Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per le Tecnologie e le Scienze

RELAZIONE FINALE

LEUCEMIA MIELOIDE ACUTA: INFLUENZA DEI GENI BAALC E ERG SULLA SOPRAVVIVENZA DEI PAZIENTI

Relatore Prof. Giuliana Cortese
Dipartimento di Scienze Statistiche

Laureando: Mauro Corbi
Matricola N 1147658

Anno Accademico 2019/2020

Ringraziamenti

A conclusione di questo percorso, vorrei ringraziare la mia relatrice, la professoressa Giuliana Cortese, che mi ha aiutato nella stesura di questo elaborato attraverso la sua guida, i suoi preziosi consigli e la disponibilità a chiarire qualsiasi mio dubbio tempestivamente.

Un ringraziamento anche ai miei genitori e mia sorella, i quali mi hanno supportato non solo durante tutto il percorso universitario, ma anche durante tutta la vita, aiutandomi ad affrontare tutte le difficoltà incontrate.

Ringrazio in particolar modo Federica, che mi è stata accanto ed è riuscita sempre a tranquillizzarmi quando l'ansia prendeva il sopravvento.

Indice

Capitolo 1

Introduzione

In Italia vengono registrati circa 8000 nuovi casi di leucemia ogni anno e un quarto di essi può essere ricondotto ad un particolare tipo di leucemia, la leucemia mieloide acuta (AML o LMA)[1]. Tale malattia è caratterizzata dalla proliferazione di cellule staminali ematopoietiche nel midollo, che finiscono con il riversarsi nel sangue provocando danni anche ad altri tessuti.

Questa malattia è più comune negli adulti di età superiore ai 60 anni. Tuttavia, non va ignorato l'impatto devastante che la leucemia mieloide acuta può avere su pazienti più giovani, e il seguente studio si focalizza proprio su individui di età inferiore ai 60 anni. Data la natura della malattia, può essere di interesse valutare quali fattori ne influenzano il decorso, in particolar modo quelli che giocano un ruolo importante nel tempo di sopravvivenza dei pazienti che ne sono affetti. Nel seguente studio viene posta l'attenzione su due geni, BAALC e ERG, perché si ritiene che il loro livello di espressione riesca ad influenzare signficativamente l'esito della malattia. I valori dell'espressione di BAALC e ERG sono stati rilevati in tre diversi istanti di tempo: al momento della diagnosi, dopo la chemioterapia a induzione e una volta che la malattia viene dichiarata in remissione. Poiché le misurazioni sono rilevate in modo longitudinale, dovranno essere studiate opportunamente come tali nei modelli di regressione.

Dopo un'attenta analisi esplorativa e la costruzione di un modello statistico di Cox per spiegare il fenomeno in base ad alcune variabili esplicative, nel seguente studio si è giunti alla conclusione che l'espressione dei geni BAALC e ERG influenza la sopravvivenza dei pazienti, con un'aspettativa di vita più lunga per

[1] https://www.airc.it/cancro/informazioni-tumori/guida-ai-tumori/leucemia-mieloide-acuta

coloro che hanno livelli bassi di tale espressione. Influenzano l'esito della malattia anche il sesso del paziente, favorendo gli uomini rispetto alle donne, e l'aver subito una ricaduta della malattia, che riduce le possiblità di sopravvivenza del 75%.

Capitolo 2

Descrizione del caso di studio

2.1 Leucemia e Leucemia Mieloide Acuta

La leucemia è un tumore del sangue causato dall'eccessiva proliferazione delle cellule staminali ematopoietiche, ovvero le cellule che una volta specializzate diventano cellule del sangue[1]. Nei pazienti che soffrono di leucemia tali cellule non completano il processo di maturazione come dovrebbero; iniziano quindi ad accumularsi nel midollo, rimpiazzando progressivamente il tessuto responsabile della produzione di cellule sanguigne alterando la produzione delle stesse. Le cellule non ancora mature si riversano poi nel sangue, raggiungendo e danneggiando anche altri organi.

Le leucemie si dividono in quattro tipi principali:

— **Leucemie acute**: sono caratterizzate da un rapido progresso e dall'accumulo di cellule giovani nel midollo osseo e nel sangue periferico

— **Leucemie croniche**: la progressione di questo tipo di leucemie è più lenta e prevedono l'accumulo di cellule mature nel midollo osseo e nel sangue periferico

— **Leucemie linfatiche**: la malattia nasce dalle cellule dei lonfoidi, le quali maturano per dare origine ai globuli bianchi

— **Leucemie mieloidi**: la malattia nasce dalle cellule mieloidi, le quali

[1] https://www.airc.it/cancro/informazioni-tumori/guida-ai-tumori/leucemia-mieloide-acuta

maturano per dare origine a globuli rossi, piastrine e anche alcuni tipi di globuli bianchi

Il caso in studio è incentrato sulla **Leucemia Mieloide Acuta (AML)**[2], malattia che origina dalle cellule mieloidi ed è caratterizzata da una rapida progressione.

Nei pazienti affetti da tale malattia si verifica una mutazione nel DNA di una cellula del midollo in fase di sviluppo. Tale cellula diventa una cellula leucemica, quindi inizia a riprodursi in maniera incontrollata moltiplicandosi anche miliardi di volte[3]. Le cellule da essa generate sono definte "blasti leucemici" e non funzionano correttamente, essendo originate da una cellula mutata e non avendo completato il processo di maturazione.

La leucemia mieloide acuta è una malattia eterogenea, il che vuol dire che durante il suo corso vengono generate delle cellule leucemiche diverse dal punto di vista genetico; tali cellule reagiscono in modo diverso al trattamento e riscono talvolta anche a resisterne maggiormente.

Data la complessità della malattia, nella maggior parte dei casi non si è ancora in grado di fornire una prognosi corretta per i pazienti che ne soffrono. Da alcuni studi, tuttavia, si ha evidenza che globuli bianchi, mutazione dei geni NPM1 e CEBPA e fenomeni ereditari influiscono sulla sopravvivenza del soggetto. Si ritiene anche che duplicazioni tandem del gene FLT3 possano interferire con la corretta prognosi.

Evidenza dell'influenza genica sull'esito della malattia è presente anche con i geni **BAALC** (Brain And Acute Leukemia, Cytoplasmatic) e **ERG** (ETS - Related Gene) che, quando vengono espressi più del dovuto, portano a conseguenze meno positive per i pazienti. Tuttavia il motivo del rapporto tra l'espressione di BAALC e ERG e l'esito della malattia è ancora incerto, data la scarsità di studi prospettici a riguardo.

[2] https://www.airc.it/cancro/informazioni-tumori/guida-ai-tumori/leucemia-mieloide-acuta

[3] https://www.lls.org/content/what-is-aml

2.2 Caso di studio

Il presente studio è volto proprio a valutare l'influenza dei valori dell'espressione di BAALC e ERG, insieme ad altri fattori, sull'esito della malattia in pazienti di età inferiore ai 60 anni affetti da leucemia mieloide acuta.
Ciascun paziente nello studio è stato sottoposto a chemioterapia a induzione dopo aver assunto una dose da 60 mg/m2 al giorno per tre giorni di daunorubicina, un antibiotico tumorale, e per altri sette giorni 100 mg/m2 di citradina, un agente chemioterapico. In seguito alla chemioterapia viene fatta un'analisi del midollo per valutarne la morfologia, utilizzando la tecnica della citometria a flusso o dei marcatori molecolari; se dopo tale analisi la malattia viene considerata in remissione, al paziente viene prescritta una terapia di consolidamento che consiste nell'assumere 12-18 mg/m2 di citradina ad alto dosaggio (Hi-DAC).
Trascorsi 21-28 giorni dalla prima terapia di consolidamento viene ripetuta l'analisi del midollo per monitorare nuovamente i parametri di interesse. La terapia si considera terminata quando i pazienti hanno ricevuto due o più terapie di citradina ad alto dosaggio oppure, ove possibile, sono stati sottoposti a trapianto di midollo allogenico.
Dal momento che lo studio è incentrato sul rapporto tra espressione genica ed esito della malattia, sono stati rilevati i valori dell'espressione genica di BAALC e ERG in tre diversi intervalli di tempo:

— Tra il giorno della diagnosi e 30 giorni dopo;

— Tra il giorno della chemioterapia e 30 giorni dopo;

— Tra il giorno in cui la malattia è stata dichiarata completamente in remissione fino all'ultimo aggiornamento sullo stato di salute del paziente.

Capitolo 3

Metodi usati nell'analisi

Il capitolo dedicato ai metodi utilizzati nel corso dell'analisi sarà strutturato in due parti, la prima relativa alle tecniche utilizzate nell'analisi esplorativa e la seconda alla creazione del modello.

3.1 Metodi nell'analisi esplorativa

Cutpoint ottimale

Dal momento che le variabili BAALC e ERG contengono valori continui è bene poter essere in grado di distinguere quando un valore sia alto o basso. Per separare in questo modo i valori delle variabili di interesse verrà utilizzata la funzione *surv_cutpoint()* del pacchetto **survminer**.

Tale funzione è basata sul concetto di **maximally selected rank statistic**[1]: si consideri un predittore X e una variabile risposta Y. Viene ipotizzata l'esistenza di un cutpoint μ tale da dividere i valori di X in due gruppi; chiamando con x il generico valore di X, quindi $x \in X$, il primo gruppo è tale da contenere i valori $x \leq \mu$, mentre il secondo contiene i valori $x > \mu$. È possibile formulare l'ipotesi di indipendenza tra X e Y come:

$$H_0 : P(Y \leq y | X \leq x) = P(Y \leq y | X > x)$$

[1] Hothorn, Torsten and Lausen, Berthold, *On the Exact Distribution of Maximally Selected Rank Statistics* (February 2002). Science Direct Working Paper No S1574-0358(04)70152-5, Available at SSRN: https://ssrn.com/abstract=3133711

e tale ipotesi può essere verificata utilizzando il valore assoluto di un'appropriata statistica a due campioni basata sui ranghi indicata con $|S_\mu|$. Il massimo di tale quantità è dato dal valore μ che meglio separa i due gruppi; come stima di μ viene considerata quindi

$$\hat{\mu} = \arg\max_{\mu} |S_\mu|$$

3.2 Modello semiparametrico di Cox con covariate tempo-dipendenti

Per modellare la funzione di rischio associata ad un evento è possibile utilizzare una delle diverse classi di modelli di sopravvivenza:

— Modello non parametrico

— Modello parametrico

— Modello semiparametrico

Spesso è preferibile usare il modello semiparametrico dal momento che gode di notevole flessibilità grazie alla parte non parametrica, adattandosi meglio ai dati, e permette di modellare l'effetto delle covariate grazie alla parte parametrica. Si assume che la censura sia a destra, che sia non informativa e che X sia la variabile aleatoria che misura il tempo all'evento. I dati sono rappresentati dalla tripla (T_i, δ_i, Z_i), dove $T_i = \min(X_i, C_i)$ cioè il minimo tra il tempo all'evento e il tempo di censura, $\delta_i = I(X_i \leq C_i)$, variabile indicatrice che indica se c'è stato un evento, e Z_i è l'insieme delle covariate misurate per l'i-esimo paziente.
La funzione di rischio viene stimata come:

$$h(X|Z) = h_0(x) \exp\{\beta^T Z\}$$

dove $h_0(x)$ è la funzione di rischio di riferimento non parametrica, mentre $\exp\{\beta^T z\}$ modella l'effetto delle covariate.

È possibile ricavare la funzione di rischio cumulato:

$$\begin{aligned} H(X|Z) &= \int_0^x h(s|t)ds \\ &= \int_0^x h_0(s)\exp\{\beta^T z\}ds \\ &= \exp\{\beta^T z\}\int_0^x h_0(s)ds \\ &= \exp\{\beta^T z\}H_0(s) \end{aligned}$$

Mentre la funzione di sopravvivenza associata è:

$$\begin{aligned} S(X|Z) &= \exp\{-H(X|S)\} \\ &= \exp\{-H_0(x)e^{\beta^T Z}\} \\ &= \exp\{-H_0(X)\}^{e^{\beta^T Z}} \\ &= S_0(x)^{e^{\beta^T Z}} \end{aligned}$$

È di interesse ricavare le stime dei parametri β relativi all'effetto delle esplicative. Per farlo si considera la funzione di verosimiglianza:

$$\begin{aligned} L(\beta, h_0(t)) &\propto \prod_{i=1}^{n} h(t_i|z_i)^{\delta_i} S(t_i|z_i) \\ &= \prod_{i=1}^{n} h_0(t_i|z_i)^{\delta_i}\exp\{\beta^T z_i\}^{\delta_i}\exp\{-H_0(t_i|z_i)e^{\beta^T z_i}\} \end{aligned}$$

Considerando $h_0(t)$ come parametro di disurbo e costruendo la verosimiglianza profilo per β si ottiene la verosimiglianza parziale di Cox:

$$PL(\beta) = \prod_{i=1}^{n} \frac{\exp\{\beta^T z_{(j)}\}}{\sum_{l\in R(t_i)}\exp\{\beta^T z_l\}}$$

Nel caso di studio corrente l'interesse è rivolto a come le variabili esplicative, che cambiano nel tempo, riescano a influenzare la risposta. Nella funzione di rischio si evidenzierà ora la dipendenza delle esplicative dal tempo:

$$h(X|Z(t)) = h_0(x)\exp\{\beta^T Z(t)\}$$

Per trattare le esplicative tempo-dipendenti è richiesta la trasformazione del dataset in una forma particolare: si considera ciascun soggetto come realizzazione di una Poisson molto lenta, e in quest'ottica i dati censurati non vengono considerati come *dati mancanti*, ma piuttosto come *conteggi non ancora avvenuti*[2]. I dati devono essere disposti nel seguente modo per consentire l'analisi:

ID paziente **(start, stop]** **status** $\boldsymbol{x_1}$ $\boldsymbol{x_2}$ $\boldsymbol{x_3}$ ....

Perciò, ad esempio, le variabili tempo-dipendenti per il paziente con codice identificativo 94 verranno codificate nel seguente modo:

ID paziente	**start**	**stop**	**status**	**BAALC**	**ERG**	...
94	0	20	0	1238.00	102.00	...
94	20	41	0	7931.00	58.00	...
94	41	284	1	56.20	69	...
⋮	⋮	⋮	⋮	⋮	⋮	

Tabella 3.1: Esempio struttura dataset

Per via di come sono definiti i calcoli interni del modello di Cox in R (o in altri programmi), ciascuna riga relativa allo stesso paziente non viene considerata singolarmente, correlata con le altre, ma viene gestita nel corretto modo automaticamente, specificando quale colonna viene considerata come codice identificativo di ciascuna unità del dataset (in questo caso la colonna è "ID Paziente").

[2] *Censoring is not "incomplete data", rather, "the geiger counter just hasn't clicked yet."* Therneau T.M.,Grambsch P.M.. *Modeling Survival Data: Extending the Cox Model. Statistics for Biology and Health*, Springer, 2000

Capitolo 4

Applicazione dei metodi

4.1 Analisi esplorativa

Il campione in esame è composto da 152 unità, i pazienti. Per ciascuno di essi sono state rilevate diverse variabili, in modo da ottenere un quadro biologico e genetico il più completo possibile:

- **Age**, età del paziente
- **Gender** , il sesso del paziente (M o F)
- **PBblast**, quantità di blasti nel sangue
- **BMofBlasts**, quantità di blasti nel midollo
- **Date of Diagnosis**, il giorno in cui al paziente è stato diagnosticata la malattia
- **Date Of Induction**, quando il paziente è stato sottoposto a chemioterapia a induzione
- **Date of CR**,il giorno in cui la malattia del paziente è stata considerata completamente in remissione
- **Date of Relapse**, il giorno in cui c'è stata la ricaduta della malattia
- **Date of Death**, il giorno in cui il paziente è deceduto
- **RFS date**, il giorno entro il quale il paziente non ha subito ricadute della malattia

— **Last followup**, ultimo aggiornamento dello stato di salute del paziente

— **OS_months**, il numero di mesi dalla diagnosi del paziente all'ultimo aggiornamento

— **OS_status**, variabile dicotomica che vale 0 se il paziente è sopravvissuto e 1 altrimenti

— **RFS months**, il numero di mesi dalla diagnosi del paziente alla prima ricaduta

— **RFS status**, variabile dicotomica che vale 0 se il paziente ha avuto una ricaduta e 1 altrimenti

— **Cytogenic Risk**, variabile in scala 1:4 che indica il rischio del paziente dal punto di vista citogenetico

— **FLT3ITD**, variabile dicotomica che identifica la presenza o l'assenza di una mutazione genetica del gene FLT3

— **NPM1 TYPE A**, variabile dicotomica che identifica la presenza o l'assenza di una mutazione genetica del gene NPM1

— **CEBPA**, variabile che identifica la presenza di una mutazione genetica del gene CEBPA

— **MLLPTD**, variabile dicotomica che identifica la presenza o l'assenza di una mutazione genetica del gene MLL

— **Genetic Risk**, indica il rischio genetico del paziente

— **BAALC e ERG**, valori dell'espressione dei geni BAALC e ERG rilevati in tre istanti: al momento della diagnosi, dopo la chemioterapia e dopo il momento in cui la malattia è stata dichiarata completamente in remissione

— **Status**, variabile dicotomica che indica con 1 se il paziente non ha subito la ricaduta della malattia, 2 altrimenti

— **Performance status**, variabile in scala 0:3 che indica lo stato di salute del paziente in generale

— **TLCCU**, variabile quantitativa continua che indica il numero totale di leucociti

Vengono riportati di seguito i riassunti delle variabili.

Variabile	**Categorie**	**n**	**%**
Gender	F	51	34%
	M	98	64%
	Censurati	3	2%
Status	0	69	45%
	1	26	18%
	2	54	35%
	Censurati	2	2%
V_cyto_risk	1	43	28%
	2	34	22%
	3	28	19%
	4	44	29%
	Censurati	3	2%
Cytogenic risk	Good	38	25%
	Intermediate	62	41%
	Poor	32	21%
	Censurati	20	13%
OS_status	0	44	29%
	1	73	48%
	Censurati	35	23%
Genetic risk	Fav	45	30%
	Int	39	26%
	Poor	48	31%
	Censurati	20	13%

Variabile	**Categorie**	**n**	%
FLT3ITD	1	25	17%
	2	107	70%
	Censurati	20	13%
NPM1typeA	1	28	19%
	2	104	68%
	Censored	20	13%
Performance status	0	5	3%
	1	92	60%
	2	28	19%
	3	6	4%
	4	1	1%
	Censurati	20	13%
CEBPA	2	51	62%
	biallelic	1	1%
	monoallelic	11	13%
	Censurati	20	24%
MLL_PTD	1	14	9%
	2	116	76%
	Censurati	22	15%
RFS_status	0	39	26%
	1	65	43%
	Censurati	48	31%

Tabella 4.1: Summary variabili qualitative

Variabile	Min	I quartile	Mediana	Media	III quartile	Max	NA
Age	15	20	27	29.88	39	58	3
Pbblast	0	29.5	51	51.62	74.50	97	33
BmofBlasts	12	39.75	66.50	60.65	82	98	33
BAALCB	10	21	65	1009.7	370	77295.0	20
BAALCI	2	18.5	58.5	627.6	211.5	9632	33
BAALCC	0	9	56.2	305.4	261.0	2624	57
ERGB	15	42.5	87.1	271.3	183.3	4261.0	20
ERGI	0	23,75	65.7	398.2	161.5	6154	32
ERGC	0	28.5	55	382.2	310.2	6087	57

Tabella 4.2: Summary variabili quantitative

Si nota che le variabili di interesse BAALCB, BAALCC, BAALCI, ERGB, ERGC e ERGI presentano valori medi superiori ai valori mediani, perciò sono asimmetriche. Volendo rappresentare graficamente tali dati è necessario ricorrere alla trasformazione logaritmica degli stessi, data la presenza di valori anomali molto elevati. Di seguito vengono riportati i boxplot del logaritmo dei valori di BAALC e ERG nei tre intervalli di tempo:

Figura 4.1: Boxplot del logaritmo dei valori di BAALC e ERG nei tre intervalli di tempo

4.1.1 Valore soglia

L'obiettivo dell'analisi è valutare se e come livelli alti e bassi di BAALC e ERG influiscano sull'aspettativa di vita del paziente; è quindi importante saper classificare ciascun valore dell'espressione genica in alto o basso, cioè è necessario conoscere un valore "soglia".

Ci sono due possibili criteri per la scelta del valore divisivo: è possibile utilizzare come valore soglia la mediana, che è la quantità per eccellenza per dividere a metà un insieme di valori specialmente quando sono presenti outliers, oppure può essere calcolato un particolare cutpoint ottimale, cioè quel valore in grado di separare più nettamente le curve di sopravvivenza nei due gruppi. Essendo le strade ugualmente valide, l'analisi verrà effettuata affiancando i risultati ottenuti con i diversi criteri per metterne in luce similiarità e differenze.

In entrambe le situazioni vengono definite variabile dicotomiche come la seguente:

$$y_i = \begin{cases} 0 & \text{se il valore è inferiore o uguale al valore soglia} \\ 1 & \text{se il valore è superiore al valore soglia} \end{cases}$$

per ciascuna misurazione di BAALC e ERG nei vari istanti.

Di seguito Viene calcolato il cutpoint ottimale[1], che verrà utilizzato successivamente per definire le variabili dicotomiche:

```
surv_cutpoint(aml, time = "OS_months", event ="OS_Status",
        variables = c("BAALCB", "ERGB","BAALCI","ERGI","BAALCC","ERGC"))
```

```
        cutpoint statistic
BAALCB       408  4.362787
ERGB         100  5.714334
BAALCI        18  4.892793
ERGI          80  4.050345
BAALCC        22  5.483164
ERGC         220  4.802476
```

[1] Il comando `surv_cutpoint` fa parte del pacchetto *survminer* di R, creato da Alboukadel Kassambara.

Di seguito vengono riportati i grafici di BAALC e ERG nei tre istanti con i valori divisi in alti e bassi secondo i due criteri:

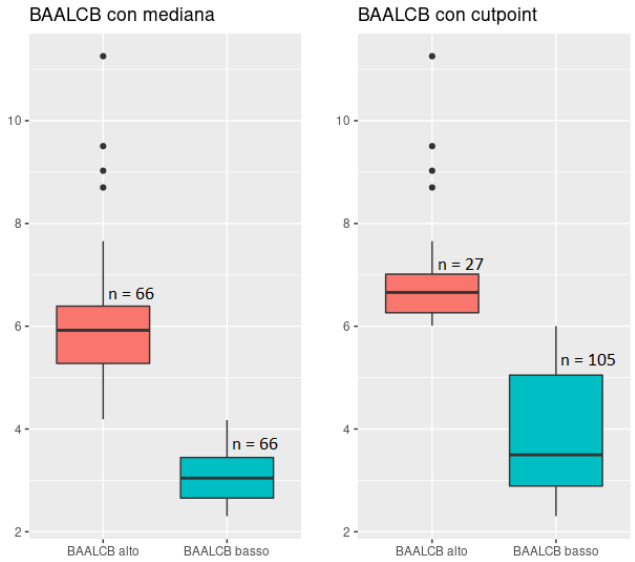

Figura 4.2: Valori di BAALCB, su scala logaritmica, divisi con i due criteri

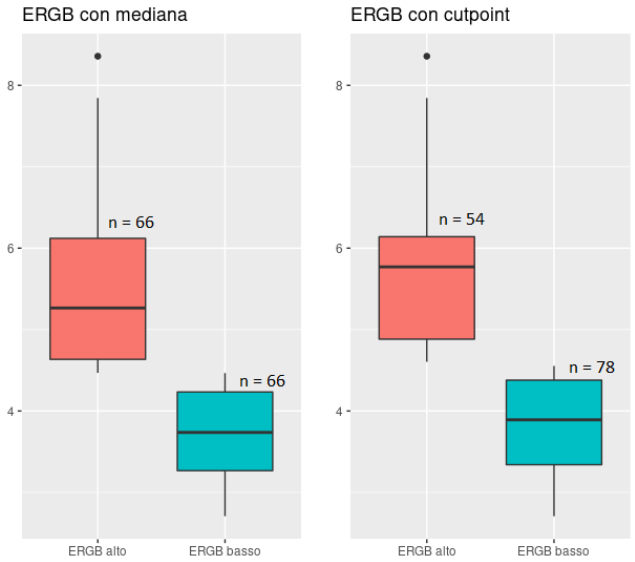

Figura 4.3: Valori di ERGB, su scala logaritmica, divisi con i due criteri

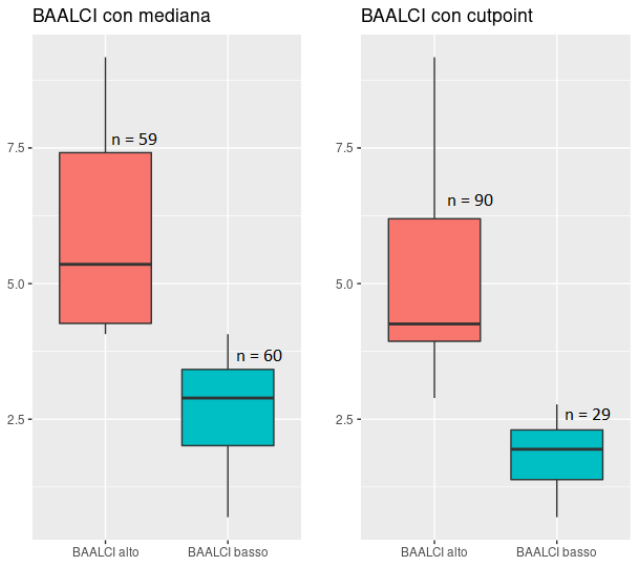

Figura 4.4: Valori di BAALCI, su scala logaritmica, divisi con i due criteri

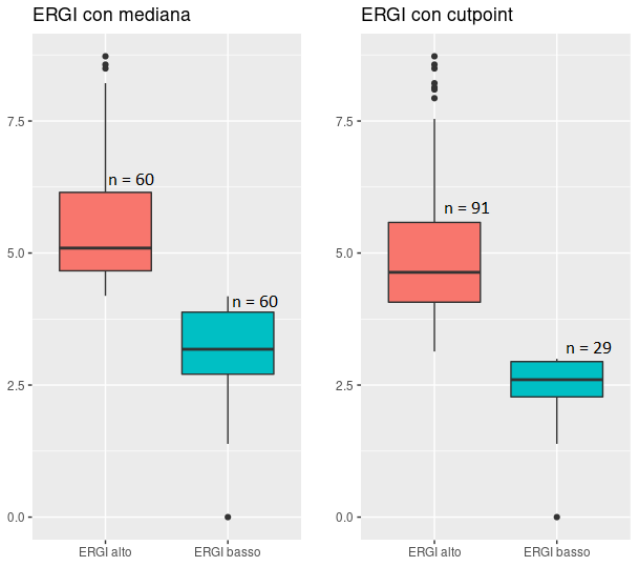

Figura 4.5: Valori di ERGI, su scala logaritmica, divisi con i due criteri

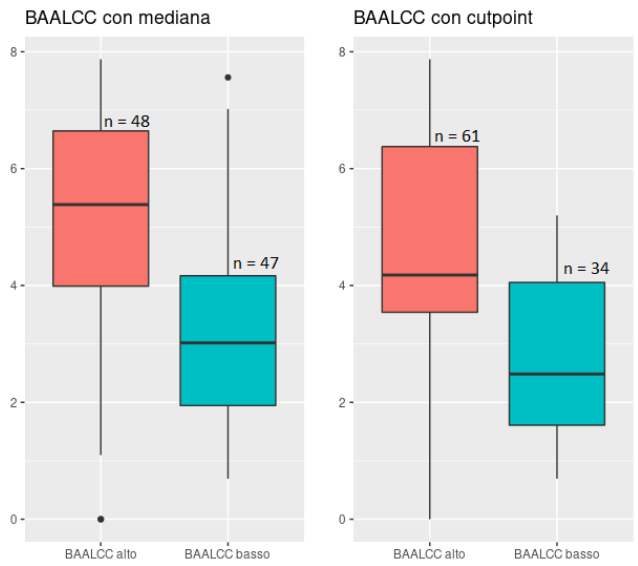

Figura 4.6: Valori di BAALCC, su scala logaritmica, divisi con i due criteri

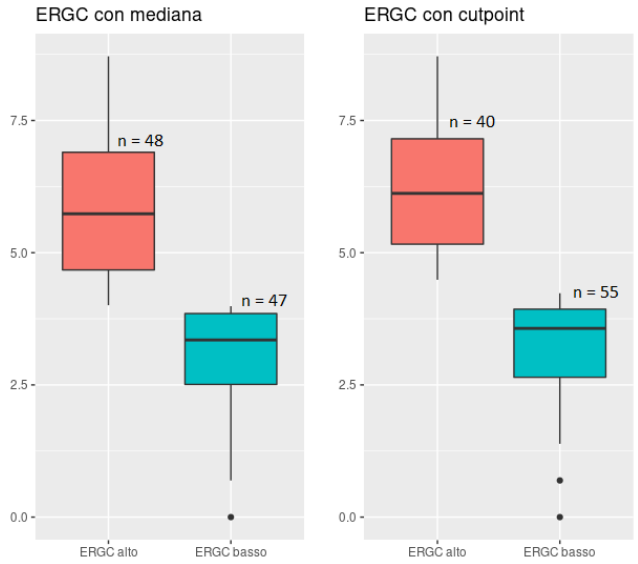

Figura 4.7: Valori di ERGC, su scala logaritmica, divisi con i due criteri

In generale è possibile notare come il cutpoint ottimale consenta di dividere in maniera più netta i valori, ma può rendere le numerosità dei due gruppi molto diverse.

4.1.2 BAALC e ERG alla diagnosi

Le seguenti tabelle identificano i valori clinici e le caratteristiche molecolari dei pazienti in base ai valori di BAALC e ERG misurati al tempo zero, ovvero alla diagnosi. Vengono riportate le tabelle calcolate in base ai due diversi valori soglia, la mediana e il cutpoint ottimale.

Ciascun confronto è affiancato da un test statistico: per le variabili qualitative viene utilizzato il test di Wilcoxon, mentre per le variabili categoriali viene affiancato il pvalue calcolato con il test di Fisher.

BAALC

BAALCB	LOW	HIGH	pvalue
GENDER			0,8552
F	24	22	
M	42	44	
AGE			0,104
Median	29,5	23,5	
Mean	30,83	28,09	
Pbblast			0,34
Median	56,5	49	
Mean	53,85	49,35	
BmofBlast			0,4997
Median	69	64	
Mean	62,39	59,01	
Cytogenic Risk			0,5755
Good Risk	18	20	
Intermediate	34	28	
Poor	14	18	
V_cyto_risk			0,2642
1	19	21	
2	19	10	
3	10	15	
4	18	20	
FLT3ITD			1
1	13	12	
2	53	54	
BAALCB			
Low			
High			
BAALCI			0,0059
Low	38	21	
High	23	37	
BAALCC			0,023
Low	31	16	
High	20	28	

	LOW	HIGH	pvalue
MLLPTD			0,1553
1	4	10	
2	61	65	
Genetic Risk			0,8202
Fav	21	24	
Intermediate	21	18	
Poor	24	24	
CEBPA			0,6306
2	53	50	
Biallelic	2	1	
Monoallelid	11	15	
TLCCU			0,2859
Median	10445	20790	
Mean	36607,73	35897,09	
NPM1 Type A			0,5235
1	16	12	
2	50	54	
Performance status			0,3705
1	48	44	
2	15	13	
3	1	5	
4	0	1	
ERGB			0,0029
Low	42	24	
High	24	42	
ERGCI			0,3611
Low	34	26	
High	28	32	
ERGC			0,02373
Low	31	16	
High	20	28	

Tabella 4.3: Confronti basati sui valori iniziali di BAALC (mediana)

I valori misurati al momento della diagnosi, divisi utilizzando la mediana, appaiono correlati con le altre misurazioni del gene a istanti differenti (BAALCI pvalue 0.005, BAALCC pvalue 0.02) e con le misurazioni di ERG ad eccezione di quella dopo la chemioterapia (ERGB pvalue 0.0029, ERGC pvalue 0.023).

BAALCB	LOW	HIGH	pvalue		LOW	HIGH	pvalue
GENDER			1	**MLLPTD**			0,7331
F	37	9		1	12	2	
M	68	18		2	91	25	
AGE			0,928	**Genetic Risk**			1
Median	26	27		Fav	36	9	
Mean	29,2952	30,111		Intermediate	31	8	
				Poor	38	10	
Pbblast			0,3624	**CEBPA**			0,4801
Median	50	57		2	83	20	
Mean	50,5	56,04		Biallelic	3	0	
				Monoallelid	19	7	
BmofBlast			0,5599	**TLCCU**			0,1073
Median	63,5	71,5		Median	12330	25600	
Mean	60,12	62,53		Mean	33726,46	46075,56	
Cytogenic Risk			0,9612	**NPM1 Type A**			0,4386
Good Risk	30	8		1	24	4	
Intermediate	50	12		2	81	23	
Poor	25	7					
V_cyto_risk			0,8899	**Performance status**			0,1
1	31	9		1	76	16	
2	24	5		2	22	6	
3	19	6		3	3	3	
4	31	7		4	0	1	
FLT3ITD			0,5926	**ERGB**			0,0001
1	19	6		Low	71	7	
2	86	21		High	34	20	
BAALCB				**ERGI**			0,2746
Low				Low	26	3	
High				High	72	19	
BAALCI			0,014	**ERGC**			0,5818
Low	28	1		Low	47	8	
High	69	21		High	32	8	
BAALCC			0,008				
Low	33	1					
High	46	15					

Tabella 4.4: Confronti basati sui valori iniziali di BAALC (cutpoint)

Se invece si considera come valore divisivo il cutpoint ottimale la situazione cambia leggermente: i valori di BAALC alla diagnosi rimangono correlati con le misurazioni dello stesso gene a istanti successivi (BAALCI pvalue 0.014, BAALCC pvalue 0.008), ma adesso sono correlati solo ai valori di ERG rilevati alla diagnosi (ERGB pvalue 0.0001).

Vengono riportate di seguito le curve di sopravvivenza rispetto ai valori di BAALCB

alla diagnosi, ottenute con entrambi i criteri divisivi:

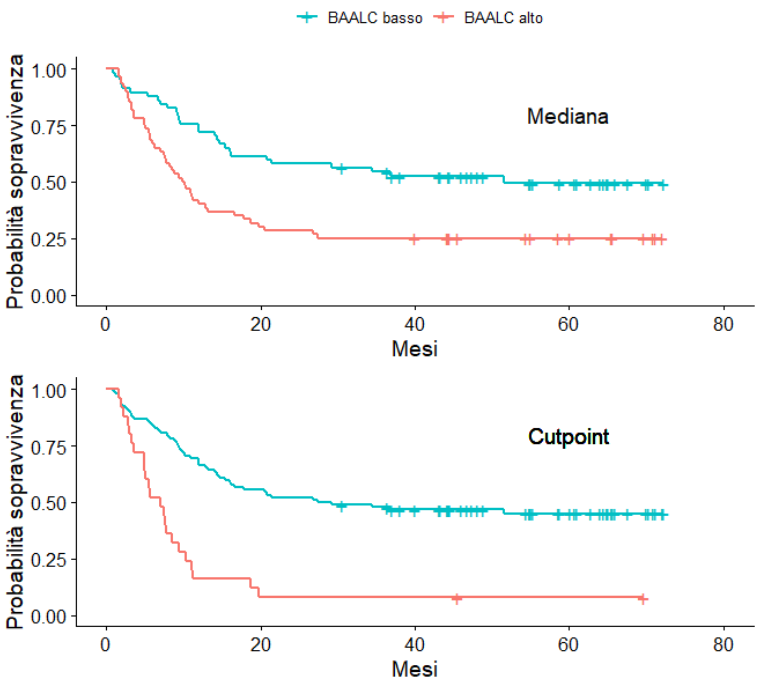

Figura 4.8: Curve di sopravvivenza rispetto al valore iniziale di BAALC

Entrambe le curve di sopravvivenza [2] mostrano un'aspettativa di vita maggiore nei pazienti con livelli bassi dell'espressione del gene BAALC. La differenza è più evidente nel gruppo diviso con il cutpoint ottimale, poiché per dividere in due gruppi i pazienti è stato utilizzato il valore che separa più nettamente le curve di sopravvivenza.

Ai grafici delle stime di Kaplan-Maier viene associato anche un test formale, basato sui ranghi logaritmici. I comandi per eseguire il test per entrambi i metodi di suddivisione dei gruppi sono:

```
> survdiff(formula = Surv(OS_months, OS_Status) ~ baalcb.med)
> survdiff(formula = Surv(OS_months, OS_Status) ~ baalcb.cut)
```

I risultati ottenuti sono riassunti nella seguente tabella:

[2] Le curve di sopravvivenza e i test formali associati sono stati calcolati con il pacchetto *survival* di R, creato da Terry Therneau.

	Valore alto	N	Osservati	Attesi	Pvalue
Mediana	No	57	28	41.9	9e-04
	Sì	60	45	31.1	
Cutpoint	No	92	50	64.06	3e-07
	Sì	25	23	8.94	

Tabella 4.5: Test dei ranghi logaritmici per la variabile BAALCB

Dai test si ricava che la differenza dell'aspettativa di vita nei pazienti con valori iniziali di BAALC diversi sia significativa, indipendentemente dal tipo di valore soglia utilizzato.

È possibile notare come la numerosità nei pazienti suddivisi con il cutpoint sia molto diversa tra i due gruppi (92 per i pazienti con BAALC basso, 25 per quelli con BAALC alto.)

ERG

ERGB	LOW	HIGH	pvalue		LOW	HIGH	pvalue
GENDER			0,3612	**MLLPTD**			0,7783
F	20	26		1	8	6	
M	46	40		2	57	59	
AGE			0,8092	**Genetic Risk**			0,7422
Median	26	26,5		Fav	23	22	
Mean	28,8	30,12		Intermediate	21	18	
				Poor	22	26	
Pbblast			0,606	**CEBPA**			0,2192
Median	51	52		2	51	52	
Mean	52,77	50,41		Biallelic	0	3	
				Monoallelid	15	11	
BmofBlast			0,4575	**TLCCU**			0,2737
Median	69	64		Median	12065	22535	
Mean	62,33	59,01		Mean	281973,73	44307,09	
Cytogenic Risk			0,322	**NPM1 Type A**			0,5235
Good Risk	23	15		1	12	16	
Intermediate	38	34		2	54	50	
Poor	15	17					
V_cyto_risk			0,7434	**Performance status**			0,4127
1	23	17		1	48	44	
2	13	16		2	14	14	
3	12	13		3	1	5	
4	18	20		4	0	1	
FLT3ITD			1	**ERGB**			
1	12	13		Low			
2	54	53		High			
BAALCB			0,0029	**ERGI**			0,043
Low	42	24		Low	38	26	
High	24	42		High	22	34	
BAALCI			0,043	**ERGC**			0,307
Low	37	22		Low	27	20	
High	26	34		High	22	26	
BAALCC			0,024				
Low	30	17					
High	19	29					

v

Tabella 4.6: Confronti basati sui valori iniziali di ERG (mediana)

La tabella mostra una relazione dei valori di ERG alla diagnosi con i valori di ERG misurati dopo la chemioterapia (ERGI pvalue 0.043) e con le misurazioni di BAALC in tutti e tre gli istanti di tempo (BAALCB pvalue 0.002, BAALCI pvalue 0.043, BAALCC pvalue 0.024).

ERGB	LOW	HIGH	pvalue		LOW	HIGH	pvalue
GENDER			0,2679	**MLLPTD**			1
F	24	22		1	8	6	
M	54	32		2	69	47	
AGE			0,6185	**Genetic Risk**			0,1479
Median	26,5	25,5		Fav	30	15	
Mean	29,69	29,12		Intermediate	25	14	
				Poor	23	25	
Pbblast			0,3697	**CEBPA**			0,5815
Median	56,5	50		2	60	43	
Mean	53,43	48,85		Biallelic	1	2	
				Monoallelid	17	9	
BmofBlast			0,9298	**TLCCU**			0,2881
Median	64	67		Median	12665	22105	
Mean	60,63	60,67		Mean	28870,13	46915,7	
Cytogenic Risk			0,1251	**NPM1 Type A**			0,2868
Good Risk	27	11		1	14	14	
Intermediate	36	26		2	64	40	
Poor	15	17					
V_cyto_risk			0,4776	**Performance status**			0,1574
1	27	13		1	56	36	
2	17	12		2	18	10	
3	12	13		3	1	5	
4	22	16		4	0	1	
FLT3ITD			0,2601	**ERGB**			
1	12	13		Low			
2	66	41		High			
BAALCB			0,00013	**ERGI**			0,0021
Low	71	34		Low	25	4	
High	7	20		High	50	41	
BAALCI			0,0003	**ERGC**			0,0197
Low	26	3		Low	39	16	
High	48	42		High	20	20	
BAALCC			<0,0001				
Low	30	4					
High	29	32					

Tabella 4.7: Confronti basati sui valori iniziali di ERG (cutpoint)

Se si considerano i valori di ERG suddivisi in base al cutpoint ottimale rimane la relazione significativa tra ERG e BAALC (BAALCB pvalue 0.00013, BAALCI pvalue 0.0003, BAALCC pvalue < 0.0001) e con il valore di ERG dopo la chemioterapia (ERGI pvalue 0.0021). In aggiunta si rileva anche una relazione significativa con il valore di ERG misurato dopo che la malattia è stata dichiarata in remissione (ERGC pvalue 0.019).

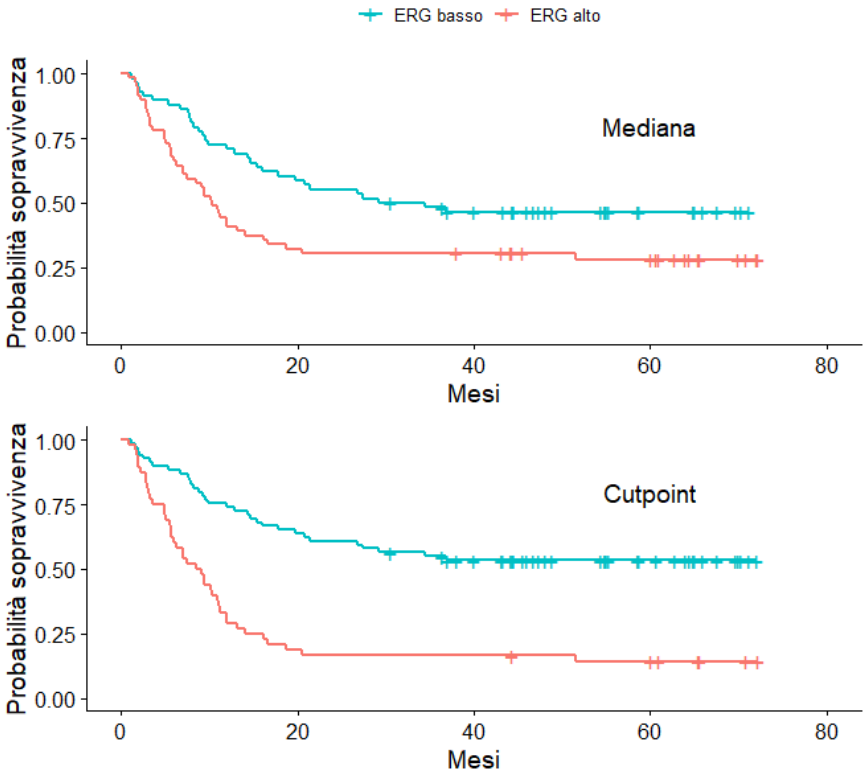

Figura 4.9: Curve di sopravvivenza rispetto al valore iniziale di ERG

I grafici delle curve mostrano una sopravvivenza maggiore nei pazienti con livelli bassi dell'espressione del gene ERG al tempo zero. Viene riportato anche il test formale:

	Valore alto	**N**	**Osservati**	**Attesi**	**Pvalue**
Mediana	No	28	25	33.4	0.01
	Sì	26	26	17.6	
Cutpoint	No	33	30	39.6	7e-04
	Sì	21	21	11.4	

Tabella 4.8: Test dei ranghi logaritmici per la variabile ERGB

Si ottiene che il valore di ERG al momento zero influenza la sopravvivenza del paziente, con aspettative di vita migliori per coloro che hanno minore espressione di tale gene.

È possibile concludere che, considerando i valori rilevati nei pazienti alla diagnosi, entrambi i geni influenzano significativamente la sopravvivenza dei pazienti. Sia con BAALC che con ERG c'è evidenza di un'aspettativa di vita superiore per coloro che hanno livelli bassi dell'espressione dei geni.

È possibile valutare se i livelli dei geni rilevati alla diagnosi abbiano un effetto congiunto sulla sopravvivenza. Di seguito viene riportato il grafico delle curve di sopravvivenza con l'effetto di entrambi i geni, considerando come valore sogllia la mediana:

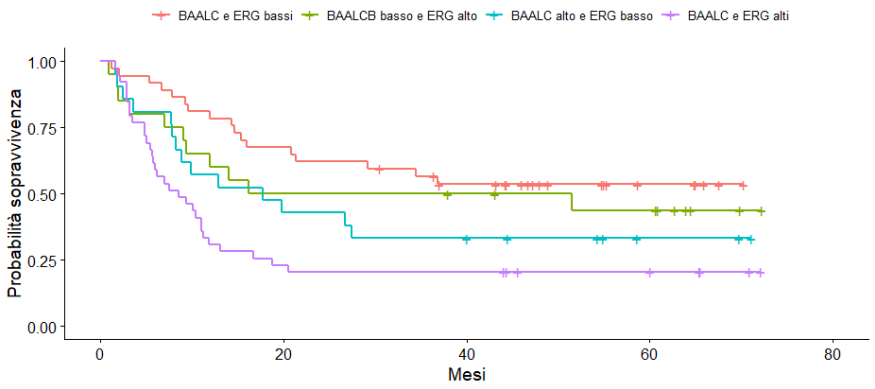

Figura 4.10: Curve di sopravvivenza con effetti congiunti (mediana)

Dal grafico si nota come la curva rossa, associata a valori bassi di entrambi i geni, sia indicativa di una probabilità di sopravvivenza superiore rispetto agli altri casi.
Utilizzando il cutpoint ottimale si ottengono le seguenti curve di sopravvivenza :

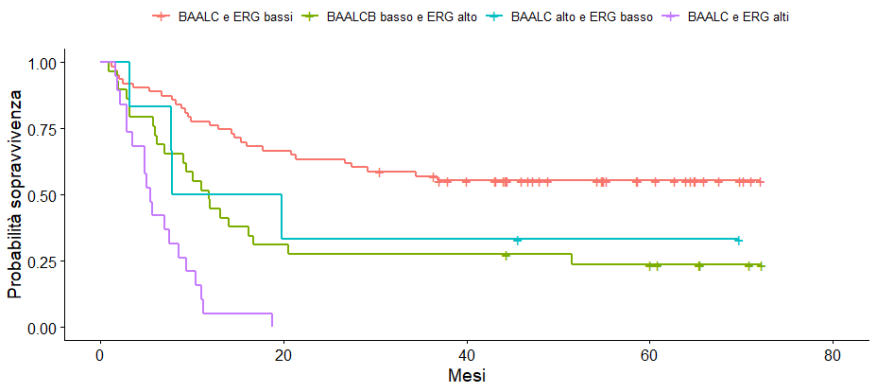

Figura 4.11: Curve di sopravvivenza con effetti congiunti (cutpoint)

Viene considerato di seguito il test formale per entrambi i casi:

	BAALC alto	**ERG alto**	**N**	**Osservati**	**Attesi**	**Pvalue**
Mediana	No	No	37	17	28.8	0.002
	No	Sì	20	11	13.1	
	Sì	Ni	21	14	13.0	
	Sì	Sì	39	31	18.1	
Cutpoint	No	No	63	28	48.12	2e-10
	No	Sì	29	22	15.94	
	Sì	No	6	4	3.55	
	Sì	Sì	19	19	5.39	

Tabella 4.9: Test dei ranghi logaritmici per l'interazione tra BAALCB e ERGB

Sia il pvalue associato al gruppo della mediana e sia quello associato al gruppo del cutpoint ottimale sono significativi (<0.05). Si conclude a favore della presenza di un effetto congiunto dei geni BAALC e ERG sulla sopravvivenza del paziente, con un'aspettativa di vita superiore per coloro che hanno sia BAALC che ERG bassi.

4.1.3 BAALC e ERG dopo la chemioterapia a induzione

Le seguenti tabelle identificano i valori clinici e le caratteristiche molecolari dei pazienti in base ai valori di BAALC e ERG misurati dopo essere stati sottoposti alla chemioterapia a induzione.

BAALC

BAALCI	LOW	HIGH	pvalue
GENDER			0,333
F	17	23	
M	42	37	
AGE			0,9258
Median	25	27	
Mean	29,18	28,76	
Pbblast			0,2224
Median	65	49	
Mean	57,65	50,92	
BmofBlast			0,4232
Median	69	59	
Mean	63,55	58,92	
Cytogenic Risk			0,9706
Good Risk	17	19	
Intermediate	29	28	
Poor	13	13	
V_cyto_risk			0,1516
1	18	20	
2	19	10	
3	10	9	
4	12	21	
FLT3ITD			1
1	12	12	
2	47	48	
BAALCB			0,00591
Low	38	23	
High	21	37	
BAALCI			
Low			
High			
BAALCC			0,0002
Low	33	13	
High	16	32	

	LOW	HIGH	pvalue
MLLPTD			1
1	6	7	
2	52	52	
Genetic Risk			0,9207
Fav	21	22	
Intermediate	16	18	
Poor	22	20	
CEBPA			0,7938
2	46	48	
Biallelic	1	2	
Monoallelid	12	10	
TLCCU			0,2317
Median	12330	21530	
Mean	29801,53	45189,63	
NPM1 Type A			1
1	13	13	
2	46	47	
Performance status			0,7157
1	45	41	
2	10	14	
3	1	2	
4	0	1	
ERGB			0,04368
Low	37	26	
High	22	34	
ERGI			0,0674
Low	35	25	
High	24	35	
ERGC			0,03839
Low	30	17	
High	19	28	

Tabella 4.10: Confronti basati sui valori di BAALC dopo la chemioterapia (mediana)

Le misurazioni di BAALC dopo la chemioterapia, divise in alto e basso con la mediana, sono correlate significativamente con le altre misurazioni di BAALC (BAALCB pvalue 0.005, BAALC pvalue 0.0002) e con i valori di ERG misurati

in tutti e tre gli istanti (ERGB pvalue 0.043, ERGI pvalue 0.06, ERGC pvalue 0.038).

BAALCI	LOW	HIGH	pvalue		LOW	HIGH	pvalue
GENDER			0,2624	**MLLPTD**			0,2976
F	7	33		1	5	8	
M	22	57		2	23	81	
AGE			0,7876	**Genetic Risk**			0,8621
Median	24	26,5		Fav	11	32	
Mean	29,75	28,7		Intermediate	7	27	
				Poor	11	31	
Pbblast			0,1517	**CEBPA**			0,4371
Median	65	51		2	21	73	
Mean	61,59	51,93		Biallelic	1	2	
				Monoallelid	7	15	
BmofBlast			0,046	**TLCCU**			0,8843
Median	73	59,5		Median	14040	17330	
Mean	69	58,65		Mean	27171,03	40907,87	
Cytogenic Risk			0,8527	**NPM1 Type A**			0,3048
Good Risk	10	26		1	4	22	
Intermediate	13	44		2	25	68	
Poor	6	20					
V_cyto_risk			0,8551	**Performance status**			0,4006
1	11	27		1	24	62	
2	6	23		2	3	21	
3	5	14		3	0	3	
4	7	26		4	0	1	
FLT3ITD			0,4293	**ERGB**			0,0003
1	4	20		Low	26	48	
2	25	70		High	3	42	
BAALCB			0,01407	**ERGI**			0,6276
Low	26	69		Low	8	21	
High	1	21		High	21	69	
BAALCI				**ERGC**			0,1021
Low				Low	19	36	
High				High	7	32	
BAALCC			<0,0001				
Low	18	15					
High	8	53					

Tabella 4.11: Confronti basati sui valori di BAALC dopo la chemioterapia (cut-point)

Se si considerano invece i valori di BAALC al secondo intervallo di tempo suddivisi con il cutpoint ottimale, le relazioni significative sono adesso con il numero di blasti nel midollo (BMofBlast pvalue 0.046), con le altre misurazioni di BAALC (BAALCB pvalue 0.01, BAALCC pvalue <0.0001) e con il valore di ERG alla diagnosi (ERGB pvalue 0.0003).
Vengono riportate di seguito le curve di sopravvivenza in base ai valori di BAALC con i relativi test formali:

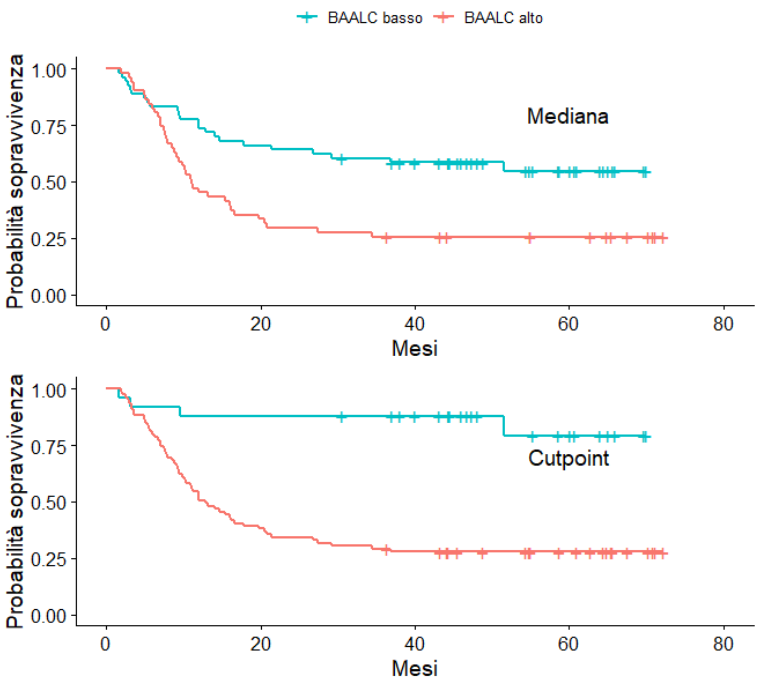

Figura 4.12: Curve di sopravvivenza rispetto al valore di BAALC dopo la chemioterapia

Il test dei ranghi logaritmici associato permette di ricavare i seguenti risultati:

	Valore alto	**N**	**Osservati**	**Attesi**	**Pvalue**
Mediana	0	53	23	35.4	0.001
	1	51	38	25.6	
Cutpoint	0	25	4	19.8	1e-05
	1	79	57	41.4	

Tabella 4.12: Test dei ranghi logaritmici per la variabile BAALCI

I grafici e i test formali mostrano che le differenza di sopravvivenza nei pazienti con valori del gene alti o bassi sono significative sia se si utilizza la mediana e sia se si utilizza il cutpoint.

ERG

ERGI	LOW	HIGH	pvalue		LOW	HIGH	pvalue
GENDER			1	**MLLPTD**			0,5589
F	21	20		1	8	5	
M	39	40		2	52	53	
AGE			0,4848	**Genetic Risk**			0,5796
Median	27,5	24,5		Fav	21	22	
Mean	29,7	28,43		Intermediate	20	15	
				Poor	19	23	
Pbblast			0,5729	**CEBPA**			0,6059
Median	58	52		2	49	46	
Mean	52,4	55,64		Biallelic	2	1	
				Monoallelid	9	13	
BmofBlast			0,3251	**TLCCU**			0,07391
Median	58	68		Median	10200	20960	
Mean	59,33	62,92		Mean	2863,97	45893,17	
Cytogenic Risk			0,6436	**NPM1 Type A**			0,825
Good Risk	18	18		1	14	12	
Intermediate	27	31		2	46	48	
Poor	15	11					
V_cyto_risk			0,6158	**Performance status**			0,9339
1	19	19		1	45	42	
2	16	13		2	12	12	
3	11	8		3	1	2	
4	14	20		4	0	1	
FLT3ITD			0,2536	**ERGB**			0,04368
1	9	15		Low	38	26	
2	51	45		High	22	34	
BAALCB			0,3611	**ERGI**			
Low	34	28		Low	30	17	
High	26	32		High	19	29	
BAALCI			0,0436	**ERGC**			0,0241
Low	35	24		Low	30	17	
High	25	35		High	19	29	
BAALCC			0,002				
Low	32	15					
High	17	31					

Tabella 4.13: Confronti basati sui valori di ERG dopo la chemioterapia (mediana)

I valori di ERG misurati dopo che i pazienti sono stati sottoposti a chiemioterapia a induzione appaiono correlati alle altre misurazioni di ERG (ERGB pvalue 0.04, ERGC pvalue 0.02), ai valori di BAALC dopo la chemioterapia e dopo la remissione della malattia (BAALCI pvalue 0.04, BAALCC pvalue 0.002). Inoltre è presente un rapporto lievemente significativo con il numero totale di leucociti nel paziente (TLCCU pvalue 0.07).

ERGI	LOW	HIGH	pvalue		LOW	HIGH	pvalue
GENDER			0,5015	**MLLPTD**			0,3028
F	8	33		1	5	8	
M	21	58		2	24	81	
AGE			0,8975	**Genetic Risk**			1
Median	27	26		Fav	11	32	
Mean	29,13	29,04		Intermediate	8	27	
				Poor	10	32	
Pbblast			0,2286	**CEBPA**			0,6465
Median	49	58		2	24	71	
Mean	48,37	55,86		Biallelic	1	2	
				Monoallelid	4	18	
BmofBlast			0,5015	**TLCCU**			0,07645
Median	55	68		Median	8800	18500	
Mean	59,15	61,67		Mean	21247,24	42400,64	
Cytogenic Risk			0,089	**NPM1 Type A**			0,6108
Good Risk	11	25		1	5	21	
Intermediate	9	49		2	24	70	
Poor	9	17					
V_cyto_risk			0,3052	**Performance status**			0,6004
1	12	26		1	22	65	
2	6	23		2	4	20	
3	6	13		3	1	2	
4	5	29		4	0	1	
FLT3ITD			1	**ERGB**			0,00213
1	6	18		Low	25	50	
2	23	73		High	4	41	
BAALCB			0,2746	**ERGI**			
Low	26	72		Low			
High	3	19		High			
BAALCI			0,0003	**ERGC**			0,1019
Low	8	21		Low	19	36	
High	21	69		High	7	33	
BAALCC			0,8119				
Low	10	24					
High	16	45					

Tabella 4.14: Confronti basati sui valori di ERG dopo la chemioterapia (cutpoint)

Considerando invece i valori suddivisi in base al cutpoint ottimale, rimangono significative le relazioni con il valore di BAALC alla chemioterapia (BAALCI pvalue 0.0003) e il valore dell'espressione di ERG alla diagnosi (ERGB pvalue 0.02). Sono inoltre lievemente significative le relazioni con il numero di leucociti (TLCCU pvalue 0.07) e la variabile "Cytogenic Risk" (pvalue 0.08).

Le curve di sopravvivenza rispetto ai valori dell'espressione del gene ERG dopo la chemioterapia sono:

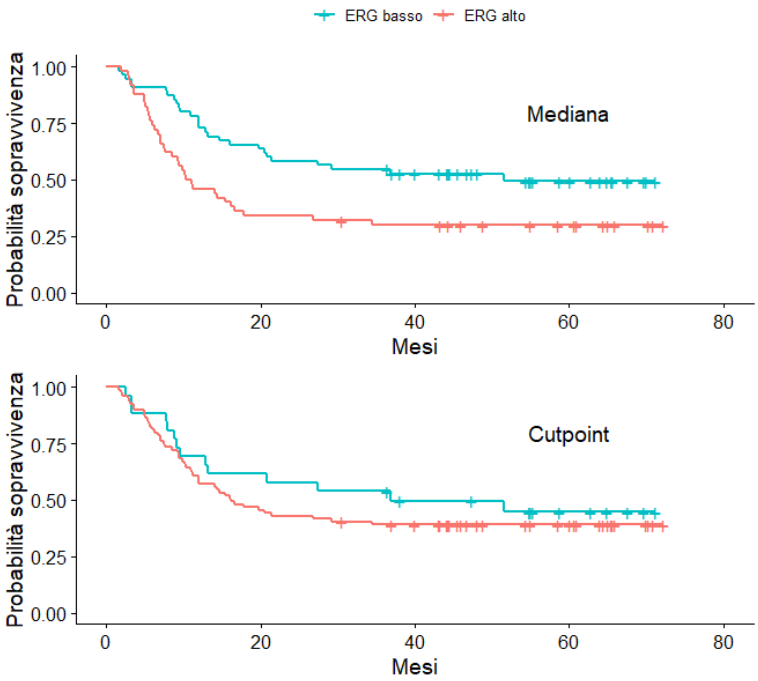

Figura 4.13: Curve di sopravvivenza rispetto al valore di ERG dopo la chemioterapia

	Valore alto	**N**	**Osservati**	**Attesi**	**Pvalue**
Mediana	No	55	27	37.4	0.007
	Sì	50	35	24.6	
Cutpoint	No	26	14	16.8	0.4
	Sì	79	48	45.2	

Tabella 4.15: Test dei ranghi logaritmici per la variabile ERGI

Le curve di sopravvivenza e i test mostrano una differenza in termini di probabilità di sopravvivenza solo per i pazienti divisi con la mediana (pvalue 0.007), mentre se si utilizza il cutpoint tale differenza sparisce (pvalue 0.4).

Si valuta ora l'effetto congiunto dei valori dei geni BAALC e ERG rilevati dopo che i pazienti sono stati sottoposti a chemioterapia a induzione. Le curve di sopravvivenza che si ottengono sono le seguenti:

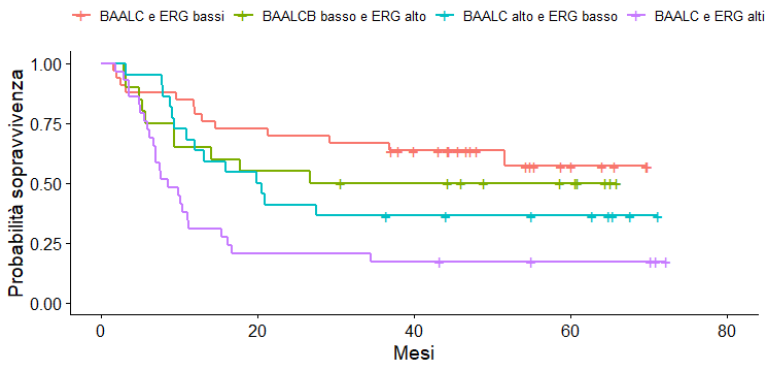

Figura 4.14: Curve di sopravvivenza con effetti congiunti (mediana)

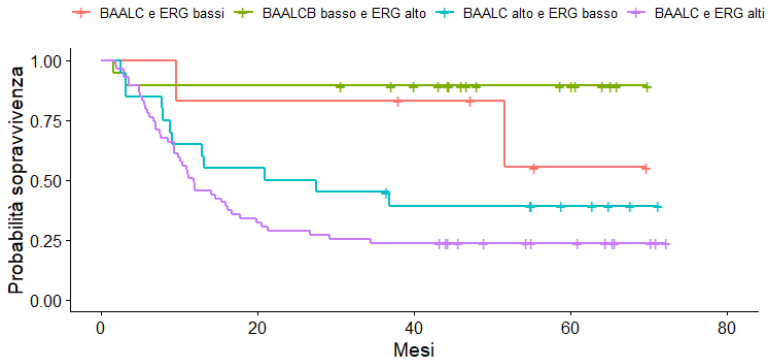

Figura 4.15: Curve di sopravvivenza con effetti congiunti (cutpoint)

Dai test formali si ottiene:

	BAALC alto	ERG alto	N	Osservati	Attesi	Pvalue
Mediana	No	No	33	13	23.4	5e-04
	No	Sì	20	10	12.0	
	Sì	No	22	13	13.7	
	Sì	Sì	29	24	11.9	
Cutpoint	No	No	6	2	4.77	8e-05
	No	Sì	19	2	14.4	
	Sì	No	20	12	11.95	
	Sì	Sì	59	45	29.45	

Tabella 4.16: Test dei ranghi logaritmici per l'interazione tra BAALCI e ERGI

L'analisi per valutare l'effetto congiunto dei valori dei geni misurati dopo la chemioterapia mostra l'evidenza di tale effetto (in entrambi i casi pvalue <0.0001), tuttavia se si osserva il grafico delle curve di sopravvivenza calcolate con il cut-point si notano delle curve con pochissime osservazioni. Questo rende difficile l'interpretazione del grafico, perché solitamente quando ci sono poche osservazioni i risultati sono poco affidabili.

4.1.4 BAALC e ERG dopo la remissione della malattia

Le seguenti tabelle identificano i valori clinici e le caratteristiche molecolari dei pazienti in base ai valori di BAALC e ERG misurati dopo che la malattia è stata dichiarata in remissione.

BAALC

BAALCC	LOW	HIGH	pvalue		LOW	HIGH	pvalue
GENDER			0,826	**MLLPTD**			0,4856
F	14	16		1	3	6	
M	33	32		2	43	41	
AGE			0,5968	**Genetic Risk**			0,2838
Median	25	24,5		Fav	20	16	
Mean	28,8	27,93		Intermediate	14	11	
				Poor	13	21	
Pbblast			0,8412	**CEBPA**			0,6416
Median	58	55,5		2	35	38	
Mean	54,09	55,09		Biallelic	1	2	
				Monoallelid	11	8	
BmofBlast			0,2982	**TLCCU**			0,8262
Median	57	71		Median	20600	17625	
Mean	60,3	64,9		Mean	32209,96	45761,04	
Cytogenic Risk			0,6667	**NPM1 Type A**			0,6331
Good Risk	17	13		1	12	10	
Intermediate	21	24		2	35	38	
Poor	9	11					
V_cyto_risk			0,3382	**Performance status**			0,3852
1	18	14		1	38	33	
2	14	10		2	7	9	
3	6	8		3	0	3	
4	9	16		4	0	1	
FLT3ITD			0,4515	**ERGB**			0,0241
1	8	12		Low	30	19	
2	39	36		High	17	29	
BAALCB			0,02373	**ERGI**			0,002
Low	31	20		Low	32	17	
High	16	28		High	15	31	
BAALCI			0,0436	**ERGC**			0,1527
Low	33	16		Low	27	20	
High	13	32		High	20	28	
BAALCC							
Low							
High							

Tabella 4.17: Confronti basati sui valori di BAALC dopo la remissione (mediana)

Una volta che la malattia viene dichiarata in remissione, i valori di BAALC nei pazienti (divisi in alti e bassi con la mediana) sono significativamente correlati al resto dei valori di BAALC (BAALCB pvalue 0.02, BAALCI pvalue 0.04) e

ai valori di ERG misurati negli istanti precedenti alla remissione (ERGB pvalue 0.02, ERGI pvalue 0.002).

BAALCC	LOW	HIGH	pvalue		LOW	HIGH	pvalue
GENDER			0,4945	**MLLPTD**			0,1513
F	9	21		1	1	8	
M	25	40		2	35	52	
AGE			0,218	**Genetic Risk**			0,1677
Median	27,5	24		Fav	16	20	
Mean	30	27,47		Intermediate	10	15	
				Poor	8	26	
Pbblast			0,5876	**CEBPA**			1
Median	60	52		2	26	47	
Mean	57,031	53,19		Biallelic	1	2	
				Monoallelid	7	12	
BmofBlast			0,079	**TLCCU**			0,6057
Median	57	71		Median	26510	15700	
Mean	57,6	65,41		Mean	35589,12	40989,64	
Cytogenic Risk			0,2969	**NPM1 Type A**			1
Good Risk	14	16		1	8	14	
Intermediate	15	30		2	26	47	
Poor	5	15					
V_cyto_risk			0,1887	**Performance status**			0,7555
1	15	17		1	28	43	
2	10	14		2	5	11	
3	3	11		3	0	3	
4	6	19		4	0	1	
FLT3ITD			0,3037	**ERGB**			<0,0001
1	5	15		Low	30	29	
2	29	46		High	4	32	
BAALCB			0,0083	**ERGI**			0,8119
Low	33	46		Low	10	16	
High	1	15		High	24	45	
BAALCI			0,0034	**ERGC**			0,083
Low	18	8		Low	24	31	
High	15	53		High	10	30	
BAALCC							
Low							
High							

Tabella 4.18: Confronti basati sui valori di BAALC dopo la remissione (cutpoint)

Dividendo i valori di BAALC con il cutpoint ottimale si ottengono relazioni significative dei valori di BAALC dopo la remissione con il numero di blasti nel midollo (BMofBlast pvalue 0.079), così come con le altre misurazioni di BAALC (BAALCB pvalue 0.008, BAALCI pvalue 0.003) e con le misurazione di ERG alla diagnosi (ERGB pvalue <0.0001) e dopo la remissione (ERGC pvalue 0.083). Analogamente a prima si considerano ora le curve di sopravvivenza per i pazienti divisi in base al valore di BAALC:

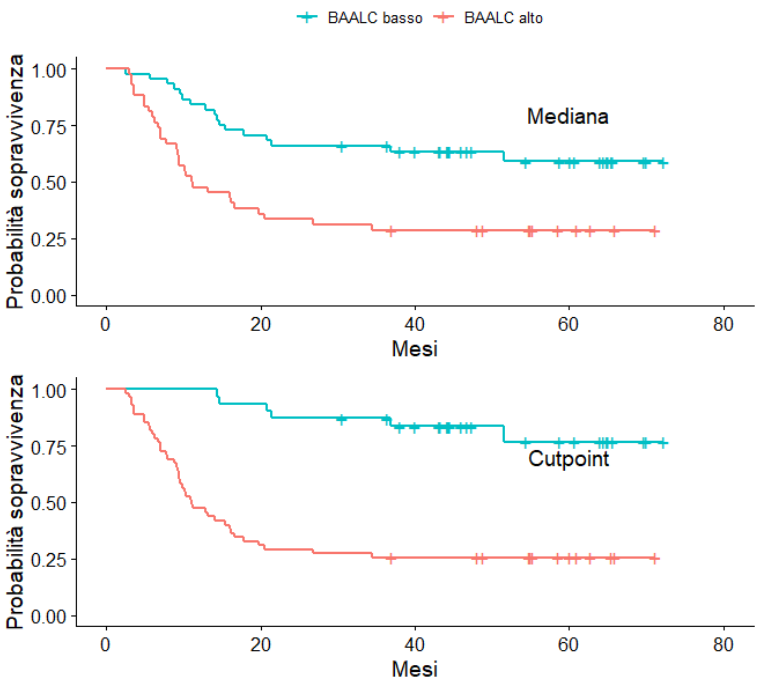

Figura 4.16: Curve di sopravvivenza rispetto al valore di BAALC dopo la remissione

I grafici concordano sul fatto che la sopravvivenza è più probabile nei pazienti con livelli bassi del gene BAALC dopo la remissione della malattia.
Viene di seguito affiancato il test dei ranghi logaritmici:

	Valore alto	**N**	**Osservati**	**Attesi**	**Pvalue**
Mediana	No	44	17	28.5	5e-04
	Sì	20	20	13.1	
Cutpoint	No	31	6	23.3	3e-07
	Sì	55	41	23.7	

Tabella 4.19: Test dei ranghi logaritmici per la variabile BAALCC

I test confermano la differenza in termini di probabilità di sopravvivenza per i pazienti, in base al valore di BAALC rilevato nell'ultimo periodo di studio per entrambi i criteri divisivi.

ERG

ERGC	LOW	HIGH	pvalue		LOW	HIGH	pvalue
GENDER			0,6628	**MLLPTD**			0,7398
F	16	14		1	5	4	
M	31	34		2	41	43	
AGE			0,6683	**Genetic Risk**			0,1721
Median	26	24,5		Fav	22	14	
Mean	28,978	27,79		Intermediate	12	13	
				Poor	13	21	
Pbblast			0,1914	**CEBPA**			0,4339
Median	53	59		2	39	34	
Mean	50,77	58,4		Biallelic	1	2	
				Monoallelid	7	12	
BmofBlast			0,4495	**TLCCU**			0,033
Median	61,5	69		Median	10130	24515	
Mean	61,11	64,04		Mean	24895,28	52923,33	1
Cytogenic Risk			0,3675	**NPM1 Type A**			
Good Risk	16	14		1	11	11	
Intermediate	24	21		2	36	37	
Poor	7	13					
V_cyto_risk			0,3619	**Performance status**			0,1941
1	18	14		1	36	35	
2	14	10		2	10	6	
3	5	9		3	0	3	
4	10	15		4	0	1	
FLT3ITD			0,8021	**ERGB**			0,307
1	9	11		Low	27	22	
2	38	37		High	20	26	
BAALCB			0,0237	**ERGI**			0,024
Low	31	20		Low	30	19	
High	16	28		High	17	29	
BAALCI			0,0436	**ERGC**			
Low	30	19		Low			
High	17	28		High			
BAALCC			0,1527				
Low	27	20					
High	20	28					

Tabella 4.20: Confronti basati sui valori di ERG dopo la remissione (mediana)

La tabella mostra che i valori di ERG dopo la remissione (con la mediana) sono significativamente correlati con le variabili BAALCB (pvalue 0.023), BAALCI (pvalue 0.04), ERGI (pvalue 0.024) e con il numero di leucociti (TLCCU 0.03).

ERGC	LOW	HIGH	pvalue		LOW	HIGH	pvalue
GENDER			0,8261	**MLLPTD**			1
F	18	12		1	5	4	
M	37	28		2	49	35	
AGE			0,8327	**Genetic Risk**			0,1711
Median	25	26		Fav	25	11	
Mean	28,27	28,52		Intermediate	14	11	
				Poor	16	18	
Pbblast			0,153	**CEBPA**			0,6354
Median	55	61		2	44	29	
Mean	51,21	59,24		Biallelic	1	2	
				Monoallelid	10	9	
BmofBlast			0,5016	**TLCCU**			0,0433
Median	64	67,5		Median	14040	26605	
Mean	61,48	61,1		Mean	26041,42	56953	
Cytogenic Risk			0,3176	**NPM1 Type A**			1
Good Risk	20	10		1	13	9	
Intermediate	26	19		2	42	31	
Poor	9	11					
V_cyto_risk			0,2514	**Performance status**			0,0614
1	22	10		1	43	29	
2	15	9		2	11	5	
3	7	7		3	0	3	
4	11	14		4	0	1	
FLT3ITD			0,4536	**ERGB**			0,0537
1	10	10		Low	39	20	
2	45	30		High	16	20	
BAALCB			0,5818	**ERGCC**			0,1019
Low	47	32		Low	19	7	
High	8	8		High	36	33	
BAALCI			0,0003	**ERGI**			
Low	19	7		Low			
High	36	32		High			
BAALCC			0,0831				
Low	24	10					
High	31	30					

Tabella 4.21: Confronti basati sui valori di ERG dopo la remissione (cutpoint ottimale)

I valori di ERG suddivisi con il cutpoint ottimale mostrano relazioni significative con i valori di BAALCB nelle ultime due fasi (BAALCI pvalue 0.0003, BAALCC pvalue 0.08) e con i valori di ERG alla diagnosi (ERGB pvalue 0.05). Sono presenti inoltre relazioni con le variabili "TLCCU" (pvalue 0.04) e con "Performance Status" (pvalue 0.06).

Le curve di sopravvivenza e i test relativi ai valori di ERGC sono i seguenti:

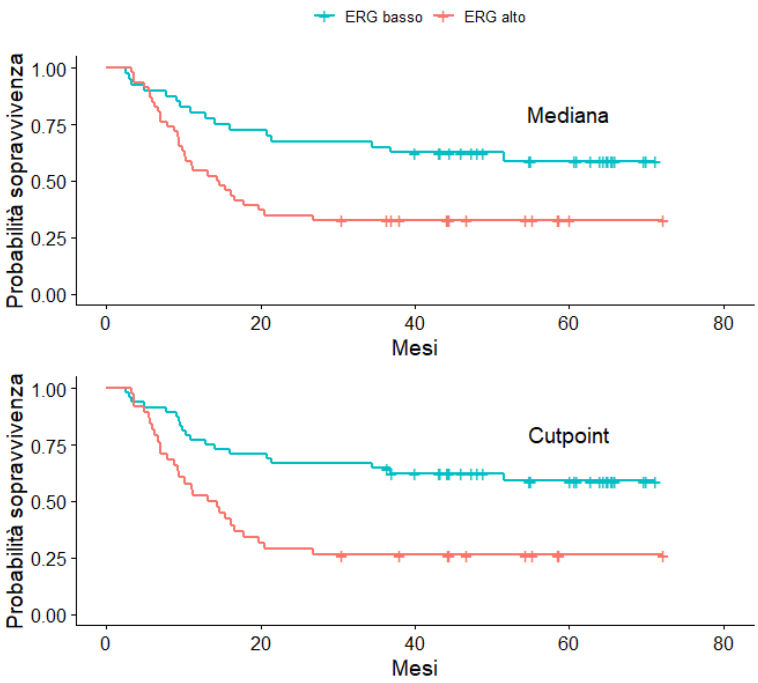

Figura 4.17: Curve di sopravvivenza rispetto al valore di ERG dopo la remissione

	Valore alto	**N**	**Osservati**	**Attesi**	**Pvalue**
Mediana	No	40	16	25.3	0.0006
	Sì	46	31	21.7	
Cutpoint	No	48	19	30.1	6e-04
	Sì	38	28	16.9	

Tabella 4.22: Test dei ranghi logaritmici per la variabile ERGI

Le curve di sopravvivenza e i grafici concordano sulla significatività dell'effetto del valore di ERG rilevato dopo la remissione della malattia, favorendo i pazienti che hanno valori bassi del gene.

Di seguito si considera la possibilità della presenza di un effetto congiunto dei geni. Viene riportato prima il grafico delle curve di sopravvivenza per il gruppo dei pazienti suddivisi secondo la mediana e successivamente il grafico dei pazienti divisi con il cutpoint ottimale.

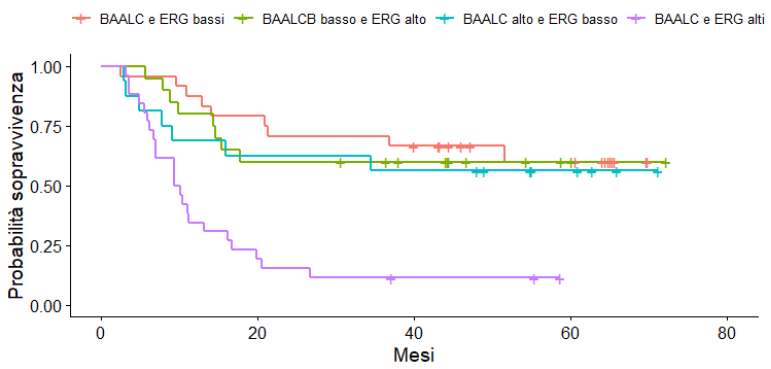

Figura 4.18: Curve di sopravvivenza con effetti congiunti (mediana)

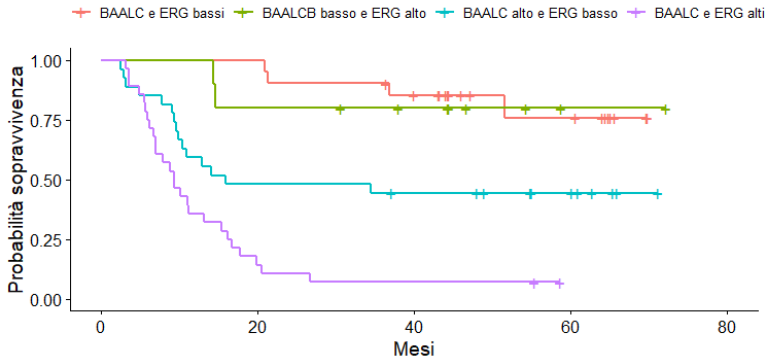

Figura 4.19: Curve di sopravvivenza con effetti congiunti (cutpoint)

	BAALC alto	ERG alto	N	Osservati	Attesi	Pvalue
Mediana	No	No	24	9	16.29	1e-05
	No	Sì	20	8	12.22	
	Sì	No	16	7	9.04	
	Sì	Sì	26	23	9.45	
Cutpoint	No	No	21	4	16.16	4e-09
	No	Sì	10	2	7.14	
	Sì	No	27	15	13.95	
	Sì	Sì	28	26	9.75	

Tabella 4.23: Test dei ranghi logaritmici per l'interazione tra BAALCC e ERGC

Sia i grafici delle curve di sopravvivenza sia i test portano a non rifiutare l'ipotesi di presenza dell'effetto congiunto dei due geni.

4.2 Modello di Cox

L'analisi esplorativa ha messo in luce come il valore dell'espressione dei geni BAALC e ERG influisca sulla sopravvivenza dei pazienti, favorendo coloro che presentano valori bassi.
È di interesse trovare un modello per spiegare in che misura le variabili riescano a influenzare l'esito della malattia nei pazienti. Una caratteristica molto importante che il modello deve avere è quella di riuscire a interpretare le misurazioni di BAALC e ERG non come tre variabili diverse, ma come tre misurazioni della stessa variabile in tre diversi istanti di tempo. Tale caratteristica va tenuta in considerazione nella costruzione del modello e per questo motivo ciascuna misurazione verrà considerata come una variabile tempo-dipendente.
Il modello che verrà di seguito interpolato è quello di Cox, che necessita di essere applicato su un dataset costruito come riportato di seguito, al fine di interpretare correttamente le misurazioni:

$$\textbf{ID paziente} \quad \textbf{(start, stop]} \quad \textbf{status} \quad \boldsymbol{x_1} \quad \boldsymbol{x_2} \quad \boldsymbol{x_3} \quad ....$$

Le misurazioni di BAALC e ERG sono relative a tre intervalli di tempo:

— Il primo intervallo è il periodo di trenta giorni successivo alla data della diagnosi

— Il secondo intervallo è il periodo di trenta giorni successivo alla data della chemioterapia a induzione

— Il terzo intervallo è il periodo di trenta giorni successivo alla data in cui la malattia è stata dichiarata in remissione

Ogni paziente avrà al più tre righe nel nuovo dataset, una per ciascun intervallo studiato. Va tenuto quindi in considerazione il momento di uscita dallo studio per ciascun paziente, che può avvenire con un evento o con una censura:

— Se il paziente i esce dallo studio nel primo periodo gli spetterà una sola riga nel dataset

— Se il paziente i esce dallo studio nel secondo periodo gli spetteranno due righe nel dataset

— Se il paziente i riesce a completare la chemioterapia e quindi iniziare la terapia di consolidamento, gli spetteranno tre righe nel dataset

Nel nuovo dataset le variabili relative all'espressione genica non saranno più divise in tre sottogruppi (BAALCB/ERGB, BAALCI/ERGI, BAALCC/ERC), ma verranno raggruppate in due sole variabili BAALC e ERG, date dal susseguirsi delle misurazioni relative allo stesso paziente nei diversi istanti di tempo.
Lo studio si concentra sulla probabilità di sopravvivenza dei pazienti: si considera una nuova variabile indicatrice chiamata "Death_Status" che vale 1 se il paziente muore durante lo studio e 0 altrimenti. "Death_Status" è costruita sul dataset principale, ma servirà da punto di partenza per definire una nuova variabile adatta al dataset necessario per applicare il modello di Cox. Nel nuovo dataset ciascun paziente ha un numero variabile di righe, e la nuova variabile "Status" dovrà adattarsi alle diverse situazioni:

— Se il paziente i esce dallo studio nel primo periodo avrà una sola riga nel dataset; "Status" avrà un solo valore e sarà pari alla variabile "Death_Status"

— Se il paziente i esce dallo studio nel secondo periodo avrà due righe nel dataset, quindi la nuova variabile status dovrà avere un valore diverso in ciascuna riga: nella prima deve valere 0, dal momento che nel primo periodo il paziente non esce dallo studio, e nella seconda riga avrà lo stesso valore della variabile "Death_Status"

— Se il paziente i esce dallo studio nel terzo periodo, similmente al caso precedente, la nuova variabile status varrà 0 nelle prime due righe e sarà uguale alla variabile "Death_Status" nella terza

Per ultimo bisogna definire le variabili *start* e *stop* che indicano gli estremi di ciascun intervallo. Si considera la data della diagnosi come tempo zero di ciascun paziente, mentre gli estremi degli intervalli vengono costruiti utilizzando il numero

di giorni che intercorrono da tale data.
Gli intervalli sono chiusi a destra, quindi quando sono presenti due o tre intervalli l'estremo iniziale di un intervallo è uguale all'estremo finale di quello precedente, ad eccezione ovviamente del primo intervallo.
Per comprendere meglio come sono stati calcolati gli intervalli, si studia ciascun gruppo di pazienti separatamente.

Gruppo 1 I pazienti appartenenti al gruppo 1 sono quelli usciti dallo studio durante il primo periodo, quindi subito dopo la diagnosi. Il loro unico intervallo prevede start pari a zero e stop pari alla data di uscita dallo studio, che sia per morte o per altre cause. Nel dataset preso in considerazione non ci sono pazienti che escono nella prima fase, perciò si passa direttamente alla definizione del gruppo 2.

Gruppo 2 I pazienti appartenenti al gruppo 2 sono stati sottoposti alla chemioterapia a induzione ma sono usciti dallo studio subito dopo di essa. Per tali pazienti sono riservate due righe nel dataset in costruzione, quindi due intervalli: il primo intervallo prevede start pari a zero e stop pari ai giorni trascorsi fino alla data della chemioterapia, mentre il secondo si chiude con i giorni alla data di morte del paziente. Coloro che appartengono al gruppo 2, infatti, escono dallo studio soltanto in tre situazioni:

— La chemioterapia non ha effetto, il che porta alla morte dei pazienti

— Complicazioni durante la chemioterapia, che portano alla morte dei pazienti

— Alcuni pazienti muoiono poco dopo aver ricevuto la chemioterapia, non riuscendo ad iniziare la terapia di consolidamento

Gruppo 3 I pazienti che si sottopongono alla chemioterapia e riescono a resistere al trattamento iniziano una terapia di consolidamento. Ciascun paziente appartenente al terzo gruppo ha quindi tre righe del dataset dedicate, con tre intervalli associati. Il primo intervallo va da zero alla data della chemioterapia, il

secondo si chiude con la data della prima seduta per la terapia di consolidamento, mentre il terzo si può chiudere in due modi:

— Il paziente riesce a concludere la terapia di consolidamento e a rispondere bene al trattamento. La data che si prende in considerazione è quella dell'ultimo aggiornamento sullo stato di salute del paziente una volta uscito dallo studio

— Il paziente subisce una ricaduta della malattia che lo porta alla morte, uscendo quindi dallo studio

— Il paziente subisce una ricaduta della malattia ma riesce a salvarsi, uscendo comunque dallo studio. In questo caso si considera di nuovo la data di ultimo followup

Seguendo i passi descritti sopra si ottiene il dataset nella giusta forma per poter applicare il modello di Cox. Si riporta di seguito un sottoinsieme del dataset per ciascun gruppo di pazienti:

	ID paziente	**start**	**stop**	**status**	**BAALC**	**ERG**	**...**
Gruppo 2	41	0	25	0	149	70	...
	41	25	835	1	284	19	..
Gruppo 3	45	0	14	0	194	448	...
	45	14	35	0	487	341	...
	45	35	307	1	1582	1816	...

Tabella 4.24: Estratto dal nuovo dataset per ciascun gruppo di pazienti descritto

Il dataset così ottenuto è composto da 388 righe, contrariamente alle 152 del dataset di partenza.
Una volta definito il nuovo dataset è possibile applicare un modello di Cox con variabili tempo-dipendenti. Nella definizione del modello in R è necessario specificare quale variabile è il codice identificativo per ciascun paziente, per permettere ad R di trattare le variabili tempo-dipendenti come tali. Ad esempio, con-

siderando solo le variabili ERG e BAALC, ovvero quelle tempo dipendenti, è possibile ottenere un modello con il seguente comando:[3]

```
>  coxph(Surv(start,stop,OS_Status)~BAALC + ERG , id = PtNo,
 data = nuovo_dataset)
```

ottenendo i seguenti risultati:

	coef	**se(coef)**	**pvalue**
BAALC	5.800e-04	1.243e-04	3.1e-06
ERG	3.559e-04	9.862e-05	0.000308

Likelihood ratio test 46.06 on 2df p = 1e-10

Tabella 4.25: Valori modello considerando solo le variabili BAALC e ERG

Per interpretare i coefficienti del modello bisogna considerare il loro elevamento a potenza, per via di come è definito il modello:

$$h(X|Z(t)) = h_0(x)\exp\{\beta^T Z(t)\}$$

Ciascun $exp(\beta_i)$ corrisponde all'hazard-ratio, ovvero al rapporto tra i rischi. In questo caso sono presenti solo variabili continue, quindi l'hazard-ratio per la variabile BAALC è definito come:

$$\frac{h_0(x)e^{\beta_1 x_1+\beta_2 x_2}}{h_0(x)e^{\beta_1(x_1+1)+\beta_2 x_2}} = e^{\beta_1}$$

Se si considerano i valori di ciascun e^{β_i} si ottengono i risultati contenuti nella seguente tabella:

BAALC	1.00058017
ERG	1.00035596

Tabella 4.26: Tabella dei valori e^{β_i}

[3]Il comando `coxph` fa parte del pacchetto *survival* di R, creato da Terry Therneau.

Per un certo t fissato, quando il valore di BAALC aumenta di 100 e il valore di ERG resta invariato, l'hazard-ratio associato è pari a

$$\frac{h_0(x)e^{\beta_1 x_1+\beta_2 x_2}}{h_0(x)e^{\beta_1(x_1+100)+\beta_2 x_2}} = e^{100\beta_1} \tag{4.1}$$

ovvero è pari a $e^{100*\beta_1} \simeq 1.06$: quando l'espressione di BAALC aumenta di 100, il rischio di morte per il paziente aumenta del 6%. Analogamente, quando l'espressione di ERG aumenta di 100, il rischio di morte per il paziente incrementa circa del 4%.

Per un'analisi più completa è possibile considerare un modello di Cox che includa più variabili. Di seguito viene riportato il modello che considera non solo l'effetto di BAALC e ERG, ma anche del sesso del paziente e se ha subito o meno una ricaduta della malattia:

```
> coxph(Surv(start,stop,OS_Status)~  Gender + Cases_RFS + BAALC + ERG ,
 id = PtNo, data = nuovo_dataset)
```

	coef	**se(coef)**	**pvalue**
GenderM	-0.8653610	0.2831571	0.002242
Cases_RFS1	-1.3599937	0.3376599	5.63e-05
BAALC	0.0004329	0.0001190	0.000274
ERG	0.0004560	0.0001084	2.58e-05

Likelihood ratio test 54.57 on 4df p=4e-11

Tabella 4.27: Valori modello con le variabili BAALC, ERG, Gender e Cases_RFS

Per valutare l'effetto di ciascuna variabile, viene riportata di seguito la tabella contenente i valori:

GenderM	0.4209
Cases_RFS1	0.2567
BAALC	1.0004
ERG	1.0005

Tabella 4.28: Tabella dei valori e^{β_i}

L'interpretazione degli HR relativi ai geni BAALC e ERG non cambia di molto rispetto al modello precedente, infatti quando l'espressione di BAALC aumenta di 100 il rischio incrementa circa del 4%, e lo stesso vale per quando l'espressione di ERG incrementa di 100. L'HR della variabile Gender mostra come negli uomini il rischio cali del 58%, mentre se si considera la variabile Cases_RFS, che vale 1 se non c'è stata una ricaduta (RFS = Relapse Free Survival), vediamo che il rischio per i pazienti che hanno subito una ricaduta scende circa del 75%.
Confrontando l'AIC dei due modelli si conclude in favore del modello che tiene conto anche delle le variabili Cases_RFS e Gender (AIC solo BAALC e ERG 423.23, AIC con RFS e Gender 405.22).

Capitolo 5

Conclusioni

L'analisi esplorativa ha messo in luce la presenza di un effetto delle variabili longitudinali BAALC e ERG sulla durata di vita dei pazienti, con maggior probabilità di sopravvvivenza nei pazienti con livelli bassi dell'espressione dei geni. L'interpolazione del modello di Cox ha confermato la presenza dell'influenza dell'espressione genica sulla sopravvivenza dei pazienti: quando l'espressione di uno dei due geni aumenta di 100 unità, il rischio di morte del paziente incrementa quasi del 4%.

Durante l'analisi esplorativa, quando è stata posta l'attenzione sui livelli di BAALC e ERG nei vari intervalli di tempo, sono emersi alcuni rapporti significativi con altre variabili. Considerando nel complesso la sopravvivenza dei pazienti nello studio è stato confermato l'effetto sull'aspettativa di vita solo delle variabili relative al sesso e l'aver subito una ricaduta della malattia; gli uomini hanno un rischio inferiore del 58% rispetto alle donne, mentre subire la ricaduta della malattia implica una diminuzione della probabilità di sopravvivenza del 75%.

L'analisi si è focalizzata anche sulla scelta di un metodo che permettesse di determinare quando un valore dell'espressione genica è alto o basso. Sono state confrontate le analisi eseguite considerando due tipi di valori soglia: la mediana e il cutpoint ottimale. Le tabelle dei confronti, i relativi test di Wilcoxon e di Fisher e le curve di sopravvivenza non mostrano significative differenze tra i due valori soglia. La differenza più importante si riscontra nelle numerosità dei gruppi generati dai valori soglia: la mediana dà origine a due gruppi con numerosità praticamente uguali, mentre il cutpoint ottimale presenta gruppi con numerosità spesso molto diverse. Non avendo riscontrato particolari vantaggi nell'usare il cutpoint ottimale, in questo caso si preferisice l'utilizzo della mediana, che gode

inoltre di una maggiore facilità di calcolo.

Bibliografia

1. Torsten Hothorn and Berthold Lausen, *On the exact distribution of maximally selected rank statistics*, Science Direct Working Paper No S1574-0358(04)70152-5, 2002.

2. Grambsch P.M. Therneau T.M., *Modeling survival data: Extending the cox model. statistics for biology and health*, Springer, 2000.

3. Laura Ventura and Walter Racugno, *Biostatistica: casi di studio in r*, Egea, 2017.

Sitografia

1. https://www.airc.it/cancro/informazioni-tumori/guida-ai-tumori/leucemia

2. https://www.airc.it/cancro/informazioni-tumori/guida-ai-tumori/leucemia-mieloide-acuta