Survival Analysis and Cox Regression

(EEBM Lecture)

INTRODUCTION

- Define the following terms: Time-to-event data, follow-up time, censoring, Kaplan-Meier curve, Cox proportional hazard regression
- Interpret a Kaplan-Meier curve: From a given curve, be able to read off survival estimates for a given time and be able to read off the median time to survival
- Explain the characteristics of study in which a log-rank test is appropriate
- Use the results of a log-rank test to draw conclusions about a scientific research question
- Interpret the value of a relative hazard (RH, sometimes also called hazard ratio, abbreviated HR)
- Given a description of a study, formulate the hypothesis of no relationship between variables as a test of a coefficient in a Cox proportional hazards regression model.
- Use the results of a hypothesis test about a coefficient in a Cox regression model to draw conclusions about a scientific research question
- Given a study and an estimated regression coefficient from a Cox model, provide an interpretation of the estimated coefficient in the context of a study
- State situations in which each type of regression analysis (linear, logistic, or Cox) is appropriate

KEY WORDS

blinding
COX proportional hazards regression
disease-free survival
effect size
follow-up
intention-to-treat analysis
Kaplan-Meier curve
log-rank test
median survival time
overall survival
power
randomization
survival analysis
Type I error
Type II error

OPTIONAL READING

Jakab, Shoenfeld, Balogh, et al "A medical nutriment has supportive value in the treatment of colorectal cancer." *British Journal of Cancer* 89: 465 – 469, 2003. Available here.

Dawson and Trapp, *Basic and Clinical Biostatistics*, 3rd Ed, Chapter 9.

Longitudinal, population-based study of racial/ethnic differences in colorectal cancer survival: impact of neighborhood socioeconomic status, treatment and comorbidity. SL Gomez, CD O'Malley, A Stroup, SJ Shema, WA Satariano. *BMC Cancer* 2007, 7:193 doi:10.1186/1471-2407-7-193 (http://www.biomedcentral.com/1471-2407/7/193). Skim the abstract and statistical analysis methods section and look over Tables 2 and 3.

NOTE: For those of you who would like more detail behind the concepts discussed in class I have included technical and computing notes in the Biostatistics course materials. These are not part of the assessed objectives.

I. WHAT IS SURVIVAL ANALYSIS?

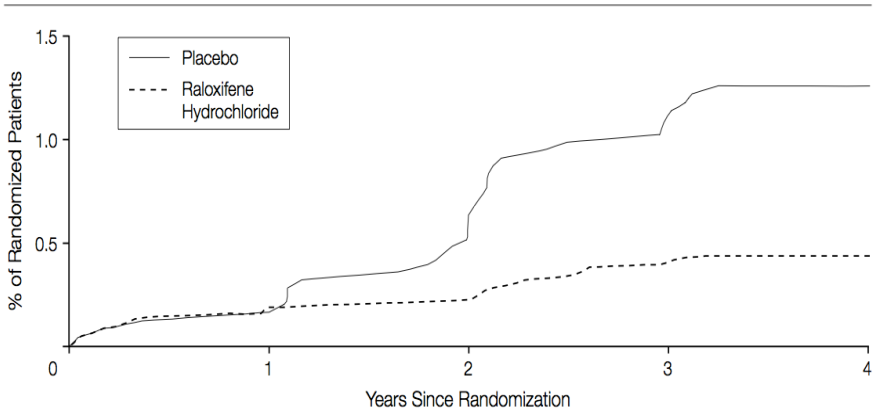
In evaluating risk factors or potential treatments for disease, it is often of interest to examine the time to an event. For example, if a treatment for breast cancer can delay death due to breast cancer, then that treatment may be clinically useful. Statistical techniques that utilize this type of analysis come under the general category of "survival analysis."

The outcome analyzed using **survival analysis** is often not death, but the time to occurrence of a non-fatal endpoint. For example, one might be interested in the time to recurrence of breast cancer among those who have been initially treated with a particular type of chemotherapy. Survival analysis techniques allow us to examine time to a fatal or non-fatal event across risk factor or treatment groups.

Example: In the MORE (Multiple Outcomes of Raloxifene Evaluation) study (JAMA, 1999; 281:2189-97) **survival curves** were used to compare the incidence of breast cancer in the placebo group to the treatment (Raloxifene) group. Raloxifene is in a class of drugs often called selective estrogen receptor modulators (SERMS) and includes tamoxifen, a drug commonly used to prevent secondary recurrence of breast cancer. The MORE study was primarily designed to examine the effect of Raloxifene on fractures, but a secondary analysis of breast cancer was performed. The data were analyzed using survival analysis techniques.

The cumulative probability of breast cancer in the placebo group at 3.5 years was estimated to be about 1%. This can be interpreted as follows: if a group of women similar to those in the study were all followed for 3.5 years, one would expect about 1% of them to develop breast cancer. In the Raloxifene group the rate was 0.5% at 3.5 years. This was a statistically significant difference.

What about time periods other than 3.5 years? At 1 year, the rates were approximately 0.2% for both groups. To get a full picture of the disease progression it is useful to plot the **cumulative incidence versus time**, a type of survival curve. (See Figure 1 which reports cumulative incidence for a raloxifene study).



The cumulative incidence of breast cancer among subjects in the placebo group and those in the combined raloxifene group are represented as a percentage of all patients randomized to either group. Statistical significance of the difference between the groups was tested by a log-rank test ($P<.001$).

Figure 1. *Survival curve depicting cumulative proportion of participants with breast cancer in the MORE trial (Raloxifine vs. placebo.)*

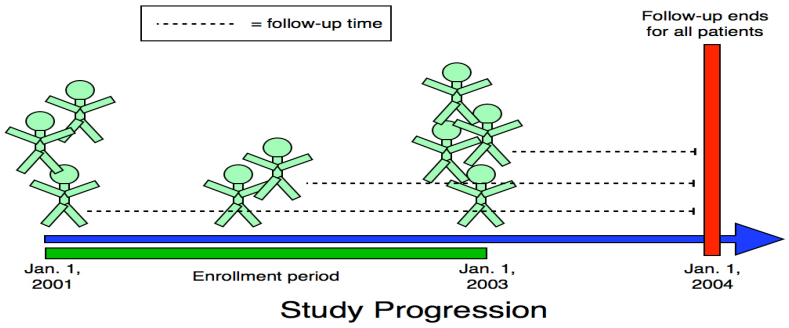


Figure 2. *Schematic showing different enrollment times and follow-up periods for a clinical trial*

II. WHY USE SURVIVAL ANALYSIS?

In the study of breast cancer, why not just count the number of women with breast cancer at each time point and calculate a relative risk or odds ratio? There are several reasons why more complicated methods are needed:

Censoring: Censoring refers to the fact that we often do not follow all the subjects until we see the event of interest. In such cases we cannot measure the time to the event. Another common reason for censored data in a study is that participants drop out of the study (follow-up is discontinued for reasons other than the end of the study). For example a participant might die of a cause unrelated to the disease in question (e.g., a motor vehicle accident) before the end of the study or simply decide to stop participating.

Different lengths of follow-up: In the context of a clinical trial, a very typical scenario is that patients are randomized into the trial over an extended period (e.g., 2 years) but follow-up for all patients ends at a fixed date. For example, suppose recruitment for a study occurred from January 1, 2001 to January 1, 2003 and all patients were followed until January 1, 2004. That would mean that those recruited early would have as much as 3 years of follow-up and those recruited later would only have 1 year of follow-up (Figure 2).

When subjects are censored we will not know their disease status at later time points and hence cannot calculate a relative risk orodds ratio. Survival analysis techniques allow us to use all the data up to dropout, accommodate different lengths of follow-up time, and allow for analysis later than the minimum follow-up time, even though this information is only available on a subset of the subjects.

Patterns of disease recurrence over time: Survival curves allow us to see the pattern of disease occurrence over time, which can often be interesting. For example, in the Women's Health Initiative (WHI), the survival curves for breast cancer show that up to about 4 years there was no discernible increase in breast cancer risk for those taking hormone replacement therapy but after 4 years, the divergence of the curves suggests increasing risk with longer-term use (see Figure 3). This pattern was an important factor in convincing the Data and Safety Monitoring Board that the study should be stopped early. The numbers at the bottom of the survival curve give the sample sizes followed for the given time and show that at the time the study was stopped, the vast majority of participants had only been followed for about 4 years but survival curve techniques allowed estimates for breast cancer incidence for as long as 7 years.

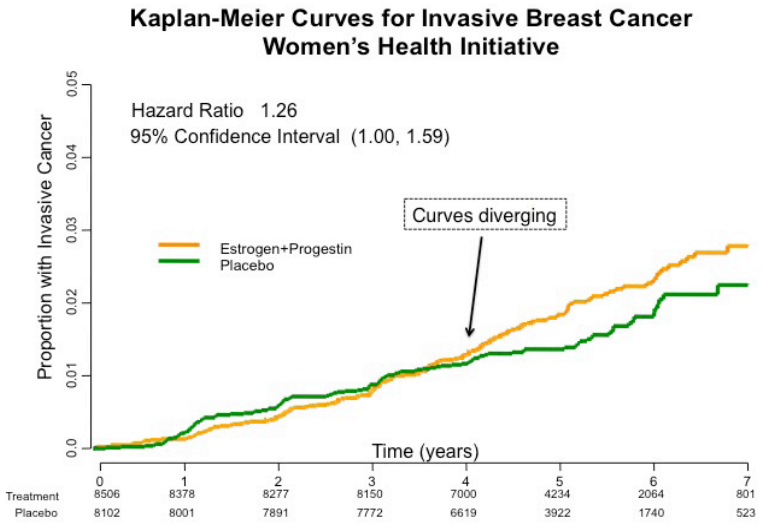


Figure 3. *Kaplan-Meier curves for invasive breast cancer in the Women's Health Initiative.*

III. HOW IS THE OUTCOME OF A SURVIVAL ANALYSIS DEPICTED GRAPHICALLY?

A **Kaplan-Meier curve** is a plot of the probability of survival that is free of an event (or equivalently, cumulative incidence) versus follow-up time.

Construction of a single survival curve: There are several techniques used to construct survival curves. The most common is the **Kaplan-Meier method** and the results are often referred to as Kaplan-Meier (KM) survival curves. KM curves are constructed by calculating the survival over intervals between points in time at which events occur, as a proportion of those patients surviving out of all those continuing in follow-up at the beginning of that interval. These estimates based on decreasing subgroups of the original participants are then combined to obtain the estimate of cumulative survival. Not surprisingly, the standard errors of the estimated cumulative incidence become larger as the numbers get smaller in the later years.

Median Survival Time is the time at which the curve intersects 0.5 (or 50%) on the Y-axis and is the time at which half the subjects have experienced the outcome. Importantly, the curve gives an estimate of the proportion at any given time of the follow-up that would have remained event-free if everyone had been followed at least to that time. Equivalently, as in the WHI example, we can plot the cumulative incidence of the disease versus time.

IV. HOW IS SURVIVAL TIME COMPARED BETWEEN TWO GROUPS OF PARTICIPANTS, AND WHAT MEASURE OF ASSOCIATION IS USED?

Comparison of two survival curves: It is often of interest to compare survival curves. In the examples above (from MORE and WHI), the curves were compared between the treated groups and placebo groups.

A formal hypothesis test can be performed to test the null hypothesis:

H_0: the cumulative incidence across all times is the same in the two populations vs.

the alternative hypothesis: H_A: the cumulative incidence is not the same
This hypothesis test is typically carried out using a test called the log rank test. The process is analogous to those described earlier for performing a two sample test or a chi-squared test: if the p-value is less than 0.05, the null hypothesis is rejected and astatistically significant result is declared, i.e., the cumulative incidence differs between groups.

Example (To be discussed further in lecture/lecture slides): Jakab, Shoenfeld, Balogh, et al.

H_0: No difference in survival between the nutriment and control populations
H_A: There is a difference in survival between the nutriment and control populations

Outcome: time disease-free
Test: log rank test
p-value = 0.018

Conclusion: try to formulate a conclusion before lecture! To be discussed further at lecture and in slides.

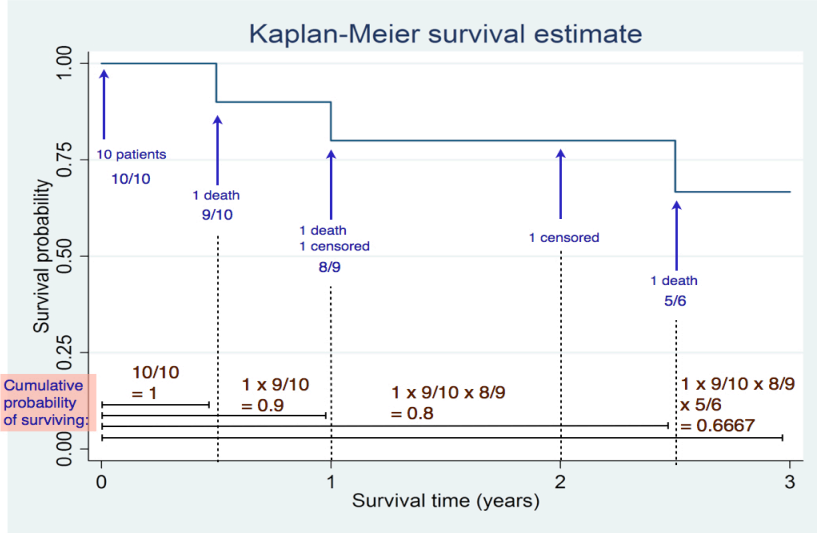
Technical and Computing Note: Here is a simplified illustration of the calculation of a KM curve. Suppose we are following 10 breast cancer survivors in a three year study. The first death occurs at 6 months, the second death occurs at one year, at which time a different individual drops out of the study, another drop out occurs at the two year mark, another death occurs at the 2.5 year mark and the remaining 5 participants are still surviving at the end of the study. Since no one died before 6 months we estimate the survival as 1 up to but not including 6 months. In the time interval from just before 6 months to just before 1 year we start by following10 individuals, one of whom dies in the time interval. So the estimated probability of surviving through the time interval from 6 months to 1 year is 0.9. To survive from the start of the study to just beyond 1 year a participant has to survive from 0 to 6 months and then from 6 months to 1 year with estimated probability 1 × 0.9 = 0.9. The next death occurs at 1 year and, by convention, dropouts are assumed to occur just after deaths. So the probability of surviving from 6 months to 1 year is 8/9 or 0.8888 and the estimated survival beyond 1 year is 0.9*0.8888 = 0.8. The next death occurs at 2.5 years. Just prior to that we are following 6 individuals (we lost two to death and two to dropout). So the estimated survival in the interval from 1 year to 2.5 years is 5/6 = 0.8333. So the probability of surviving from 0 to beyond 2.5 years is 1 × 0.9 × 0.888× 0.8333 = 0.6667. The Stata output below lists these calculations.

```
. sts list

      failure _d:  failure

  analysis time _t:  time
```

Time	Total	Fail	Beg. Lost	Net Survivor Function	Std. Error	[95% Conf.	Int.]
.5	10	1	0	0.9000	0.0949	0.4730	0.9853
1	9	1	1	0.8000	0.1265	0.4087	0.9459
2	7	0	1	0.8000	0.1265	0.4087	0.9459
2.5	6	1	0	0.6667	0.1610	0.2717	0.8815
3	5	0	5	0.6667	0.1610	0.2717	0.8815



In the comparison of two proportions, the measurement of association that is usually used (for cohort studies or randomized trials) is the relative risk. For comparing two survival curves, the most common measure of association is the r**elative hazard (RH)** (or **hazard ratio (HR)**), which can be thought of as an overall, averaged estimate of the relative risk across the entire follow-up period. Just like the relative risk, under the null hypothesis of no difference, RH=1. As noted in the branch Finding Significance, there is a correspondence between confidence intervals and hypothesis tests. In comparing two populations using the relative hazard, we could construct a confidence interval. If the confidence interval does not contain 1, then the association is statistically significant. From the WHI example, the RH was 1.26 with a 95% confidence interval from slightly above 1.00 (rounds off to 1.00) to 1.59. One method used to estimate the relative hazard (or hazard ratio) is Cox Proportional Hazards regression models, discussed next.

V. WHAT TYPE OF REGRESSION MODELS TIME-TO-EVENT OUTCOMES, AND HOW ARE THE RESULTS CHARACTERIZED?

Cox Proportional Hazards Models

The most commonly used regression model for survival analysis is the Cox proportional hazards model, which models time to an event (i.e., survival analysis) as a function of a set of predictors. This method can accommodate the common issues of censoringand differing lengths of follow-up time. The log rank test (which we studied in the context of comparing survival curves for two groups) is equivalent to a proportional hazards model in which there is only a single binary predictor (e.g., treatment vs. placebo). The coefficient from proportional hazards models can be used to calculate a relative hazard associated with each variable, much like the coefficient in the logistic model is used to calculate an odds ratio. A relative hazard is analogous to a relative risk or oddsratio. Formulae for converting the coefficient in a proportional hazards model to a relative hazard are identical to those used inlogistic regression for calculating an odds ratio, i.e., exponentiating the coefficient multiplied by the size of the change in the predictor gives the hazard ratio.

HR = $e^{(b)}$ for one unit change in predictor
HR = $e^{(10*b)}$ for 10 unit change in predictor

In a Cox proportional hazards regression, p-values for testing the null hypothesis that the coefficient is zero (or equivalently that the relative hazard is 1) are used as in logistic regression, as are 95% confidence intervals for the coefficients or relative hazards.

VI. HOW ARE STATISTICAL INFERENCES MADE FROM A COX PROPORTIONAL HAZARD MODEL?

Example (To be discussed further in lecture/lecture slides): Jakab, Shoenfeld, Balogh, et al, Table 3.

In this example, the first two variables, staging and MSC treatment are significantly associated with the outcome, while the others are not. We say they are "independently" associated or that "after adjustment for confounders," staging and MSC remains associated with the outcome. The column titled "Exp B" is actually the hazard ratio (HR). An increase of 1 stage doubles the risk of colorectal cancer. Receiving MSC treatment (y/n) decreases the risk of colorectal cancer by 67%.

Example: (from Optional Reading, Gomez,et al BMC Cancer 2007). Gomez et al studied racial-ethnic differences in survival among colorectal cancer patients. They used Kaplan-Meier curves to estimate the 5 year survival estimates in Table 2. They then used Cox proportional hazards regression in Table 3 to try to understand why the rates were different by considering a variety of predictors of survival.

Computing Note: Here is an explanation of how the log rank test works. For more detail than presented here, see
http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/12-survival-analysis

The log rank test works by comparing the observed number of events in each time interval (between events) with the "expected" number that would have occurred if the null hypothesis were true. Suppose there were no events between 6 months and one year in a study, but that, at one year, there were 4 events. Further assume that 70 people were being followed from 6 months to 1 year (i.e., 70 participants had not yet had an event and were not censored) with 50 in the first group and 20 in the second group. Then the "expected" number of events for the first group would be 4*(50/70) = 2.86 and the "expected" number for the second group would be 4*(20/70) = 1.14. The difference between the observed and "expected" numbers for one of the groups is summed over all event times and normalized with its standard error to form a formal hypothesis test. The calculations for the Cox regression model are quite complicated and invariably performed using a statistics package.

Below are analyses for data similar to that of Jakab, Shoenfeld, Balogh, et al on time remaining disease-progression-free analyzed using the log rank test and via Cox regression in Stata. The log rank test gives a p-value of 0.0181 and the Cox regression estimates the hazard ratio to be .4522148 (the hazard of disease progression in the nutriment group is estimated to be only 45% that of the control group) with a p-value of 0.021 and a 95% confidence interval from 0.2301036 to 0.8887223. Generally the log rank test and the Cox model with a single categorical predictor give very similar p-values.

Log rank test:

```
sts test msc

        failure _d:    death
  analysis time _t:    disfreetime

Log-rank test for equality of survivor functions

       | Events     Events
msc    | observed   expected
-------+--------------------
0      | 36         28.06
1      | 11         18.94
-------+--------------------
Total  | 47         47.00

         chi2 (1) =     5.58
          Pr>chi2 =     0.0181
```

Cox regression analysis:

```
stcox msc

           failure _d:  death
analysis time _t:       disfreetime

Cox regression - - no ties

No. of subjects =         168          Number of obs =        168
No. of failures =          47
Time at risk    = 17010.67049
                                       LR chi2 (1) =    5.99
Log likelihood  = -230.70224           Prob > chi2 = 0.0144

------------------------------------------------------------------------------
  _t   Haz. Ratio   Std. Err.     z    P>|z|   [95% Conf. Interval]
-------------------------------------------------------------------------------
 msc    .4522148    .1558849   -2.30  0.021    .2301036    .8887223
------------------------------------------------------------------------------
```

VII. WHICH TYPES OF REGRESSION MODELS SHOULD BE USED FOR PARTICULAR OUTCOME TYPES?

Summary of Types of Multiple Regression Models

Type of outcome	Type of regression analysis	Examples of outomes
Continuous	Linear	BP, cholesterol, stress score
Binary	Logistic	CHD (yes/no), fracture (yes/no)
Time to event	Cox Proportional Hazards	Time to breast cancer, time to death