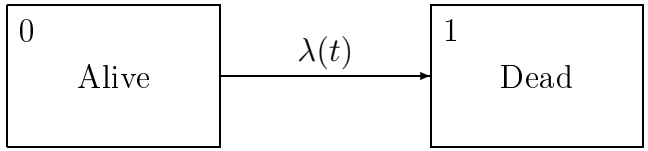
- Epidemiology: "study of distributions and determinants of disease in populations"
- Disease: (often) binary events occurring in time
- survival analysis = analysis of cohort studies

Survival data

- time from "zero" to event (death)
- right-censoring, delayed entry
- basic quantity:
  - death intensity
  - = mortality rate
  - = hazard function
  - = $\lambda(t) \approx \text{Prob}(\text{die before } t+\Delta \mid \text{alive } t)/\Delta$



Two-state model for survival data



$$\lambda(t) \approx \text{Prob}(\text{state } 1 \text{ time } t+\Delta \mid \text{state } 0 \text{ time } t)/\Delta$$

$$S(t) = \text{Prob}(\text{state } 0 \text{ time } t), \text{ survival function.}$$

Survival data

The classical relation:

risk (= "cumulative incidence")

$$= 1 - S(t)$$

$$= 1 - \exp(-\int_0^t \lambda(u)\mathrm{d}u)$$

=1 - exp(-"cumulative rate")

between risk and rate holds when there are no competing risks.

Approximation when the risk is low:

1 - exp(-"cumulative rate")

$\approx$ "cumulative rate".

This (partly) justifies the name.

Survival data

Frequently used regression models for $\lambda(t)$:

- The Cox regression model
- Poisson regression

With covariates z_i for individual i, both can be written in the form:

$$\lambda_i(t \mid z) = \lambda_0(t) \exp(\beta^\mathsf{T} z_i).$$

- Choice of t?
- Choice of $\lambda_0(t)$?
- z_i may be time-dependent, $z_i(t)$

What is time?

Which time origin should we use when defining the time variableś

Starting point	Time scale
Birth	Age
Any fixed date	Calendar time
First exposure	Time exposed
Entry into study	Time in study
Disease onset	Time since onset
Start of treatment	Time on treatment

Poisson regression.

According to a chosen time variable, suitable intervals are selected, $I_1, \ldots, I_k$, and the model has piecewise constant hazards:

$$\lambda_0(t) = \lambda_{0j} \text{ when } t \in I_j.$$

Likelihood contribution for individual i observed from entry time v_i to exit time t_i:

$$l_i = (\lambda_i(t_i))^{d_i} \exp(-\int_{v_i}^{t_i} \lambda_i(t)\mathrm{d}t),$$

where $d_i = 1$ if individual i fails at t_i, $d_i = 0$ if i is censored at t_i.

NB: independent censoring.

Poisson regression: categorical covariates.

When all z are categorical it is not hard to show that a sufficiency reduction is possible:

Create two "time intervals by covariates" tables with cells, say,

$$C_1, \ldots, C_m$$

One table with numbers of events in each cell: $D_1, \ldots, D_m$,

one table with numbers of person-years in each cell: $T_1, \ldots, T_m$.

Then the *total likelihood* is

$$L = \prod_r (\lambda_r)^{D_r} \exp(-\lambda_r T_r)$$

where λ_r is the rate in cell C_r, i.e. the product of the proper λ_{0j} and $\exp(\beta^\mathsf{T} z)$, i.e. the two tables are sufficient.

Poisson regression and the Poisson distribution?

L is proportional to the likelihood we would get by considering $D_1, \ldots, D_m$ as independent Poisson variates with parameters $\lambda_r T_r$.

This may be a useful fact when choosing computer programs to analyse the data.

However, this interpretation is ackward since the T_r are random and deriving the likelihood from the piecewise constant hazard model is much more obvious and satisfactory.

The Cox regression model.

In the Poisson model we work with prespecified time intervals and assume the baseline rate to be constant within these intervals.

In the Cox model the baseline rate $\lambda_0(t)$ is left completely arbitrary corresponding to letting the time intervals ("bands") shrink into "clicks" (Clayton & Hills).

Start with **Poisson model** with two explanatory variables A and B and with **time divided into bands**:

$$\underbrace{\text{Rate}}_{} = \underbrace{\text{Time}}_{} \times \underbrace{A \times B}_{}$$

or

$$\lambda_i^t = \lambda_{0t} \cdot \theta_i$$

$t \sim$ time band (j), $i \sim$ subject, $\theta_i = \exp(\beta^\mathsf{T} z_i)$.

Log likelihood contribution from i in band t:

$$d_i^t \log\left(\lambda_i^t\right) - y_i^t \lambda_i^t$$

where $d_i = I(i \text{ fails in band } t), y_i^t =$ total time spent by i in band t.

Total log likelihood:

$$\sum_{i,t} \left[d_i^t \log\left(\lambda_{0t} \cdot \theta_i\right) - y_i^t \lambda_{0t} \cdot \theta_i \right].$$

Profile log likelihood for θ-parameters:

$$\sum_{j,t} d_j^t \log\left(\frac{\theta_j}{\sum_i y_i^t \theta_i} \right).$$

Now: consider **clicks** instead of time bands.

Length h, redefine y_i^t as indicator, i.e. $y_i^t = 1$ if i is **at risk** at time $t, y_i^t = 0$ ($y_i^{t,old} = y_i^{t,new} h$). The profile log likelihood is then:

$$\sum_{j,t} d_j^t \log\left(\frac{\theta_j}{\sum_i y_i^{t,old}\theta_i}\right)$$

$$= \underbrace{\sum_{j,t} d_j^t \log\left(\frac{\theta_j}{\sum y_i^{t,new}\theta_i}\right)} - D\log(h).$$

Here, first term equals: (since $d_j^t = \binom{0}{1}$)

$$= \sum_{\text{Failures}} \log\left(\frac{\theta_{(\text{for case})}}{\sum_{\text{Risk Set}}\theta}\right)$$

$=$ **Cox's (partial) likelihood.**

Turns out to work fine in spite of the large number of nuisance parameters (λ_c^t).

Risk set (at time t) $=$ set of subjects who could have been the case (at time t)

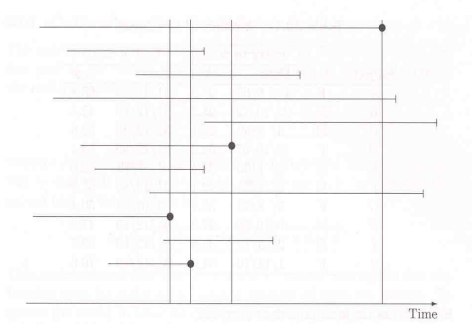


Fig. 30.1. Composition of risk sets.

Depends on how **TIME** is defined (age, calendar time,...)

Time-dependent explanatory variables

In the Cox regression model:

$$\begin{array}{ccccc} \text{Rate} & = & \text{Time} & \times & A \times B \\ \lambda_i^t & = & \lambda_{0t} & \cdot & \theta_i \end{array}$$

and the θ-parameters (effects of explanatory variables $A, B, \cdots$) are estimated from the profile log likelihood:

$$\sum_{\text{failures}} \log\left(\frac{\theta_{(\text{for case})}}{\sum_{\text{Risk set}}\theta}\right). \qquad (*)$$

Some times (often!) the interesting explanatory variables **depend on time** (exposure, treatment etc.), i.e.,

$$\theta \text{ depends on time}$$

One can still use (*) for estimation but computing time increases.

Reduction of computing time may be achieved by sampling from the risk set (Nested case-control study)

$$\sum_{\text{failures}} \log\left(\frac{\theta_{(\text{for case})}}{\sum_{\text{Case-control set}} \theta}\right)$$

(May save other resources then computing time)

Several time origins ("scales").

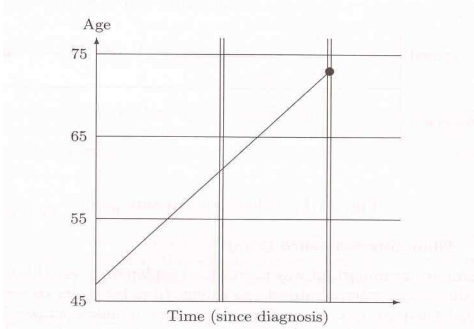


Fig. 31.2. Follow-up by age and time.

Possible Cox model

$$\begin{array}{rcccc} \log(\text{Rate}) & = & \underbrace{\text{Time}} & + & \underbrace{\text{Age} + A + B} \\ \log(\lambda_i^t) & = & \log(\lambda_{0t}) & + & \log(\theta_i) \end{array}$$

or vice versa! Which time scale should be considered "basic"? Or should we use a Poisson model?

- parametric/non-parametric
- effects of interest?

The bottom line is that Cox and Poisson models are very similar and give nearly identical results.

Repeated binary variates.

Another model which is strongly related to Cox and Poisson is the cloglog link model for repeated binary variates.

Return to the time intervals $I_1, \ldots, I_k$, and the data:

$$d_i^j = I(i \text{ fails in } I_j).$$

Then $\pi_{ij} = \text{Prob}(i \text{ fails in } I_j \mid i \text{ is alive at the end of } I_{j-1})$ may be related to covariates using, e.g. the cloglog link:

$$\log(-\log(1-\pi_{ij})) = \alpha_j + \beta^{\mathsf{T}} z_i.$$

This is in fact the Cox model for grouped survival data with $\alpha_j = \log(\int_{I_j} \lambda_0(u)\mathrm{d}u)$. The cloglog link may, in principle be replaced by other links for binary data like the logit. However, then β no longer has a log rate ratio interpretation.

Relative Risk, Odds Ratio, Rate ratio

When can they be interchanged?

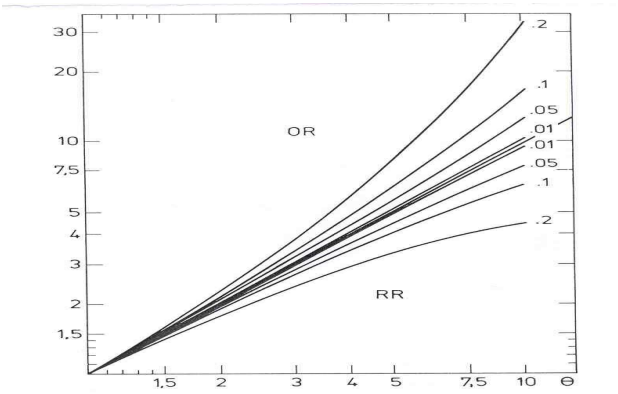
Fix time interval, $(0, T), T = 1$, and let:

- π_0 =Prob(diseased before T) (ought to be $\pi_0(T)$)
- λ_0 corresponding *rate*, i.e. $\pi_0 = 1 - e^{-\lambda_0 \cdot T}$
- fix θ =rate ratio= $\frac{\lambda_1}{\lambda_0}$,
- i.e., $\pi_1 = 1 - e^{-\theta\lambda_0 \cdot T}$
- and relative risk=$RR = \frac{\pi_1}{\pi_0}$,
- then odds=$\Omega_0 = \frac{\pi_0}{1-\pi_0}, \Omega_1 = \frac{\pi_1}{1-\pi_1}$
- and finally odds ratio= $OR = \frac{\Omega_1}{\Omega_0}$

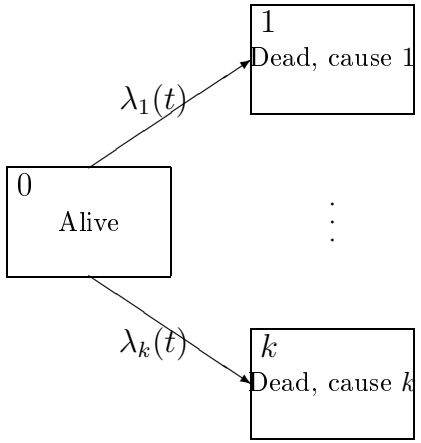
Relative Risk, Odds Ratio, Rate ratio: examples.

π_0	θ	π_1	RR	OR
0.1	5	0.41	4.1	$\frac{0.41/0.59}{0.1/0.9}$ =6.2
0.01	5	0.049	4.9	$\frac{0.049/0.951}{0.01/0.99}$ =5.1
0.1	1.5	0.146	1.46	$\frac{0.146/0.854}{0.1/0.9}$ =1.54

Relative Risk, Odds Ratio, Rate ratio: all combinations.



Generalisations: competing risks



Competing risks

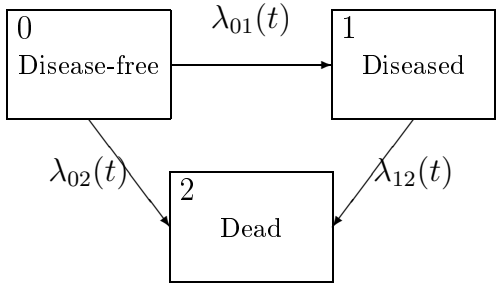
E.g., $k = 3$ causes:

- cancer
- cardio-vascular diseases
- other causes

Cause-specific intensities (e.g. cause 1)

$$\lambda_1(t) \approx \text{Prob}(\text{state } 1 \text{ time } t+\Delta \mid \text{state } 0 \text{ time } t)/\Delta$$

Generalisations: illness-death model



The illness-death or disability model (chronic disease).

Illness-death model

Transition intensities

Disease incidence rate:

$$\lambda_{01}(t) \approx \text{Prob}(\text{state } 1 \text{ time } t+\Delta \mid \text{state } 0 \text{ time } t)/\Delta$$

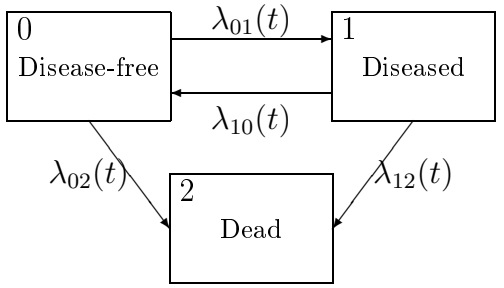
Mortality rate among disease-free (e.g. standard mortality)

$$\lambda_{02}(t) \approx \text{Prob}(\text{state } 2 \text{ time } t+\Delta \mid \text{state } 0 \text{ time } t)/\Delta$$

Mortality rate among diseased ("fatality rate")

$$\lambda_{12}(t) \approx \text{Prob}(\text{state } 2 \text{ time } t+\Delta \mid \text{state } 1 \text{ time } t)/\Delta$$

Generalisations: illness-death model



The illness-death or disability model (recurrent disease).

Illness-death model, recurrent disease

Transition intensities

As above, but also "cure rate":

$$\lambda_{10}(t) \approx \text{Prob}(\text{state } 0 \text{ time } t+\Delta \mid \text{state } 1 \text{ time } t)/\Delta$$

Bone marrow transplantations

States:

- Transplanted
- Graft versus host disease
- Relapse
- Death

etc. etc.

Common features

Models are given by intensities:

$$\lambda_{ij}(t) \approx \text{Prob}(\text{state } j \text{ time } t+\Delta \mid \text{state } i \text{ time } t)/\Delta$$

Intensities may be modelled using:

Cox regression, Poisson regression

In order to use Cox regression, a data file must be created *for each transition* including:

- entry time (some times 0)
- exit time
- exit status (relevant transition or not)
- covariates

Common features

In order to use Poisson regression, a data file must be created *for each transition* and for each combination of covariates, including:

- time spent in state
- number of transitions

e.g.

	Age 1	Age 2	Age 3
Exposed	T_{01}, D_{01}	T_{02}, D_{02}	T_{03}, D_{03}
Non-exposed	T_{11}, D_{11}	T_{12}, D_{12}	T_{13}, D_{13}

This enables us to analyse the intensities (rates) and estimate rate ratios.

Alternatively: individual records.

Probabilities (risks)?

In survival analysis:

$$S(t) = \exp(-\int_0^t \lambda(u)\mathrm{d}u).$$

In more general multi-state models:

- transition probabilities are more complex functions of the intensities
- no general computer programs exist

In the competing risks model:

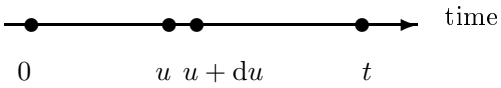
(2 causes of death:)

Survival probability: $P_{00}(t) = \text{Prob(alive time } t)$

$$= \exp(-\int_0^t (\lambda_1(u) + \lambda_2(u))\mathrm{d}u).$$

Cumulative incidence:

$$P_{01}(t) = \text{Prob(dead from cause 1 before time } t) = \int_0^t P_{00}(u)\lambda_1(u)\mathrm{d}u.$$



This means that:

- $P_{01}(t)$ (and similarly $P_{02}(t)$) may be estimated from $\lambda_1(t)$ *and* $\lambda_2(t)$.
- SE's may also be estimated.
- R functions in `cmprsk` can do it.
- SAS macros exist.

In the competing risks model:

What does

$$1 - \exp(-\int_0^t \lambda_1(u)\mathrm{d}u)$$

estimate?

$$\text{Prob(Dead from cause 1 before } t)$$

IF $\lambda_2(t) = 0$!

i.e., if the competing risk does not exist.

This hypothetical probability is rarely of interest. However, it is used frequently anyhow!

"Relapse survival curve" in clinical cancer studies.

Censoring in survival studies

Can this be treated as a competing risk?

When, in survival studies, we draw the Kaplan-Meier estimator only the death intensity is taken into account - NOT the censoring intensity.

This makes sense if the population without censoring makes sense.

Example: event = death due to cancer, what about censoring due to

- end of study
- emigration
- death due to traffic accidents
- death due to cardiovascular disease
- loss to follow-up

Competing risks example: bone marrow transplantation.

1715 leukemia patients with BMT:

- 537 ALL, 340 AML, 838 CML
- 1026 early stage, 410 intermediate stage, 279 advanced stage
- 1224 HLA-identical sibling, 383 HLA-matched unrelated donor, 108 HLA-mismatched unrelated donor

Analysis:

- Cox regression models for cause-specific hazards of "relapse" and "death in remission"
- Estimation of cumulative incidences

Cox regression models for cause-specific hazards

	Relapse		Death	
Covariate	$\hat{\beta}$	(SE)	$\hat{\beta}$	(SE)
HLA-id. sibling	0	-	0	-
HLA-matched donor	0.011	0.15	0.811	0.097
HLA-mismatched donor	-0.944	0.36	1.118	0.14
ALL	0	-	0	-
AML	-0.271	0.15	-0.195	0.14
CML	-0.721	0.16	0.291	0.117
Early stage	0	-	0	-
Intermed. stage	0.640	0.15	0.474	0.10
Advanced stage	1.848	0.15	0.781	0.13
Karnofsky> 90	-0.118	0.14	-0.504	0.11

Cumulative incidences

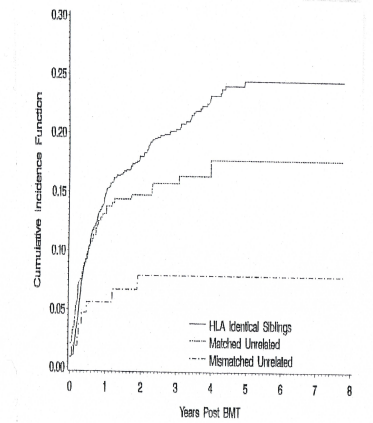


Figure 1. Cumulative incidence of relapse.

Cumulative incidences

Methods to study how the cumulative incidences depend on covariates are needed:

- plug-in
- Fine and Gray, *JASA* 1999.
- pseudo-values Klein and Andersen, *Biometrics* 2005
- Scheike and Zhang, Res. reports, 2005

References.

- Clayton, D.G., Hills, M. (1993) *Statistical Models in Epidemiology.* Oxford Univ. Press.
- Andersen, P.K., Keiding, N. (2002). Multi-state models for event history analysis. *Stat. Meth. Med. Res.* **11**, 91-115.
- Andersen, P.K., Abildstrom, S.Z., Rosthøj, S. (2002). Competing risks as a multi-state model. *Stat. Meth. Med. Res.* **11**, 203-215.