# Survival analysis

# Phân tích sống còn

## Overview

## Tổng quan

Survival analysis focuses on describing for a given individual or group of individuals, a defined point of event called **the failure** (occurrence of a disease, cure from a disease, death, relapse after response to treatment…) that occurs after a period of time called **failure time** (or **follow-up time** in cohort/population-based studies) during which individuals are observed. To determine the failure time, it is then necessary to define a time of origin (that can be the inclusion date, the date of diagnosis…).

*Phân tích sống còn* tập trung mô tả cho một cá thể hay một nhóm cá thể nhất định. Một điểm xác định của một sự kiện được gọi là ***failure*** (có nghĩa là sự kiện xảy ra như xảy ra bệnh, chữa khỏi bệnh, tử vong, tái phát sau khi đáp ứng với điều trị,.v.v.) mà xảy ra sau một khoảng thời gian được gọi là ***thời gian dẫn đến sự kiện*** (hay còn gọi là ***thời gian theo dõi*** trong nghiên cứu đoàn hệ/nghiên cứu dựa vào dân số) trong suốt thời gian các cá thể được quan sát. Để xác định thời gian dẫn đến sự kiện, chúng ta cần xác định thời điểm bắt đầu (có thể là ngày nhận vào, ngày chẩn đoán, .v.v.).

The target of inference for survival analysis is then the time between an origin and an event. In current medical research, it is widely used in clinical studies to assess the effect of a treatment for instance, or in cancer epidemiology to assess a large variety of cancer survival measures.

Mục tiêu suy luận đối với phân tích sống còn là khoảng thời gian giữa thời điểm bắt đầu và thời điểm sự kiện xảy ra. Trong nghiên cứu y học hiện nay, phân tích sống còn được sử dụng rộng rãi trong các nghiên cứu lâm sàng để đánh giá hiệu quả của một phương phương điều trị hoặc để đánh giá tình trạng sống còn của một số các biện pháp điều trị ung thư.

It is usually expressed through the **survival probability** which is the probability that the event of interest has not occurred by a duration t.

Nó thường được diễn đạt qua “***xác suất sống sót”***: là xác xuất mà sự hiện đang quan tâm ***không xảy ra*** trong khoảng thời gian t.

**Censoring**: Censoring occurs when at the end of follow-up, some of the individuals have not had the event of interest, and thus their true time to event is unknown. We will mostly focus on right censoring here but for more details on censoring and survival analysis in general, you can see references.

***Censoring***: Censoring xảy ra khi sự kiện đang quan tâm không xảy ra trong một số cá thể ở cuối quá trình theo dõi, và vì thế, thời gian thật dẫn đến sự kiện của những cá thể này là không biết. Trong chương này, chúng tôi tập trung chủ yếu vào sự kiện không xảy ra về phía bên phải. Để biết thêm chi tiết về các loại censoring và phân tích sống còn nói chung, xem thêm các tài liệu tham khảo.

## Preparation

### Load packages

### Tải các gói lệnh

To run survival analyses in R, one the most widely used package is the **survival** package. We first install it and then load it as well as the other packages that will be used in this section:

In this handbook we emphasize p\_load() from **pacman**, which installs the package if necessary and loads it for use. You can also load installed packages with library() from **base** R. See the page on [R basics](#r-basics) for more information on R packages.

This page explores survival analyses using the linelist used in most of the previous pages and on which we apply some changes to have a proper survival data.

**Survival** là gói lệnh được sử dụng rộng rãi nhất để phân tích sống còn bằng R. Đầu tiên, chúng ta cài đặt và sau đó tải gói lệnh này cũng như các gói lệnh khác sẽ được sử dụng trong phần này:

Trong cuốn sổ tay này, chúng tôi nhấn mạnh đến hàm p\_load() từ gói lệnh **pacman**, mà hàm này sẽ cài đặt gói lệnh khi cần thiết và tải nó để sử dụng. Có thể tải các gói lệnh đã cài đặt bằng hàm library() trong **base** R. Xem thêm thông tin về các gói lệnh của R trong chương [Cơ bản về R](#r-basics).

Chương này sẽ tìm hiểu phân tích sống còn bằng cách dùng bộ số liệu linelist đã được sử dụng trong hầu hết các chương trước và thay đổi một vài điểm để tạo ra bộ số liệu phù hợp cho phân tích sống còn.

### Import dataset

### Nhập bộ số liệu

We import the dataset of cases from a simulated Ebola epidemic. If you want to follow along, [click to download the “clean” linelist](https://github.com/epirhandbook/Epi_R_handbook/raw/master/data/case_linelists/linelist_cleaned.rds) (as .rds file). Import data with the import() function from the **rio** package (it handles many file types like .xlsx, .csv, .rds - see the [Import and export](#import-and-export) page for details).

Chúng tôi nhập bộ số liệu của các ca bệnh được mô phỏng từ một vụ dịch Ebola. Để tiện làm theo, [nhấp tải số liệu linelist](https://github.com/epirhandbook/Epi_R_handbook/raw/master/data/case_linelists/linelist_cleaned.rds) “đã được làm sạch” (dưới dạng .rds). Nhập số liệu này bằng hàm import() trong gói lệnh **rio** (hàm này chấp nhận nhiều loại tập tin như .xlsx, .rds, .csv – xem chi tiết trong chương [Nhập và xuất số liệu](#import-and-export)).

### Data management and transformation

### Quản lý và chuyển đổi số liệu

In short, survival data can be described as having the following three characteristics:

1. the dependent variable or response is the waiting time until the occurrence of a well-defined event,
2. observations are censored, in the sense that for some units the event of interest has not occurred at the time the data are analyzed, and
3. there are predictors or explanatory variables whose effect on the waiting time we wish to assess or control.

Thus, we will create different variables needed to respect that structure and run the survival analysis.

Nói ngắn gọn, số liệu cho phân tích sống còn có ba đặc điểm sau:

1. biến phụ thuộc hay đáp ứng là khoảng thời gian từ thời điểm bắt đầu đến thời điểm một sự kiện (được xác định rõ) xảy ra
2. quan sát censored là các quan sát mà sự kiện quan tâm không xảy ra tại thời điểm phân tích số liệu, và
3. các biến dự đoán hay giải thích có ảnh hưởng đến thời gian dẫn đến sự kiện mà chúng ta muốn đánh giá hoặc kiểm soát.

We define:

* a new data frame linelist\_surv for this analysis
* our event of interest as being “death” (hence our survival probability will be the probability of being alive after a certain time after the time of origin),
* the follow-up time (futime) as the time between the time of onset and the time of outcome in days,
* censored patients as those who recovered or for whom the final outcome is not known ie the event “death” was not observed (event=0).

Chúng tôi xác định:

* một bộ số liệu mới linelist\_surv cho phân tích này
* sự kiện quan tâm là “tử vong” (vì thế xác suất sống sót sẽ là xác suất sống trong một khoảng thời gian nhất định sau thời đểm bắt đầu),
* thời gian theo dõi (futime) là *số ngày* giữa thời điểm khởi phát bệnh và thời điểm có kết cục,
* bệnh nhân censored là những người đã hồi phục hoặc những người không biết kết cục tức là sự kiện “tử vong” không được nhận thấy (event=0).

**CAUTION:** Since in a real cohort study, the information on the time of origin and the end of the follow-up is known given individuals are observed, we will remove observations where the date of onset or the date of outcome is unknown. Also the cases where the date of onset is later than the date of outcome will be removed since they are considered as wrong.

THẬN TRỌNG: Vì trong một nghiên cứu đoàn hệ thật sự, thông tin về thời điểm bắt đầu và thời điểm kết thúc theo dõi của các ca thể được biết, chúng tôi sẽ loại bỏ các quan sát không có ngày bắt đầu và ngày có kết cục. Ngoài ra, các trường hợp có ngày khởi phát bệnh trễ hơn ngày có kết cục cũng bị loại bỏ vì các trường hợp này được xem là sai.

**TIP:** Given that filtering to greater than (>) or less than (<) a date can remove rows with missing values, applying the filter on the wrong dates will also remove the rows with missing dates.

MẸO: Bởi vì lọc đến giá trị lớn hơn (>) hoặc nhỏ hơn (<) một ngày có thể loại bỏ các hàng không có một/vài giá trị biến (missing values), nên áp dụng việc lọc vào các ngày sai cũng sẽ loại bỏ các hàng có ngày bị thiếu.

We then use case\_when() to create a column age\_cat\_small in which there are only 3 age categories.

Sau đó, chúng tôi sử dụng hàm case\_when() để tạo ra một cột age\_cat\_small mà chỉ có 3 giá trị của nhóm tuổi.

**TIP:** We can verify the new columns we have created by doing a summary on the futime and a cross-tabulation between event and outcome from which it was created. Besides this verification it is a good habit to communicate the median follow-up time when interpreting survival analysis results.

**MẸO:** Chúng ta có thể kiểm tra lại các cột mới đã được tạo ra bằng cách thực hiện tóm tắt đối với biến số futime và bảng chéo giữa biến số event và biến kết cục outcome. Bên cạnh việc kiểm tra này, đây là một thói quen tốt để biết được thời gian theo dõi trung vị khi giải thích kết quả của phân tích sống còn.

Now we cross-tabulate the new age\_cat\_small var and the old age\_cat col to ensure correct assignments

Bây giờ, chúng ta tạo bảng chéo giữa biến nhóm tuổi mới age\_cat\_small và biến nhóm tuổi cũ age\_cat để đảm bảo tính chính xác của việc chuyển đổi số liệu.

Now we review the 10 first observations of the linelist\_surv data looking at specific variables (including those newly created).

Bây giờ, chúng ta xem lại 10 quan sát đầu tiên của bộ số liệu bằng cách xem xét các biến cụ thể (bao gồm cả các biến mới được tạo ra).

We can also cross-tabulate the columns age\_cat\_small and gender to have more details on the distribution of this new column by gender. We use tabyl() and the adorn functions from **janitor** as described in the [Descriptive tables](#descriptive-tables) page.

Chúng ta cũng có thể tạo bảng chéo giữa cột biến age\_cat\_small và gender để biết thêm chi tiết về sự phân bố của biến số mới này theo giới tính. Chúng ta sử dụng hàm tabyl() và hàm adorn từ gói lệnh **janitor** như được mô tả trong chương [Bảng mô tả số liệu](#descriptive-tables).

## Basics of survival analysis

## Cơ bản về phân tích sống còn

### Building a surv-type object

### Tạo ra đối tượng surv

We will first use Surv() from **survival** to build a survival object from the follow-up time and event columns.

Đầu tiên, chúng ta dùng hàm Surv()từ gói lệnh **survival** để tạo ra một ‘đối tượng *Surv’* từ cột thời gian theo dõi và cột sự kiện.

The result of such a step is to produce an object of type Surv that condenses the time information and whether the event of interest (death) was observed. This object will ultimately be used in the right-hand side of subsequent model formulae (see [documentation](https://cran.r-project.org/web/packages/survival/vignettes/survival.pdf)).

Kết quả của bước này tạo ra một ‘đối tượng *Surv’* bao gồm thông tin thời gian và có hay không sự kiện quan tâm (tử vong) được nhận thấy. Đối tượng này sẽ được sử dụng sau này ở phía bên phải trong công thức của những mô hình tiếp theo ( xem [tài liệu](https://cran.r-project.org/web/packages/survival/vignettes/survival.pdf))

To review, here are the first 10 rows of the linelist\_surv data, viewing only some important columns.

Để xem lại số liệu, đây là 10 hàng đầu tiên của bộ số liệu linelist\_surv, chỉ xem các cột quan trọng.

And here are the first 10 elements of survobj. It prints as essentially a vector of follow-up time, with “+” to represent if an observation was right-censored. See how the numbers align above and below.

Và đây là 10 thành phần đầu tiên của ‘đối tượng *Surv’* survobj. Về bản chất, nó xuất ra đưới dạng một véc tơ của biến số thời gian theo dõi, dấu “+” là đại điện cho một quan sát censored ở phía bên phải. Xem cách các con số sắp xếp bên trên và bên dưới.

### Running initial analyses

### Thực hiện các phân tích ban đầu

We then start our analysis using the survfit() function to produce a survfit object, which fits the default calculations for **Kaplan Meier** (KM) estimates of the overall (marginal) survival curve, which are in fact a step function with jumps at observed event times. The final survfit object contains one or more survival curves and is created using the Surv object as a response variable in the model formula.

Sau đó, chúng ta bắt đầu phân tích bằng cách sử dụng hàm survfit() để tạo ra một đối tượng survfit phù hợp với các tính toán mặc định cho các ước tính **Kaplan Meier** (KM) của đường cong sống sót chung (cận biên), mà thực tế là một hàm bước với các bước nhảy tại thời điểm sự kiện được quan sát. Đối tượng survfit cuối cùng chứa đựng một hoặc nhiều đường cong sống sót và được tạo ra bằng cách sử dụng đối tượng Surv làm biến đáp ứng trong công thức của mô hình.

**NOTE:** The Kaplan-Meier estimate is a nonparametric maximum likelihood estimate (MLE) of the survival function. . (see resources for more information).

**LƯU Ý:** Ước tính Kaplan-Meier là một ước tính khả dĩ tối đa phi tham số của hàm sống còn (xem mục nguồn để biết thêm thông tin).

The summary of this survfit object will give what is called a life table. For each time step of the follow-up (time) where an event happened (in ascending order):

* the number of people who were at risk of developing the event (people who did not have the event yet nor were censored: n.risk)
* those who did develop the event (n.event)
* and from the above: the probability of not developing the event (probability of not dying, or of surviving past that specific time)
* finally, the standard error and the confidence interval for that probability are derived and displayed

Tóm tắt của đối tượng survfit này sẽ cho một bảng được gọi là bảng sống còn (life table). Đối với mỗi bước thời gian theo dõi (time) là nơi một sự kiện xảy ra (theo thứ tự tăng dần):

* số người có nguy cơ dẫn đến sự kiện (người chưa có sự kiện cũng như sự kiện chưa xảy ra: n.risk)
* những người có sự kiện (n.event)
* và từ những dữ kiện trên tính xác suất không đưa đến sự kiện (xác suất không tử vong, hoặc sống sót sau khoảng thời gian cụ thể đó)
* cuối cùng, sai số chuẩn và khoảng tin cậy cho xác suất đó được tính toán và trình bày

We fit the KM estimates using the formula where the previously Surv object “survobj” is the response variable. “~ 1” precises we run the model for the overall survival.

Chúng ta thực hiện các ước tính KM bằng cách sử dụng công thức với đối tượng Surv trước đó “survobj” làm biến đáp ứng. “~ 1” diễn đạt rẳng chúng ta đang thực hiện mô hình sống còn chung.

While using summary() we can add the option times and specify certain times at which we want to see the survival information

Khi sử dụng hàm summary(), chúng ta có thể thêm tùy chọn times và cụ thể các thời điểm nhất định mà chúng ta muốn xem các thông tin sống còn

We can also use the print() function. The print.rmean = TRUE argument is used to obtain the mean survival time and its standard error (se).

Chúng ta cũng có thể sử dụng hàm print(). Đối số print.rmean = TRUE được sử dụng để có được giá trị trung bình của thời gian sống sót và sai số chuẩn (se).

**NOTE:** The restricted mean survival time (RMST) is a specific survival measure more and more used in cancer survival analysis and which is often defined as the area under the survival curve, given we observe patients up to restricted time T (more details in Resources section).

**LƯU Ý:** Thời gian sống sót trung bình giới hạn là một đo lường sống còn cụ thể ngày càng được sử dụng trong phân tích sống còn của bệnh ung thư và thường được định nghĩa là khu vực dưới đường cong, khi chúng ta quan sát bệnh nhân cho đến thời gian giới hạn T (xem phần Nguồn để biết thêm chi tiết).

**TIP:** We can create the surv object directly in the survfit() function and save a line of code. This will then look like: linelistsurv\_quick <- survfit(Surv(futime, event) ~ 1, data=linelist\_surv).

**MẸO:** Chúng ta có thể tạo ra đối tượng surv trực tiếp trong hàm survfit() và tiết kiệm một dòng lệnh. Thực hiện điều này như sau: linelistsurv\_quick <- survfit(Surv(futime, event) ~ 1, data=linelist\_surv).

### Cumulative hazard

### Mối nguy tích lũy

Besides the summary() function, we can also use the str() function that gives more details on the structure of the survfit() object. It is a list of 16 elements.

Bên cạnh hàm summary(), chúng ta có thể sử dụng hàm str(), hàm này cho biết chi tiết hơn vể cấu trúc của các đối tượng trong hàm survfit(). Cấu trúc này là một danh sách của 16 thành phần.

Among these elements is an important one: cumhaz, which is a numeric vector. This could be plotted to allow show the **cumulative hazard**, with the **hazard** being the **instantaneous rate of event occurrence** (see references).

Một thành phần quan trọng trong số những thành phần này là cumhaz, một véc tơ số. Thành phần này có thể được vẽ để hiển thị **mối nguy tích lũy**, với **mối nguy** là **tỷ suất xảy ra sự kiện tức thời** (xem tài liệu tham khảo).

### Plotting Kaplan-Meir curves

### Vẽ đường cong Kaplan-Meir

Once the KM estimates are fitted, we can visualize the probability of being alive through a given time using the basic plot() function that draws the “Kaplan-Meier curve”. In other words, the curve below is a conventional illustration of the survival experience in the whole patient group.

Sau khi ước tính KM đã được hợp chỉnh, chúng ta có thể hình dung xác suất sống sót qua một thời gian nhất định bằng cách dùng hàm để vẽ “đường cong Kaplan-Meier”. Nói cách khác, đường cong bên dưới là một minh họa đường cong sống sót quy ước của toàn bộ nhóm bệnh nhân.

We can quickly verify the follow-up time min and max on the curve.

Chúng ta có thể nhanh chóng xác minh thời gian theo dõi tối thiểu và tối đa bên cạnh đường cong.

An easy way to interpret is to say that at time zero, all the participants are still alive and survival probability is then 100%. This probability decreases over time as patients die. The proportion of participants surviving past 60 days of follow-up is around 40%.

Một cách dễ dàng để giải thích là phát biểu rằng, tại thời điểm 0, tất cả người tham đều sống và xác suất sống sót khi đó là 100%. Xác suất ngày giảm dần theo thời gian khi có bệnh nhân tử vong. Tỷ lệ người tham gia sống sót sau 60 ngày là khoảng 40%.

The confidence interval of the KM survival estimates are also plotted by default and can be dismissed by adding the option conf.int = FALSE to the plot() command.

Khoảng tin cậy của các ước tính KM cũng được vẽ mặc định trên biểu đồ và có thể bị loại bỏ bằng cách thêm tùy chọn conf.int = FALSE vào trong lệnh plot().

Since the event of interest is “death”, drawing a curve describing the complements of the survival proportions will lead to drawing the cumulative mortality proportions. This can be done with lines(), which adds information to an existing plot.

Vì sự kiện quan tâm là “tử vong”, việc vẽ một đường cong mô tả phần bù tỷ lệ sống sót sẽ đưa đến việc vẽ tỷ lệ tử vong tích lũy. Điều này có thể được thực hiện với hàm lines(), bổ sung thông tin trên biểu đồ hiện có.

## Comparison of survival curves

## So sánh các đường cong sống sót

To compare the survival within different groups of our observed participants or patients, we might need to first look at their respective survival curves and then run tests to evaluate the difference between independent groups. This comparison can concern groups based on gender, age, treatment, comorbidity…

Để so sánh các đường cong sống sót của những nhóm người tham gia hoặc bệnh nhân khác nhau, đầu tiên chúng ta có thể cần xem xét các đường cong tương ứng của các nhóm và sau đó thực hiện các kiểm định để lượng giá sự khác biệt giữa các nhóm độc lập. So sánh này có thể liên quan đến các nhóm dựa vào giới tính, tuổi tác, điều trị, bệnh đi kèm,.v.v.

### Log rank test

### Kiểm định log rank

The log rank test is a popular test that compares the entire survival experience between two or more independent groups and can be thought of as a test of whether the survival curves are identical (overlapping) or not (null hypothesis of no difference in survival between the groups). The survdiff() function of the **survival package** allows running the log-rank test when we specify rho = 0 (which is the default). The test results gives a chi-square statistic along with a p-value since the log rank statistic is approximately distributed as a chi-square test statistic.

Kiểm định log rank là một kiểm định phổ biến để so sánh toàn bộ quá trình sống sót giữa hai hay nhiều nhóm *độc lập* và có thể xem xét các đường cong sống sót có tương đồng (chồng chéo) hay không (giả thuyết vô hiệu là không có sự khác biệt về sự sống sót giữa các nhóm). Hàm survdiff() trong gói lệnh **survival** cho phép thực hiện kiểm định log-rank khi chúng ta cụ thể rho = 0 (mặc định). Kết quả kiểm định cho ra một thống kê chi bình phương cùng với giá trị p vì log-rank phân phối gần giống như thống kê của kiểm định chi bình phương.

We first try to compare the survival curves by gender group. For this, we first try to visualize it (check whether the two survival curves are overlapping). A new survfit object will be created with a slightly different formula. Then the survdiff object will be created.

Đầu tiên, chúng ta thử so sánh các đường cong sống sót theo giới tính. Đối với điều này, chúng ta thử hình dung nó (kiểm tra xem hai đường cong sống sót có chồng lên nhau không). Một đối tượng survfit sẽ được tạo ra với một công thức hơi khác một chút. Sau đó, đối tượng survdiff sẽ được tạo ra.

By supplying ~ gender as the right side of the formula, we no longer plot the overall survival but instead by gender.

Bằng cách cung cấp ~ gender ở phía bên phải của công thức, chúng ta không còn vẽ biểu đồ đường cong sống sót chung mà thay vào đó là theo giới tính

Now we can plot the survival curves by gender. Have a look at the order of the strata levels in the gender column before defining your colors and legend.

Bây giờ, chúng ta có thể vẽ các đường cong sống sót theo giới tính. Hãy xem thứ tự giá trị biến của giới tính trước khi xác định màu sắc và chú giải.

And now we can compute the test of the difference between the survival curves using survdiff()

Và bây giờ, chúng ta có thể tính toán kiểm định sự khác biệt giữa các đường cong bằng cách sử dụng hàm survdiff()

We see that the survival curve for women and the one for men overlap and the log-rank test does not give evidence of a survival difference between women and men.

Chúng ta thấy rằng đường cong sống sót cho nữ và đường cong cho nam chồng lên nhau và kiểm định log-rank không đưa ra bằng chứng về sự khác biệt sống sót giữa nam và nữ.

Some other R packages allow illustrating survival curves for different groups and testing the difference all at once. Using the ggsurvplot() function from the **survminer** package, we can also include in our curve the printed risk tables for each group, as well the p-value from the log-rank test.

Một số gói lệnh trong R cho phép minh họa các đường cong sống sót cho các nhóm và kiểm định sự khác biệt cùng một lúc. Sử dụng hàm ggsurvplot() từ gói lệnh **survminer**, chúng ta cũng có thể bao gồm các bảng nguy cơ với các đường cong này cũng như giá trị p từ kiểm định log-rank.

**CAUTION:** **survminer** functions require that you specify the survival object and again specify the data used to fit the survival object. Remember to do this to avoid non-specific error messages.

**THẬN TRỌNG:** Các hàm từ gói lệnh **survminer** đòi hỏi chúng ta cụ thể đối tượng sống sót và cụ thể bộ số liệu để so sánh đối tượng sống sót. Hãy nhớ làm điều này để tránh thông báo lỗi do không cụ thể.

We may also want to test for differences in survival by the source of infection (source of contamination).

Chúng ta có thể cũng muốn kiểm định sự khác biệt về sống còn theo nguồn lây (nguồn ô nhiễm).

In this case, the Log rank test gives enough evidence of a difference in the survival probabilities at alpha= 0.005. The survival probabilities for patients that were infected at funerals are higher than the survival probabilities for patients that got infected in other places, suggesting a survival benefit.

Trong trường hợp này, kiểm định log-rank cho thấy có đủ bằng chứng về sự khác biệt trong xác suất sống sót với alpha= 0.005. Xác suất sống sót cho những bệnh nhân bị nhiễm tại các đám tang cao hơn xác suất sống sót ở những nơi khác mà gợi ý về lợi ích sống sót.

## Cox regression analysis

## Phân tích bằng hồi quy Cox

Cox proportional hazards regression is one of the most popular regression techniques for survival analysis. Other models can also be used since the Cox model requires important assumptions that need to be verified for an appropriate use such as the proportional hazards assumption: see references.

Hồi quy mối nguy theo tỷ lệ Cox (sau này sẽ gọi ngắn gọn là hồi quy Cox) là một trong những kỹ thuật hồi quy phổ biến nhất cho phân tích sống còn. Các mô hình khác cũng có thể được sử dụng, vì để sử dụng thích hợp mô hình Cox, *các giả định quan trọng* cần phải được xác minh: xem tài liệu tham khảo.

In a Cox proportional hazards regression model, the measure of effect is the **hazard rate** (HR), which is the risk of failure (or the risk of death in our example), given that the participant has survived up to a specific time. Usually, we are interested in comparing independent groups with respect to their hazards, and we use a hazard ratio, which is analogous to an odds ratio in the setting of multiple logistic regression analysis. The cox.ph() function from the **survival** package is used to fit the model. The function cox.zph() from **survival** package may be used to test the proportional hazards assumption for a Cox regression model fit.

Trong một mô hình hồi quy Cox, đo lường ảnh hưởng là **tỷ suất mối nguy/nguy cơ** (HR), là nguy cơ xảy ra sự kiện (hay là nguy cơ tử vong trong ví dụ này), của người tham gia sống sót đến một thời điểm cụ thể. Thông thường, chúng ta quan tâm đến việc so sánh giữa các nhóm độc lập về nguy cơ của nó, và sử dụng tỷ số nguy cơ mà tương tự như tỷ số chênh khi thực hiện hồi quy logistic đa biến. Hàm cox.ph() từ gói lệnh **survival** được sử dụng hợp chỉnh mô hình. Hàm cox.zph() từ gói lệnh **survival** có thể được sử dụng để kiểm tra tính phù hợp của giả định nguy cơ theo tỷ lệ với mô hình hồi quy Cox.

**NOTE:** A probability must lie in the range 0 to 1. However, the hazard represents the expected number of events per one unit of time.

* If the hazard ratio for a predictor is close to 1 then that predictor does not affect survival,
* if the HR is less than 1, then the predictor is protective (i.e., associated with improved survival),
* and if the HR is greater than 1, then the predictor is associated with increased risk (or decreased survival).

**LƯU Ý:** Xác suất phải nằm trong khoảng từ 0 đến 1. Tuy nhiên, nguy cơ đại điện cho số sự kiện dự đoán trên một đơn vị thời gian.

* Nếu tỷ số nguy cơ cho một yếu tố dự đoán gần bằng 1, thì yếu tố dự đoán đó không ảnh hưởng đến sự sống sót,
* Nếu HR nhỏ hơn 1, thì yếu tố dự đoán là yếu tố bảo vệ (tức là yếu tố liên quan đến cải thiện khả năng sống sót),
* Và nếu HR lớn hơn 1, thì yếu tố dự đoán kết hợp với tăng nguy cơ (hay là giảm khả năng sống sót).

### Fitting a Cox model

### Hợp chỉnh (fitting) một mô hình Cox

We can first fit a model to assess the effect of age and gender on the survival. By just printing the model, we have the information on:

* the estimated regression coefficients coef which quantifies the association between the predictors and the outcome,
* their exponential (for interpretability, exp(coef)) which produces the hazard ratio,
* their standard error se(coef),
* the z-score: how many standard errors is the estimated coefficient away from 0,
* and the p-value: the probability that the estimated coefficient could be 0.

Đầu tiên, chúng ta có thể hợp chỉnh một mô hình để đánh giá ảnh hưởng của tuổi và giới lên sự sống sót. Chỉ cần xuất mô hình, chúng ta có những thông tin sau:

* các ước số hồi quy coef để xác định mối liên hệ giữa các biến dự đoán và biến kết cục
* lũy thừa của các ước số (exp(coef)) để có được tỷ số nguy cơ,
* các sai số chuẩn se(coef),
* chỉ số z-score: bao nhiêu sai số chuẩn là hệ số ước tính khác biệt từ giá trị 0,
* và giá trị p: xác suất mà ước số có thể là 0.

The summary() function applied to the cox model object gives more information, such as the confidence interval of the estimated HR and the different test scores.

Áp dụng hàm summary() cho các đối tượng của mô hình Cox cho biết thêm thông tin như là khoảng tin cậy của HR và các chỉ số kiểm định khác.

The effect of the first covariate gender is presented in the first row. genderm (male) is printed, implying that the first strata level (“f”), i.e the female group, is the reference group for the gender. Thus the interpretation of the test parameter is that of men compared to women. The p-value indicates there was not enough evidence of an effect of the gender on the expected hazard or of an association between gender and all-cause mortality.

Kết quả của hiệp biến đầu tiên gender được trình bày ở hàng đầu tiên. genderm (nam) được in ra có ngụ ý rằng vị trí tầng đầu tiên (“f”), tức là nhóm nữ, là nhóm tham chiếu cho biến số giới tính. Vì thế, giải thích các thông số kiểm định là của nam so với nữ. Giá trị p chỉ ra rằng không có đủ bằng chứng về ảnh hưởng của giới tính lên mối nguy hay là không có đủ bằng chứng về mối liên quan giữa giới và tử vong (do tất cả các nguyên nhân).

The same lack of evidence is noted regarding age-group.

Cũng thiếu bằng chứng như vậy đối với biến số nhóm tuổi.

It was interesting to run the model and look at the results but a first look to verify whether the proportional hazards assumptions is respected could help saving time.

Thật rất thú vị để thực hiện mô hình và xem kết quả, nhưng quan sát đầu tiên để xác minh xem có thỏa các giả định về nguy cơ theo tỷ lệ hay không mà có thể giúp tiết kiệm thời gian.

**NOTE:** A second argument called method can be specified when computing the cox model, that determines how ties are handled. The default is “efron”, and the other options are “breslow” and “exact”.

**CHÚ Ý:** đối số thứ hai được gọi là method có thể được định rõ khi tính toán mô hình Cox để xác định cách ràng buộc được vận dụng. Phương pháp mặc định là “efron”, và các tùy chọn khác là “breslow” và “exact”.

In another model we add more risk factors such as the source of infection and the number of days between date of onset and admission. This time, we first verify the proportional hazards assumption before going forward.

Trong một mô hình khác, chúng tôi thêm nhiều yếu tố nguy cơ hơn như nguồn lây và số ngày giữa ngày khởi phát và ngày nhập viện. Điều trước tiên vào lúc này là xác minh các giả định nguy cơ theo tỷ lệ trước khi thực hiện các bước tiếp theo.

In this model, we have included a continuous predictor (days\_onset\_hosp). In this case we interpret the parameter estimates as the increase in the expected log of the relative hazard for each one unit increase in the predictor, holding other predictors constant. We first verify the proportional hazards assumption.

Trong mô hình này, chúng ta bao gồm một biến dự đoán liên tục (days\_onset\_hosp). Trong trường hợp này, chúng ta giải thích các ước tính của thông số như là sự gia tăng theo lôgarít kỳ vọng của nguy cơ tương đối cho mỗi mức tăng của một đơn vị trong biến dự đoán, bằng cách giữ các yếu tố dự đoán khác cố định. Đầu tiên chúng ta xác minh giả định nguy cơ theo tỷ lệ.

The graphical verification of this assumption may be performed with the function ggcoxzph() from the **survminer** package.

Việc xác minh bằng đồ thị của giả định này có thể được thực hiện bằng hàm ggcoxzph() từ gói lệnh **survminer.**

The model results indicate there is a negative association between onset to admission duration and all-cause mortality. The expected hazard is 0.9 times lower in a person who who is one day later admitted than another, holding gender constant. Or in a more straightforward explanation, a one unit increase in the duration of onset to admission is associated with a 10.7% (coef \*100) decrease in the risk of death.

Kết quả mô hình chỉ ra rằng có mối liên quan nghịch giữa khoảng thời gian từ khởi phát bệnh đến nhân viện và tử vong do tất cả các nguyên nhân. Nguy cơ dự đoán là bằng 0.9 lần ở một người nhập viện trễ hơn một ngày so với người khác, khi giữ biến giới tính cố định. Hay giải thích một cách dễ hiểu hơn, tăng một đơn vị thời gian từ lúc khởi phát đến nhập viện thì có liên quan đến giảm 10.7% (coef \*100) nguy cơ tử vong.

Results show also a positive association between the source of infection and the all-cause mortality. Which is to say there is an increased risk of death (1.21x) for patients that got a source of infection other than funerals.

Kết quả cũng cho thấy một mối liên quan thuận giữa nguồn lây và tử vong. Điều này có nghĩa là nguy cơ tử vong của bệnh nhân có nguồn lây bằng 1.21 lần so với bệnh nhân có nguồn lây là đám tang.

We can verify this relationship with a table:

Chúng ta có thể xác minh mối quan hệ này bằng một bảng:

We would need to consider and investigate why this association exists in the data. One possible explanation could be that patients who live long enough to be admitted later had less severe disease to begin with. Another perhaps more likely explanation is that since we used a simulated fake dataset, this pattern does not reflect reality!

Chúng ta cần phải xem xét và điều tra tại sao sự kết hợp này tồn tại trong số liệu. Một giải thích có thể chấp nhận được là bệnh nhân nhập viện trễ hơn vì có bệnh lúc đầu ít nghiêm trọng hơn. Một giải thích khác có lẽ dễ chấp nhận hơn là vì chúng ta sử dụng một bộ số liệu mô phỏng, mẫu này không phản ánh đúng thực tế!

### Forest plots

### Biểu đồ chứa nhiều hệ số

Sau đó, chúng ta có thể vẽ kết quả của mô hình Cox bằng cách sử dụng hàm ggforest()từ gói lệnh **survminer** để vẽ biểu đồ chứa nhiều hệ số.

We can then visualize the results of the cox model using the practical forest plots with the ggforest() function of the **survminer package**.

## Time-dependent covariates in survival models

## Các hiệp biến phụ thuộc vào thời gian trong mô hình sống còn

Some of the following sections have been adapted with permission from an excellent [introduction to survival analysis in R](https://www.emilyzabor.com/tutorials/survival_analysis_in_r_tutorial.html) by [Dr. Emily Zabor](https://www.emilyzabor.com/)

Một số phần sau được điều chỉnh từ phần [Giới thiệu phân tích sống còn bằng R](https://www.emilyzabor.com/tutorials/survival_analysis_in_r_tutorial.html) với sự cho phép của [TS. Emily Zabor](https://www.emilyzabor.com/)

In the last section we covered using Cox regression to examine associations between covariates of interest and survival outcomes.But these analyses rely on the covariate being measured at baseline, that is, before follow-up time for the event begins.

Trong phần trước, chúng ta đã đề cập đến việc sử dụng hồi quy Cox để kiểm tra mối liên quan giữa hiệp biến và biến kết cục sống còn. Nhưng những phân tích này dựa trên hiệp biến được đo lường ở thời điểm ban đầu, tức là trước thời gian theo dõi sự kiện bắt đầu.

What happens if you are interested in a covariate that is measured **after** follow-up time begins? Or, what if you have a covariate that can change over time?

Điều gì sẽ xảy ra nếu chúng ta quan tâm đến một hiệp biến được đo **sau khi** thời gian theo dõi bắt đầu? Hoặc điều gì sẽ xảy ra nếu chúng ta có một hiệp biến có thể thay đổi theo thời gian.

For example, maybe you are working with clinical data where you repeated measures of hospital laboratory values that can change over time. This is an example of a **Time Dependent Covariate**. In order to address this you need a special setup, but fortunately the cox model is very flexible and this type of data can also be modeled with tools from the **survival** package.

Ví dụ: có thể chúng ta đang làm việc với các số liệu lâm sàng mà chúng lặp lại đo lường các giá trị xét nghiệm mà có thể thay đổi theo thời gian. Đây là một ví dụ về hiệp biến phụ thuộc vào thời gian. Để nhấn mạnh vào vấn đề này, chúng ta cần một thiết lập đặc biệt, nhưng may là mô hình Cox rất linh động và loại số liệu này cũng có thể được mô hình hóa bằng các công cụ từ gói lệnh **survival**.

### Time-dependent covariate setup

### Thiết lập hiệp biến phụ thuộc vào thời gian

Analysis of time-dependent covariates in R requires setup of a special dataset. If interested, see the more detailed paper on this by the author of the **survival** package [Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model](https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf).

Phân tích các hiệp biến phụ thuộc vào thời gian trong R đòi hỏi thiết lập một bộ dữ liệu đặc biệt. Nếu quan tâm, hãy xem chi tiết hơn về vấn đề này được viết bởi tác giả của gói lệnh **survival** [Sử Dụng Hiệp Biến Phụ Thuộc Vào Thời Gian và Hệ Số Phụ Thuộc Vào Thời Gian trong Mô Hình Cox](https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf).

For this, we’ll use a new dataset from the SemiCompRisks package named BMT, which includes data on 137 bone marrow transplant patients. The variables we’ll focus on are:

* T1 - time (in days) to death or last follow-up
* delta1 - death indicator; 1-Dead, 0-Alive
* TA - time (in days) to acute graft-versus-host disease
* deltaA - acute graft-versus-host disease indicator;
  + 1 - Developed acute graft-versus-host disease
  + 0 - Never developed acute graft-versus-host disease

Đối với vấn đề này, chúng ta sử dụng một bộ số liệu mới từ gói lệnh SemiCompRisks có tên là BMT, bộ số liệu này bao gồm 137 bệnh nhân cấy ghép tủy xương. Các biến chúng ta sẽ tập trung vào là:

* T1 – thời gian (tính bằng ngày) đến khi tử vong hoặc đến lần theo dõi cuối cùng
* delta1 – chỉ số tử vong; 1-Tử vong, 0-Còn sống
* TA – thời gian (tính theo ngày) đến khi phát bệnh GVHD cấp tính (bệnh tế bào ghép tấn công vật chủ)
* deltaA – chỉ số của bệnh GVHD cấp tính;
  + 1 – Có bệnh GVHD cấp tính
  + 0 – Không có bệnh GVHD cấp tính

We’ll load this dataset from the **survival** package using the **base** R command data(), which can be used for loading data that is already included in a R package that is loaded. The data frame BMT will appear in your R environment.

Chúng ta tải bộ số liệu ngày từ gói lệnh **survival** bằng cách sử dụng lệnh data() trong **base** R, mà có thể được sử dụng để tải số liệu mà đã có trong gói lệnh R đã được tải. Khung số liệu BMT sẽ hiển thị trong môi trường R.

#### Add unique patient identifier

#### Thêm mã nhận dạng bệnh nhân

There is no unique ID column in the BMT data, which is needed to create the type of dataset we want. So we use the function rowid\_to\_column() from the **tidyverse** package **tibble** to create a new id column called my\_id (adds column at start of data frame with sequential row ids, starting at 1). We name the data frame bmt.

The dataset now looks like this:

Không có cột ID nhân dạng trong bộ số liệu BMT, cột này cần thiết để tạo ra bộ số liệu mà chúng ta muốn. Vì vậy, chúng ta sử dụng hàm rowid\_to\_column() từ gói lệnh **tibble** trong **tidyverse** để tạo một biến id mới gọi là my\_id (thêm cột ở vị trí đầu tiên của bộ số liệu với việc đánh số hàng tuần tự theo số nhận dạng bắt đầu từ số 1). Chúng ta đặt tên bộ số liệu này là bmt.

Bộ số liệu mới bây giờ trông như thế này:

#### Expand patient rows

#### Mở rộng hàng của các bệnh nhân

Next, we’ll use the tmerge() function with the event() and tdc() helper functions to create the restructured dataset. Our goal is to restructure the dataset to create a separate row for each patient for each time interval where they have a different value for deltaA. In this case, each patient can have at most two rows depending on whether they developed acute graft-versus-host disease during the data collection period. We’ll call our new indicator for the development of acute graft-versus-host disease agvhd.

* tmerge() creates a long dataset with multiple time intervals for the different covariate values for each patient
* event() creates the new event indicator to go with the newly-created time intervals
* tdc() creates the time-dependent covariate column, agvhd, to go with the newly created time intervals

Tiếp theo, chúng ta sử dụng hàm tmerge() với các hàm hỗ trợ event() và tdc() để tái cấu trúc bộ số liệu. Mục tiêu của chúng ta là tái cấu trúc bộ số liệu để tạo một hàng riêng biệt cho từng bệnh nhân trong mỗi khoảng thời gian mà họ có giá trị deltaA khác nhau. Trong bộ số liệu này, mỗi bệnh nhân có thể có nhiều nhất là hai hàng tùy thuộc vào việc họ có phát bệnh GVHD hay không trong giai đoạn thu thập số liệu. Chúng ta sẽ gọi chỉ số mới cho việc phát bệnh GVHD là agvhd.

* tmerge() tạo một bộ số liệu dài với nhiều khoảng thời gian cho các giá trị hiệp biến khác nhau cho từng bệnh nhân
* event() tạo chỉ số cho sự kiện mới để đi cùng với khoảng thời gian mới được tạo ra
* tdc() tạo cột hiệp biến phụ thuộc vào thời gian agvhd để đi cùng với các khoảng thời gian mới được tạo ra

To see what this does, let’s look at the data for the first 5 individual patients.

Để xem điều này thực hiện thế nào, hãy xem số liệu của 5 bệnh nhân đầu tiên.

The variables of interest in the original data looked like this:

Các biến quan tâm trong bộ số liệu gốc trông như sau:

The new dataset for these same patients looks like this:

Bộ số liệu mới cho cùng bệnh nhân trông như sau:

Now some of our patients have two rows in the dataset corresponding to intervals where they have a different value of our new variable, agvhd. For example, Patient 1 now has two rows with a agvhd value of zero from time 0 to time 67, and a value of 1 from time 67 to time 2081.

Bây giờ, một số bệnh nhân có hai hàng trong bộ số liệu tương ứng với khoảng thời gian mà họ có giá trị khác của biến mới, agvhd. Ví dụ như Bệnh nhân số 1 hiện có hai hàng có giá trị của biến agvhd bằng 0 từ thời điểm 0 đến 67 và giá trị bằng 1 từ thời điểm 67 đến 2081.

### Cox regression with time-dependent covariates

### Hồi quy Cox với hiệp biến phụ thuộc vào thời gian

Now that we’ve reshaped our data and added the new time-dependent aghvd variable, let’s fit a simple single variable cox regression model. We can use the same coxph() function as before, we just need to change our Surv() function to specify both the start and stop time for each interval using the time1 = and time2 = arguments.

Bây giờ, chugns ta đã định hình lại số liệu và thêm biến mới aghvd phụ thuộc vào thời gian, hãy hợp chỉnh mô hình cox đơn biến. Chúng ta có thể sử dụng cùng hàm coxph() như trước, chỉ cần thay đổi trong hàm Surv() để chỉ rõ thời gian bắt đầu và thời gian kết thúc cho mỗi khoảng thời gian bằng cách sử dụng các đối số cho time1 = và time2 =.

Again, we’ll visualize our cox model results using the ggforest() function from the **survminer package**.:

Một lần nữa, chúng ta hình dung kết quả mô hình cox bằng cách sử dụng hàm ggforest() từ gói lệnh **survminer**:

As you can see from the forest plot, confidence interval, and p-value, there does not appear to be a strong association between death and acute graft-versus-host disease in the context of our simple model.

Khoảng tin cậy và giá trị p từ biểu đồ cho thấy rằng không có mối liên hệ chặt chẽ giữa tử vong và bệnh GVHD cấp tính trong mô hình hồi quy đơn biến này.

## Resources

## Nguồn

[Survival Analysis Part I: Basic concepts and first analyses](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2394262/)

[Survival Analysis in R](https://www.emilyzabor.com/tutorials/survival_analysis_in_r_tutorial.html)

[Survival analysis in infectious disease research: Describing events in time](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2954271/)

[Chapter on advanced survival models Princeton](https://data.princeton.edu/wws509/notes/c7.pdf)

[Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model](https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf)

[Survival analysis cheatsheet R](https://publicifsv.sund.ku.dk/~ts/survival/survival-cheat.pdf)

[Survminer cheatsheet](https://paulvanderlaken.files.wordpress.com/2017/08/survminer_cheatsheet.pdf)

[Paper on different survival measures for cancer registry data with Rcode provided as supplementary materials](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6322561/)