Untitled

Jorge Portugal

12/1/2021

```
## Warning in eval(substitute(list(...)), `_data`, parent.frame()): NAs introduced
## by coercion
```

Analysis

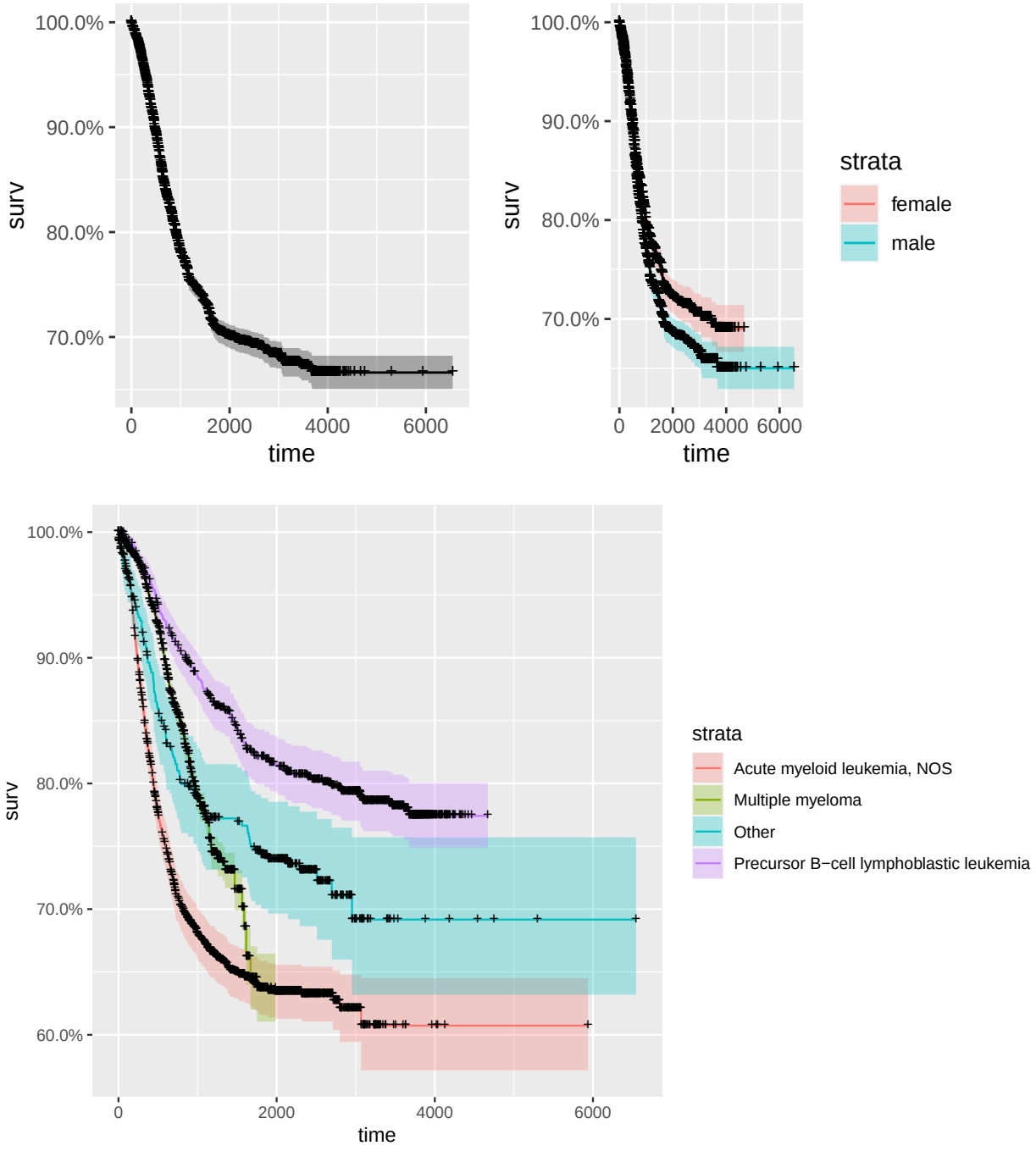
(Exploratory) Survival Analysis

For the first part of this project, we decided to do some exploratory survival analysis. However, due to the nature of how the data was collected, as well as how vastly incomplete the data was, some assumptions had to have been made. After looking at the dataset dictionary, the definition of the variable Vital Status is defined as "the survival state of the person registered on the protocol". That is the only clear definition given out for that variable. As a result, we make the assumption that even though a patient's info is censored, we consider them dead.

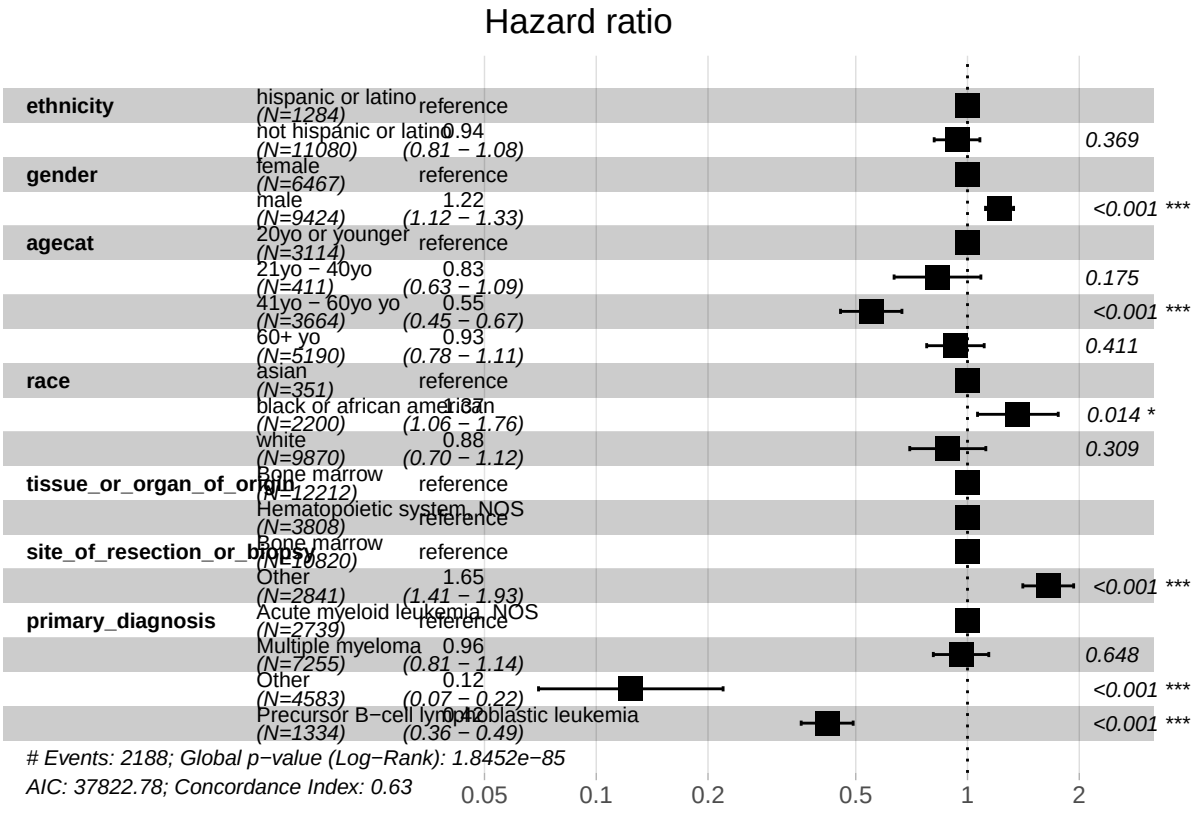
Additionally, our time event variable is "days to last follow-up" . We were recommended to go over the "days to death" variable and perhaps use that as our time event variable, but upon further examination, that variable had a large proportion of missing values and highly skewed the data.

Our main goal was to obtain results from machine learning methods. While the Survival Analysis portion may not follow the traditional guidelines (due to the vastly incompleteness of the dataset). It is mostly meant to be an exploratory analysis to give us and the readers a feel for the data at hand.

Initially, Kaplan-Meier plots were fit on the data for visualization purposes, the first plot other than the average survival rate for everyone drops to about 77% after 1000 days, but the dropoff levels off afterwards. For better visualization, KM plots were generated for gender and for primary disease diagnosis. From here, we can see that females have a higher survival rate than males; and that Precursor B-cell lymphoblastic leukemia was disease with the highest survival rate, whole Acute myeloid leukemia had the lowest survival rate for the entire time period



Some of the more complete variables that had a relation to Vital Status were: age, gender, ethnicity, race, tissue or organ of origin, site of resection or biopsy, and primary disease diagnosis. Using these variables, A Cox proportional Hazards model was constructed to see the effect these variables had on Vital Status, as well as to check if said effects were significant. A forest plot was used to depict these results. From this forest plot, we can see that the most significant factors were: if the individual was a male (p-value $<.001$), if they were in the 41-60 age range (p-value $<.001$), if they were black/African-American (p-value $=.014$), if the site of resection or autopsy was anywhere other than bone marrow (p-value $<.001$), or if their primary disease diagnosis was Precursor B-cell lymphoblastic leukemia (p-value $<.001$).


Hazard ratio
ethnicity
hispanic or latino (N=1284) reference
not hispanic or latino (N=11080) 0.94 (0.81 − 1.08) 0.369
gender
female (N=6467) reference
male (N=9424) 1.22 (1.12 − 1.33) <0.001 ***
agecat
20yo or younger (N=3114) reference
21yo − 40yo (N=411) 0.83 (0.63 − 1.09) 0.175
41yo − 60yo yo (N=3664) 0.55 (0.45 − 0.67) <0.001 ***
60+ yo (N=5190) 0.93 (0.78 − 1.11) 0.411
race
asian (N=351) reference
black or african american (N=2200) 1.37 (1.06 − 1.76) 0.014 *
white (N=9870) 0.88 (0.70 − 1.12) 0.309
tissue_or_organ_of_origin
Bone marrow (N=12212) reference
Hematopoietic system, NOS (N=3808) reference
site_of_resection_or_biopsy
Bone marrow (N=10820) reference
Other (N=2841) 1.65 (1.41 − 1.93) <0.001 ***
primary_diagnosis
Acute myeloid leukemia, NOS (N=2739) reference
Multiple myeloma (N=7255) 0.96 (0.81 − 1.14) 0.648
Other (N=4583) 0.12 (0.07 − 0.22) <0.001 ***
Precursor B−cell lymphoblastic leukemia (N=1334) 0.42 (0.36 − 0.49) <0.001 ***
Events: 2188; Global p−value (Log−Rank): 1.8452e−85
AIC: 37822.78; Concordance Index: 0.63
0.05
0.1
0.2
0.5
1
2