Survival Analysis with AML -Cancer Data

Time-to-event - (Overall-)Survival-probability estimation of patients with Acute Myeloid Leukaemia

Karolina Saegner & Melissa Muszelewski

Data

QRT Challenge provided

- Data of a total of 4,516 patients with clinical (Blood) and molecular (e.g genetic) data
- https://challengedata.ens.fr/challenges/162

Data

Content of dataset

Clinical data:

Title	Parameter	Data Type
CENTER	Clinical centre	Categorical
BM_BLAST	Bone marrow blasts in % (blasts are abnormal blood cells)	Numerical
WBC	White Blood Cell count in Giga/L	Numerical
ANC	Absolute Neutrophil count in Giga/L	Numerical
MONOCYTES	Monocyte count in Giga/L	Numerical
HB	Haemoglobin in g/dL	Numerical
PLT	Platelets count in Giga/L	Numerical
CYTOGENETICS	Karyotypes	Numerical and categorical

Risk analysis

Title	Parameter	Data Type
OS_YEARS	Overall survival in time in years	numerical
OS_STATUS	1 (death),) (alive at the last follow-up	numerical

Molecular data:

Title	Parameter	Data Type
ID	unique identifier per patient	Numerical
CHR, START, END	Chromosomal position of the mutation on the human genome	Numerical
REF, ALT	The reference nucleotide and the alternate (mutant) nucleotide	Categorical
GENE	The gene affected by the mutation	Categorical
PROTEIN_CHANGE	he impact of the mutation on the protein produced by the gene	categorical
EFFECT	Broad classification of the mutation’s impact on gene function	categorical
VAF	Variant Allele Fraction, representing the proportion of cells carrying the mutation	Numerical
DEPHT	the average number of times a particular nucleotide in the DNA sequence is read during sequencing	Numerical

Data

Risk score analysis

Clinical data:

Title	Parameter	Data Type
CENTER	Clinical centre	Categorical
BM_BLAST	Bone marrow blasts in % (blasts are abnormal blood cells)	Numerical
WBC	White Blood Cell count in Giga/L	Numerical
ANC	Absolute Neutrophil count in Giga/L	Numerical
MONOCYTES	Monocyte count in Giga/L	Numerical
HB	Haemoglobin in g/dL	Numerical
PLT	Platelets count in Giga/L	Numerical
CYTOGENETICS	Karyotypes	Numerical and categorical

Molecular data:

Title	Parameter	Data Type
ID	unique identifier per patient	Numerical
CHR, START, END	Chromosomal position of the mutation on the human genome	Numerical
REF, ALT	The reference nucleotide and the alternate (mutant) nucleotide	Categorical
GENE	The gene affected by the mutation	Categorical
PROTEIN_CHANGE	he impact of the mutation on the protein produced by the gene	categorical
EFFECT	Broad classification of the mutation's impact on gene function	categorical
VAF	Variant Allele Fraction, representing the proportion of cells carrying the mutation	Numerical
DEPHT	the average number of times a particular nucleotide in the DNA sequence is read during sequencing	Numerical

Risk analysis

Title	Parameter	Data Type
OS_YEARS	Overall survival in time in years	numerical
OS_STATUS	1 (death),) (alive at the last follow-up	numerical

Selected AML Features :

BM_BLAST
HB
PLT

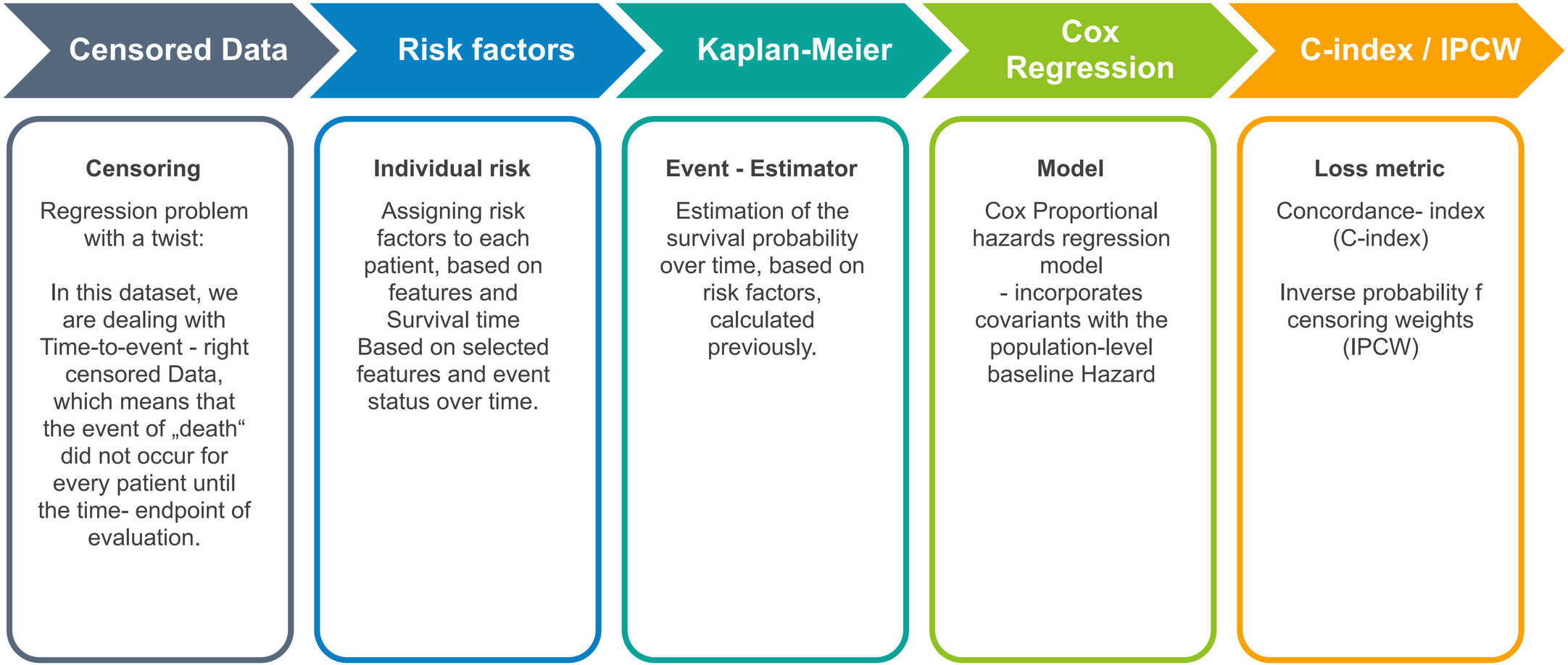
Risk score analysis based on:

OS_YEARS
OS_STATUS

BM_BLAST
HB
PLT

Survival Analysis

With Scikit-Survival, which is built on Scikit-learn



Sources

Sources I relied on:

- Scikit-Survival documentation
  - https://scikit-survival.readthedocs.io/en/stable/index.html
- Medium article: „Roman Emperors“
  - https://medium.com/@josephgeorgelewis2000/last-man-standing-survival-analysis-in-python-c7d7132f8471
- Medium article „Survival Analysis simplified“
  - https://medium.com/@zynp.atlii/survival-analysis-simplified-explaining-and-applying-with-python-7efacf86ba32
- Paper: Clark et. al.:Survival Analysis Part I: Basic concepts and first analyses
  - https://pmc.ncbi.nlm.nih.gov/articles/PMC2394262/
- Proportional Hazards Model Wikipedia
  - https://en.wikipedia.org/wiki/Proportional_hazards_model