Survival Analysis

Time-to-event -
(Overall-) Survival-probability
of patients with Acute Myeloid Leukemia (AML)

Karolina Saegner & Melissa Muszelewski

Data provider

QRT-challenge

ChAllenge By MathA Data

https://challengedata.ens.fr/challenges/162

Overall Survival Prediction for patients diagnosed with Myeloid Leukemia
by QRT

- Predictive models in healthcare – Onkology
  - Aim: To predict Overall Survival
- Cancer Data
  - 24 clinical centers
- 4 516 patients with AML
  - Content Data
    - Clinical (blood, cell-count)
    - Molecular (e.g genetic)

https://www.qube-rt.com/

Acute Myeloid Leukaemia (AML)

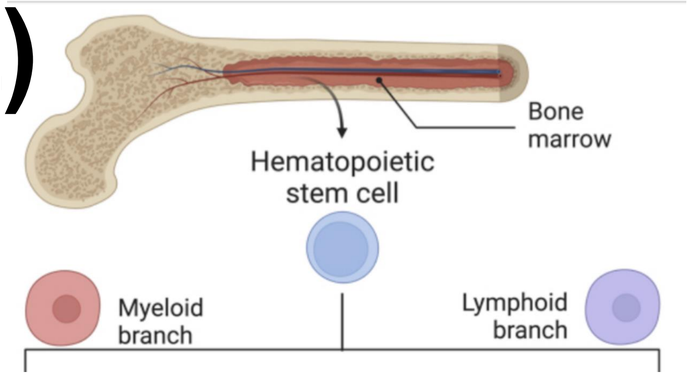
Blood cancer type

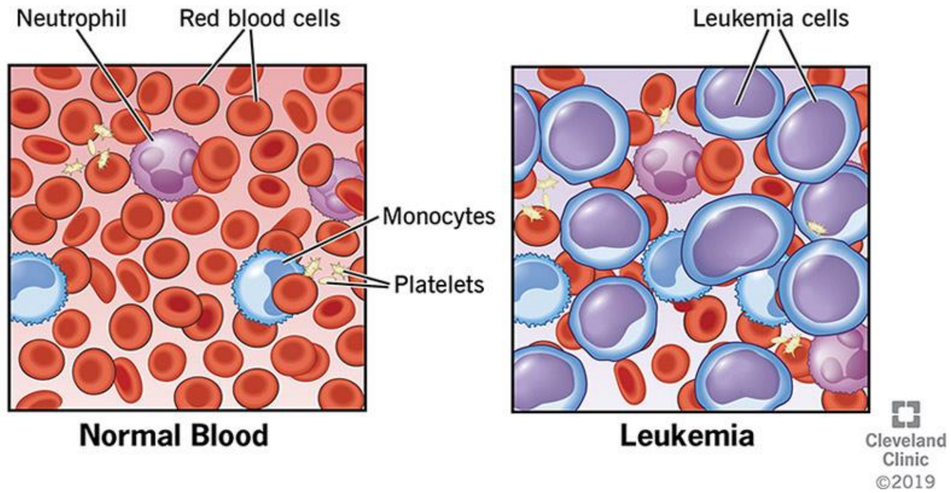
- Leukemia type - more common in adults
  - Rapid growth
    - Abnormal cells:
      - Bone marrow
      - Blood

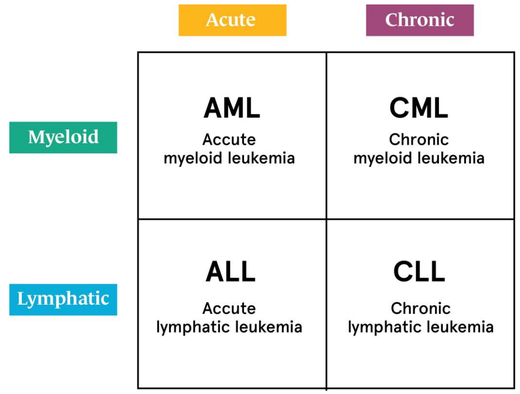
--> Myeloid "Blast" cells

  - Immature (not fully developed)

- Intereferences
  - Normal blood cell production







Data

Content of the 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

Parameter Selection

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

Selected AML Features :

BM_BLAST
HB
PLT

Dataset characteristics

```
clinical_df['BM_BLAST'].head()
0    14.0
1     1.0
2    15.0
3     1.0
4     6.0
Name: BM_BLAST, dtype: float64
```

```
clinical_df['HB'].head()
0     7.6
1    11.6
2    14.2
3     8.9
4    11.1
Name: HB, dtype: float64
```

```
clinical_df['PLT'].head()
0    119.0
1     42.0
2     81.0
3     77.0
4    195.0
Name: PLT, dtype: float64
```

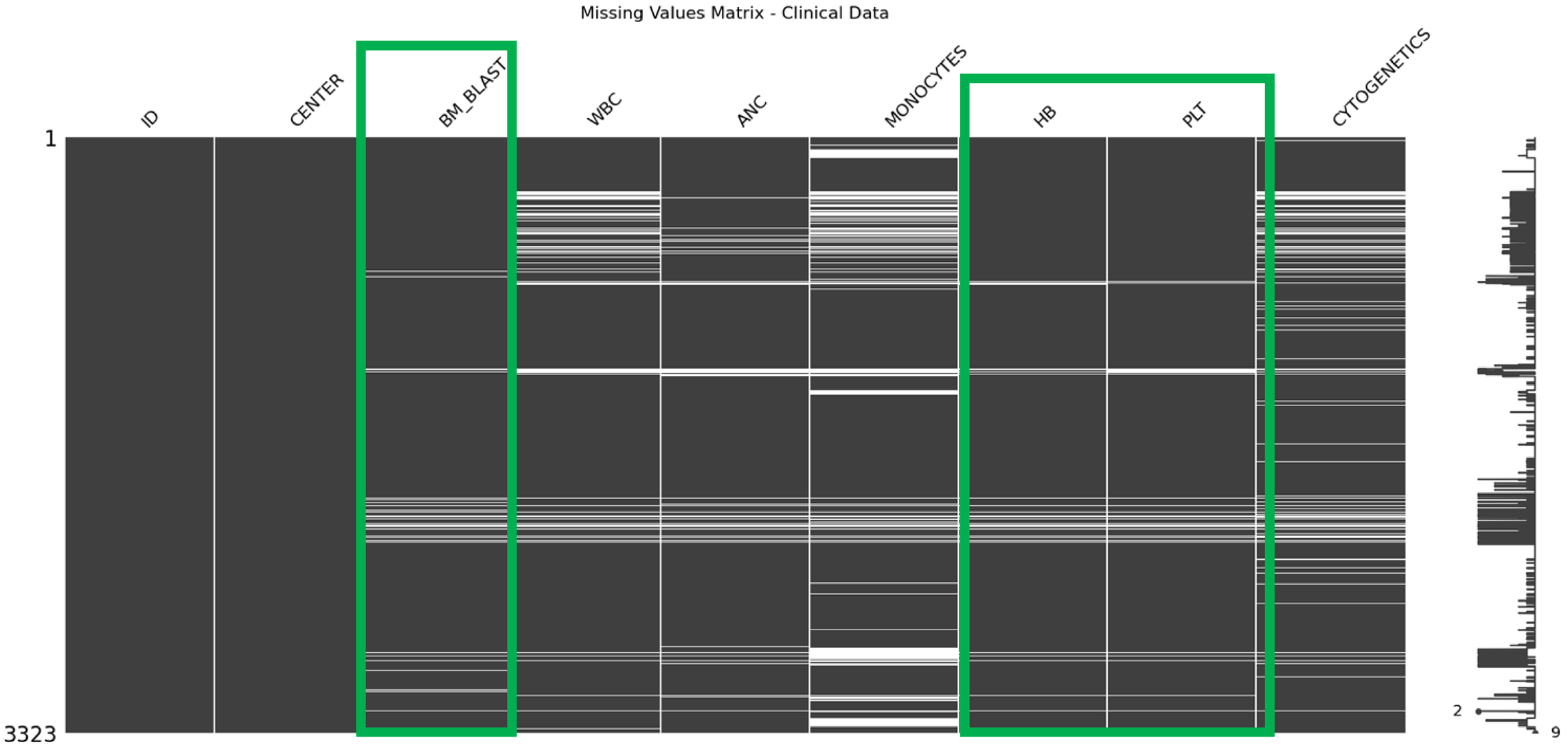
```
missing_values_clinical
ID                0
CENTER            0
BM_BLAST        109
WBC             272
ANC             193
MONOCYTES       601
HB              110
PLT             124
CYTOGENETICS    387
```

```
outcome_df['OS_YEARS'].head()
0    1.115068
1    4.928767
2    2.043836
3    2.476712
4    3.145205
Name: OS_YEARS, dtype: float64
```

```
outcome_df['OS_STATUS'].head()
0    1.0
1    0.0
2    0.0
3    1.0
4    0.0
Name: OS_STATUS, dtype: float64
```

```
missing_values_outcome
ID              0
OS_YEARS      150
OS_STATUS     150
```

Missing values (clinical data)

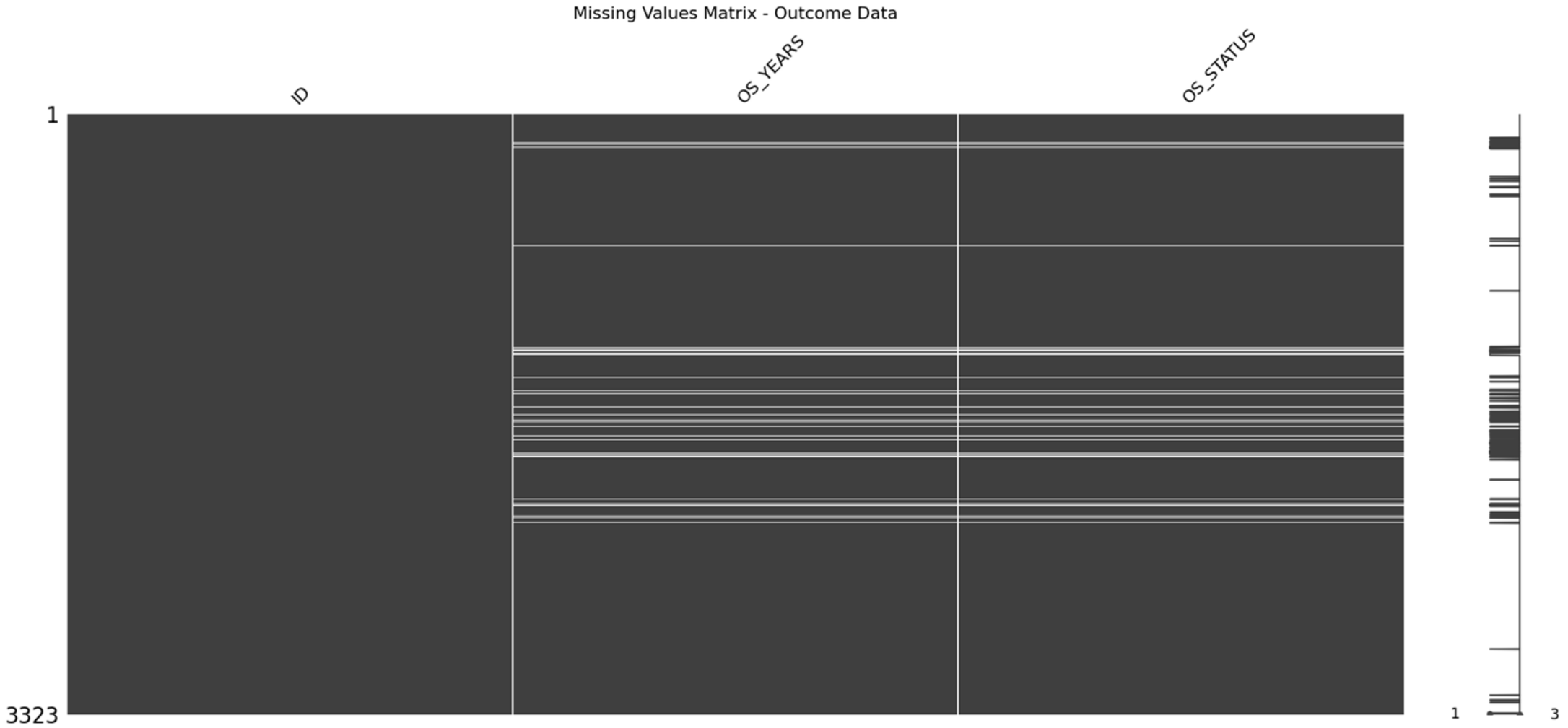


Missing value imputation

K Nearest Neighbours (KNN) for numerical variables:

- Scaling to have mean = 0 and std = 1 before KNN (`scaler = StandardScaler(); scaler.transform`)
- Finding the k most similar patients based on scaled Euclidean distance
- Imputing missing values using the average of those k neighbours (`imputer = KNNImputer(n_neighbors=min(n_neighbors, complete_cases))`)
- Transforming back to the original scale (`scaler.inverse_transform`)

Missing target values



Missing value imputation

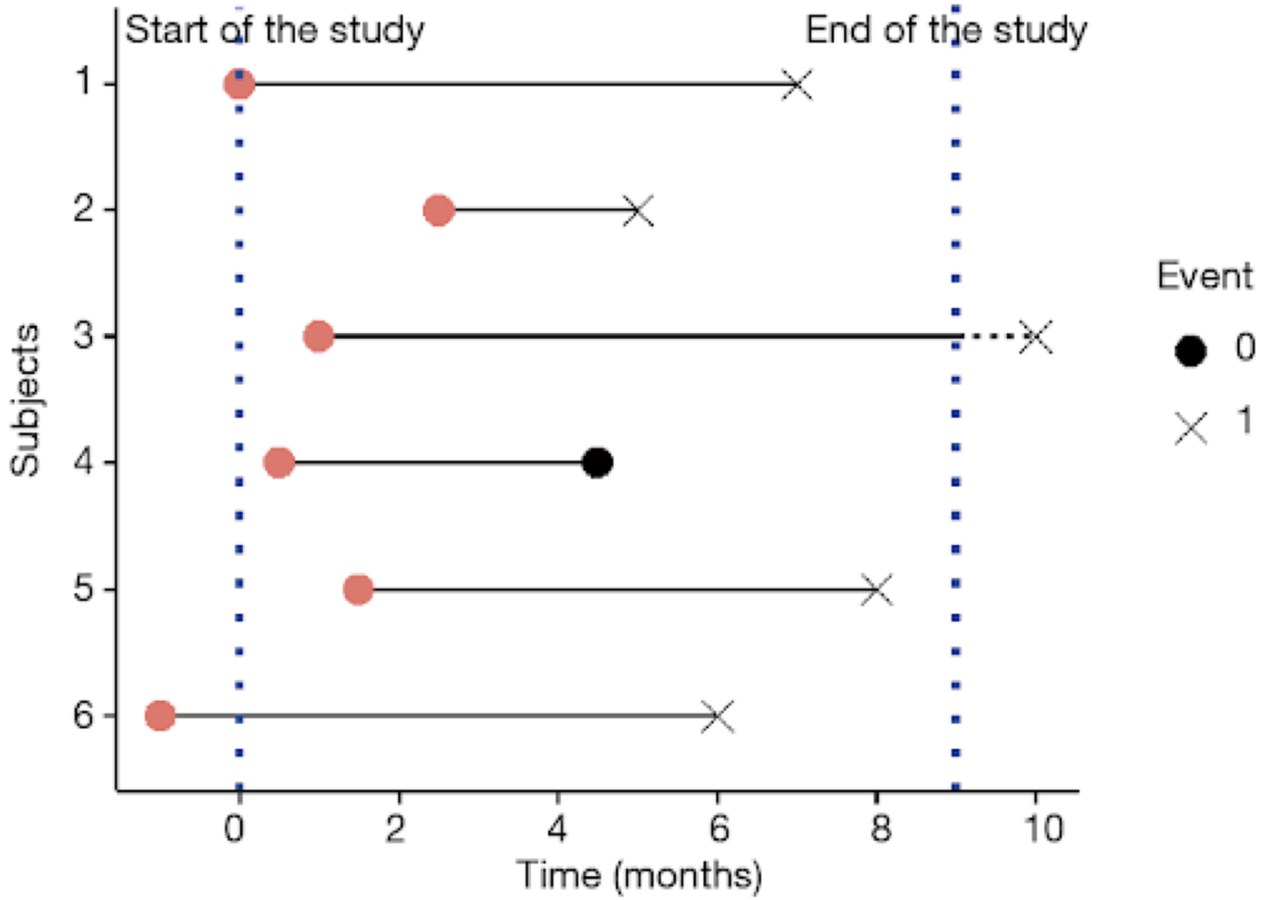
- OS_YEARS (number of years survived after the diagnosis) and OS_STATUS (alive or dead) at the time of data collection --> NaN if the patient was still alive at that point
- Any data imputation for such data is meaningless, and since we should not have any missing data --> removal of those patients with no data

Data censoring

Censored Data in Survival Analysis

Time-to-event Data

- Right censoring
- Status
  - Event = death
  - Data of last appointment (alive)
    - No total survival time



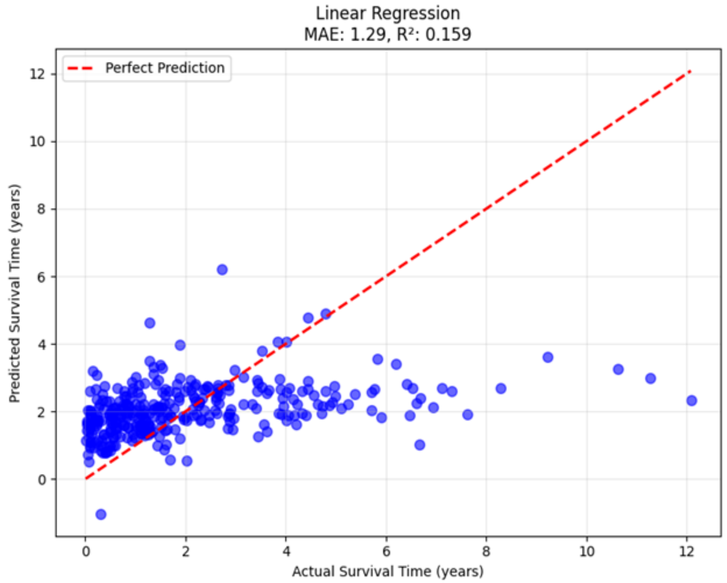
https://projector-video-pdf-converter.datacamp.com/25923/chapter2.pdf

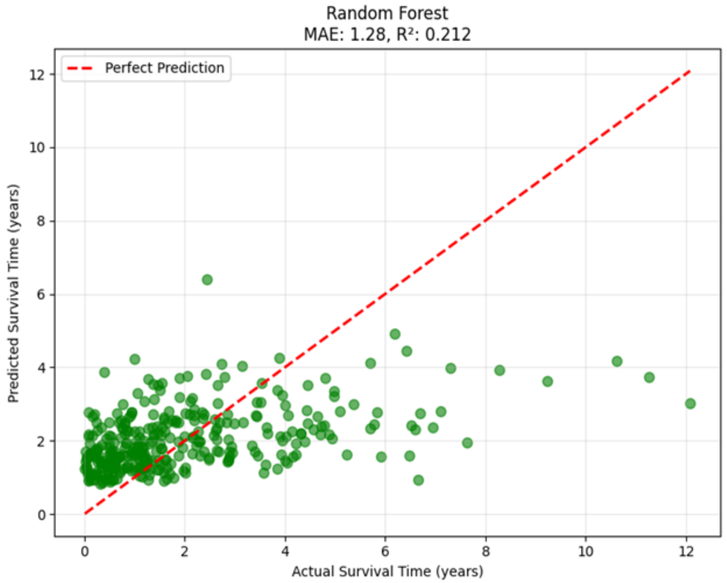
Baseline models

Linear regression and Random Forest

```python
# Build simple Linear Regression baseline
lr_model = LinearRegression()
lr_model.fit(X_train_events, y_train_time)

# Make predictions
y_pred_lr = lr_model.predict(X_test_events)
```

```python
# Build Random Forest baseline
rf_model = RandomForestRegressor(n_estimators=100, random_state=42, max_depth=5)
rf_model.fit(X_train_events, y_train_time)

# Make predictions
y_pred_rf = rf_model.predict(X_test_events)
```





Metric	Linear Regression	Random Forest
MAE (years)	1.295	1.284
RMSE (years)	1.796	1.739
R^2 Score	0.159	0.212

Survival Analysis with right–censored data

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

Selected AML Features :

BM_BLAST
HB
PLT

Risk score analysis based on:

OS_YEARS
OS_STATUS

BM_BLAST
HB
PLT

Cytogenetics

Proportional Hazard analysis:

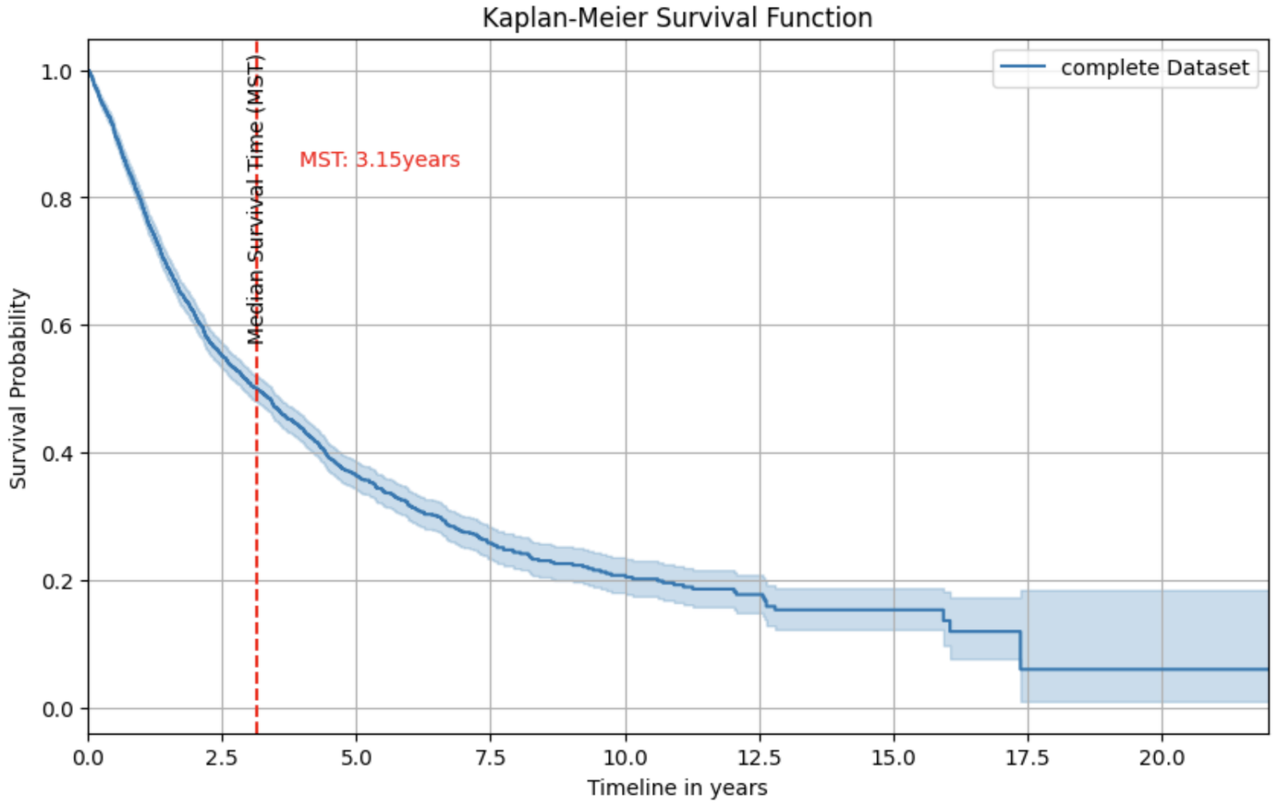
BM_BLAST
HB
PLT

Censored Data → **Kaplan - Meier** → **Risk Factors** → **Cox proportional hazards** → **Inverse probability censoring weightsIPCW**

Kaplan-Meier

Unconditional Survival estimate

- Non- Parametric
- Confidence intervals change over time
  - Sample size
- Median survival Time
  - **When half of the patients are expected to be alive**
- KM-Estimate
  - "Risk Factor"
  - Prediction
- Survival probability at time (t)



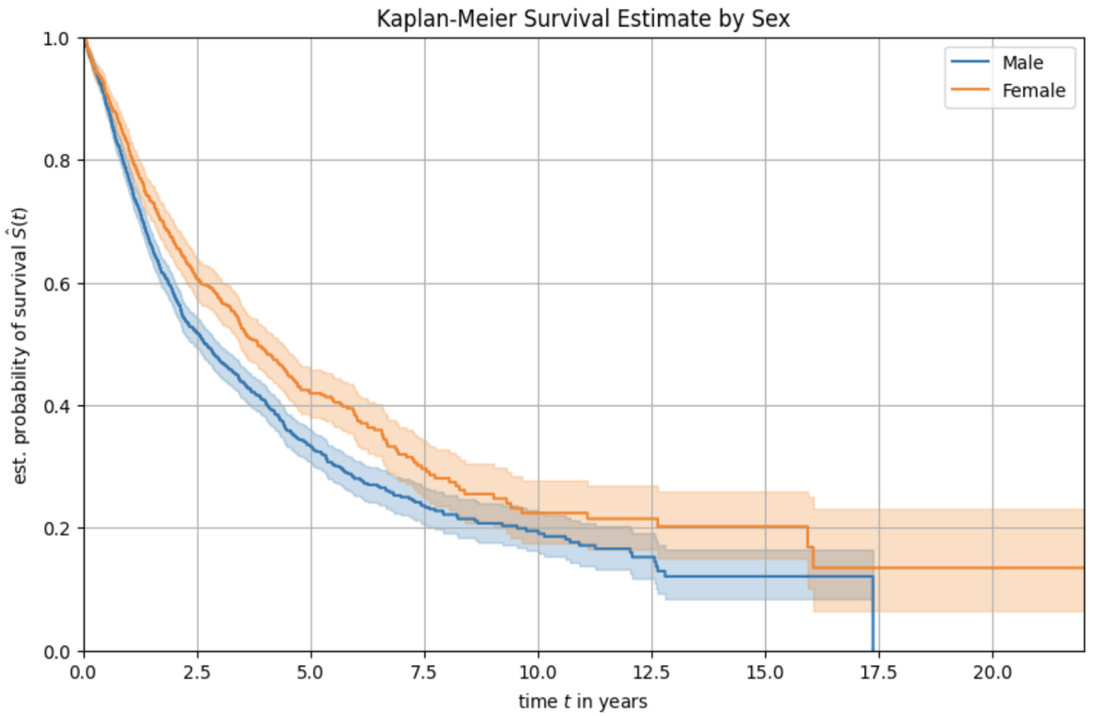

```
3.14520547945205
           KM_estimate
timeline
0.000000      1.000000
0.002740      0.999679
0.005479      0.999679
0.008219      0.999037
0.010959      0.999037
...                ...
16.389041     0.118853
16.567123     0.118853
17.147945     0.118853
17.375342     0.059427
22.043836     0.059427

[1653 rows x 1 columns]
```

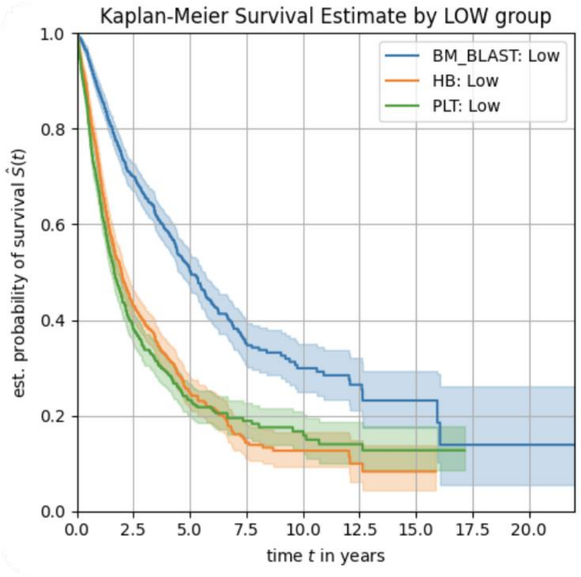
	1 year	3 years	5 years
Survival Probability	79 %	51 %	36 %

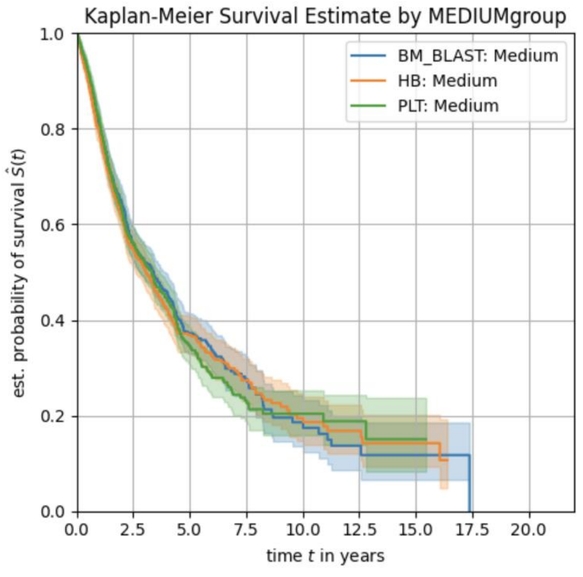
Kaplan-Meier

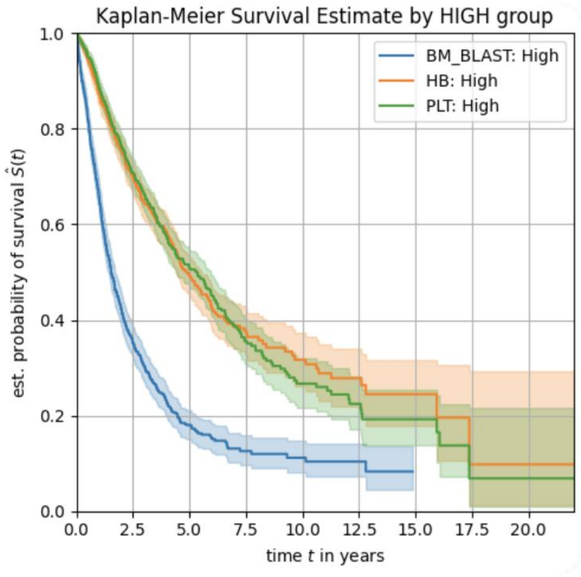
Groups survival distribution

- Impact on survival of
  - different Parameters
    - grouped





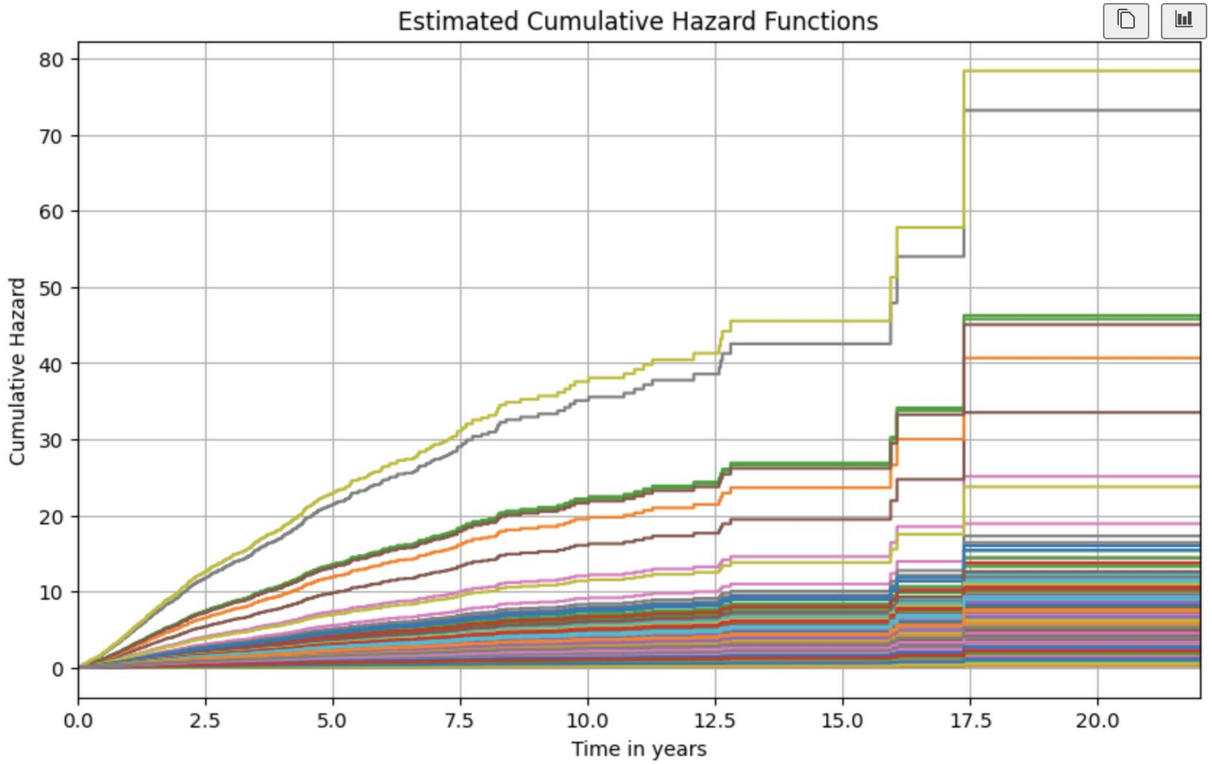




Cox proportional hazard

Multivariate approach

- Semi-parametric
- Compareable to linear regression
- Relation ratio
  - Event incidence
    - Hazard function
      - Cumulative hazard
  - Covariates
- High value = high risk

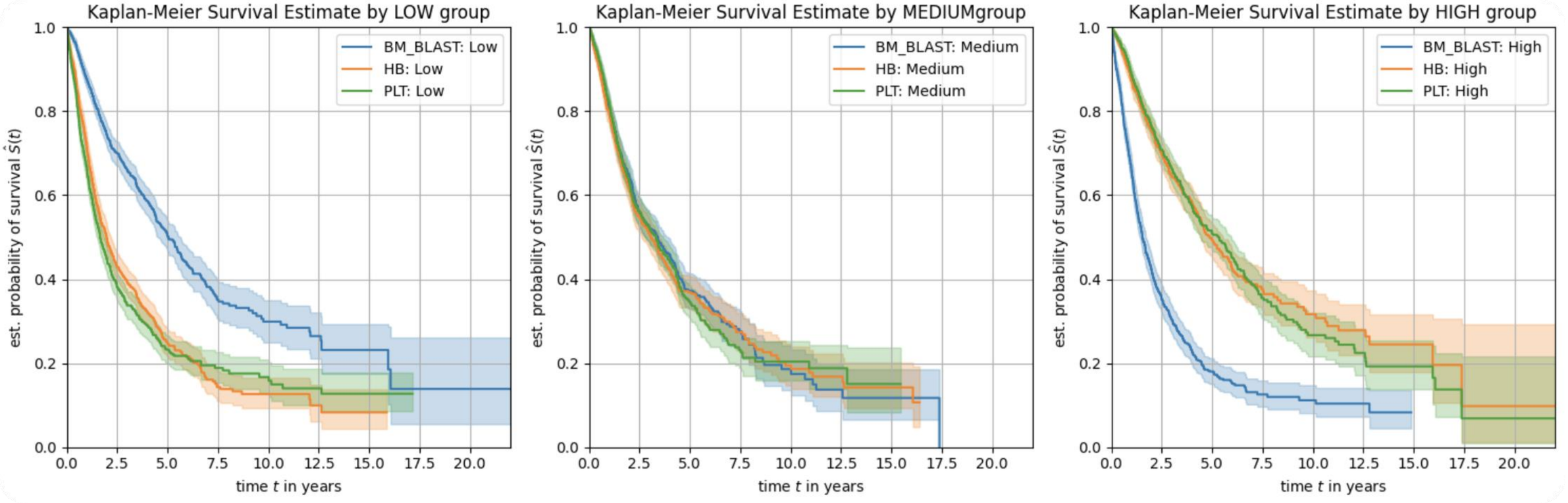



```
Cox Proportional Hazards Model Coefficients:
                coef  exp(coef)
HB         -0.162277   0.850206
PLT        -0.002008   0.997994
BM_BLAST    0.031534   1.032036
```

Kaplan-Meier

Comparison with CPH Coefficients

```
Cox Proportional Hazards Model Coefficients:
                coef  exp(coef)
HB         -0.162277   0.850206
PLT        -0.002008   0.997994
BM_BLAST    0.031534   1.032036
```



Deep Learning Cox Proportional Hazards Model

Survival analysis with pytorch https://github.com/havakv/pycox

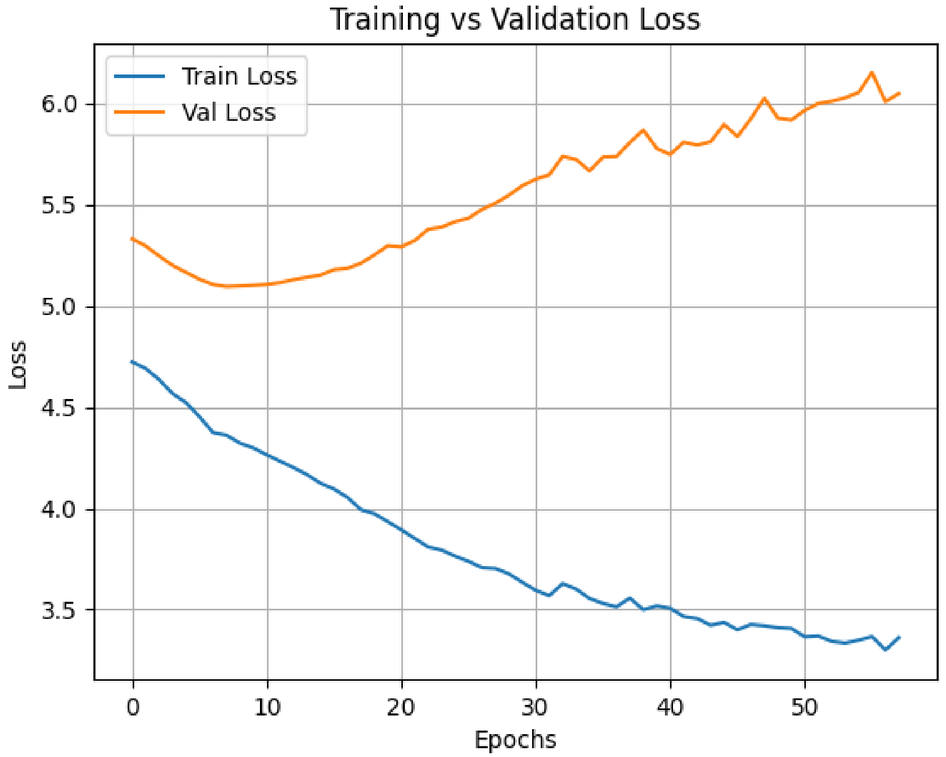
```
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numerical_cols),
        ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_cols)
    ],
    remainder='drop'
)
```

```
###### model no. 3
# Define MLP model
class MLP(torch.nn.Module):
    def __init__(self, in_features, num_nodes, out_features):
        super().__init__()
        self.net = torch.nn.Sequential(
            torch.nn.Linear(in_features, num_nodes[0]), # first hidden layer
            torch.nn.ReLU(), # activation function
            torch.nn.Dropout(0.2), # Regularization
            torch.nn.Linear(num_nodes[0], num_nodes[1]), # second hidden layer
            torch.nn.ReLU(),
            torch.nn.Dropout(0.2),
            torch.nn.Linear(num_nodes[1], out_features) # output layer
        )

    def forward(self, x):
        return self.net(x)
```

```
# Parameters
in_features = X_train.shape[1]
num_nodes = [64, 32]
out_features = 1

# Instantiate network
net = MLP(in_features, num_nodes, out_features)

# Optimizer with small LR and some weight decay
optimizer = tt.optim.Adam(lr=1e-3, weight_decay=1e-4)

# Model
model = CoxPH(net, optimizer)
```

```
batch_size = 256
epochs = 500
callbacks = [EarlyStopping(patience=50)]

print("Training CoxPH model...")
log = model.fit(train_data[0], train_data[1],
                batch_size=batch_size, epochs=epochs,
                val_data=val_data,
                callbacks=callbacks,
                verbose=True)
```

Deep Cox



Baseline models

Metric	Linear Regression	Random Forest
MAE (years)	1.295	1.284
RMSE (years)	1.796	1.739
R^2 Score	0.159	0.212

Concordance index = .73 (0.5 - random prediction, 1.0 - perfect prediction)

```
RMSE (Expected Survival vs True Duration): 2.11
MAE  (Expected Survival vs True Duration): 1.49
```

If we had more time...

BMC Med Res Methodol. 2013 Dec 7:13:152. doi: 10.1186/1471-2288-13-152.

Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome

Patrick Royston [1], Mahesh K B Parmar

Affiliations + expand

PMID: 24314264 PMCID: PMC3922847 DOI: 10.1186/1471-2288-13-152

Conclusions: We conclude that the hazard ratio cannot be recommended as a general measure of the treatment effect in a randomized controlled trial, nor is it always appropriate when designing a trial. Restricted mean survival time may provide a practical way forward and deserves greater attention.

If we had more time...

Possible further steps/future applications

- Molecular data
  - Genetics
  - Abnormalities in metabolism
- Feature clustering
- Competing risks

Limitations

Model / Prediction- limitations

- Impact of censoring in data
  - e.g, seen in kaplan meier survival plots with different max years data
- Library functions
- Output of values, exact context has to be researched in e.g. documentation
- Kernel limitations
- Age dependencies not shown

Reality Check

Comparison

- Kaplan Meier
  - Survival probability
  - Groups survival distribution
- Cox proportional Hazards

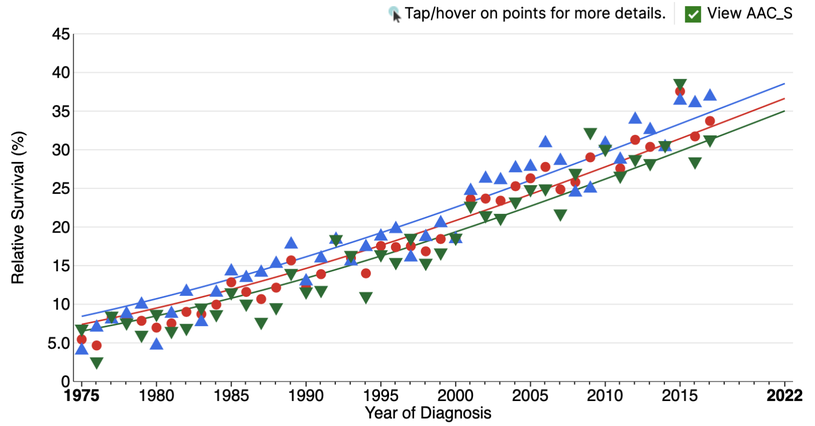
```
Cox Proportional Hazards Model Coefficients:
                coef  exp(coef)
HB         -0.162277   0.850206
PLT        -0.002008   0.997994
BM_BLAST    0.031534   1.032036
```

	1 year	3 years	5 years
Ours	79 %	51 %	36 %
UK *	34 %	20 %	16 %
USA **	48 %	32 %	28 %

Types of Leukemia	ALL	AML	CLL	CML
5- year survival rate*	69.9%	29.5%	87.2%	70.6%
Number of deaths per 100,000 persons	0.4	2.7	1.1	0.3
Death is highest among those aged	65-84	65+	75+	75+

1975-2022
All Races / Ethnicities By Sex, 5-year Relative Survival, All Ages

Graph | Data Table



* https://hmrn.org/statistics/survival

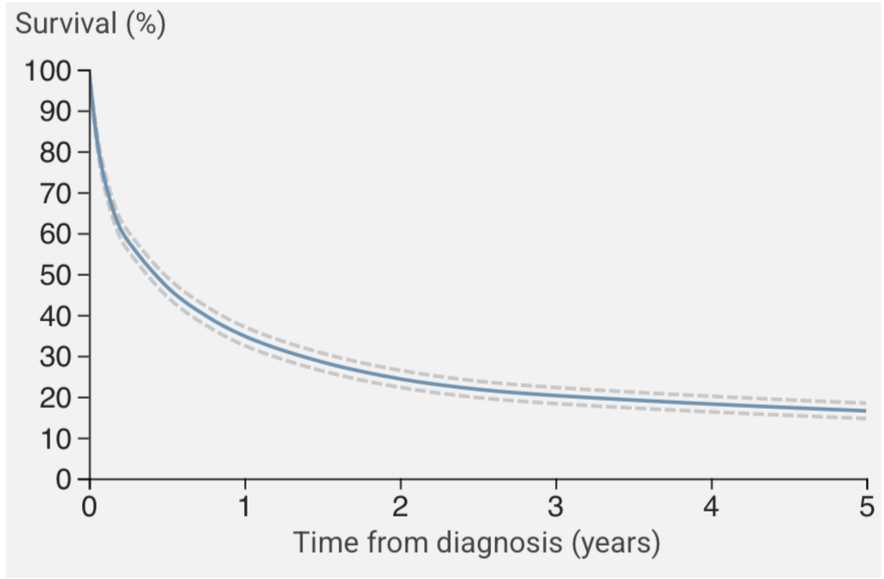
* https://seer.cancer.gov/statfacts/html/amyl.html

Up-to-date research data

CANCER RESEARCH UK

Acute myeloid leukaemias

	Annual incidence rate per 100,000 [1]			Median age at diagnosis (years)	Net survival % (95% confidence interval) [2]		
	Total	Male	Female		1 year	3 year	5 year
Acute myeloid leukaemias	4.4	5.2	3.7	72.1	34.2 (32.0 - 36.5)	20.2 (18.2 - 22.1)	16.5 (14.6 - 18.4)
Acute myeloid leukaemia	4.1	4.9	3.4	73.0	32.0 (29.7 - 34.3)	17.5 (15.6 - 19.4)	13.6 (11.8 - 15.4)
Acute promyelocytic leukaemia	0.3	0.3	0.3	54.6	68.8 (59.9 - 77.7)	61.4 (51.6 - 71.0)	61.7 (51.9 - 71.5)



https://hmrn.org/statistics/survival

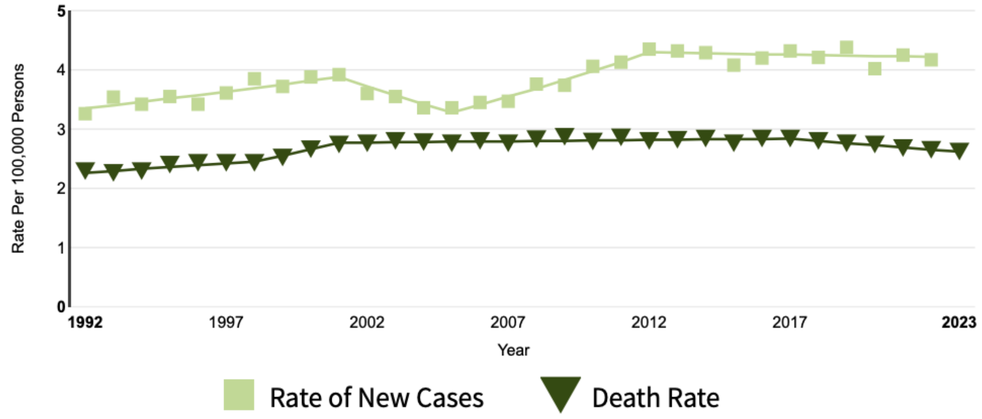
NIH NATIONAL CANCER INSTITUTE
Surveillance, Epidemiology, and End Results Program

Cancer Stat Facts: Leukemia — Acute Myeloid Leukemia (AML)

Estimated New Cases in 2025	22,010
% of All New Cancer Cases	1.1%

Estimated Deaths in 2025	11,090
% of All Cancer Deaths	1.8%

5-Year Relative Survival
32.9%
2015–2021



https://seer.cancer.gov/statfacts/html/amyl.html

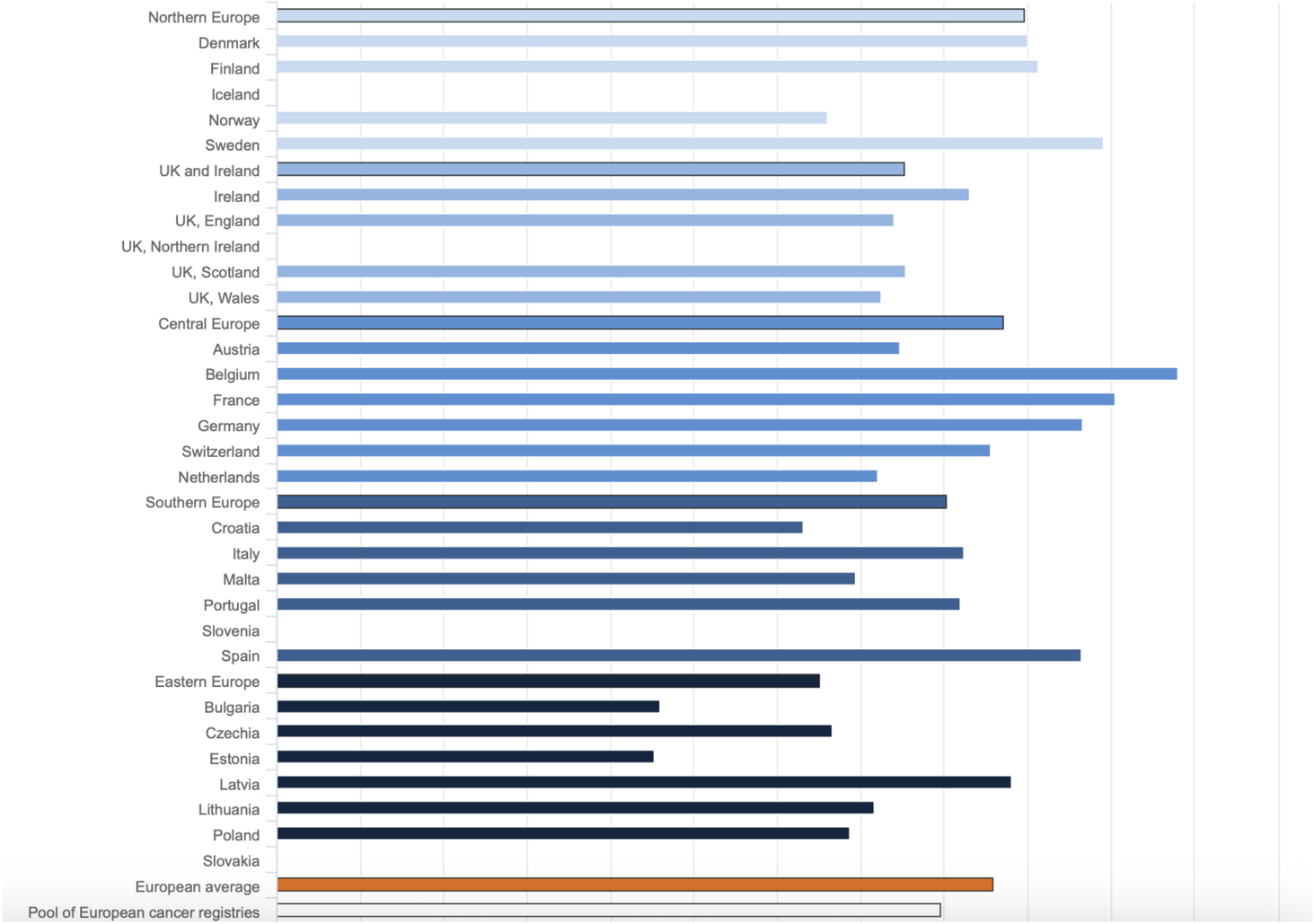
Up-to-date research data

European Commission

ECIS - European Cancer Information System

Age-standardised 5-year relative survival by country

Both sexes, AML, 15+ years, 2000-2007

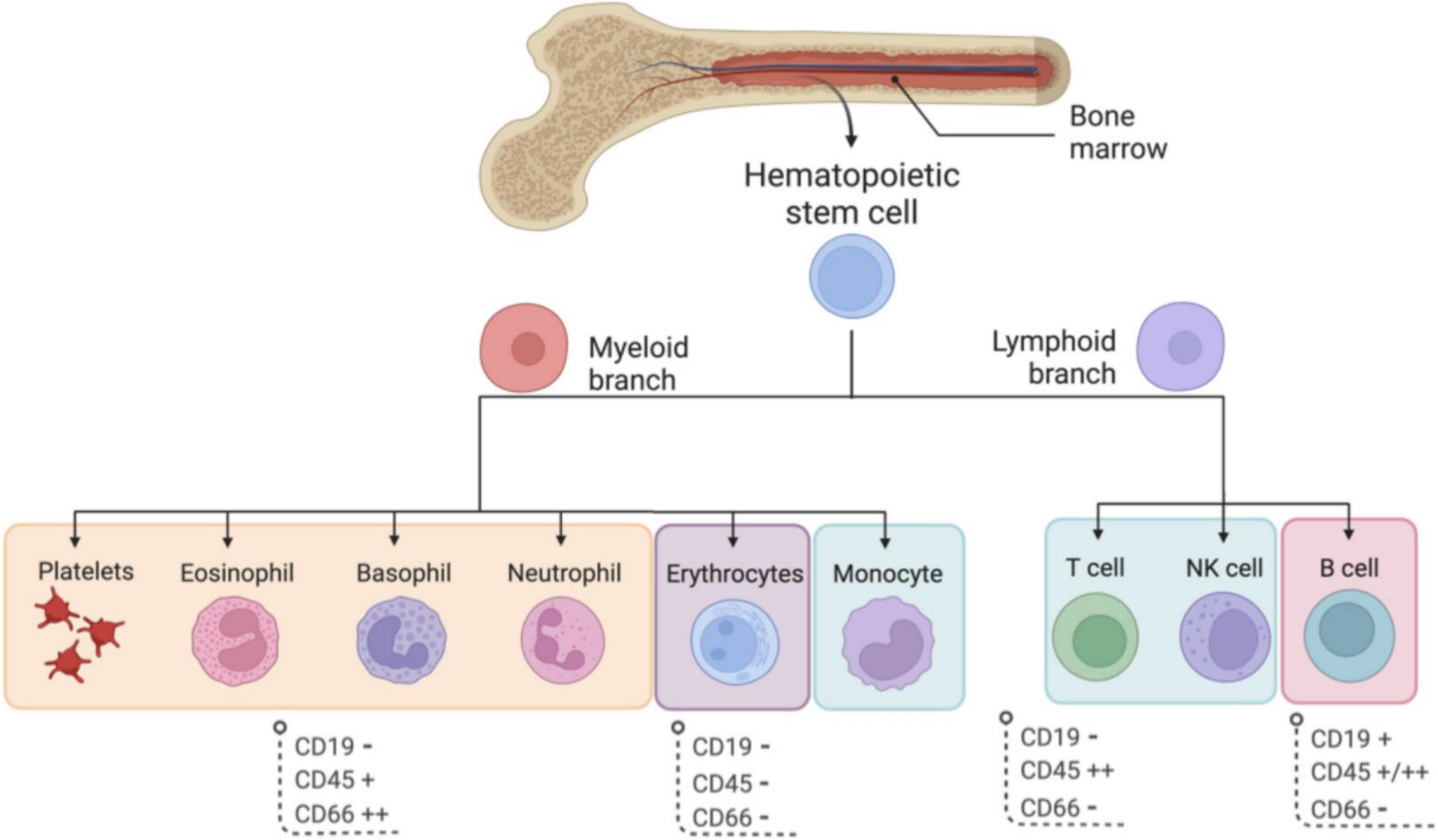


https://ecis.jrc.ec.europa.eu/explorer.php?$0-2$1-All$2-All$4-1,2$3-65$6-0,14$5-2000,2007$7-1$CRelativeSurvivalCountry$X0_15-RSC

Thank you!

Karolina Saegner & Melissa Muszelewski

Additional Material



https://www.themoonlight.io/fr/review/biorxiv/automatic-computational-classification-of-bone-marrow-cells-for-b-cell-pediatric-leukemia-using-umap

Literature

Literature review

- Application of machine learning in breast cancer survival prediction using a multimethod approach (Nature, 2024, Hamedi *et al.*)
    - https://www.nature.com/articles/s41598-024-81734-y
- A novel perspective on survival prediction for AML patients: Integration of machine learning in SEER database applications (Heliyon Elsevier, 2025, Jia *et al.)*
    - https://pmc.ncbi.nlm.nih.gov/articles/PMC11795080
- Identification of relevant features using SEQENS to improve supervised machine learning models predicting AML treatment outcome (BMC Springer, 2025, Pons-Suñer *et al.)*
    - https://link.springer.com/article/10.1186/s12911-025-03001-y

Sources

Survival Analysis

- Scikit-Survival documentation
  - https://scikit-survival.readthedocs.io/en/stable/index.html
- Lifelines documentation
  - https://lifelines.readthedocs.io/en/latest/
- 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

Additional Sources

Cancer, Leukemia, Databases

- Wikpedia Leukemia
  - https://en.wikipedia.org/wiki/Leukemia
- Leukemia Overview, Cleveland Clinic
  - https://my.clevelandclinic.org/health/diseases/4365-leukemia
- Cancer Reaearch UK
  - https://www.cancerresearchuk.org/
- DocCheck Flexikon (German)
  - https://flexikon.doccheck.com/de/Leuk%C3%A4mie
- Krebs einfach erklärt
  - https://simpleclub.com/lessons/biologie-krebs
- Deutsche Krebsgesellschaft
  - https://www.krebsgesellschaft.de/onko-internetportal/basis-informationen-krebs/krebsarten/leukaemie.html
- Deutsche Krebshilfe
  - https://www.krebshilfe.de/infomaterial/Blaue_Ratgeber/Leukaemie_BlaueRatgeber_DeutscheKrebshilfe.pdf

Code

Examples of changes/Methods

- Stratify in train_test_split
  - Even distribution of Event occurence

```
# Split the data into train and test sets
# usage of stratify to ensure balanced distribution of censored data in both train and test sets
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y["OS_STATUS"], random_state=0
)
```

```
Number of events in train data (y_train): 1400 , Distribution: , 52.67 %
Number of events in test data (y_test): 350 , Distribution: , 52.63 %
```