Bressan Manuel – Carrara Davide – Ghiglione Carlo – Maddaloni Federica – Zanotti Daniela

Applied Statistics – A.A. 20/21
Tutors: Prof. Ieva, Dr. Spreafico, Dr. Gregorio

MDS
Genomic Clustering

POLITECNICO
MILANO 1863

HUMANITAS
RESEARCH HOSPITAL


PROBLEM
RECAP
GROUPS
SURVIVAL
ANALYSIS

PROBLEM RECAP

CLUSTERING

GROUPS ANALYSIS

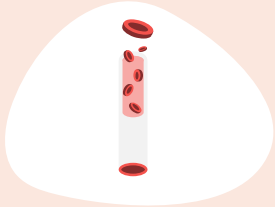
SURVIVAL ANALYSIS

PROBLEM RECAP

MDS – What it is

- Blood disease (**morphologic abnormalities** of myeloid cells)
- Possible **AML** (Acute Myeloid Leukemia) evolution
- Only curative treatment: **transplant**, but additional risks (rejection, disease relapse)
- **Risk stratification** based on patient's genomics

DATASET



*Patient's clinical data (N=**2876**)*

Transplanted/Not transplanted

*Clinical **development** of the disease and of the transplant:*

- *AML evolution*
- *Rejection*
- *Survival*

***Genes** mutations (47)*

***Chromosomes** anomalies (13)*

IPSS-R

Bersanelli (2021)

*Classification with respect to **international standards***

*Classification in **8 genomic groups** (from a previous study)*

Data provided by: HUMANITAS RESEARCH HOSPITAL

OBJECTIVES

Find a new clustering of patients in order to check **robustness** of given groups

Build a model for **survival** using our groups and clinical features

CLUSTERING

DISSIMILARITY MEASURE

$$d_{i,j} = \frac{n + n_D^{i,j} - n_C^{i,j}}{n}$$

n : total number of genes considered

$n_D^{i,j}$: total number of mutations present in i and not in j and viceversa

$n_C^{i,j}$: total number of mutations in common between i and j

ROADMAP TO OUR CLUSTERING

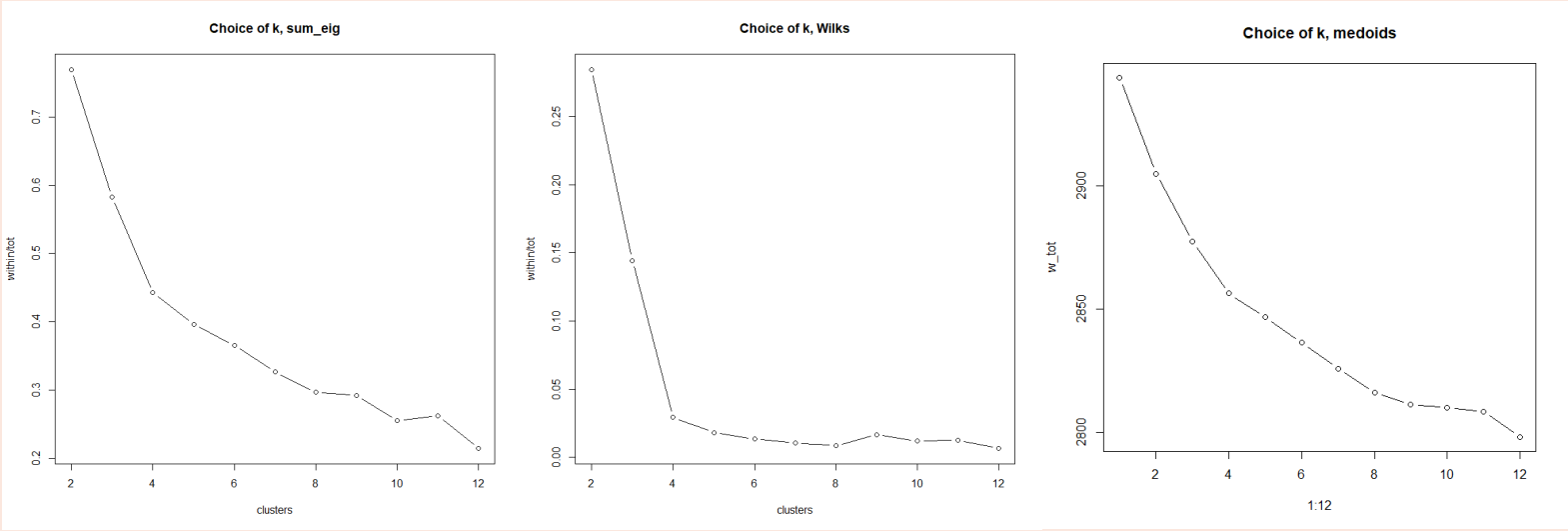
Hierarchical methods

Ward clustering

Isomap projection

K-Medoids

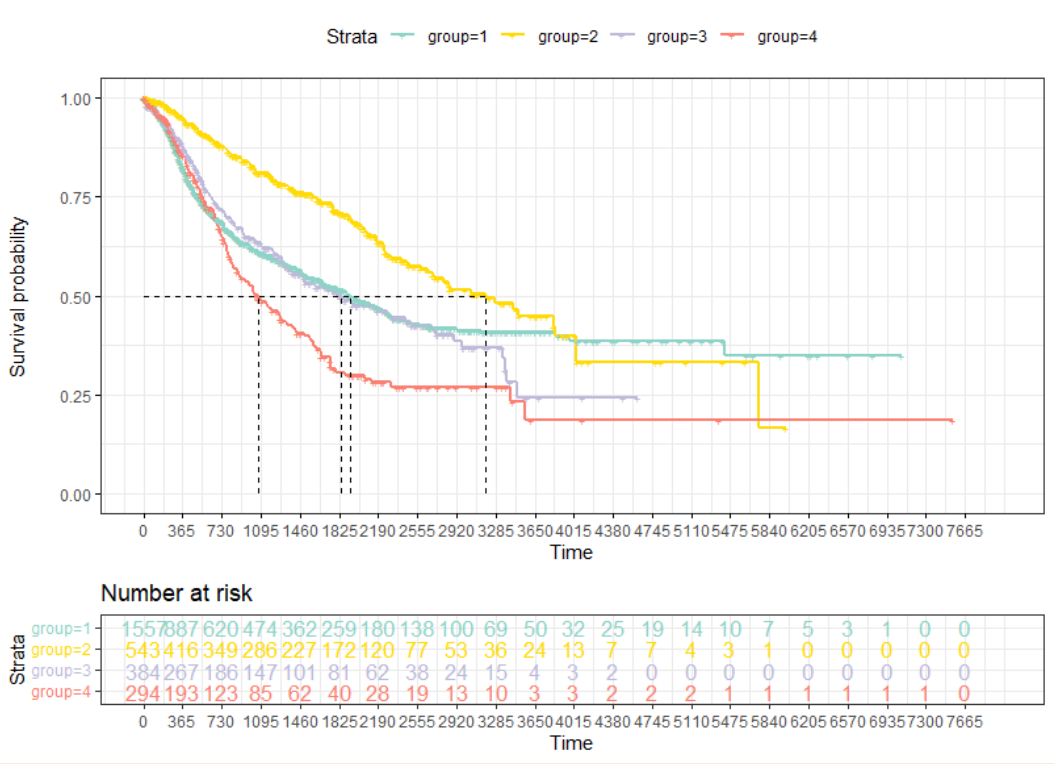
VARIABILITY ANALYSIS



We tried **different approaches** in order to account for the variability of the groups

The proper k seems to be **4**

K = 4



WHAT IS A KAPLAN-MEIER SURVIVAL CURVE ?

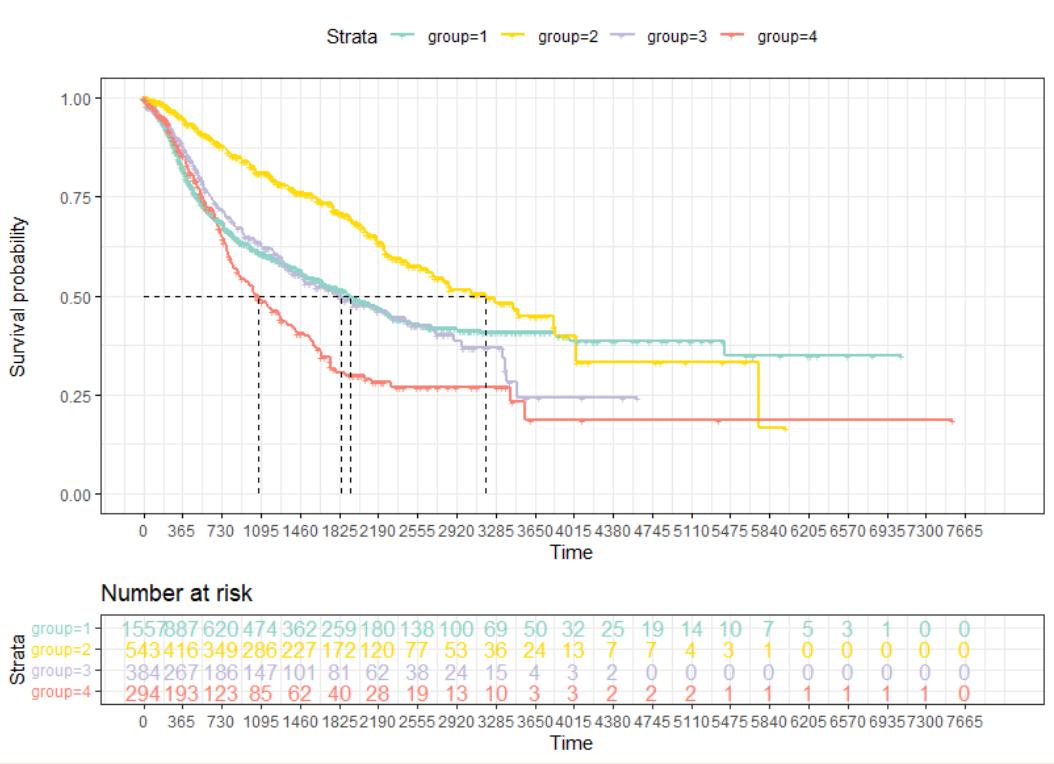
Fraction of patients **living** for a certain amount of time after diagnosis.

WHAT IS THE LOG-RANK TEST?

Non-parametric statistical test for comparing the **survival distributions** of two (or more) groups.

H0: Equal survival distributions
H1: Different survival distribution

K = 4



Unbalance between the numerosity of the various clusterings.

Groups 1 and 3 have **almost identical survival curves**

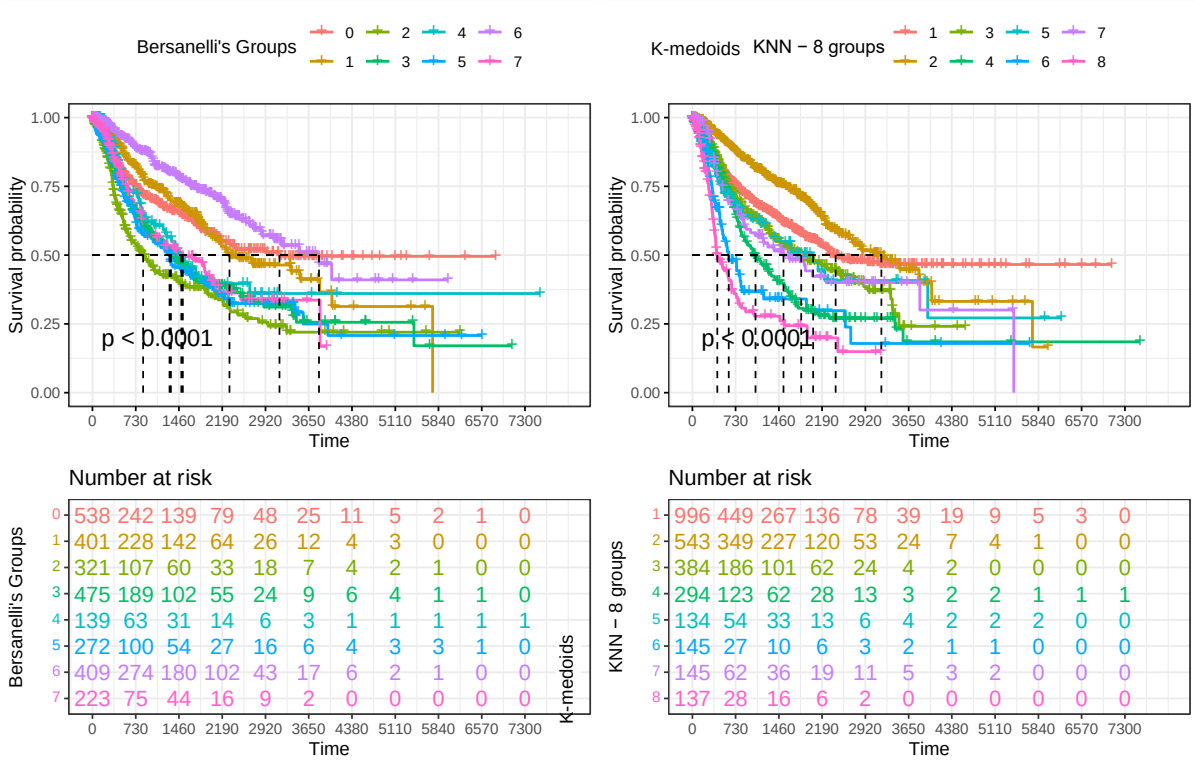
A NEW APPROACH

K = 4 → K = 8

Adopt Bersanelli's ***«clinical rationale»:*** start with a clustering of **8 groups**.

Compare the results with the original genomic groups.

K = 8 vs BERSANELLI



Better separation of our survival curves.

Unbalanced numerosity of certain groups.

K = 8 vs BERSANELLI

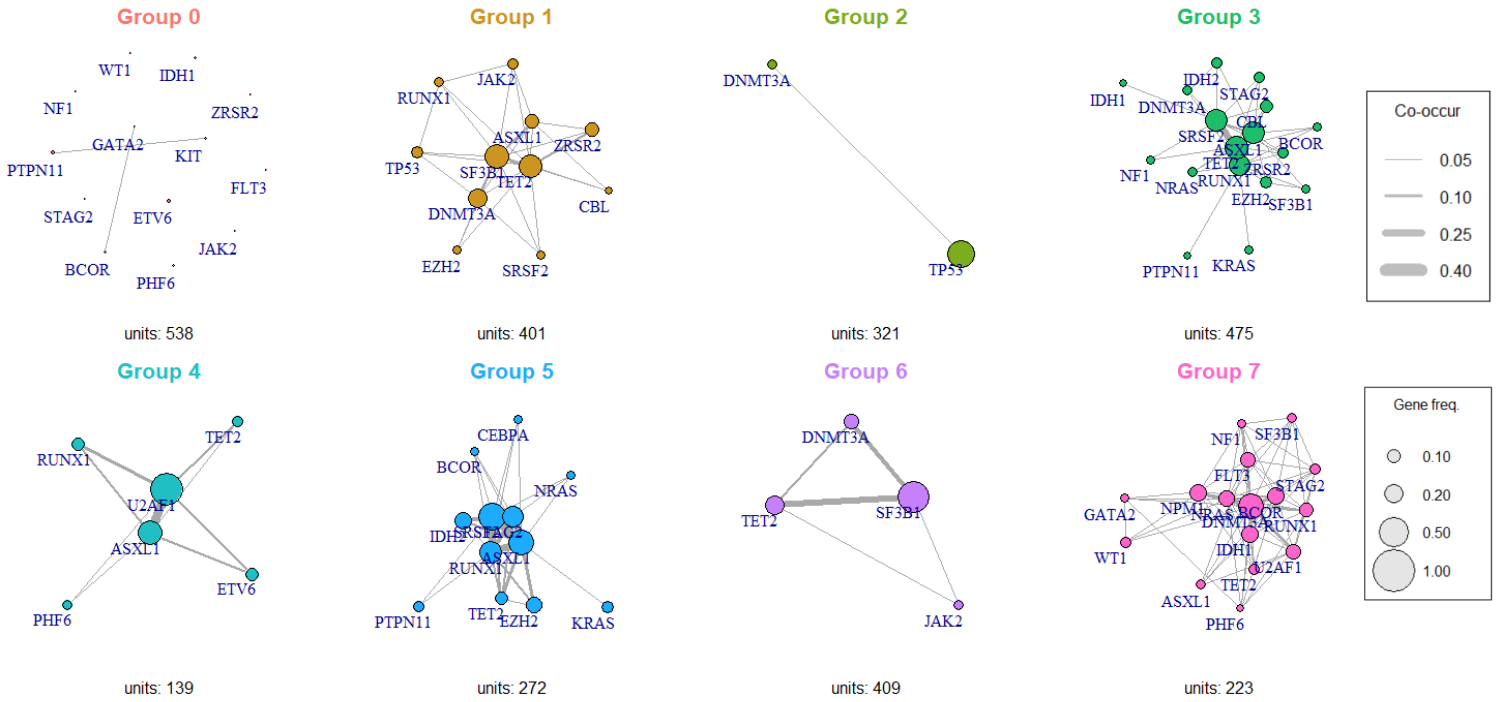
Our Groups \ Bersanelli Groups	Group 0	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Total
Group 1	538	27	115	58	59	58	61	80	996
Group 2	0	168	12	28	0	0	318	17	543
Group 3	0	110	14	171	13	32	26	18	384
Group 4	0	17	14	94	57	98	0	14	294
Group 5	0	5	7	39	4	71	0	8	134
Group 6	0	7	15	76	6	13	0	28	145
Group 7	0	60	18	6	0	0	4	57	145
Group 8	0	7	126	3	0	0	0	1	137
Total	538	401	321	475	139	272	409	223	

Some groups have a **good correspondence** with Bersanelli's ones.

Some Bersanelli's groups are mixed into our **group 1**.

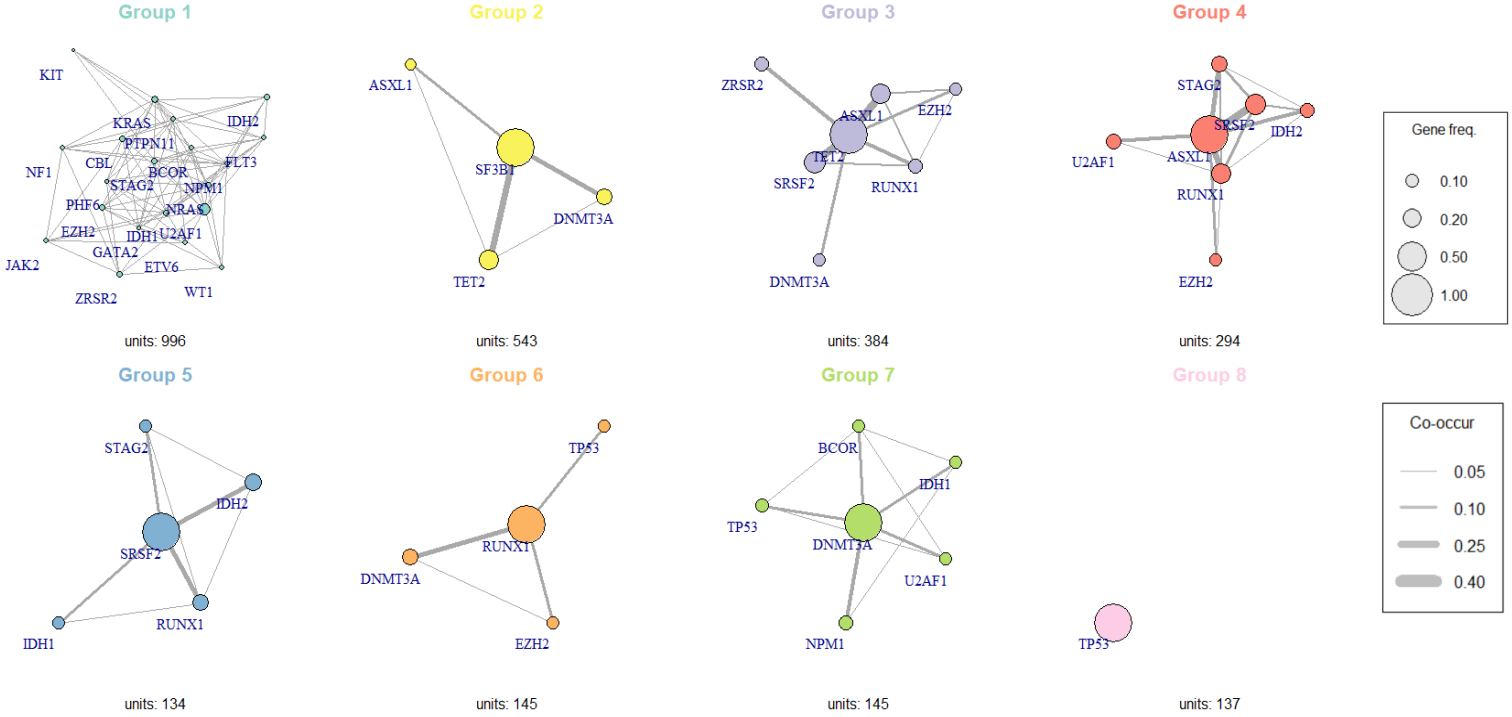
PATTERN ANALYSIS

BERSANELLI

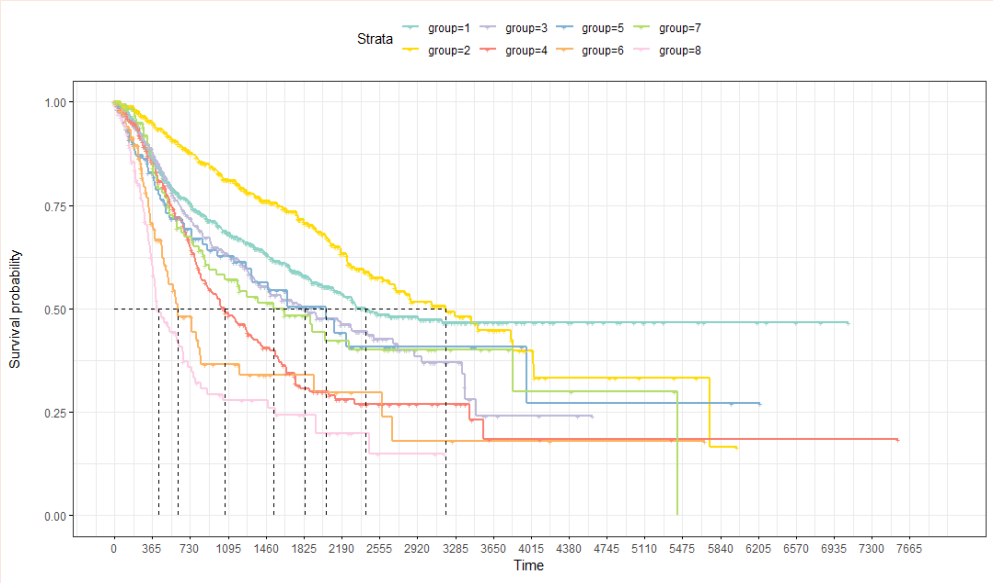


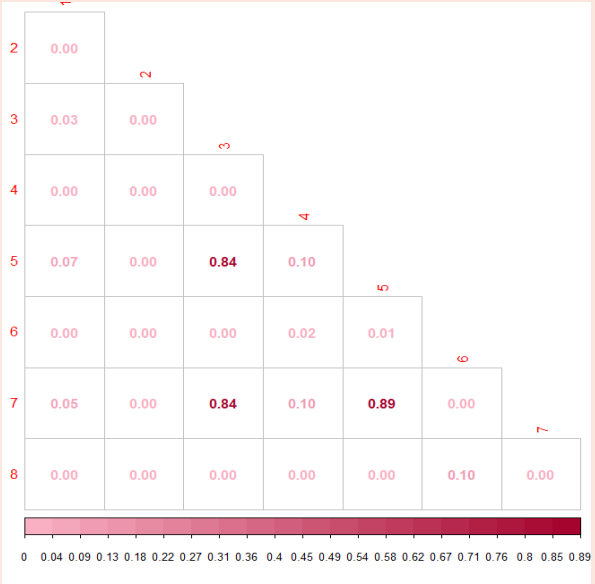
PATTERN ANALYSIS

K = 8



GROUPS AGGREGATION

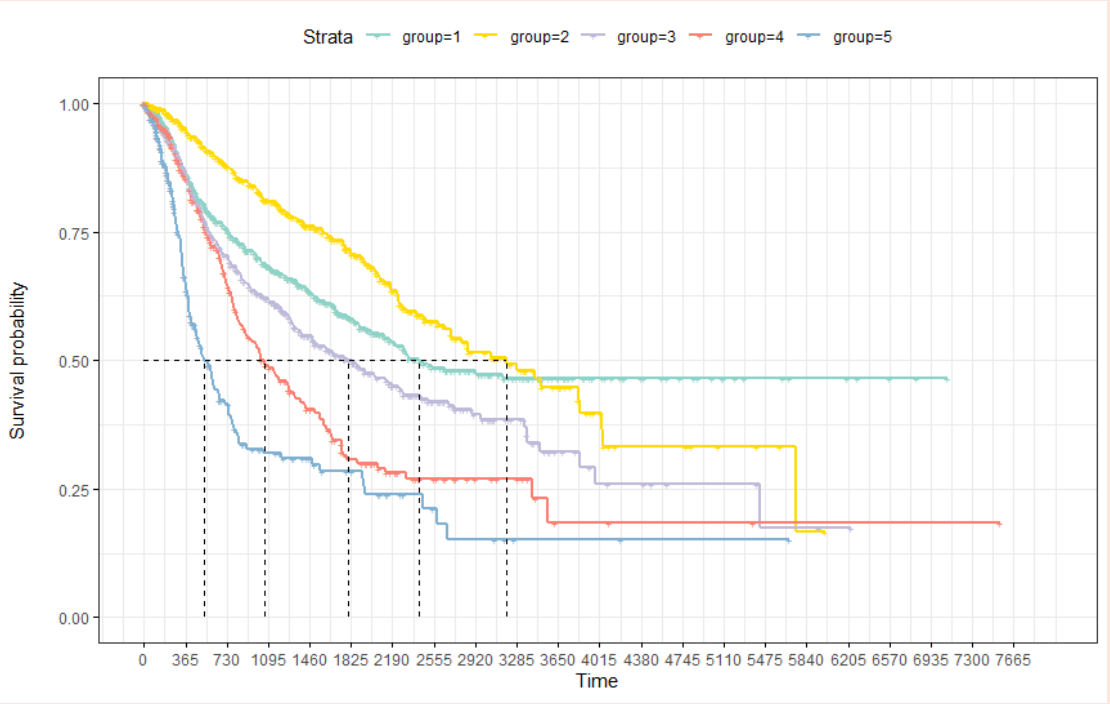




Many survival curves are **overlapping.**

High p-values of **log-rank test** for many pairs.

K = 5



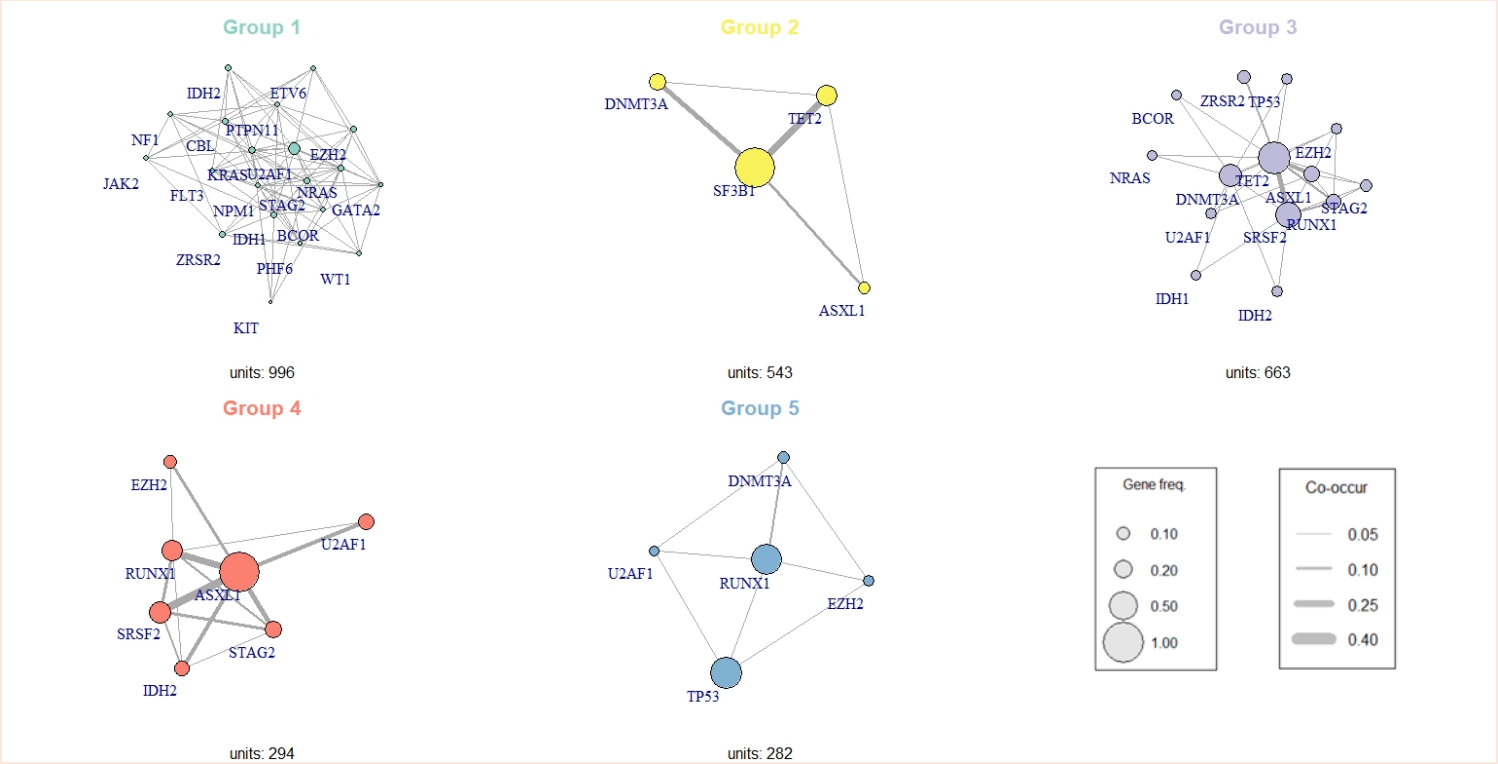
Survival curves are **very well separated**.

No significant evidence to further aggregate the groups.

Group 1: 996 patients
Group 2: 543 patients
Group 3: 663 patients
Group 4: 294 patients
Group 5: 282 patients

PATTERN ANALYSIS

K = 5



GROUPS ANALYSIS

1

- BIGGEST GROUP (~1000 PATIENTS)
- PATIENTS WITH FEW GENES MUTATIONS (usually 0 or 1)
- ASSOCIATED WITH BERSANELLI'S GROUP 0

2

- NOT TRASPLANTED PATIENTS
- NO COMPLEX KARYOTYPE
- HIGH VALUES OF PLATELETS
- LOW LEVEL OF RISK
- ASSOCIATED WITH BERSANELLI'S GROUPS 1 AND 6

3

- SECOND HIGHEST NUMEROSITY (663 PATIENTS)
- LESS EVIDENT GENOMIC PATTERN
- ASSOCIATED WITH BERSANELLI'S GROUPS 1 AND 3

4

- HIGH NUMBER OF GENES MUTATIONS
- INTERMEDIATE-HIGH LEVEL OF RISK
- LOW SURVIVAL EXPECTATIONS
- ASSOCIATED WITH BERSANELLI'S GROUP 3, 4 AND 5

5

- SIGNIFICANT PRESENCE OF COMPLEX KARYOTYPE
- LOW VALUES OF PLATELETS
- VERY HIGH LEVEL OF RISK
- ASSOCIATED WITH BERSANELLI'S GROUP 2

BERSANELLI VALIDATION

Using a more **direct** and **simple** approach, we obtain **comparable results** in terms of grouping with respect to Bersanelli.

The grouping structure can be further **simplified** (*Occam's razor*) still obtaining coherent results.

SURVIVAL ANALYSIS

COX SURVIVAL MODEL

The Cox model explores the relationship between the **survival** of a patient *i* and some **covariates** $\boldsymbol{X_i}$

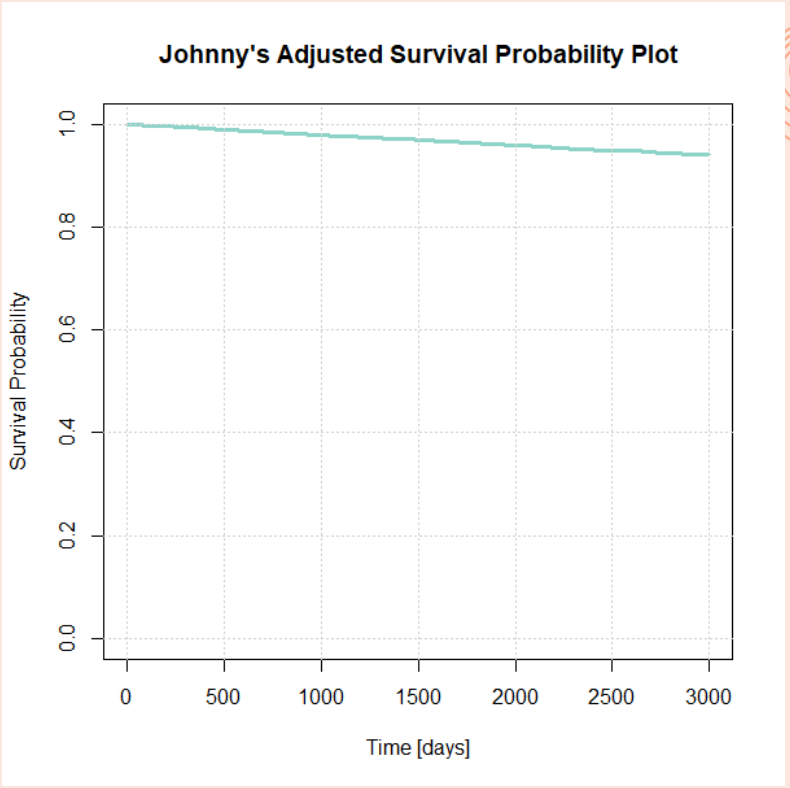
$$h_i(t|\,\boldsymbol{X_i}) = h_0(t)\, exp\{\boldsymbol{X_i^T \beta}\}$$

From the **hazard**/risk function $h(t)$ we can estimate the **survival curve** of each patient, given its features.

$\boldsymbol{\beta}$ coefficients are estimated using **partial likelihood** maximization, similarly to linear regression

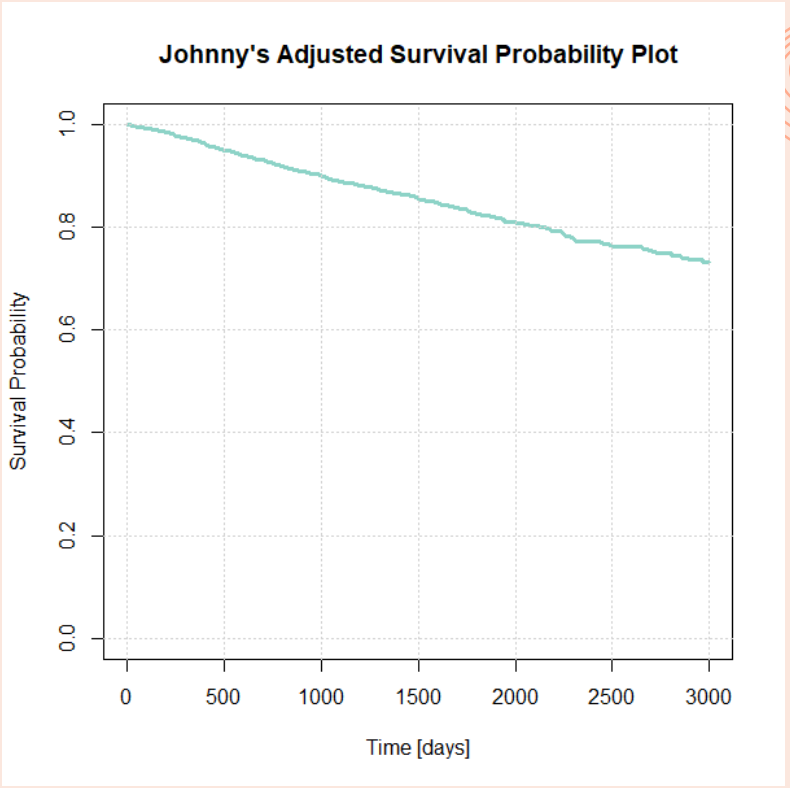
CONSIDER JOHNNY

- **Age at diagnosis:** 25
- **Gender:** Male
- **Karyotype:** Simple
- **AML evolution:** No
- **Transplanted:** No



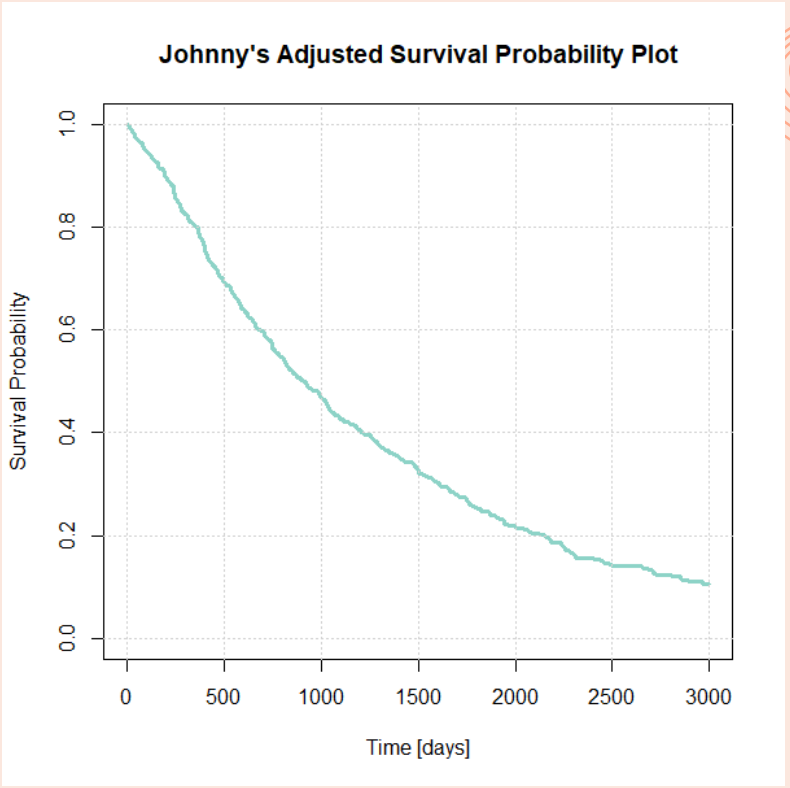
CONSIDER JOHNNY

- **Age at diagnosis:** 25
- **Gender:** Male
- **Karyotype:** Complex
- **AML evolution:** No
- **Transplanted:** No



CONSIDER JOHNNY

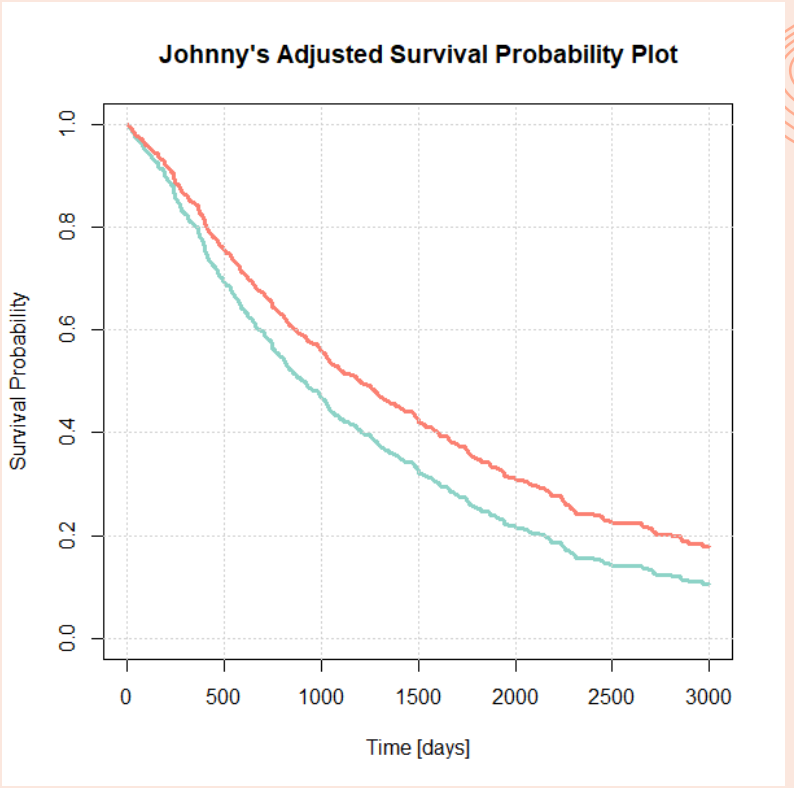
- **Age at diagnosis:** 25
- **Gender:** Male
- **Karyotype:** Complex
- **AML evolution:** Yes
- **Transplanted:** No



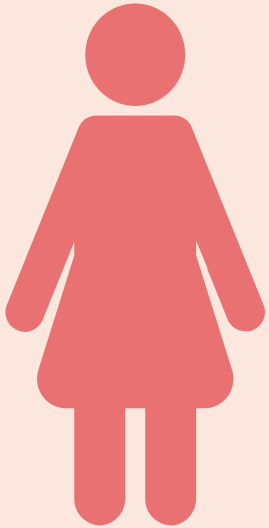
CONSIDER JOHNNY

- **Age at diagnosis:** 75
- **Gender:** Male
- **Karyotype:** Simple
- **AML evolution:** Yes
- **Transplanted:** No

- **Age at diagnosis:** 25
- **Gender:** Male
- **Karyotype:** Complex
- **AML evolution:** Yes
- **Transplanted:** No



CONSIDER EMILY

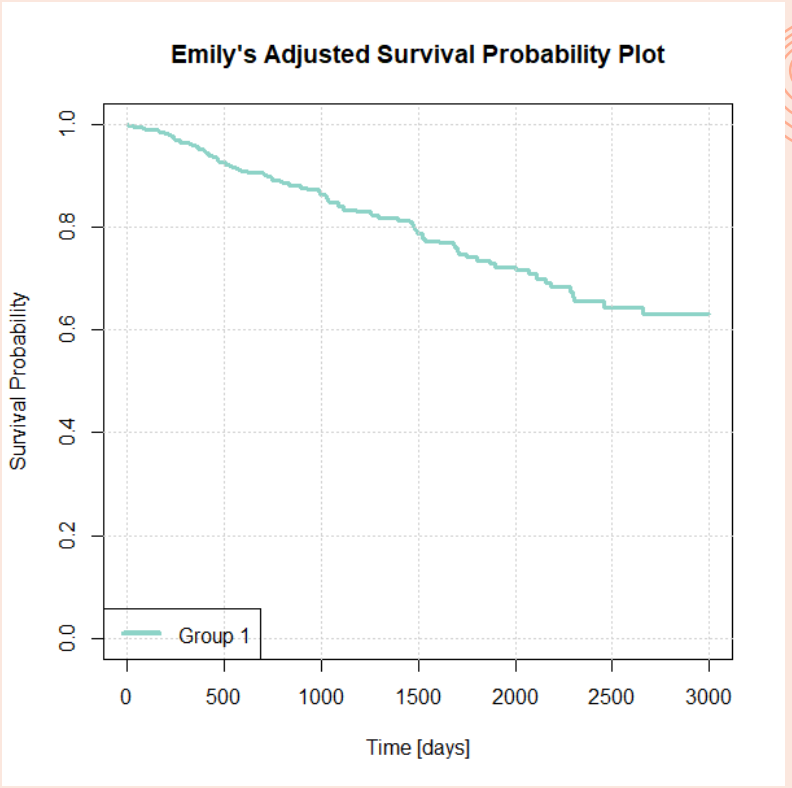


- **Age at diagnosis:** 55
- **Gender:** Female
- **Karyotype:** Complex
- **AML evolution:** No
- **Transplanted:** No
- **Blood values:** Average

CONSIDER EMILY

GROUP 1

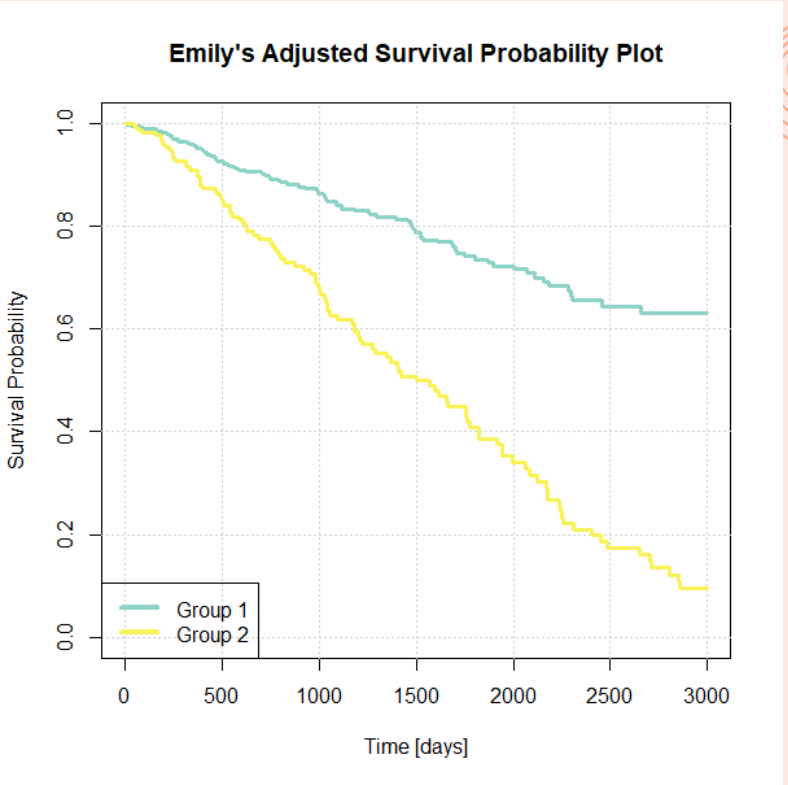
- **Age:** + 7% risk per year
- **Gender:** Not significant
- **Complex Karyotype:** + 136% risk
- **AML evolution:** > + 1000% risk
- **AML • Age:** - 5% risk per year
- **Blood values:** Not significant



CONSIDER EMILY

GROUP 2

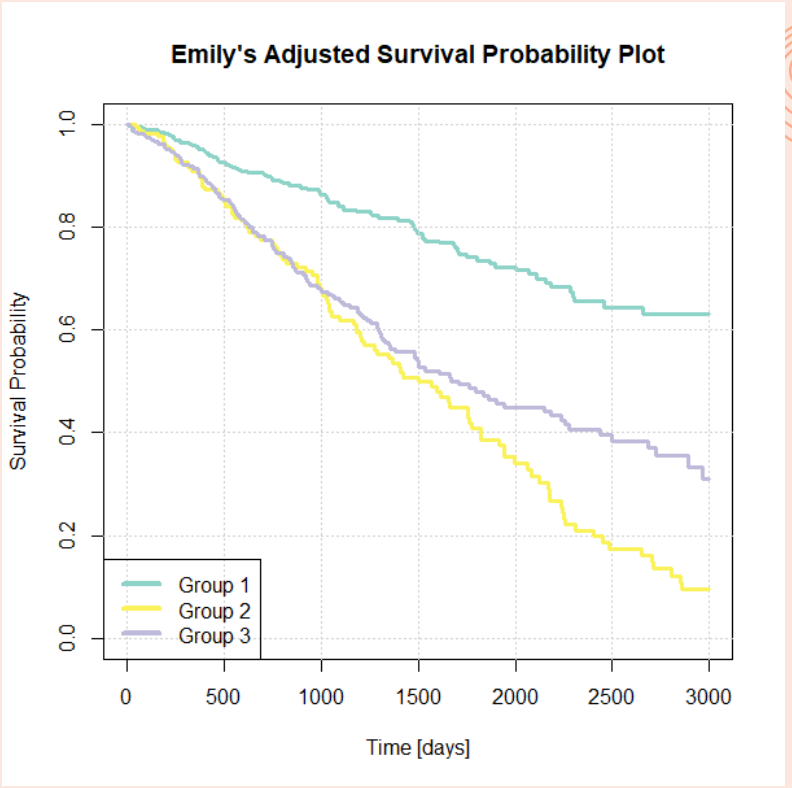
- **Age:** + 9% risk per year
- **Gender:** +73% risk if male
- **Complex Karyotype:** > + 1000% risk
- **AML evolution:** + 112% risk
- **Blood values:** Significant



CONSIDER EMILY

GROUP 3

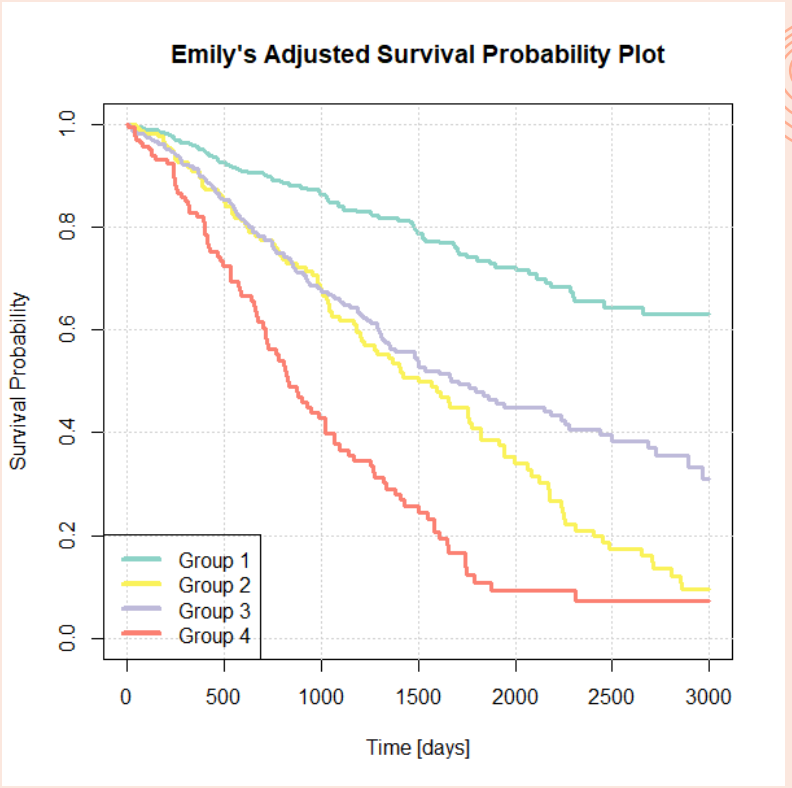
- **Age:** + 5% risk per year
- **Gender:** Not significant
- **Complex Karyotype:** + 182% risk
- **AML evolution:** 48% risk
- **Blood values:** Significant



CONSIDER EMILY

GROUP 4

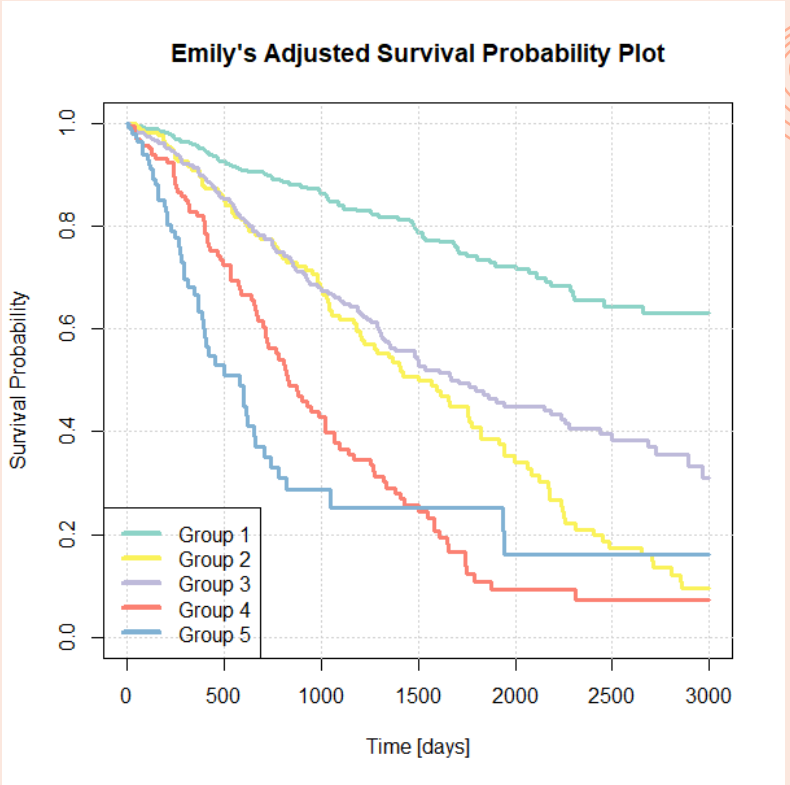
- **Age:** + 4% risk per year
- **Gender:** Not significant
- **Complex Karyotype:** + 190% risk
- **AML evolution:** Not significant
- **Blood values:** Significant



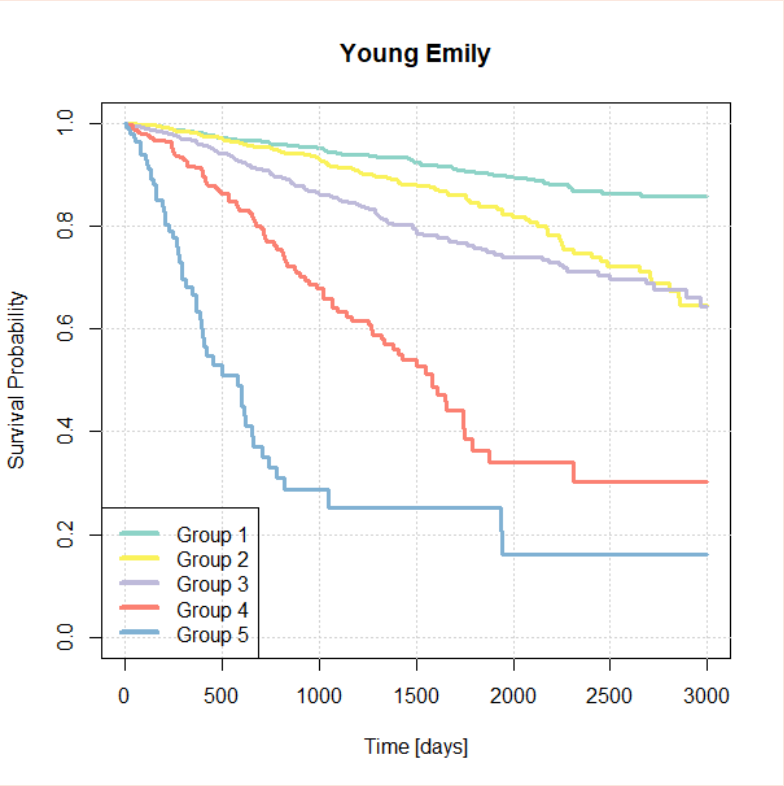
CONSIDER EMILY

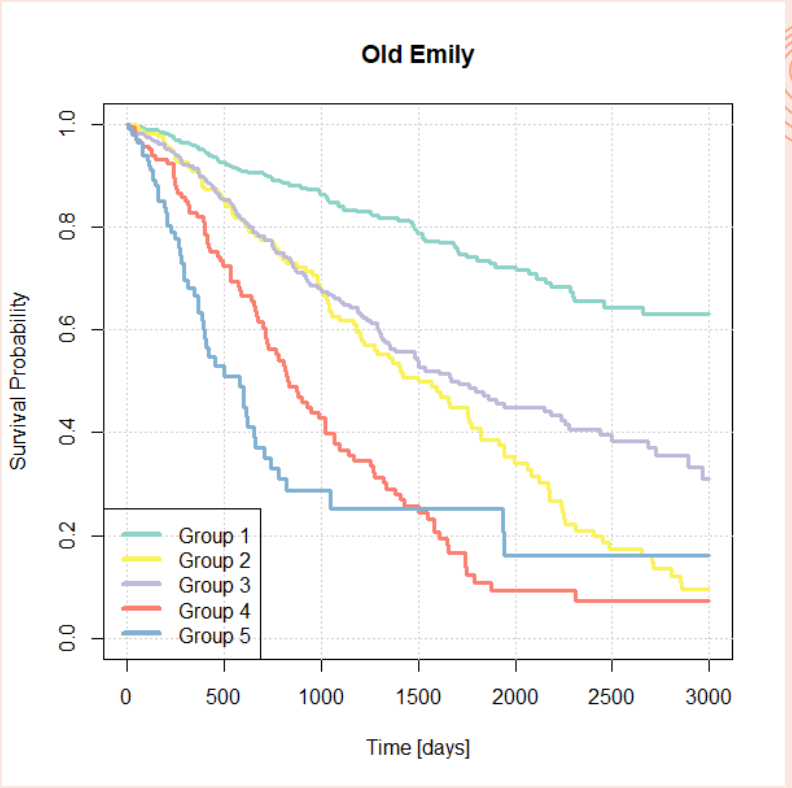
GROUP 5

- **Age:** Not significant
- **Gender:** Not significant
- **Complex Karyotype:** + 196% risk
- **AML evolution:** Not significant
- **Blood values:** Significant



YOUNGER EMILY





CONCLUSIONS

Our objective was to create **clusters** of patients based on their **genomic** features and exploit them to better understand **survival dynamics**.

As we can see from Emily case, **risk factors** impact differently on patients depending on their **genomic group**.

Knowing the genomic profile of a **new patient** can be of enormous importance in developing an effective **customized treatment**.