BIOST/EPI 537
SURVIVAL DATA ANALYSIS IN EPIDEMIOLOGY

Winter 2025
Instructor: Ting Ye
GROUP PROJECT
Due by 11:59pm on Friday 14th March, 2025

OBJECTIVE

The objective of this project is to put into practice the skills learned in this course through the analysis of a medical dataset. In particular, you are expected to:

(a) use descriptive techniques and regression methods appropriate for survival data to answer the scientific questions outlined below,

(b) organize your findings in a scientific report that you would present to your scientific collaborators.

Please include you R code as a separate file, but don't put R code in the report.

DESCRIPTION

Bone marrow transplant is a standard treatment for acute leukemia. Recovery following bone marrow transplantation is a complex process. Prognosis for recovery may depend on risk factors known at the time of transplantation, such as patient and/or donor age and sex, stage of initial disease, and time from diagnosis to transplantation. The ultimate prognosis may change as the patient's post-transplantation experience unfolds, with the occurrence of events at random times during the recovery process, including the development of acute graft-versus-host disease (aGVHD) and the return of platelet counts to a normal level. Transplantation can be considered a failure when a patient's leukemia returns (relapse) or when he or she dies while in remission.

A multicenter study was conducted to evaluate whether patient and donor characteristics as well as unfolding clinical events are predictive of death in patients receiving allogeneic marrow transplantation. All patients enrolled were prepared with a radiation-free conditioning regimen consisting of a combination of oral Busulfan and intravenous cyclophosphamide. Enrollment took place at four different hospitals in the US and in Australia. The investigators were interested in whether the following characteristics measured at the time of transplantation (baseline) were associated with the risk of a negative outcome (death or relapse): patient age and sex, donor age and sex, patient and donor cytomegalovirus (CMV) immune status, the wait time from diagnosis to transplantation, disease group, French-American-British (FAB) classification based on standard morphological criteria, and prophylactic use of methotrexate to prevent aGVHD. A total of 137 patients were enrolled between March 1, 1984 and June 30, 1989. Patients were followed until death or end of the study.

You may consider this study to essentially be an exploratory investigation with the goal of identifying factors that could be used for informing patient prognosis. Previous studies have suggested that the patient's CMV status may be associated with morbidity and mortality. The importance of the donor's CMV status has been controversial.

A description of available variables is on the last page.

DIRECTIVES

Utilizing the available dataset, address the following scientific questions as appropriately as possible, and describe your findings in a scientific report.

1. Provide an estimate of disease-free survival time for patients enrolled in this study. What are the main characteristics of this summary?
2. How do patients in different disease groups or in different FAB classifications compare to each other with respect to other available baseline measurements?
3. Are any of the measured baseline variables associated with differences in disease-free survival?
4. It is generally thought that aGVHD has an anti-leukemic effect. Based on the available data, is occurrence of aGVHD after transplantation associated with improved disease-free survival? Is it associated with a decreased risk of relapse? In view of this, do you consider aGVHD as an important prognostic event?
5. Among the patients who develop aGVHD, are any of the measured baseline factors associated with differences in disease-free survival?
6. Is prophylactic use of methotrexate associated with an increased or decreased risk of developing aGVHD? Provide an estimate of the survival function of time from transplant until onset of aGVHD separately for patients either administered methotrexate or not. In doing so, consider the importance of accounting for relevant confounding factors.
7. Based on the available data, is recovery of normal platelet levels associated with improved disease-free survival? Is it associated with a decreased risk of relapse?

If you perform any test of hypothesis, you may use significance level 0.10 to compensate for the relatively small sample size of this dataset, provided you clearly state that you are doing so.

Report your findings in a scientific report targeted toward an audience of peer researchers who are not necessarily experts in statistical methods. Your report should be organized into sections, including:

1. Introduction

Motivate the importance of the problem at hand, introduce the scientific questions of interest and their relevance, and provide brief background on the available dataset.

2. Methods

Summarize the statistical methods that you have used to answer the scientific questions of interest. This should include a brief justification for the choice of estimand, model and/or method employed, and should clearly state the analysis choices for each of the questions.

3. Results and Discussion

Summarize the results of your analyses, explicitly answer the scientific questions of interest and discuss the implication of your findings as well as the limitations of your analysis. Throughout, provide a careful interpretation for key quantities that address the scientific questions.

4. Contributions

State the contributions of each member in your group.

Tables and Figures

Include these at the end of your report.

Use of AI tools (does not count against the page maximum)
A reminder that using these tools for this project is allowed as long as their use is cited. This citation should include a description of which AI tool was used and what it was used for. However, it is never permissible enter datasets into AI tools and it is never permissible to copy text output verbatim into your report. If you never used AI tools in your report, you should state this in this section.

R Code (does not count against the page maximum)
Provide a copy of the code for your analysis. Your code should have adequate documentation so that a TA or instructor could replicate the reported analysis.

Length
No more than 10 pages of single-spaced written material (not including tables, figures, and statement of use of AI tools) in Times New Roman font size 11 and margins of at least one inch on all sides. Please note that a longer report is not necessarily a better report - only use the space that you need and no more.

EVALUATION

Your report will be evaluated on the basis of:

- Scientific appropriateness and presentation (50%).
- Statistical appropriateness (50%).

The following scale will be used for each of the two evaluation components:

10: outstanding paper in every respect
9: excellent paper - no issues
8: good paper - some minor issues
7: fair paper - several minor issues or some major issues
6 or less: major issues

Each group will hand in a single project. Although groups will collaborate on the project, only one person per group should submit the final paper on behalf of the group.

The final project is worth 40% of the grade. Direct questions to the instructor or the TA. Use the course discussion board to ask questions if appropriate.

The variables represented in the dataset are as follows, in order of columns:

- `id`: patient ID number
- `ts`: time until death or end of study (in days)
- `deltas`: death indicator 1 = dead 0 = alive at last follow-up time (td)
- `tdfs`: time until relapse, death or end of study (in days)
- `deltar`: relapse indicator 1 = relapse 0 = disease-free
- `deltadfs`: disease-free survival indicator 1 = dead or relapsed 0 = alive and disease-free
- `ta`: time until acute graft-versus-host disease (aGVHD) (in days)
- `deltaa`: acute graft-versus-host disease (aGVHD) indicator 1 = acute graft-versus-host disease at time ta 0 = never experienced acute graft-versus-host Disease
- `tp`: time until recovery of normal platelet levels (in days)
- `deltap`: recovery of normal platelet levels indicator 1 = recovered 0 = not recovered
- `disgroup`: 1 = ALL ALL refers to "acute lymphoblastic leukemia". 2 = AML Low Risk AML refers to "acute myelocytic leukemia". 3 = AML High Risk
- `age`: patient age (in years)
- `male`: indicator of patient being male 1 = patient is male 0 = patient is female
- `cmv`: patient CMV status 1 = patient is CMV positive 0 = patient is CMV negative
- `donorage`: age of donor (in years)
- `donormale`: indicator of donor being male 1 = donor is male 0 = donor if female
- `donorcmv`: donor CMV status 1 = donor is CMV positive 0 = donor is CMV negative
- `waittime`: waiting time until transplant (in days)
- `fab`: disease subtype 1 = FAB grade 4 or 5 and AML 0 = otherwise
- `hospital`: recruitment center 1 = The Ohio State University (Columbus, OH) 2 = Alfred (Melbourne, Australia) 3 = St. Vincent (Sydney, Australia) 4 = Hahnemann (Philadelphia, PA)
- `mtx`: prophylactic use of methotrexate 1 = yes 0 = no

NOTE: The FAB (French-American-British) classification is a subtyping of leukemia.

From www.cancer.org:
Several years ago, an international conference of prominent hematologists and oncologists specializing in leukemia treatment and pathologists specializing in laboratory tests for blood disease diagnosis was held to decide upon the best system of classification of acute leukemias. This group of French, American and British doctors decided that acute leukemias should be divided into eight subtypes of AML and three subtypes of ALL.
Here, we are using a simplified disease group variable and a dichotomous FAB variable.