Intro to survival analysis

(from an epidemiologists perspective)

Gustaf Edgren, MD PhD
Associate professor of epidemiology
Karolinska Institutet

Outline

- What is survival? Why do we care?
- Regression models for survival and prognosis studies
- Time dependent covariates
- Standardized mortality ratios (SMR:s)
- Relative survival and cure rates

Risk vs. survival models?

- Thematically, standard risk models are based on the two-by-two table (i.e. binary outcomes)

Exposure	Dead: Yes	Dead: No	
+	13	42	55
-	2	53	55

- In survival (which, face it, is 0 in everyone) analysis, time to the binary event is also of interest
- Other conceptual models are therefore necessary

Expression of survival

- In clinical settings, patient survival is often expressed in terms of X% 5-year survival or that the average survival was Y years
- This is fine and dandy – and very intuitive – but such expressions really measure different things and are frequently setting-dependent:
  - When did they die?
  - Keep in mind age, sex, SES, etc.
  - Death from other causes?

Risk in survival

- the probability of an event (complication, death)
- time to the event

Time-to-event measurements

- Time from diagnosis of cancer to death due to the cancer
- Time from diagnosis of localized cancer to metastases
- Time from randomization to death in a myocardial infarction clinical trial
- Time from HIV infection to AIDS
- ……

Survival analysis – outcome

→ survival proportion (cumulative incidence)

→ event rate (hazard rate ~ incidence rate)

Risk estimation

- Directly from cumulative incidence (closed cohorts)
- Indirectly from incidence rates (open cohorts)

Indirect risk estimation

- From rate to risk:
- Short follow-up, low risk: CI = IR x time
- Long follow-up, changing IR ?

Simplistic survival analysis

- Divide time into short bands
- Calculate the period-specific death or survival proportions
- Multiply the probabilities

Life table

Period	# at risk	Died	Survived	Period-specific death	Period-specific survival	Cumulative survival	Cumulative death
1	100	8	92	8/100=**0.08**	92/100=**0.92**	**0.92**	**0.08**
2	92	7	85	7/92=**0.076**	85/92=**0.92**	0.92x0.92=**0.85**	1.0-0.85=**0.15**
3	85	5	80	5/85=**0.06**	80/85=**0.95**	0.92x0.92x0.95=**0.80**	1.0-0.80=**0.20**

Problems

- Competing risks
  - Death from other causes
- Loss to follow-up
  - Censoring

”Actuarial” life table

Period	# at start	Censored	# at risk	Died	Survived	Period-specific death	Period-specific survival	Cumulative survival	Cumulative death
1	100	5	97.5	8	89.5	8/97.5 =**0.082**	89.5/97.5 =**0.92**	**0.92**	1-0.92 =**0.08**
2	87	4	85	7	78	7/85 =**0.082**	78/85 =**0.92**	0.92x0.92 =**0.85**	1-0.85 =**0.15**
3	76	6	73	5	68	5/73 =**0.068**	68/73 =**0.93**	0.85x0.93 =**0.79**	1-0.79 =**0.21**

Presentation

- Cumulative survival at (until) a certain point in time:

  eg. 5-year survival, 10-year survival

- Graphical (survival function)

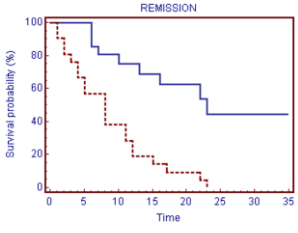
The Kaplan-Meier method *(product limit method)*

- A new survival proportion is calculated for each event that occurs
- → more precise

Kaplan-Meier method

Interval to event, months	# at risk	Survived	Period-specific survival	Cumulative survival
7	20	17	17/20=**0.85**	**0.85**
8	17	16	16/17=**0.94**	0.85x0.94=**0.80**
10	16	15	15/16=**0.94**	0.85x0.94x0.94=**0.75**

Kaplan-Meier curve - example



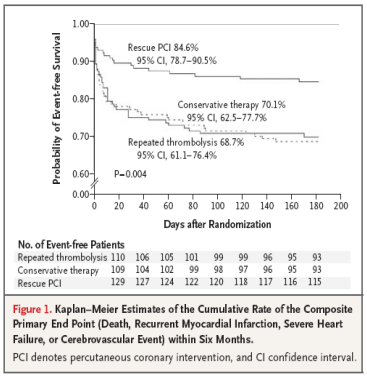
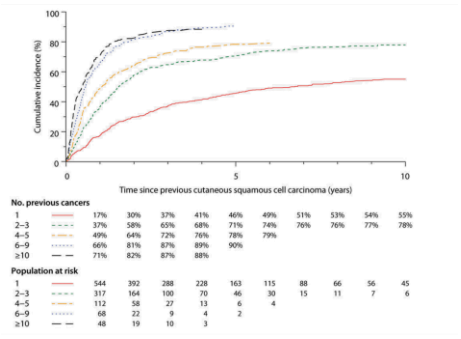


Figure 1. Kaplan–Meier Estimates of the Cumulative Rate of the Composite Primary End Point (Death, Recurrent Myocardial Infarction, Severe Heart Failure, or Cerebrovascular Event) within Six Months.

PCI denotes percutaneous coronary intervention, and CI confidence interval.

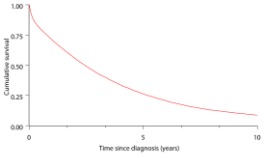


Comparing survival in two groups

- Easy way:
  - Compare survival at some point in time (should be a priori defined)
  - Compare median survival time
- Sophisticated way (accounting for the total survival experience)
  - the logrank test (hypothesis test)
  - the hazard ratio (point estimate of relative risk)

Modeling survival

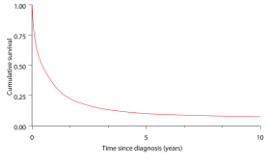
Myeloma



Median survival=2.44 years

10-year survival = 7.7%

AML



Median survival=0.44 years

10-year survival = 7.3%

Wish-list for survival regression

- A suitable regression model for survival modeling needs to be able to:
  - Handle censoring
  - Adjust for confounders
  - Handle time-dependence
  - Handle baseline hazard functions with very varying shapes
  - Ideally, should be able to produce estimates of both relative and absolute risk

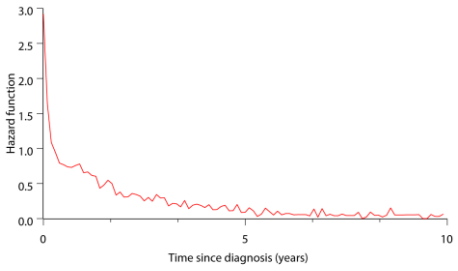
Regression models for survival

- The most commonly used model for survival regression is probably the Cox model – often referred to the proportional hazards model
- It was proposed by Sir David Cox, and is based on the modeling of (not survival), but rather of hazard functions
- Recognizing the varying shapes of hazard functions, Cox ingeniously introduced an assumption under which the baseline function could be ignored

Cox regression

- Regression model that is often used for survival (and other situations where time to an event is under study)
- Often referred to as ”proportional hazard model”
- Based on modeling of hazards (i.e. the instantaneous risk of death)
- Requires no assumptions about the shape of the survival curve…

Hazard function for AML



Cox regression basics

- The Cox model models all risks in relation to an baseline hazard (which is not estimated, but factored out)
- The Cox model assumes that all covariates act on this baseline risk multiplicatively
  - I.e. it is relative risk model – cannot calculate absolute risks
- The Cox model assumes that all covariates act equally (proportionally) on the baseline hazard over time
  - I.e. the effects of the covariates do not change with time
- The Cox model handles both right-censured and interval-type data
  - I.e. it is able to handle time-dependent covariates

Hazards

- Theoretically, the hazard function is a theoretical measure of the instantaneous mortality ratio
- It relates to survival as:

  $S(t) = 1 - \int h(t)$ i.e. Survival is 1 minus the sum of all hazard

  $h(t) = -dS(t)$ i.e. The hazard is the negative instantaneous change in survival

Cox regression assumptions

- Main assumptions for Cox regression:
  - Proportional hazards*
  - Non-informative censoring
  - Sufficient sample size**
  - *Not too many ties (relative)*

Cox model – pro and con

- Advantages
  - Quick to converge
  - Robust and forgiving model
  - Commonly used (!)
- Disadvantages
  - The proportional hazards assumption
  - No absolute risks can be computed
  - Does not handle truly aggregated data
  - Really requires access to reliable cause-specific death rates

More survival models…

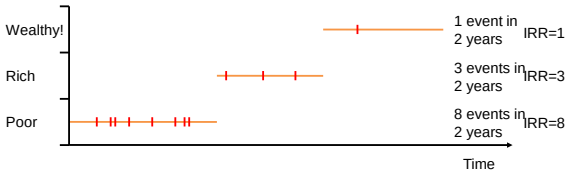
- Other methods for survival analysis:
  - Poisson regression, models the risk of a particular event per unit time (on a group level)
  - Accelerated failure time models, models the time to death and how your covariates modify this (on a multiplicative scale)*
  - You can also model the proportion that is alive afeter X years as a binary outcome using, for example, logistic regression

Time-dependent covariates

- In many situations, covariates change in a meaningful way
- Depending on the situation, this can and can not be accounted for in the analysis
- Typical examples of time-dependent covariates are:
  - Age
  - Calendar period
  - Cumulative exposure to (some environmental agent)
  - Income
  - Etc.

Time-dependent covariates, cont.

- The principal fashion for managing time-dependent covariates is to "split" follow-up time and events according to which exposure stratum they contributed:



Time-dependent covariates, cont.

- While conceptually simple:
  - Time-dependent covariates is difficult to execute,
  - May give you results that are difficult to interpret
  - Requires CAREFUL thought
- Some general rules:
  - Never know what you don't know (at that time)
  - Never condition on the future

Time-dependent covariates, cont.

- For time-dependent covariates with a fixed "origin," there are very standardized solutions:
  - Lexis (implemented in Stata, macro for SAS)
  - Fstpyrs macro for SAS (http://sourceforge.net/projects/pyrsstep)
- These can chop time up by age and calendar period relatively easily

What is prognosis?

- Prognosis is a prediction of the future outcome based on the current state
- Typically deals with binary outcomes (where time is of the essence)
- Prognosis is typically dealt with using various multi-variate models

Relative survival

- Background:
  - While SMR:s give us the opportunity to study relative mortality, it says little about survival
  - Likewise, Kaplan-Meier methods fail to consider the expected mortality (due to age, etc.)
- Thus, we want to be able to assess survival relative to the general population

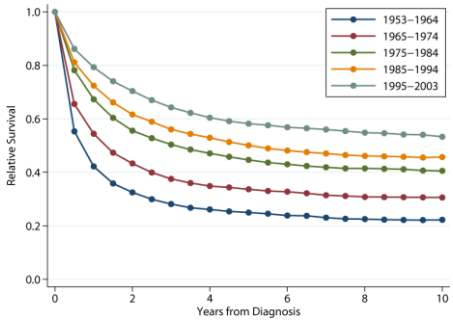
Relative survival, cont.

$$\text{Relative survival} = \frac{\text{Observed survival}}{\text{Expected survival}}$$

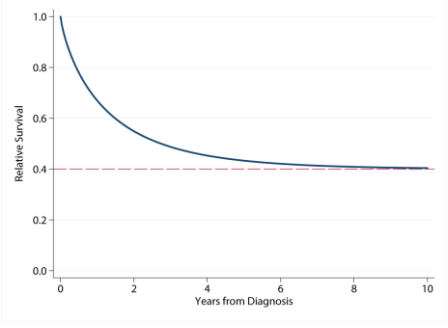
- Expected survival obtained from national population life tables stratified by age, sex and other covariates
- Estimate of mortality associated with a disease without requiring information on cause of death. Can also be expressed on hazard scale:

Excess mortality=observed mortality - expected mortality

Relative survival, colon cancer



Cure?



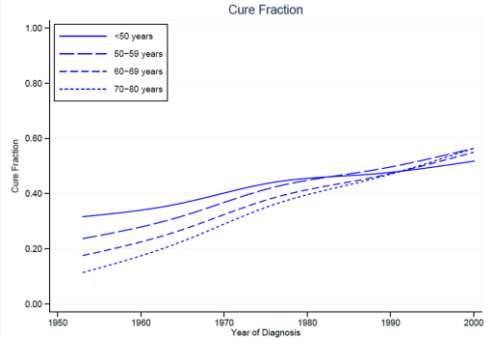
Statistical cure?

- From a relative survival model, it is possible to estimate the point at which the survival of your population returns to that of the background population
- This point is generally referred to as the cure point, and the relative survival at that point is referred to as cure fraction

Cure analysis

- Cure rate analysis is a relatively new field
- It assesses the occurrence of statistical cure
  - I.e. the population cure rate
- It does NOT say anything about individual cure
- With this comes some advantages:
  - The cure fraction is not affected by things like lead time
  - It allows the reliable comparison of calendar period effects with

Cure rate



Summary survival

- Survival analysis methods are required when the outcome of interest has a time dimension
- Result is presented as survival proportion or hazard rate
- Graphical presentation with KM curve
- Comparison of survival or hazard between exposure groups with logrank test or regression analysis (hazard ratio)

Summary survival (2)

- Several different methods exist for the multivariate regression of survivial: Cox, Poisson, etc
- All use slightly different approaches, but should give equivalent results (if executed correctly)
- Relative survival give us an alternative approach, where survival is compared with the expected in the ”background” population