Survival analysis

Wan Nor Arifin

Biostatistics and Research Methodology Unit, Universiti Sains Malaysia.

Email: wnarifin@usm.my

Section 1

Kaplan-Meier Method

Example

Hepatitis C Virus Infection Among Injection Drug Users

Survival Analysis of Time to Seroconversion

Holly Hagan,[*] Hanne Thiede,[†] and Don C. Des Jarlais[‡]

Background: Time to hepatitis C virus (HCV) seroconversion in initially seronegative injection drug users has not been directly measured, and public health planning would benefit from specifying the window of opportunity for prevention of infection, and factors that affect timing of infection.

Methods: Four hundred eighty-four HCV antibody-negative injection drug users in Seattle, Washington were followed a median of 2.1 years to observe seroconversion. We examined time to HCV seroconversion in relation to subject characteristics using the Kaplan-Meier method and Cox proportional hazards regression. A weighted-average time to HCV seroconversion was calculated among new injectors (injecting ≤2 years) using seroprevalence and seroincidence data.

Many HIV prevention programs for injection drug users are expanding their goals to include prevention of hepatitis C virus (HCV) infections.[1–6] However, although both HCV and HIV could be transmitted through parenteral exposure, and HIV prevention programs would typically teach or provide materials for safe injection, studies have shown a modest or no effect of HIV prevention programs on HCV transmission.[7–11] This could be because HCV is transmitted more efficiently than HIV,[12] and also because there could be more potential sources of HCV exposure in the injection setting (such as drug cookers or filtration cotton).[13–15] In addition, there could be more potential HCV

Figure 1: Kaplan-Meier Survival analysis

Introduction

1. A statistical method to analyze:
   - outcome: time to event (e.g. death, recurrence etc).
   - (comparison) predictors/independent variables: categorical variables.
2. It is concerned with survival probability at specific time points over a time interval (follow-up period), e.g. five year survival etc.

Interval survival

Basically, the *interval survival* at time t is as follows,

$$Interval\ survival, p_t = \frac{Survivors,\ n_t - Deaths,\ e_t}{Survivor,\ n_t}$$

Cumulative survival

The survival is usually represented by *cumulative survival* until time t,

$$\textit{Cumulative survival}, s_t = p_0 p_1 p_2 ... p_{t-1}$$

Censoring

In follow-up study over a period of time, not all subjects will experience event (e.g. not everyone die in 5 years). The subjects are called *censored* observations. In survival analysis, this censored observations are taken into account.

Censoring

Table 1: Acute myeloid leukemia data [1].

time	status
9	1
13	1
13	0
18	1
23	1
28	0
31	1
34	1
45	0
48	1

[1] Miller, R. (1997). Survival analysis. John Wiley & Sons.

Section 2

Cox proportional hazards regression method

Example

Hepatitis C Virus Infection and the Risk of Coronary Disease

Adeel A. Butt,[1,2,3] Wang Xiaoqiang,[2,3] Matthew Budoff,[5] David Leaf,[6,7] Lewis H. Kuller,[4] and Amy C. Justice[8,9]

[1]University of Pittsburgh School of Medicine, [2]Center for Health Equity Research and Promotion, [3]VA Pittsburgh Healthcare System, and [4]Department of Epidemiology, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania; [5]Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, [6]VA Greater Los Angeles Healthcare System, and [7]David Geffen School of Medicine at UCLA, Los Angeles, California; and [8]VA Connecticut Healthcare System, West Haven, and [9]Yale University School of Medicine and Public Health, New Haven, Connecticut

Background. The association between hepatitis C virus (HCV) infection and coronary artery disease (CAD) is controversial. We conducted this study to determine and quantify this association.

Methods. We used an established, national, observational cohort of all HCV-infected veterans receiving care at all Veterans Affairs facilities, the Electronically Retrieved Cohort of HCV Infected Veterans, to identify HCV-infected subjects and HCV-uninfected control subjects. We used the Cox proportional-hazards model to determine the risk of CAD among HCV-infected subjects and control subjects.

Results. We identified 82,083 HCV-infected and 89,582 HCV-uninfected subjects. HCV-infected subjects were less likely to have hypertension, hyperlipidemia, and diabetes but were more likely to abuse alcohol and drugs and to have renal failure and anemia. HCV-infected subjects had lower mean ($\pm$ standard deviation) total plasma cholesterol (175 $\pm$ 40.8 mg/dL vs. 198 $\pm$ 41.0 mg/dL), low-density lipoprotein cholesterol (102 $\pm$ 36.8 mg/dL vs. 119 $\pm$ 38.2 mg/dL), and triglyceride (144 $\pm$ 119 mg/dL vs. 179 $\pm$ 151 mg/dL) levels, compared with HCV-uninfected subjects ($P < .001$ for all comparisons). In multivariable analysis, HCV infection was associated with a higher risk of CAD (hazard ratio, 1.25; 95% confidence interval, 1.20–1.30). Traditional risk factors (age, hypertension, chronic obstructive pulmonary disease, diabetes, and hyperlipidemia) were associated with a higher risk of CAD in both groups, whereas minority race and female sex were associated with a lower risk of CAD.

Conclusions. HCV-infected persons are younger and have lower lipid levels and a lower prevalence of hypertension. Despite a favorable risk profile, HCV infection is associated with a higher risk of CAD after adjustment for traditional risk factors.

Figure 2: Sample paper

Introduction

1. A statistical method to model:
   - outcome: time to event (e.g. death, recurrence etc).
   - predictors/independent variables: numerical, categorical variables.
2. In contrast to KM approach, it is concerned with hazard of event.

Hazard

Basically, the (interval) *hazard* at time t is as follows,

$$Hazard,\ h_t = \frac{Deaths,\ e_t}{Survivors,\ n_t \times Interval,\ u_t}$$

where interval u_t is the time interval from present time t until the next event time $t + 1$.

Cumulative hazard function

cumulative hazard function, $H(t)$ is calculated from the *estimated cumulative survival function*, $S(t)$ as follows,

$$H(t) = -log_e S(t)$$

Cox Proportional Hazards Model

The formula for Cox proportional hazards (PH) model,

$$log_e\left(\frac{hazard\ at\ time,\ t}{baseline\ hazard\ at\ time,\ t}\right) =$$

$$log_e(hazard\ ratio,\ HR) = coefficients \times numerical\ predictors + coefficients \times categorical\ predictors$$

Cox Proportional Hazards Model

or in notational form,

$$log_e\left(\frac{h(t)}{h_0(t)}\right) =$$

$$log_e HR = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

where we have k predictors. Notice there is something missing in the equations above, which is the intercept (β_0). It is because the intercept $= 0$ at time $= 0$, i.e. nobody experiences the event at the start of the followup period, everyone is still alive!

Whenever the predictor is a categorical variable with more than two levels, remember to consider dummy (binary) variable(s).

Hazard ratio

Hazard ratio (HR) is the ratio of hazards of two levels. HR for a predictor is easily calculated from a Cox PH model,

$$HR_i = e^{\beta_i}$$