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

Analisi di dati di sopravvivenza per l'identificazione di biomarcatori prognostici a partire da dati di RNA-Seq in pazienti con Adenocarcinoma del Colon (COAD)

Relatore Prof. Davide Risso
Dipartimento di Scienze Statistiche

Laureanda Sophie Grace Parolin
Matricola 2089980

Anno Accademico 2023/2024

Ringrazio la mia famiglia, i miei amici
e i professori per il supporto ricevuto.
Senza di voi non sarei chi sono oggi.

Abstract

0.1 Versione italiana

In questa tesi, è stato esplorato il potenziale delle informazioni genomiche e cliniche per migliorare la predizione della sopravvivenza nei pazienti affetti da adenocarcinoma del colon (COAD). Utilizzando i dati di *The Cancer Genome Atlas* (TCGA), sono state implementate diverse metodologie di analisi di sopravvivenza basate su modelli di Cox penalizzati (lasso, elastic net e group lasso) e non penalizzati, con l'obiettivo di valutare l'impatto dei geni e delle variabili cliniche sulla prognosi dei pazienti.

L'analisi si è inizialmente focalizzata sulla verifica della riproducibilità dei risultati dell'articolo "*Genome-wide Identification and Analysis of Prognostic Features in Human Cancer*" di Smith e Sheltzer (2022), nel quale vengono individuati potenziali biomarcatori prognostici in diversi tipi di tumore. Questa fase ha evidenziato limiti e criticità che possono emergere nel contesto dell'analisi genomica su larga scala, a cui si è cercato di ovviare attraverso un'analisi più approfondita.

Nonostante l'impiego di metodi avanzati, come il *group lasso*, il quale consente di considerare più *pathway* simultaneamente, in questo contesto, è emerso che le informazioni genomiche non hanno contribuito in modo significativo alla predizione della sopravvivenza. Questo suggerisce che potrebbero essere necessarie strategie o dati differenti per evidenziare il ruolo dei geni nella prognosi del COAD.

0.2 English Version

In this thesis, the potential of genomic and clinical information to improve survival prediction in patients with colon adenocarcinoma (COAD) was explored. Using data from The Cancer Genome Atlas (TCGA), various survival analysis methodologies were implemented, based on both penalized Cox models (lasso, elastic net, and group lasso) and non-penalized models, with the aim of assessing the impact of genes and clinical variables on patient prognosis.

The analysis initially focused on verifying the reproducibility of the results from the article "Genome-wide Identification and Analysis of Prognostic Features in Human Cancer" by Smith e Sheltzer (2022), which identifies potential prognostic biomarkers across various cancer types. This phase highlighted limitations and challenges that may arise in large-scale genomic analysis, which were addressed through a more in-depth examination.

Despite the use of advanced methods, such as group lasso, which allows the simultaneous consideration of multiple pathways, it was found that genomic information did not significantly contribute to survival prediction in this context. This suggests that different strategies or data might be needed to better capture the role of genes in COAD prognosis.

Indice

Introduzione

Il cancro rappresenta una delle principali cause di mortalità a livello globale, suscitando un crescente interesse scientifico nella comprensione dei meccanismi molecolari che influenzano lo sviluppo e la progressione dei tumori. La ricerca in questo ambito è stata notevolmente potenziata da iniziative come il *The Cancer Genome Atlas* (TCGA), che hanno permesso una mappatura dettagliata degli effetti genetici sulla crescita tumorale, generando dati genomici e clinici di grande rilevanza.

Questa tesi si propone di valutare il contributo effettivo delle informazioni genomiche nella predizione della prognosi e della sopravvivenza nei pazienti oncologici, integrandole con le caratteristiche cliniche e demografiche. L'obiettivo principale è quello di analizzare il potenziale predittivo delle variabili genomiche nell'offrire informazioni aggiuntive rispetto ai dati clinici, al fine di identificare biomarcatori prognostici rilevanti e sviluppare modelli statistici capaci di supportare decisioni cliniche personalizzate.

Nonostante i progressi nel campo dell'oncologia e della genomica, l'integrazione di queste discipline rimane una sfida complessa. L'uso diretto dei dati genomici nei modelli di sopravvivenza è soggetto a numerose difficoltà, dovute in parte alla loro complessità e all'alta dimensionalità rispetto ai dati clinici. Questo lavoro adotta, pertanto, diversi metodi di regressione di Cox penalizzati, tra cui il lasso, l'*elastic net* e il *group lasso*, per affrontare il problema dell'alta dimensionalità e valutare l'impatto di gruppi di geni, strutturati in *pathway*, sulla predizione della sopravvivenza.

Un'ulteriore sfida tecnica riscontrata è stata la gestione integrata e robusta dei dati genomici e clinici. Sarebbe auspicabile, in futuro, lo sviluppo di strumenti avanzati per la gestione dei metadati e dei dati genomici in un formato unificato, al fine di minimizzare il rischio di errori di sincronizzazione tra i diversi *dataset*.

In questa tesi, tutte le analisi sono state condotte utilizzando il *software* R, con una soglia di significatività fissata a 0.05 per l'identificazione dei geni rilevanti. I risultati evidenziano che, nel contesto analizzato, l'inclusione delle sole variabili cliniche nel modello fornisce una capacità predittiva sufficiente, rendendo non necessaria l'aggiunta

delle variabili genomiche. Per maggiori dettagli sull'ambiente di lavoro, si rimanda al *Listing* A.1 in Appendice.

Capitolo 1

Approcci e criticità nell'analisi genomica di dati di sopravvivenza

Questa tesi parte dalla verifica della riproducibilità dei risultati presentati nell'articolo "*Genome-wide identification and analysis of prognostic features in human cancer*" di Smith e Sheltzer (2022). Gli autori, mediante un'analisi genomica su larga scala, individuano migliaia di biomarcatori prognostici, i quali identificano i pazienti maggiormente a rischio di progressione di diversi tipi di cancro. Tuttavia, questa tesi si propone anche di evidenziare limiti e criticità di tale analisi, cercando di colmare eventuali lacune attraverso un'analisi maggiormente approfondita.

1.1 Confronti

Nell'ambito dell'analisi genomica, spesso si riscontrano una varietà di approcci metodologici che possono portare a risultati contrastanti o addirittura errati. Un esempio emblematico è rappresentato dall'articolo di Uhlen et al. (2017) "*A pathology atlas of the human cancer transcriptome*", pubblicato su *Science*. Le metodologie utilizzate in questo studio sono oggetto di ampie contestazioni nell'articolo "*Pitfalls in re-analysis of observational omics studies: a post-mortem of the human pathology atlas*" di Gilis et al. (2020), nel quale gli autori ricostruiscono le analisi di Uhlen et al. (2017), mettendone in evidenza le limitazioni e apportando significativi miglioramenti alle stesse. In particolare, Uhlen et al. (2017) conducono un'analisi di sopravvivenza avvalendosi esclusivamente dello stimatore di *Kaplan-Meier* (KM) della funzione di sopravvivenza per ciascun gene (per ulteriori dettagli, si rimanda a Kaplan e Meier (1958)), senza ricorrere a modelli maggiormente avanzati che consentirebbero, tra le altre cose, l'inclusione di covariate continue. Un'altra criticità di questa analisi, che dipende dall'impossibilità di includere

covariate continue, risiede nella dicotomizzazione dei dati di conteggio in due gruppi basati sul livello di espressione del gene analizzato, il che porta a una notevole perdita di informazione. In aggiunta, l'analisi non tiene conto di eventuali variabili confondenti, tuttavia l'omissione di tali fattori di confondimento dal modello può comportare una sovrastima o una sottostima dell'effetto reale e non vengono, inoltre, considerati i problemi associati all'esecuzione di test multipli, portando alla presenza di notevoli falsi positivi, ovvero geni identificati come associati alla sopravvivenza, quando in realtà non lo sono, come mostrato da Gilis et al. (2020) (per ulteriori dettagli in merito alle problematiche lagate ai test multipli si rimanda al Paragrafo 2.1).

Rispetto a Uhlen et al. (2017), Smith e Sheltzer (2022) seguono un approccio metodologico maggiormente rigoroso. In primo luogo, utilizzano il modello di regressione semi-parametrico di Cox a rischi proporzionali (si rimanda al Paragrafo 1.2 per ulteriori dettagli in merito a questo modello). A differenza dell'analisi basata su *Kaplan-Meier*, il modello di Cox consente di includere covariate continue, evitando così la necessità di dicotomizzare le misurazioni dell'espressione genica. Inoltre, effettuano correzioni per test multipli e tengono conto di possibili variabili confondenti, anche se, avendo riscontrato un elevato grado di concordanza tra gli z-score associati ai geni ottenuti dai modelli univariati e dai modelli multivariati, proseguono l'analisi e formulano le loro conclusioni basandosi esclusivamente sui modelli privi di confondenti. Tuttavia, soprattutto in ambito medico, trascurare i confondenti può avere gravi conseguenze sia nella comprensione delle relazioni tra trattamenti e risultati clinici sia nella formulazione di raccomandazioni terapeutiche. I confondenti sono variabili che possono influenzare sia l'esposizione (ad esempio, a un farmaco) sia l'esito (ad esempio, la guarigione o l'insorgenza di complicazioni). Se non vengono controllati, i confondenti possono distorcere le associazioni osservate, portando a conclusioni errate.

Sebbene Smith e Sheltzer (2022) abbiano analizzato la relazione tra ogni singolo gene e la sopravvivenza, non hanno considerato l'effetto combinato di più geni simultaneamente. È importante sottolineare che, analizzando ogni gene singolarmente, si potrebbero rilevare delle associazioni spurie con la sopravvivenza, dovute più alla correlazione tra i geni che a una vera e propria causalità . Di conseguenza, un gene che sembra predittivo della prognosi potrebbe essere fuorviante a causa della sua correlazione con altri geni.

Riassumendo, gli approcci metodologici nell'analisi genomica possono variare significativamente, influenzando i risultati ottenuti. L'articolo di Smith e Sheltzer (2022) rappresenta un passo in avanti rispetto a studi precedenti, ma presenta comunque delle limitazioni.

Nella sezione successiva, verrà approfondito il modello di Cox, strumento cruciale per l'analisi di sopravvivenza.

1.2 Il modello di Cox

Nel campo dell'analisi dei dati di sopravvivenza, uno degli strumenti più utilizzati è il modello di regressione semi-parametrico noto come modello di Cox a rischi proporzionali (Cox, 1972; Cox e Oakes, 1987). Questo modello è particolarmente rilevante in contesti biomedici e clinici, dove è essenziale comprendere come diverse covariate influenzino il tempo fino al verificarsi di un evento di interesse, come la morte o la recidiva di una malattia.

Il modello di Cox permette di analizzare quali relazioni intercorrono tra covariate, sia categoriali che continue (come età, sesso, trattamenti ricevuti, espressione genica, ecc.) e il rischio di un determinato evento nel tempo. Inoltre, il modello consente di gestire dati censurati a destra, di includere una vasta gamma di variabili esplicative e ha la capacità di fornire stime dei rapporti di rischio (*hazard ratios*) associati alle covariate, facilitando l'identificazione dei fattori di rischio. Infatti, per la r-ma covariata, $exp(\beta_r)$ esprime di quanto varia moltiplicativamente il rischio dell'evento di interesse per ogni variazione unitaria della covariata x_r, a parità di tutte le altre covariate, dove β_r è il coefficiente della covariata x_r. Tutte queste caratteristiche lo rendono uno strumento fondamentale, estremamente flessibile e potente per l'analisi dei dati di sopravvivenza.

Tale modello descrive il rischio in funzione del tempo t e delle covariate $(x_1, \ldots, x_p)$, senza però esplicitare la dipendenza dal tempo. In particolare, si assume che la funzione di rischio sia espressa come

$$\lambda_T(t; \boldsymbol{x}) = \lambda_0(t) \exp(\beta_1 x_1 + \ldots + \beta_p x_p),$$

dove $\lambda_0(t)$ rappresenta il rischio di base, una funzione ignota che dipende solamente dal tempo ed è la stessa per tutti i soggetti. La dipendenza da t è quindi inglobata interamente nella funzione $\lambda_0(t)$, senza alcuna assunzione specifica per il suo andamento nel tempo.

Si tratta di un modello a rischi proporzionali (PH, *proportional hazards*), in cui le costanti di proporzionalità sono determinate dai termini esponenziali che dipendono solo dalle variabili esplicative e non dal tempo. Infatti, il rapporto tra i rischi di due soggetti, indicando con $\boldsymbol{x}^1$ e $\boldsymbol{x}^2$ i vettori delle rispettive covariate, è dato da

$$HR = \frac{\lambda_T(t; \boldsymbol{x}^1)}{\lambda_T(t; \boldsymbol{x}^2)} = exp(\beta_1(x_1^1 - x_1^2) + \ldots + \beta_p(x_p^1 - x_p^2)),$$

che è indipendente dal tempo. Quindi, sebbene le funzioni di rischio varino nel tempo, il loro rapporto si mantiene costante.

La funzione di sopravvivenza può essere ricavata dalla funzione di rischio cumulato $\Lambda_T(t; \boldsymbol{x}) = \Lambda_0(t) \exp(\beta_1 x_1 + \ldots + \beta_p x_p)$ da cui segue che la funzione di sopravvivenza è

$$S_T(t; \boldsymbol{x}) = P(T > t|\mathbf{x}) = S_0(t)^{exp(\beta_1 x_1 + \ldots + \beta_p x_p)}.$$

La funzione di sopravvivenza stimata $\hat{S}_T(t; \boldsymbol{x})$ può essere utilizzata per fare previsioni sulla probabilità di sopravvivenza dei pazienti nel tempo, in base alle loro covariate. Ad esempio, in un contesto clinico, è possibile utilizzare questa funzione per stimare la probabilità che un paziente con determinate caratteristiche sopravviva oltre un certo periodo di tempo.

Nel modello di Cox, la funzione di verosimiglianza completa dipende sia dal vettore dei parametri di regressione $\boldsymbol{\beta}$ sia dal parametro ignoto $\lambda_0(t)$. Cox (1972) ha proposto di fattorizzare la verosimiglianza completa come

$$L(\lambda_0, \boldsymbol{\beta}) = L(\lambda_0) L(\boldsymbol{\beta}|\lambda_0)$$

e di considerare per l'inferenza su $\boldsymbol{\beta}$ soltanto la seconda componente, ignorando in tale modo il parametro di disturbo $\lambda_0(t)$. Si considera in pratica la sola verosimiglianza parziale $L_p(\beta) = L(\boldsymbol{\beta}|\lambda_0)$, che permette di stimare la componente parametrica $\boldsymbol{\beta}$ senza dover specificare la forma di $\lambda_0(t)$.

Siano $t_{(1)}, t_{(2)}, \ldots, t_{(J)}$ i tempi all'evento ordinati e sia $R(t_{(j)})$ l'insieme dei soggetti a rischio al tempo $t_{(j)}$, con $j = 1, \ldots, J$. La verosimiglianza parziale può essere espressa come

$$L_P(\beta) = \prod_{j=1}^{J} \frac{\lambda_T(t_{(j)}; x_{(j)})\, dt}{\sum_{i \in R(t_{(j)})} \lambda_T(t_{(j)}; x_i)\, dt}. \tag{1.1}$$

Sostituendo la funzione di rischio $\lambda_T(t; x) = \lambda_0(t) \exp(\beta_1 x_1 + \cdots + \beta_p x_p)$, si ottiene

$$\begin{aligned} L_P(\beta) &= \prod_{j=1}^{J} \frac{\lambda_0(t_{(j)}) \exp(\beta_1 x_{1j} + \cdots + \beta_p x_{pj})}{\sum_{i \in R(t_{(j)})} \lambda_0(t_{(j)}) \exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})} \\ &= \prod_{j=1}^{J} \frac{\exp(\beta_1 x_{1j} + \cdots + \beta_p x_{pj})}{\sum_{i \in R(t_{(j)})} \exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})}. \end{aligned}$$

Massimizzando (1.1), si ottengono le stime $\hat{\beta}_1, \ldots, \hat{\beta}_p$, che quantificano l'associazione tra ciascuna covariata e il rischio di evento. Tali stime $\hat{\boldsymbol{\beta}}$ vengono solitamente calcolate tramite algoritmi iterativi, come il metodo di *Newton-Raphson* (Press et al., 2007).

Questa tecnica consente di ottenere stime efficienti per i coefficienti β senza dover conoscere la forma della funzione di rischio di base, il che rende il modello di Cox particolarmente flessibile.

In sintesi, il modello di Cox a rischi proporzionali rappresenta uno strumento molto importante nel contesto dell'analisi dei dati di sopravvivenza, consentendo di valutare l'influenza di variabili esplicative sul rischio di un evento nel tempo e di gestire la censura dei dati. La sua capacità di fornire stime dei rapporti di rischio e di ricavare funzioni di sopravvivenza lo rende particolarmente utile in contesti biomedici e clinici.

Nella sezione successiva, verrà fornita una panoramica dettagliata dei dati utilizzati in questo studio, evidenziando le caratteristiche principali del *dataset*.

1.3 I dati

I dati utilizzati in questa tesi sono i medesimi utilizzati da Smith e Sheltzer (2022) nell'articolo in esame e provengono dal programma *The Cancer Genome Atlas* (TCGA) del *National Institutes of Health* (NIH) (Tomczak et al., 2015). Questo progetto, avviato nel 2006, ha raccolto profili molecolari multi-piattaforma e dati clinico-patologici relativi a oltre 11,000 tumori appartenenti a 33 diversi tipi di cancro, con l'obiettivo di facilitare la caratterizzazione molecolare dei principali tipi di cancro presenti negli Stati Uniti. Sebbene i dati clinici e patologici siano stati raccolti per ogni paziente, l'analisi genomica è stata prioritaria rispetto al *follow-up* clinico.

I dati TCGA sono stati acquisiti dal TCGA - *PanCanAtlas* (The Cancer Genome Atlas Research Network et al., 2013). Maggiormente nello specifico, si è utilizzato il *file EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv*, il quale contiene il numero di *reads* allineate per ciascun gene, calcolate con il *software* STAR. Sebbene i dati siano stati preventivamente normalizzati, il metodo di normalizzazione utilizzato non è esplicitamente dichiarato. È stato utilizzato anche il *file TCGA-CDR-SupplementalTableS1.xlsx*, contenente le variabili cliniche. Per ulteriori dettagli sui dati, si rimanda al seguente link: https://gdc.cancer.gov/about-data/publications/pancanatlas.

Si noti che, l'analisi di sopravvivenza sulle coorti di TCGA è considerata appropriata e affidabile. Studi come quello di Smith e Sheltzer (2022) e Liu et al. (2018) hanno

dimostrato che le coorti di TCGA forniscono dati sufficientemente robusti e dettagliati per condurre analisi approfondite sulla sopravvivenza dei pazienti.

In particolare, per questa tesi, per ragioni computazionali, si è deciso di analizzare un sottoinsieme dei dati disponibili. Infatti, gli stessi autori, Smith e Sheltzer (2022), affermano che per eseguire l'intero codice sviluppato da loro in Python, applicato a tutti i tipi di tumore, a tutte le piattaforme e a tutti i geni presenti nei *dataset*, è necessaria una notevole potenza di calcolo, richiedendo almeno 30 *core* e diverse centinaia di GB di RAM. Per tale motivo e poiché nell'articolo di riferimento hanno fornito la maggior parte delle informazioni prognostiche, insieme ai dati sulla metilazione del DNA e ai dati sulle alterazioni del numero di copie (*Copy Number Alterations*, CNAs), sono stati utilizzati i dati di *RNA-seq*. Si è inoltre deciso di focalizzare l'analisi su pazienti affetti da adenocarcinoma del colon (COAD). Questi dati sono stati scelti come esempio rappresentativo all'interno del vasto *dataset* del TCGA. Le diverse analisi e conclusioni, tuttavia, possono essere estese ad altri tipi di cancro e alle altre piattaforme presenti nel *database*.

Inoltre, l'analisi, inizialmente, si è concentrata su un singolo *pathway* considerato biologicamente rilevante: hsa05230 (*Central carbon metabolism in cancer - Homo sapiens*). Questo *pathway* è stato selezionato per la sua rilevanza nel metabolismo centrale del carbonio nei tumori ed è stato tratto dalla *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (Kanehisa e Goto, 2000). In Figura 1.1 si può osservare come questo *pathway* subisca sostanziali modifiche nelle cellule maligne; inoltre, emerge che i geni Ras, PI3K, Akt e c-Myc sono oncogeni, mentre SIRT3, SIRT6 e p53 agiscono come soppressori tumorali. Infatti, la trasformazione maligna delle cellule richiede adattamenti specifici del metabolismo cellulare per sostenere la loro crescita e sopravvivenza (per ulteriori dettagli sul *pathway* selezionato e sul suo legame con il cancro si rimanda a Warburg (1956)). L'approfondimento di questo *pathway* fornisce una prospettiva importante per comprendere come le cellule tumorali modifichino il loro metabolismo per facilitare la propria crescita e sopravvivenza. Lo studio di questi meccanismi, infatti, può rivelare potenziali target terapeutici e migliorare la comprensione delle dinamiche molecolari che guidano la progressione del cancro.

Le covariate utilizzate nell'analisi sono le stesse adottate da Smith e Sheltzer (2022), ovvero le variabili cliniche fornite da Liu et al. (2018). Infatti, per garantire un uso appropriato dell'ampio *dataset* clinico del TCGA associato alle caratteristiche genomiche, Liu et al. (2018) hanno sviluppato un *dataset* standardizzato chiamato TCGA

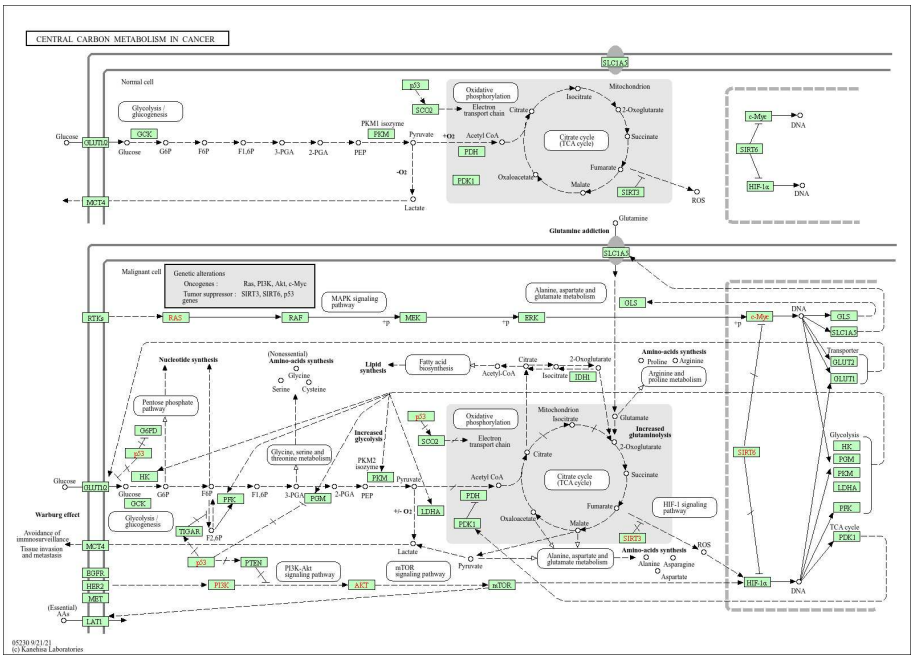

FIGURA 1.1: *Metabolismo centrale del carbonio nei tumori - Homo sapiens (human).* Nel pannello in alto è possibile visualizzare il normale funzionamento di questo *pathway*, mentre nel pannello in basso il suo funzionamento in una cellula maligna. *Tratto da KEGG.*

Pan-Cancer Clinical Data Resource (TCGA-CDR). Questo *dataset* include quattro principali *endpoint* di sopravvivenza clinica: sopravvivenza globale (OS), intervallo libero da progressione (PFI), intervallo libero da malattia (DFI) e sopravvivenza specifica per malattia (DSS).

I diversi *endpoint* clinici sono definiti come segue:

- **sopravvivenza globale (OS)**: il periodo che intercorre dalla data di diagnosi fino alla data di morte per qualsiasi causa. Il tempo censurato è calcolato dalla data di diagnosi iniziale fino alla data dell'ultimo contatto;

- **intervallo libero da progressione (PFI)**: il periodo che va dalla data di diagnosi fino alla prima insorgenza di un nuovo evento tumorale, tra cui: progressione della malattia, recidiva locoregionale, metastasi a distanza, nuovo tumore primario o morte con tumore. I pazienti vivi senza aver sperimentato uno di questi eventi o deceduti senza tumore sono censurati;

- **intervallo libero da malattia (DFI)**: il periodo che va dalla data di diagnosi fino alla data del primo evento di progressione tumorale successivo alla determinazione di uno stato libero dalla malattia dopo la diagnosi e il trattamento iniziali. Questo include: recidiva locoregionale, metastasi a distanza, sviluppo di un nuovo tumore

primario nello stesso organo o morte per avanzamento dello stesso tumore. Il tempo censurato è calcolato dalla data di diagnosi iniziale alla data dell'ultimo contatto o alla data di morte senza aver sperimentato l'evento di interesse;

- **sopravvivenza specifica per malattia (DSS)**: il periodo che va dalla data di diagnosi iniziale fino alla data di morte per la malattia. Il tempo censurato va dalla data di diagnosi iniziale fino alla data dell'ultimo contatto o fino alla data di morte per altra causa.

Quindi, nel determinare gli *endpoint* PFI, DFI e DSS, Liu et al. (2018) hanno censurato i pazienti che sono deceduti senza aver sperimentato l'evento di interesse. Di conseguenza, si è ipotizzato che, se un paziente non fosse deceduto per altre cause, avrebbe sperimentato l'evento di interesse.

L'uso della sopravvivenza globale (OS) è comune, ma presenta delle limitazioni poiché, includendo decessi per cause non legate al cancro, manca di specificità e non riflette necessariamente la biologia del tumore, la sua aggressività o la risposta alla terapia. Al contrario, la sopravvivenza specifica per malattia (DSS), che considera unicamente la morte per il tipo di cancro diagnosticato, garantisce una maggiore rilevanza per comprendere la biologia del cancro e l'impatto terapeutico. Tuttavia, ricavare la DSS è complesso poiché solo 6 dei 33 tipi di cancro inclusi nel *database* TCGA comprendono anche informazioni riguardanti la causa della morte. Inoltre, l'uso della OS o della DSS richiede lunghi tempi di *follow-up*; pertanto, in numerosi studi clinici si preferisce utilizzare *endpoint* come DFI o PFI che fungono da surrogati per indicare una futura mortalità per cancro, evitando in questo modo la necessità di un *follow-up* prolungato. PFI, a differenza di DFI, è associato a meno ambiguità, poiché non è necessario determinare se un paziente abbia mai raggiunto uno stato libero dalla malattia dopo la diagnosi e il trattamento iniziali (Liu et al., 2018).

La sopravvivenza globale (OS) è un *endpoint* appropriato per quei tumori che sono maggiormente aggressivi, come COAD, in cui il tempo mediano per un evento di OS è di 13.3 mesi e di conseguenza il *follow-up* mediano di 22 mesi ha permesso di rilevare eventi nel 22.2% dei casi (102 eventi di OS su 459 casi). Al contrario, per un tipo di cancro meno aggressivo come il carcinoma della prostata (PRAD), dove si sono verificati solo 10 eventi di OS su 500 casi (2% dei casi), la sopravvivenza globale (OS) non è, chiaramente, un *endpoint* adeguato allo studio. È pratica comune nelle sperimentazioni cliniche quella di scegliere *endpoint* intermedi, come la progressione della malattia o gli eventi di recidiva, quando si lavora con tumori meno aggressivi, poichè gli esiti di

sopravvivenza complessiva richiederebbero altrimenti eccessivi tempi di *follow-up* per ottenere delle analisi significative (Liu et al., 2018).

Sulla base delle raccomandazioni sull'uso degli *endpoint* per ciascun tipo di cancro fornite da Liu et al. (2018) e per coerenza con l'articolo di Smith e Sheltzer (2022), si è scelto di utilizzare la sopravvivenza globale (OS) come *endpoint* clinico per l'analisi di COAD. Le variabili confondenti considerate sono le stesse utilizzate da Smith e Sheltzer (2022) e fornite nel TCGA-CDR, in quanto ritenute ragionevoli, ovvero: età alla diagnosi patologica iniziale, genere e stadio patologico del tumore secondo la classificazione AJCC (Amin et al., 2017):

- **Livello 1**: Stadio I, Stadio IA;
- **Livello 2**: Stadio II, Stadio IIA, Stadio IIB, Stadio IIC;
- **Livello 3**: Stadio III, Stadio IIIA, Stadio IIIB, Stadio IIIC;
- **Livello 4**: Stadio IV, Stadio IVA, Stadio IVB.

Si noti che, il trattamento può fungere da potenziale fattore di confondimento e si dovrebbe tenerne adeguatamente conto quando disponibile. Tuttavia, i dati sul trattamento non sono stati inclusi nel TCGA-CDR poichè la storia del trattamento potrebbe non essere completa per vari motivi: non tutte le istituzioni da cui sono stati prelevati i campioni di tessuto per il progetto TCGA erano luoghi di trattamento, gli aggiornamenti erano limitati per i pazienti arruolati prospetticamente e vi era una notevole eterogeneità nei dati di trattamento (Liu et al., 2018).

Liu et al. (2018) hanno selezionato campioni di tumore primario, non metastatico, per la caratterizzazione molecolare, fatta eccezione per lo studio sul melanoma cutaneo (SKCM), che consentiva l'inclusione di entrambi i tipi di campioni. Infatti, l'inclusione di sottogruppi molecolari come predittori potrebbe potenzialmente compromettere la significatività statistica delle apparenti differenze di esito. Pertanto, raccomandano di utilizzare solo i dati molecolari provenienti dai tumori primari, poiché i dati clinici corrispondenti, inclusi importanti dettagli temporali, sono stati raccolti in modo completo solamente al momento della diagnosi iniziale. Per tale motivo, si è ritenuto opportuno limitare l'analisi ai soli campioni di tumore primario.

Nel capitolo successivo verranno descritte maggiormente nel dettaglio le metodologie adottate da Smith e Sheltzer (2022), con l'obiettivo di replicarne il procedimento.

Capitolo 2

Riproducibilità

La riproducibilità è un principio fondamentale della ricerca scientifica, essenziale per confermare la validità e l'affidabilità dei risultati. La riproduzione degli esperimenti consente di verificare se le scoperte siano effettivamente generalizzabili e non frutto di artefatti metodologici. In particolare, un'analisi è detta riproducibile se, partendo dagli stessi dati grezzi e dal codice, un altro analista è in grado di riprodurre gli stessi risultati dell'analisi originale. Si noti, però, che la riproducibilità è una condizione necessaria, ma non sufficiente, per affermare che un'analisi sia corretta.

In questo capitolo, si descriverà lo studio di Smith e Sheltzer (2022) presentandone successivamente una replica. Attraverso questo approccio, ci si propone di evidenziare eventuali discrepanze nei risultati e di testare la robustezza delle conclusioni raggiunte. L'analisi presterà particolare attenzione ai metodi utilizzati per affrontare il problema relativo alla molteplicità dei test.

2.1 Correzioni per test multipli

Nelle circostanze in cui si eseguono numerosi test, è cruciale considerare il problema della molteplicità dei test. In questo contesto, si effettuano dei confronti multipli, eseguendo un test per ciascun gene, testando, dunque, numerose ipotesi simultaneamente. È necessario tenere conto del fatto che, per puro caso, si potrebbero generare dei falsi positivi. Pertanto, risulta essenziale applicare correzioni appropriate per mantenere il livello di significatività desiderato e ridurre il rischio di trarre conclusioni errate.

Infatti, nel contesto di un test d'ipotesi, si è interessati a confrontare un'ipotesi nulla, H_0, contro un'ipotesi alternativa, H_1. Utilizzando una statistica test, si verifica se nei dati esiste sufficiente evidenza per rifiutare l'ipotesi nulla. A tal fine, si fissa un livello di significatività α e, se si osserva un p-value inferiore a α, si rifiuta l'ipotesi nulla.

La probabilità di commettere un errore di primo tipo è $Pr(\text{Rifiuto } H_0|H_0 \text{ è vera}) = \alpha$. Tuttavia, se si effettuano N test indipendenti con livello di significatività α, la probabilità di rifiutare almeno un'ipotesi nulla quando tutte sono vere è pari a $1-(1-\alpha)^N$. In tal caso, senza correzioni, si osserva un aumento atteso della percentuale di falsi positivi, motivo per cui è importante considerare tecniche di aggiustamento.

Nel contesto dello studio in esame, Smith e Sheltzer (2022) eseguono 3,091,782 test, se non si controllasse per la molteplicità dei test, considerando un livello di significatività pari a 0.05, ci si aspetterebbe che circa il 5% dei test risultino significativi semplicemente per caso. Pertanto, assumendo che l'ipotesi nulla sia vera per tutti i geni, ci si aspetterebbero circa 154,589 test significativi. Analogamente, considerando un livello di significatività dell'1%, ci si aspetterebbe che circa 30,918 test risultino significativi per caso.

Per affrontare questo problema, si adottano diverse procedure di controllo. In particolare, in questa tesi si adotterà la procedura descritta da Benjamini e Hochberg (1995), così come fatto da Smith e Sheltzer (2022) e la tecnica descritta da Efron, Tibshirani et al. (2001), utilizzata da Gilis et al. (2020).

2.1.1 *Benjamini-Hochberg*

La procedura di Benjamini e Hochberg (1995) è una procedura per il controllo del *False Discovery Rate* (FDR) o tasso di falsi positivi, ovvero la frazione attesa di falsi positivi tra le ipotesi che sono state dichiarate significative.

Si tratta di una procedura sequenziale. Si supponga di avere J ipotesi nulle $H_1, \ldots, H_J$ e i corrispettivi p-value $p_1, \ldots, p_J$. La procedura consiste nell'ordinare i p-value dal più basso al più alto: $p_{(1)}, \ldots, p_{(J)}$ e per un livello di significatività α, considera il più grande k tale che $p_{(k)} \leq \frac{k}{J}\alpha$. Dunque, si rifiutano le ipotesi $H_{(1)}, \ldots, H_{(k)}$ e non si rifiutano le altre.

Con la correzione di *Benjamini-Hochberg* (BH), il p-value aggiustato viene calcolato moltiplicando il p-value per il numero di test effettuati e diviso per il rango: $\tilde{p}_{(i)} = \min\left(1, \min_{j \geq i}\left\{\frac{J}{j}p_{(j)}\right\}\right)$.

2.1.2 *Efron*

Nell'ambito dell'analisi di dati genomici, è comunemente assunto che la maggior parte dei geni possa essere considerata nulla, ovvero non associati alla sopravvivenza. In

assenza di effetti reali che indichino un'associazione tra l'espressione genica e la sopravvivenza, si assume che la distribuzione dei p-value, i quali quantificano la significatività dell'associazione tra espressione genica e sopravvivenza, sia uniformemente distribuita nell'intervallo [0, 1]. Tuttavia, la presenza di un numero relativamente ridotto di geni effettivamente associati alla sopravvivenza provoca un'inflazione dei p-value prossimi a zero. Pertanto, in un contesto ideale, la distribuzione dei p-value si presenta come una mistura di valori molto bassi, indicativi di geni associati alla sopravvivenza, e di una distribuzione uniforme su [0, 1], rappresentativa dei geni non associati alla sopravvivenza.

Come illustrato nella Figura 2.1, la distribuzione osservata dei p-value nel caso in esame suggerisce una deviazione dalla distribuzione nulla attesa delle statistiche del test.

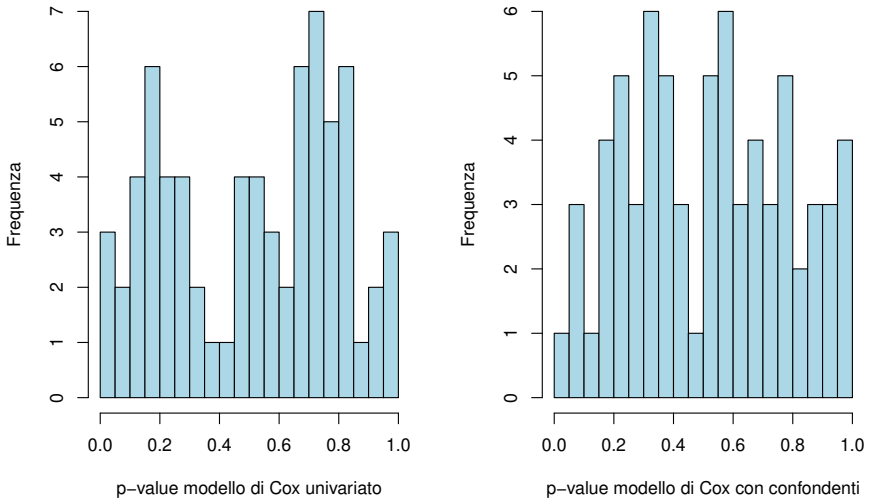

FIGURA 2.1: *Istogramma della distribuzione dei p-value per il modello a rischi proporzionali di Cox.*
Pannello di sinistra: p-value per i modelli univariati. **Pannello di destra**: p-value per i modelli con i confondenti: sesso, età alla diagnosi e stadio del tumore.

Secondo quanto riportato da Efron, Tibshirani et al. (2001) e ribadito da Gilis et al. (2020), vi sono quattro ragioni principali per cui la distribuzione nulla teorica può fallire: (I) il fallimento delle assunzioni matematiche, (II) la correlazione tra geni, (III) la correlazione tra pazienti e (IV) la presenza di confondenti non osservati negli studi osservazionali. Per ovviare a questi problemi, Efron, Tibshirani et al. (2001) propongono una tecnica che stima empiricamente la distribuzione nulla.

Per implementare questa tecnica, è stato utilizzato il pacchetto `locfdr` (Efron, Turnbull et al., 2011), seguendo l'approccio descritto da Gilis et al. (2020). Come indicato da Efron, Tibshirani et al. (2001), i p-value sono stati convertiti in z-value mediante la seguente formula:

$$z_i = \Phi^{-1}(1 - p_i),$$

dove p_i rappresenta il p-value originale che indica la significatività dell'associazione tra l'espressione del gene i e la sopravvivenza, mentre Φ è la funzione di distribuzione cumulativa per la distribuzione normale standard e z_i è lo z-value risultante per il gene i.

Il punto di partenza è che esistono due tipi di ipotesi: le ipotesi nulle vere e le ipotesi alternative vere. Tale modello, noto come modello a due gruppi, formalizza la presenza di questi due gruppi. Efron, Tibshirani et al. (2001) assumono una mistura per gli z-score:

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z),$$

dove $f_0(z)$ rappresenta la distribuzione degli z-score per i geni non associati alla sopravvivenza (sotto H_0), $f_1(z)$ è la distribuzione degli z-score per i geni associati alla sopravvivenza (sotto H_1) e π_0 è la proporzione di geni non associati alla sopravvivenza.

Il tasso locale di falsi positivi, noto come *local false discovery rate* (lfdr), è definito come

$$lfdr(z) = \frac{\pi_0 f_0(z)}{f(z)},$$

dove $lfdr(z)$ rappresenta la probabilità a posteriori che un gene specifico con un determinato punteggio z sia un falso positivo.

Efron, Tibshirani et al. (2001) hanno anche mostrato il legame tra il lfdr e il FDR. In particolare, si può esprimere il FDR come segue: $FDR(z) = P(Null \mid Z \leq z) = \frac{\pi_0 F_0(z)}{F(z)}$. Pertanto, è possibile affermare che $FDR(z) = E[lfdr(Z) \mid Z \leq z]$.

È importante notare che, mentre il FDR rappresenta una proprietà globale di un insieme di ipotesi, il fdr è una proprietà locale che si applica esclusivamente a un'ipotesi individuale. Nel contesto del FDR, supponendo di utilizzare un α pari a 0.05, si stabilisce che in media solo il 5% dei geni identificati come associati alla sopravvivenza si riveleranno essere falsi positivi. Al contrario, con il lfdr si sostiene che, per ognuno dei geni identificati come associati alla sopravvivenza, la probabilità di derivare dall'ipotesi

nulla è del 5%. Pertanto, si tratta di una probabilità locale che non fornisce informazioni sul numero complessivo di falsi positivi presenti nel *dataset*, nè sulla probabilità di osservare falsi positivi nel contesto analizzato.

Efron, Tibshirani et al. (2001) sostengono inoltre che l'utilizzo di un livello di significatività del lfdr pari a 0.2 corrisponda approssimativamente a un livello di significatività del FDR pari a 0.05 in molte applicazioni pratiche.

Per calcolare il lfdr, è necessario stimare le due distribuzioni, nulla e marginale. Attraverso l'utilizzo del pacchetto `locfdr`, i parametri π_0, $f_0(z)$ e $f(z)$ vengono stimati empiricamente.

2.2 Lo studio di Smith e Sheltzer (2022)

Smith e Sheltzer (2022) eseguono una trasformazione logaritmica in base due dei dati di espressione genica e dei *microRNA*, limitando tali dati a un valore minimo di zero e utilizzandoli successivamente come *input* per i modelli. Come evidenziato nel paragrafo 1.1, gli autori utilizzano il modello a rischi proporzionali di Cox per l'analisi della sopravvivenza, con l'obiettivo di identificare biomarcatori prognostici in pazienti oncologici. In particolare, implementano un modello univariato di Cox per ogni gene, in cui le singole caratteristiche genomiche vengono associate all'esito dei pazienti e un modello di Cox con variabili confondenti, che include le informazioni relative alle variabili cliniche insieme ai dati genomici, sempre per ogni gene. Questo procedimento è eseguito per ogni caratteristica genomica disponibile nel progetto TCGA (CNAs, metilazione, mutazioni, espressione genica, espressione *miRNA* ed espressione proteica) e si estende a ogni tipo di cancro.

L'analisi si basa su un campione le cui dimensioni variano in funzione del numero di pazienti con ciascun tipo di cancro e per ciascuna piattaforma genomica. Nel caso specifico dell'adenocarcinoma del colon (COAD), considerando i dati di *RNA-Seq*, sono disponibili 448 casi.

In totale, Smith e Sheltzer (2022) hanno generato 3,091,782 modelli di Cox univariati, riportando, per ogni modello, lo z-value associato al gene. Gli autori dichiarano di aver identificato 112,303 coppie di caratteristiche genomiche e tipo di cancro significativamente associate al tempo di sopravvivenza dei pazienti. In alcune occasioni, affermano di aver controllato il FDR tramite il metodo di BH, fissato all'1%, mentre altre volte affermano che uno z-value maggiore di 1.96 o inferiore a -1.96 è stato considerato significativo.

Tuttavia, gli autori riportano una mediana di 2,145 caratteristiche genomiche significative per ogni tipo di cancro, rispetto a una mediana di 93,000 caratteristiche genomiche analizzate e per quanto riguarda il COAD, hanno identificato 151 caratteristiche genomiche significative su 99,050 esaminate, corrispondenti a circa lo 0.2%. L'applicazione della correzione di BH agli z-value ottenuti da Smith e Sheltzer (2022) tramite la funzione R `p.adjust`, utilizzando il metodo `BH` o, equivalentemente, `fdr`, ha evidenziato che, considerando un FDR fissato al 5%, si ottengono 151 caratteristiche genomiche significative sulle 99.050 indagate, mentre, riducendolo il livello del FDR all'1%, il numero di caratteristiche significative si riduce a 30. In particolare, per l'espressione genica sono state identificate 145 e 27 caratteristiche genomiche significative, mentre per l'espressione dei *miRNA* 6 e 3, rispettivamente, per un FDR del 5% e dell'1%. Per quanto concerne le CNAs, la metilazione del DNA, le mutazioni e l'espressione proteica, non sono emerse caratteristiche genomiche significative né a un FDR del 5% né a un FDR dell'1%. Ne consegue che, le diverse conclusioni presentate nell'articolo sono attribuibili a un tasso di falsi positivi del 5%.

In aggiunta, si evidenzia come gli z-value ottenuti dai modelli univariati e dai modelli con variabili confondenti risultino altamente concordanti sia all'interno dei singoli tipi di dati, con un coefficiente di correlazione di *Pearson* mediano pari a 0.96, sia all'interno dei singoli tipi di cancro, con un coefficiente di correlazione di *Pearson* mediano pari a 0.95. Questo risultato suggerisce che un numero limitato di marcatori prognostici è stato influenzato dall'inclusione di variabili cliniche, portando gli autori a proseguire le analisi successive utilizzando i modelli univariati in virtù di questo alto grado di concordanza. Tuttavia, nel caso specifico del tumore al colon-retto (COAD), tale correlazione non risulta altrettanto elevata (69.3%).

2.3 Replica dello studio

Le variabili cliniche per i pazienti affetti da adenocarcinoma del colon (COAD) sono relative a 459 individui; tuttavia, è stata osservata la presenza di un valore mancante a livello della risposta "sopravvivenza complessiva", portando all'esclusione di un paziente, consentendo un'analisi su 458 individui. Per quanto riguarda l'espressione genica, sono disponibili dati su 20,531 geni e 11,070 pazienti, sebbene questi dati siano relativi a tutti e 33 i tipi di tumore presenti nel *database*. Si è, quindi, proceduto a mantenere solo i pazienti presenti sia nel *dataset* clinico sia in quello di espressione genica.

Nel campione di 458 individui analizzato, sono stati identificati 216 pazienti di sesso femminile e 242 di sesso maschile. Per quanto riguarda lo stadio della malattia, sono

stati registrati 76 tumori di stadio I, 177 di stadio II, 129 di stadio III e 65 di stadio IV; i restanti casi presentano valori mancanti. Si registrano 102 decessi e 356 censure. Infine, l'età alla diagnosi varia da un minimo di 31 anni a un massimo di 90, con un'età media di circa 67 anni.

Le ricerche condotte da Liu et al. (2018) e Smith e Sheltzer (2022) evidenziano che, come previsto, i tumori di stadio III e IV nel *dataset* TCGA presentano esiti significativamente peggiori rispetto ai tumori di stadio I e II. Inoltre, è stato osservato che i pazienti più anziani tendono ad avere esiti sfavorevoli rispetto ai pazienti più giovani.

Per effettuare il confronto tra le curve di sopravvivenza di maschi e femmine, dei quattro stadi del tumore e dell'età alla diagnosi raggruppata nelle classi considerate da Smith e Sheltzer (2022) per l'analisi non parametrica della sopravvivenza ($< 40, 40-59, 60-79, 80+$), è stata impiegata la funzione `survdiff()` del pacchetto R `survival`, la quale consente di eseguire il test del *log-rank*. Questo metodo non parametrico permette di testare l'ipotesi di uguaglianza delle funzioni di rischio tra due o più gruppi in presenza di dati censurati (per ulteriori dettagli, si veda Kleinbaum e Klein (2012)).

Le curve di *Kaplan-Meier* (KM), riportate in Figura 2.2, ottenute tramite la funzione `ggsurvplot()` del pacchetto R `survminer`, mostrano il risultato del test del *log-rank*. Queste curve suggeriscono che la sopravvivenza non dipende nè dal sesso del paziente (p-value $= 0.56$), nè dall'età alla diagnosi raggruppata in classi (p-value $= 0.12$). Tuttavia, la sopravvivenza è significativamente diversa per i diversi stadi del tumore (p-value < 0.0001), confermando le osservazioni dei precedenti articoli riguardanti lo stadio del tumore per i pazienti con tumore COAD.

Nel *dataset* di espressione genica disponibile, erano presenti 70 dei 71 geni del *pathway* di interesse (hsa05230). Si è, quindi, utilizzato un *dataset* composto da 448 casi per 70 geni.

Per replicare i risultati dell'articolo originale, sono stati implementati modelli di Cox univariati che valutassero la sopravvivenza in relazione a ciascun gene, senza includere covariate e successivamente gli stessi modelli con l'inclusione delle covariate di interesse. Questo dopo aver eseguito la trasformazione logaritmica dei valori di espressione genica. Le analisi menzionate sono state effettuate tramite la funzione `coxph` del pacchetto R `survival`, utilizzata anche da Smith e Sheltzer (2022). La Tabella A.1 in Appendice riporta gli z-value ottenuti da Smith e Sheltzer (2022) e quelli ottenuti nella presente replica. Si osserva che i risultati dell'analisi univariata sono identici, a eccezione di alcune lievi differenze nei valori decimali e nei geni LDHAL6A e RET. In particolare, si

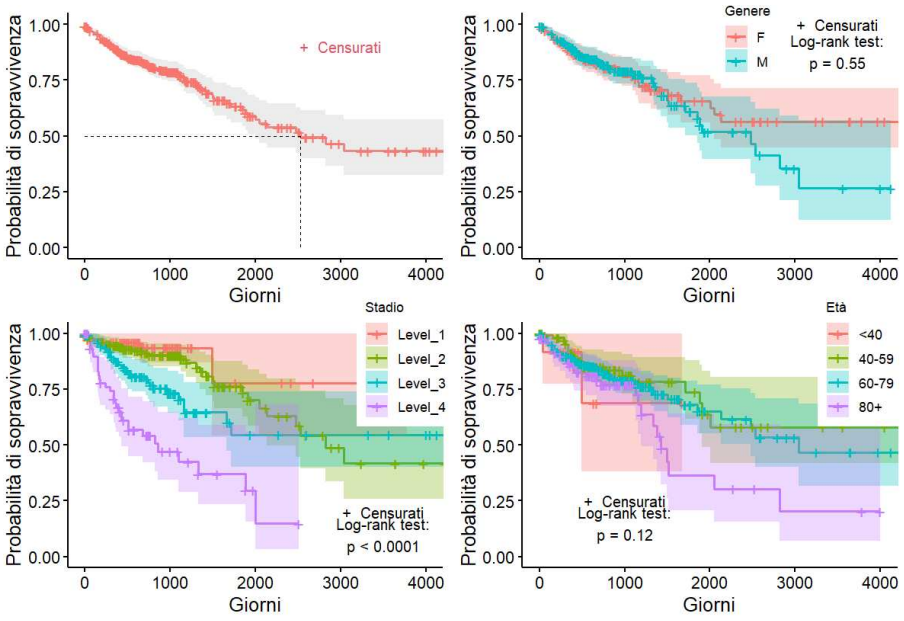

FIGURA 2.2: *Curve sopravvivenza di Kaplan-Meier per i pazienti affetti da COAD e test del log-rank.*
Panello in alto a sinistra: Curva di sopravvivenza globale. La linea tratteggiata rappresenta il tempo di sopravvivenza mediano. **Panello in alto a destra:** sopravvivenza dei pazienti suddivisi in base al sesso. **Panello in basso a sinistra:** sopravvivenza dei pazienti suddivisi in base allo stadio del tumore. **Panello in basso a destra:** sopravvivenza dei pazienti suddivisi in base all'età.

ottiene una correlazione del 99.2% tra i risultati. Mentre, per quanto riguarda l'analisi con i confondenti, la correlazione è del 95.7%. Nella Figura 2.3, è evidente questo comportamento: nel caso univariato solo due geni si discostano in modo evidente, mentre nel caso multivariato emerge una maggiore dispersione.

Nelle Tabelle A.2 e A.3 in Appendice vengono riportati i p-value associati agli z-value ottenuti con il modello di Cox, rispettivamente per il modello univariato e per il modello con confondenti. Oltre ai p-value non corretti, sono riportati i p-value corretti per test multipli mediante il metodo del *false discovery rate* di *Benjamini e Hochberg*, nonchè i p-value corretti per (i) deviazioni dalla distribuzione nulla teorica sottostante e (ii) per test multipli attraverso la procedura lfdr di *Efron*.

In seguito alla correzione per test multipli, effettuata tramite il metodo di *Benjamini-Hochberg*, non sono emersi geni significativi, il che indica l'assenza di evidenze statistiche a sostegno della presenza di geni *marker* associati alla sopravvivenza complessiva per COAD e per il *pathway* considerato. Tuttavia, applicando il metodo proposto da Efron, Turnbull et al. (2011) e utilizzando il livello di significatività di 0.2 consigliato per il lfdr, sono emersi tre geni significativi dai modelli univariati: PFKM ($lfdr = 0.04$), PGAM2 ($lfdr = 0.03$) e SIRT3 ($lfdr = 0.04$). Queste associazioni con la sopravvivenza, tuttavia,

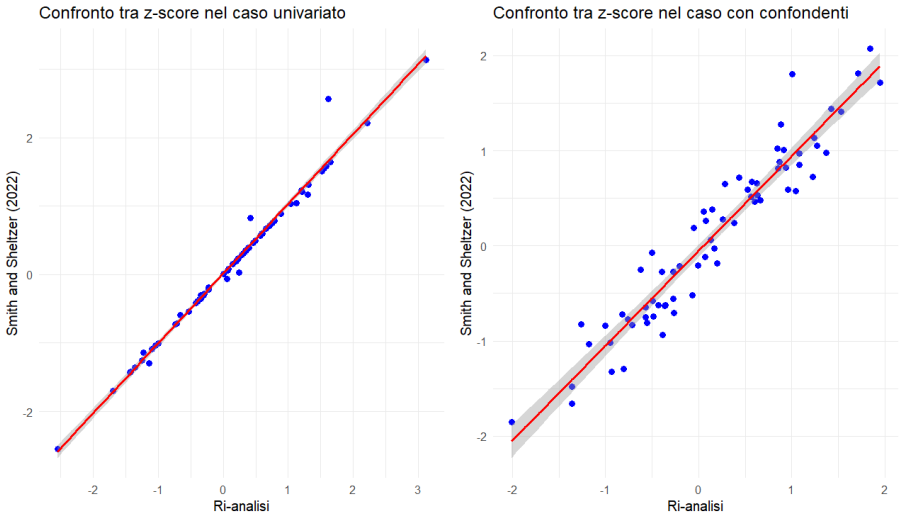

FIGURA 2.3: *Confronto tra gli z-score della replica e dello studio originale di Smith e Sheltzer (2022).* **Pannello di sinistra:** modello di Cox univariato. **Pannello di destra:** modello di Cox con confondenti.
I punti blu rappresentano le coppie di z-score. La linea rossa rappresenta il modello di regressione lineare che meglio si adatta ai dati, mostrando la relazione lineare tra i valori degli z-score della replica e quelli dello studio originale. L'area ombreggiata in grigio intorno alla linea di regressione rappresenta l'intervallo di confidenza e fornisce un'indicazione della precisione della stima della linea di regressione.

scompaiono con l'introduzione delle covariate confondenti, suggerendo una presumibile sovrastima dell'effetto.

Si noti che non è stato possibile stabilire con certezza la corrispondenza dei geni significativi tra i risultati di questo studio e quelli di Smith e Sheltzer (2022), poiché questi ultimi hanno riportato solo gli z-value, senza esplicitare quali geni risultino significativi. Tuttavia, la concordanza tra il numero di geni indicati come significativi dagli autori e gli z-value ottenuti suggeriscono una sostanziale coerenza tra i risultati. Inoltre, questi risultati sono in linea con quelli ottenuti da Gilis et al. (2020). Infatti, l'utilizzo dell'approccio del *local false discovery rate* di *Efron* non ha evidenziato alcuna associazione significativa tra l'espressione genica e la sopravvivenza per i tumori considerati.

2.4 Considerazioni

In questo capitolo, sono stati replicati e convalidati i risultati ottenuti da Smith e Sheltzer (2022) riguardanti l'analisi della sopravvivenza nei pazienti affetti da adenocarcinoma del colon (COAD).

I risultati dell'analisi hanno mostrato un elevato grado di concordanza con quelli

riportati nell'articolo originale, con differenze minime e trascurabili. Tuttavia, l'analisi ha anche evidenziato che le conclusioni di Smith e Sheltzer (2022) possono essere influenzate da un tasso di falsi positivi superiore rispetto a quanto indicato.

Nel prossimo capitolo, ci si concentrerà su modelli maggiormente complessi, considerando simultaneamente più geni per ottenere una comprensione più accurata delle relazioni tra genoma e prognosi.

Capitolo 3

Analisi di sopravvivenza con dati ad alta dimensionalità

Per indagare la relazione tra l'espressione genica di tutti i geni, le variabili cliniche considerate e la sopravvivenza nei pazienti affetti da COAD, si ripercorrono prevalentemente le metodologie illustrate nell'articolo "*Tutorial on survival modeling with applications to omics data*" di Zhao et al. (2024).

In particolare, per condurre quest'analisi, si sono adattati due modelli di Cox senza penalizzazione: uno impiegando solo le covariate cliniche e l'altro con l'inclusione dei geni, selezionati sulla base del criterio di informazione di *Akaike* (AIC; per maggiori dettagli si rimanda ad Akaike (1974)). La selezione dei geni è stata effettuata mediante un approccio di *forward selection*. Inoltre, sono stati adattati due modelli di Cox con penalizzazione: uno con penalizzazione lasso e l'altro con penalizzazione *elastic net*.

Come illustrato nel Paragrafo 1.3, per motivi di efficienza computazionale, si è scelto, inizialmente, di limitare l'analisi ai geni appartenenti al *pathway* di interesse, riducendo notevolmente la dimensionalità dei dati.

Nei prossimi paragrafi verranno esplicitate le modifiche effettuate al *dataset*, spiegate le metodologie utilizzate e illustrati i diversi risultati ottenuti.

3.1 Analisi esplorative

L'analisi esplorativa dei dati rappresenta una fase cruciale in qualsiasi studio scientifico, in particolare quando si tratta di dati ad alta dimensionalità, come quelli derivanti dall'espressione genica. Questa fase permette di comprendere le principali caratteristiche del *dataset* e di individuare eventuali problemi o anomalie.

In questa sezione, verranno descritti i passaggi dell'analisi esplorativa che hanno condotto a una modifica del *dataset* in esame, a partire dal filtraggio dei geni poco espressi, passando per la gestione dei dati mancanti, fino ad arrivare alle modifiche necessarie per l'applicazione di tecniche di penalizzazione come il lasso.

3.1.1 Filtraggio dei geni poco espressi

La fase del filtraggio dei geni poco espressi rappresenta una fase molto importante nell'analisi genomica, in quanto consente di rimuovere i geni poco o per nulla informativi. In questo contesto, sono stati eliminati i geni che presentavano, in media, meno di 10 *reads* tra tutti i campioni. Poiché l'analisi si basa sul logaritmo dell'espressione genica, è stato fissato come valore di discrimine 2.3, corrispondente a un valore esponenziale di circa 10. Questo procedimento ha portato all'eliminazione di 9 geni, riducendo il numero totale da 70 a 61. In questo modo, si è garantito che fossero mantenuti solo i geni con un livello di espressione sufficientemente alto da essere considerati affidabili per l'analisi, migliorando la stabilità e la precisione nella selezione finale delle caratteristiche.

3.1.2 Gestione dei dati mancanti

In precedenza, lavorando con modelli univariati e per coerenza con il lavoro di Smith e Sheltzer (2022), non sono stati utilizzati metodi per l'imputazione dei dati mancanti. Tuttavia, se nell'analisi univariata i pazienti con valori mancanti venivano semplicemente esclusi, nell'analisi multivariata ci sarebbero stati problemi di incoerenza nella dimensione del *dataset*.

Dopo la fase di filtraggio, sono rimaste solamente 11 osservazioni mancanti relative allo stadio del tumore e rappresentando solo il 2.5% del numero totale di osservazioni, si è deciso di imputare questi valori con la mediana. Come si vedrà maggiormente nel dettaglio nel Paragrafo 3.2, i dati sono stati suddivisi in *dataset* di stima e *dataset* di verifica. È fondamentale gestire con attenzione l'imputazione dei valori mancanti per evitare di introdurre informazioni dall'insieme di verifica durante la fase di addestramento, il che potrebbe portare a una stima eccessivamente ottimistica delle prestazioni del modello. Pertanto, l'imputazione è stata effettuata separatamente per ciascun sottoinsieme dei dati anche se, in entrambi i casi, la mediana corrispondeva allo stadio 2 del tumore.

3.1.3 Ulteriori modifiche al *dataset*

Al fine di adattare i dati ai requisiti della tecnica di penalizzazione lasso, la quale necessita esclusivamente di variabili numeriche, le categorie dello stadio di malattia sono state trasformate in tre variabili *dummy*, utilizzando lo stadio 1 come categoria di riferimento.

Inoltre, mentre la regressione lineare è invariante rispetto a trasformazioni di scala delle covariate, le regressioni penalizzate sono sensibili a tali trasformazioni, con potenziali impatti sulle stime dei coefficienti. Poiché il termine di penalità agisce uniformemente su tutti i coefficienti, è necessario che essi siano sulla stessa scala. Pertanto, si è proceduto alla standardizzazione delle variabili, trasformando i dati relativi all'espressione genica e all'età alla diagnosi affinché presentassero media pari a zero e varianza unitaria.

Infine, è stato necessario sostituire i tempi di sopravvivenza pari a zero con il valore 0.0001, poiché le funzioni che implementano il modello di Cox con penalizzazione lasso non accettano tempi non positivi. Tale correzione è stata adottata per garantire una stima accurata del modello ed evitare errori durante l'analisi. Dato che il valore 0.0001 è estremamente vicino a zero, questa modifica non incide in modo significativo sui risultati.

3.2 Insieme di stima e di verifica

Nel contesto dell'analisi statistica, è fondamentale considerare il compromesso tra distorsione e varianza nella selezione del modello. Quando la complessità del modello, rappresentata in questo studio dal numero di covariate, è bassa, si assiste a un incremento della distorsione e una diminuzione della varianza; viceversa, all'aumentare della complessità del modello, la distorsione tende a ridursi, mentre la varianza aumenta. Oltre un certo livello di complessità, la varianza aumenta senza un corrispondente miglioramento della distorsione, comportando un rischio di sovra-adattamento ai dati, è quindi necessario operare una scelta strategica che bilanci queste due componenti in conflitto.

Poiché la distorsione è legata al reale processo generatore dei dati e non può essere calcolata direttamente, una tecnica comune per individuare un equilibrio tra queste due quantità consiste nella suddivisione del *dataset* in due insiemi distinti: insieme di stima e insieme di verifica. L'insieme di stima, selezionato casualmente, viene utilizzato per la stima dei vari modelli candidati, mentre l'insieme di verifica, costituito dalla restante parte dei dati, serve a valutare le prestazioni dei modelli e a identificare quello più

adeguato in un'ottica di previsione. È di fondamentale importanza che la valutazione di un modello prognostico si basi su dati di verifica completamente indipendenti da quelli utilizzati per la stima, altrimenti si rischierebbe di ottenere risultati eccessivamente ottimistici.

Per ulteriori approfondimenti sul compromesso tra distorsione e varianza, nonché sulla tecnica di suddivisione dei dati in insiemi di stima e verifica, si rimanda a Hastie et al. (2009).

Nel presente studio, è stata effettuata una suddivisione casuale dei 448 pazienti affetti da COAD del TCGA, assegnandone il 70% (313 pazienti) all'insieme di stima e il 30% (135 pazienti) all'insieme di verifica.

3.3 Indici per valutare le prestazioni del modello nell'insieme di verifica

L'inclusione delle caratteristiche omiche è giustificata solo se queste nuove covariate apportano un valore prognostico aggiuntivo rispetto ai soli fattori clinici consolidati. In altri termini, il nuovo modello prognostico deve dimostrare un miglioramento significativo delle capacità predittive rispetto al modello di riferimento. In questo contesto, il modello di sopravvivenza basato esclusivamente sulle variabili clinico-demografiche (genere, età alla diagnosi e stadio del tumore) sarà utilizzato come modello di *benchmark*.

Come evidenziato da Zhao et al. (2024), nel contesto dei dati di sopravvivenza, per valutare le capacità predittive di un modello nell'insieme di verifica è necessario considerare sia il potere discriminatorio, ovvero la capacità di classificare correttamente i pazienti in categorie di alto e basso rischio di evento, sia la calibrazione, la concordanza tra le probabilità di sopravvivenza previste e gli esiti osservati. Metriche come la curva ROC (*Receiver Operating Characteristic*), l'area sotto la curva ROC (AUC, *Area Under the Curve*) e l'indice di concordanza sono indicatori di discriminazione, mentre il punteggio di *Brier* viene utilizzato per valutare le prestazioni di calibrazione.

Per ulteriori dettagli sulla validazione dei modelli predittivi nel contesto di dati di sopravvivenza, si rimanda a Rahman et al. (2017) e a Royston e Altman (2013).

È opportuno sottolineare anche l'importanza della verifica delle assunzioni del modello dopo la sua stima. Il modello di Cox, ad esempio, presuppone la proporzionalità dei rischi e una relazione lineare tra il logaritmo del rischio e le covariate. Tuttavia, tali controlli sono ritenuti adeguati principalmente per modelli a bassa dimensionalità.

Come osservato da Zhao et al. (2024), ulteriori sviluppi metodologici sono necessari per il controllo delle assunzioni in contesti ad alta dimensionalità.

Nei paragrafi successivi verranno descritte le metriche utilizzate per valutare il potere prognostico dei modelli nell'insieme di verifica.

3.3.1 Bontà di adattamento

Per valutare il potere discriminatorio di un modello di Cox, i punteggi prognostici, rappresentati dal predittore lineare $\mathbf{X}_i\beta$, possono essere dicotomizzati utilizzando il valore mediano. Questa procedura consente di suddividere i pazienti dell'insieme di verifica in due gruppi: a basso e ad alto rischio di evento. Le curve di sopravvivenza per ciascun gruppo possono quindi essere stimate mediante lo stimatore di *Kaplan-Meier* e confrontate utilizzando il test del *log-rank*.

In modo analogo, i punteggi prognostici possono essere suddivisi in tre o più gruppi mediante i quantili, permettendo di classificare i pazienti in categorie a basso, medio e alto rischio. Anche in questo caso, il test del *log-rank* può essere utilizzato per esaminare le differenze tra le curve di sopravvivenza relative ai diversi gruppi.

3.3.2 Curva ROC e AUC

La curva ROC (*Receiver Operating Characteristic*) rappresenta uno strumento utile per la valutazione della capacità discriminatoria di un modello di sopravvivenza quando si considera un punto temporale fisso.

La costruzione della curva ROC si basa sulla variazione della soglia utilizzata per suddividere i pazienti in due gruppi, consentendo di valutare come la capacità del modello di distinguere tra pazienti a basso e alto rischio di evento cambi al variare del *cut-off* scelto per la classificazione. Per ciascun valore della soglia, si calcolano due metriche fondamentali: la sensibilità e la specificità. La sensibilità, definita come il rapporto tra i veri positivi e la somma di veri positivi e falsi negativi, misura la proporzione di eventi correttamente identificati dal modello. La specificità, invece, è il rapporto tra i veri negativi e la somma di falsi positivi e veri negativi e misura la proporzione di non-eventi correttamente identificati dal modello. Tracciando la sensibilità contro il complemento della specificità ($1 -$ specificità) per ogni valore della soglia, si ottiene la curva ROC.

Questa curva viene sintetizzata attraverso una misura complessiva delle prestazioni del modello, l'AUC (*Area Under the Curve*), ovvero l'area sotto la curva ROC. Un'AUC

pari a 0.5 indica una capacità discriminatoria equivalente a una classificazione casuale, mentre un'AUC vicina a 1 riflette una capacità discriminatoria perfetta.

Sebbene la curva ROC e l'AUC siano utili per valutare la capacità discriminatoria complessiva di un modello, esse considerano tutti i possibili valori di soglia, fornendo così una valutazione globale delle prestazioni. Tuttavia, nella pratica clinica, viene spesso selezionato un singolo valore di soglia per la classificazione. Di conseguenza, le prestazioni del modello possono variare in base alla soglia scelta; pertanto l'AUC e la curva ROC vanno interpretate con cautela, tenendo conto del *cut-off* specifico utilizzato nell'analisi.

Si noti che, essendo il modello di Cox un modello semiparametrico, non fornisce previsioni dirette sulla distribuzione di sopravvivenza. Tuttavia, è possibile trasformare i predittori lineari in previsioni di sopravvivenza utilizzando lo stimatore di *Breslow* per il rischio cumulativo di base (Breslow, 1972). Questa trasformazione consente di ottenere le probabilità di sopravvivenza necessarie per la costruzione della curva ROC e il calcolcolo dell'AUC.

Nel presente studio, è stata considerata la probabilità di sopravvivenza a 10 anni per valutare la capacità del modello di classificare correttamente i pazienti nell'insieme di verifica e per implementare questa metodologia è stato utilizzato il pacchetto `risksetROC` in R (Heagerty e Zheng, 2005).

3.3.3 *C-index*

L'indice di concordanza (Harrell, Lee et al., 1996), o *C-index*, valuta l'abilità di un modello nel distinguere tra pazienti con tempi di sopravvivenza differenti. In particolare, un modello risulta migliore se, confrontando due pazienti, assegna un punteggio prognostico più elevato (indicativo di un rischio maggiore) al paziente con un tempo di sopravvivenza inferiore.

Formalmente, l'indice di concordanza è definito come segue:

$$C = \mathbb{P}\{S(t|\mathbf{X}_i(t)) < S(t|\mathbf{X}_j(t)) \mid T_i < T_j \text{ e } \delta_i = 1\},$$

dove $S(t|\mathbf{X}_i(t))$ rappresenta la probabilità di sopravvivenza al tempo t per il paziente i, dato il vettore di covariate $\mathbf{X}_i(t)$. L'indicatore δ_i assume valore 1 se l'evento di interesse (ad esempio, il decesso) si è verificato per il paziente i e 0 altrimenti.

L'indice di concordanza considera tutte le coppie di pazienti osservati nel *dataset* e, in presenza di censura, include solo le coppie in cui almeno uno dei due pazienti ha sperimentato l'evento.

Questo indice misura la proporzione di coppie di pazienti che sono concordanti. Un valore $C = 0.5$ indica che il modello ha una capacità discriminatoria pari a quella di una classificazione casuale, mentre un valore $C = 1$ rappresenta una perfetta capacità discriminatoria, in cui il modello ordina correttamente tutti i pazienti in base al rischio di sopravvivenza. In generale, un *C-index* maggiore di 0.7 è considerato indicativo di una buona capacità predittiva.

Se per una coppia di pazienti i e j, con $T_i < T_j$, il modello assegna un punteggio prognostico maggiore a i, la coppia è considerata concordante. Al contrario, se il punteggio prognostico assegnato a j è superiore, la coppia è considerata discordante.

Nel contesto del modello di Cox, il quale fornisce una stima del rischio proporzionale basato sulle covariate, la concordanza tra due pazienti si traduce nel confronto dei punteggi prognostici, ovvero dei predittori lineari $\mathbf{X}_i\beta$. In particolare, il modello è concordante se il paziente con un tempo di sopravvivenza inferiore ha un predittore lineare maggiore, ovvero se $\mathbf{X}_i\beta > \mathbf{X}_j\beta$ quando $T_i < T_j$ (Rahman et al., 2017).

L'indice di concordanza fornisce dunque una valutazione globale della capacità di un modello di distinguere tra pazienti con diverse prognosi, senza essere influenzato da un punto temporale specifico. Tuttavia, essendo una misura globale, non fornisce indicazioni dettagliate sulla capacità del modello di predire eventi a tempi specifici, aspetto che potrebbe essere invece catturato da misure temporali più specifiche come l'AUC tempo-dipendente.

Esistono diversi tipi di indici di concordanza utilizzati nella modellazione della sopravvivenza, tra cui l'indice di *Harrell* (Harrell, Califf et al., 1982) e quello di *Uno* (Uno et al., 2011). La scelta tra questi due indici dipende principalmente dal livello di censura presente nei dati. L'indice di *Harrell* (C_H) è ideale per *dataset* con una bassa percentuale di censura, poiché considera solo le coppie in cui almeno uno dei due tempi è osservato. Al contrario, l'indice di *Uno* ($C_U(\tau)$) è adatto in presenza di censura elevata o non indipendente, poiché include pesi basati sulla probabilità di censura per le coppie. Nel dataset analizzato, con 102 decessi e 356 censure, risulta quindi più opportuno utilizzare l'indice di *Uno* per ridurre l'impatto della censura sulla stima della concordanza.

3.3.3.1 C_H di Harrell, Califf et al. (1982)

L'indice di concordanza C_H proposto da *Harrell* considera tutte le coppie di pazienti in cui almeno un evento è osservato e stima C_H come la proporzione di tali coppie in cui il paziente con il tempo di sopravvivenza inferiore ha anche un rischio predetto più elevato.

Nella pratica, il modello di Cox viene utilizzato per calcolare i punteggi prognostici dei pazienti e successivamente, lo stimatore C_H valuta la capacità del modello di ordinare correttamente i pazienti rispetto ai loro tempi di sopravvivenza. L'indice di *Harrell* può essere interpretato come la probabilità che, in una coppia casuale di pazienti, il modello classifichi correttamente i loro rischi di evento.

L'implementazione è stata effettuata tramite la funzione `Cindex()` del pacchetto `glmnet` in R.

3.3.3.2 C_U di Uno et al. (2011)

In presenza di censura, l'indice C_H può risultare distorto, poiché esclude le coppie di pazienti per le quali uno dei due tempi osservati è censurato, un problema che può emergere anche nel caso di censura indipendente. Per ovviare a questa limitazione, Uno et al. (2011) hanno proposto uno stimatore alternativo, $C_U(\tau)$, il quale utilizza un metodo di pesatura basato sulla probabilità di censura, migliorando così la stima della concordanza in presenza di dati censurati.

Un parametro chiave di C_U è τ, l'orizzonte temporale entro il quale viene calcolata la concordanza. La scelta di τ consente di focalizzare l'analisi su un intervallo di tempo di maggiore interesse, evitando di considerare dati scarsamente informativi a tempi più lontani. Tuttavia, per questa applicazione, non è si è specificato τ, permettendo alla stima della concordanza di utilizzare l'intero arco temporale disponibile.

L'implementazione è stata realizzata tramite la funzione `UnoC()` del pacchetto `survAUC` in R.

3.3.4 Punteggio di *Brier*

Il punteggio di *Brier* (Brier, 1950) è una metrica utile per valutare la qualità delle previsioni di un modello di sopravvivenza, in quanto considera sia la discriminazione, ovvero la capacità del modello di distinguere tra eventi e non eventi, sia la calibrazione, ossia la capacità delle probabilità previste di riflettere accuratamente le probabilità osservate.

Il punteggio di *Brier* dipendente dal tempo rappresenta l'errore quadratico medio atteso nella previsione della probabilità di sopravvivenza ed è calcolato come segue:

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{S}(t|X_i)^2\mathbb{I}\{T_i \leq t, \delta_i = 1\}}{\hat{G}(t_i)} + \frac{(1-\hat{S}(t|X_i))^2\mathbb{I}\{T_i > t\}}{\hat{G}(t)}\right]$$

dove:

- t_i è il tempo di sopravvivenza osservato dell'individuo i-esimo;
- $\hat{S}(t|X_i)$ è la probabilità di sopravvivenza predetta dal modello per l'individuo i-esimo al tempo t.
- $\mathbb{I}\{\cdot\}$ è una funzione indicatrice che assume valore 1 se la condizione all'interno delle parentesi graffe è vera, 0 altrimenti;
- δ_i è l'indicatore di evento, che vale 1 se l'evento è osservato e 0 in caso di censura;
- $\hat{G}(t)$ rappresenta la stima della distribuzione di censura, calcolata tramite il metodo di *Kaplan-Meier*.

Questo punteggio dipendente dal tempo incorpora quindi sia i casi censurati sia gli eventi osservati, offrendo una misura del grado di accuratezza delle previsioni di sopravvivenza in funzione del tempo.

È inoltre possibile calcolare il punteggio di *Brier* integrato (IBS), il quale fornisce una misura complessiva e riassuntiva dell'accuratezza predittiva del modello su un intero periodo di *follow-up*, sintetizzando l'errore predittivo medio del modello sull'intero periodo di osservazione. L'IBS viene ottenuto integrando il punteggio di *Brier* dipendente dal tempo $BS(t)$ sull'intervallo di tempo considerato:

$$IBS = \frac{1}{\tau}\int_0^{\tau} BS(t)\,dt$$

dove τ rappresenta la durata del periodo di *follow-up*.

Il punteggio di *Brier* varia tra 0 e 1, dove un valore più basso indica una migliore accuratezza predittiva del modello. In generale, un punteggio inferiore a 0.1 è indicativo di un buon livello di accuratezza, mentre un punteggio superiore a 0.25 suggerisce una bassa capacità di previsione e calibrazione.

Per l'implementazione questo metodo si è utilizzata la funzione `Score()` del pacchetto `riskRegression` in R. Nella stima dell'IBS, è stato impostato come orizzonte

temporale τ il tempo massimo disponibile nel *dataset* di validazione, consentendo di considerare l'intero periodo di osservazione disponibile per ottenere una misura riassuntiva dell'errore predittivo medio del modello.

3.4 Modello di Cox multivariato

È stato implementato un modello di Cox multivariato utilizzando la funzione `coxph` in R, partendo da un modello di base che includeva esclusivamente le covariate clinico-demografiche, ossia genere, età alla diagnosi e stadio della malattia.

Successivamente, mediante una procedura di *forward selection*, i predittori genomici sono stati progressivamente inseriti nel modello, valutando l'impatto di ciascun gene sulle prestazioni del modello. In particolare, le variabili sono state aggiunte una alla volta, includendo di volta in volta il gene che determinava il maggiore decremento del *criterio di informazione di Akaike* (AIC) (Akaike, 1974), il quale penalizza i modelli eccessivamente complessi per evitare il rischio di sovra-adattamento.

Nella Tabella 3.1 è riportato l'ordine di inserimento dei geni, il loro contributo al modello in termini di AIC, il valore del test di verosimiglianza associato al gene inserito a ciascuno *step* e il relativo p-value, mentre la Tabella 3.2 mostra l'*output* del modello selezionato, con i coefficienti stimati, gli *hazard ratio* e i valori di significatività associati ai coefficienti.

3.5 Regressione penalizzata

La regressione penalizzata (Hastie et al., 2009) è una tecnica ampiamente utilizzata per migliorare le capacità predittive dei modelli statistici, specialmente in presenza di un elevato numero di variabili. Questa metodologia trova particolare applicazione nella medicina personalizzata per il cancro, dove l'obiettivo è identificare un sottoinsieme parsimonioso di caratteristiche rilevanti legate agli esiti di sopravvivenza (Mohr et al., 2024).

Infatti, in contesti in cui il numero di covariate è molto elevato e supera il numero di osservazioni, la stima dei parametri può risultare impossibile senza qualche forma di sparsità. In questi casi, la penalizzazione può essere utilizzata per ottenere modelli più semplici e interpretabili. Tuttavia, anche nei contesti non sparsi, metodi come la regressione *ridge* possono essere utili per ridurre la varianza delle stime, pur senza annullare i coefficienti. La tecnica lasso, invece, è particolarmente adatta quando si

TABELLA 3.1: *Ordine di inserimento dei geni nel modello di Cox basato sulla selezione stepwise in avanti.*
Colonna 1: Passo di inserimento del gene nel modello. **Colonna 2:** Gene aggiunto al modello a ciascun passo. **Colonna 3:** Valore dell'AIC del modello dopo l'aggiunta del gene. **Colonna 4:** Valore del test di verosimiglianza (*Likelihood Ratio Test*) per il gene aggiunto. **Colonna 5:** p-value associato al test di verosimiglianza del gene aggiunto.

Step	**Gene**	**AIC**	**LRT**	**Pr**$(>\chi^2)$
Modello Base	-	732.32	-	-
1	MET	729.62	4.70	0.03 *
2	PIK3CA	727.47	4.16	0.04 *
3	PIK3R2	727.06	2.41	0.12
4	PTEN	726.87	2.19	0.14
5	C12orf5	726.36	2.51	0.11
6	MAPK3	725.83	2.52	0.11
7	PGAM2	724.77	3.06	0.08
8	HK1	723.38	3.39	0.07 .
9	MYC	723.15	2.23	0.14
10	LDHB	720.70	4.45	0.03 *

presume che il modello reale sia sparso, come spesso accade in ambito genomico, poiché consente di selezionare un sottoinsieme di variabili rilevanti portando a zero gli altri coefficienti.

In situazioni caratterizzate da dati ad alta dimensionalità, i modelli tradizionali di regressione possono risultare inadeguati per due motivi principali: il rischio di eccessivo adattamento (*overfitting*) e la presenza di collinearità tra le variabili. Inoltre, i metodi tradizionali diventano inutilizzabili quando il numero di covariate supera il numero di osservazioni. L'introduzione di un termine di penalizzazione nel processo di stima dei coefficienti contribuisce non solo a prevenire l'*overfitting*, ma migliora anche l'interpretabilità del modello. In particolare, la penalizzazione può ridurre a zero i coefficienti di alcune variabili, permettendo di concentrare l'attenzione sulle covariate più influenti. Metodi come lasso ed *elastic net* sono ideali per gestire queste situazioni, con l'*elastic net* che si rivela particolarmente utile quando esistono forti correlazioni tra le variabili, come spesso accade in ambito genomico, poiché combina le proprietà della regressione *ridge* e del lasso.

Nonostante i vantaggi, i modelli penalizzati presentano alcune limitazioni nell'ambito dell'inferenza statistica. A causa del termine di penalizzazione, i risultati ottenuti non seguono le distribuzioni asintotiche abituali, rendendo difficile la costruzione di intervalli di confidenza e l'esecuzione di test di ipotesi affidabili per i coefficienti. Di conseguenza,

Tabella 3.2: *Output del modello di Cox selezionato.*
Colonna 1: Variabile inclusa nel modello. **Colonna 2:** Coefficiente stimato dal modello di Cox. **Colonna 3:** Esponenziale del coefficiente stimato (*hazard ratio*). **Colonna 4:** Esponenziale dell'opposto del coefficiente stimato. **Colonna 5:** Errore standard del coefficiente stimato. **Colonna 6:** Valore del test z. **Colonna 7:** p-value associato al test z.

Variabile	**Coef**	**exp(Coef)**	**exp(-Coef)**	**se(Coef)**	z	**Pr**$(>\|z\|)$
Età	0.55	1.74	0.58	0.14	4.07	4.64e-05 ***
Genere	0.15	1.16	0.86	0.24	0.60	0.55
Stadio 2	0.25	1.29	0.78	0.51	0.50	0.62
Stadio 3	0.81	2.26	0.44	0.51	1.59	0.11
Stadio 4	1.96	7.08	0.14	0.52	3.74	0.00 ***
MET	-0.29	0.75	1.34	0.13	-2.23	0.03 *
PIK3CA	0.36	1.44	0.70	0.16	2.26	0.02 *
PIK3R2	0.29	1.34	0.75	0.14	2.08	0.04 *
PTEN	0.26	1.30	0.77	0.14	1.84	0.07 .
C12orf5	0.18	1.20	0.83	0.11	1.69	0.09 .
MAPK3	0.28	1.32	0.76	0.15	1.87	0.06 .
PGAM2	0.29	1.34	0.75	0.14	2.15	0.03 *
HK1	-0.30	0.74	1.34	0.14	-2.18	0.03 *
MYC	-0.31	0.73	1.36	0.14	-2.19	0.03 *
LDHB	0.30	1.36	0.74	0.15	1.98	0.05 *

le tecniche tradizionali di inferenza non sono direttamente applicabili e l'interpretazione statistica dei risultati richiede approcci alternativi, tra questi, si possono citare i metodi di *bootstrap*, gli approcci bayesiani e le tecniche di *post-selection inference*, che permettono di valutare la stabilità e la significatività delle variabili selezionate. Recenti sviluppi teorici hanno inoltre proposto metodi per la costruzione di intervalli di confidenza specifici per modelli penalizzati, tenendo conto del processo di selezione delle variabili (Lockhart et al., 2014).

Per questa tesi, si utilizzeranno due tra i metodi di penalizzazione più comuni, ovvero la penalizzazione lasso (*Least Absolute Shrinkage and Selection Operator*) (Tibshirani, 1997) e la penalizzazione *elastic net* (Simon, Friedman, Hastie et al., 2011). Il pacchetto `glmnet` di R (Friedman et al., 2010) rappresenta uno strumento computazionalmente efficiente per l'implementazione di modelli di regressione penalizzati, supportando sia il lasso che l'*elastic net*, consentendo di selezionare automaticamente le variabili rilevanti e di regolarizzare i coefficienti.

Nelle sezioni successive verranno descritti nel dettaglio sia la penalizzazione lasso sia la penalizzazione *elastic net*, illustrando come questi metodi possano essere applicati ai modelli di Cox per l'analisi di dati di sopravvivenza.

3.5.1 Lasso

La regressione penalizzata lasso (Tibshirani, 1997) applicata ai modelli di Cox rappresenta una metodologia efficace per la selezione delle variabili e la stima di modelli di sopravvivenza. Il lasso introduce una penalizzazione basata sulla norma ℓ_1 dei coefficienti di regressione, favorendo soluzioni sparse, ovvero soluzioni in cui molti coefficienti risultano esattamente pari a zero.

La stima del modello di Cox con penalizzazione lasso si ottiene massimizzando la seguente funzione di *log*-verosimiglianza parziale penalizzata:

$$\frac{2}{n} \cdot \ell(\beta|\mathcal{D}) - \lambda||\beta||_1,$$

dove $\frac{2}{n}$ è un fattore di scala introdotto per migliorare la stabilità numerica, normalizzando la funzione di *log*-verosimiglianza parziale rispetto alla dimensione del campione. Il *dataset* $\mathcal{D} = \{(T_i, \delta_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ comprende n osservazioni, con $\mathbf{X}_i$ che rappresenta le p caratteristiche del i-esimo paziente. Il parametro di regolazione $\lambda \geq 0$ controlla la forza della penalizzazione, mentre la norma ℓ_1 dei coefficienti è definita come $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. La *log*-verosimiglianza parziale è data da:

$$\ell(\beta|\mathcal{D}) = \log \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{X}_i\beta)}{\sum_{l \in \mathcal{R}_k} \exp(\mathbf{X}_l\beta)} \right)^{\delta_i},$$

dove $\mathcal{R}_k = \{l : Y_l(T_k) = 1\}$ rappresenta l'insieme di individui a rischio al tempo T_k.

Il parametro di regolazione λ è fondamentale nel determinare il grado di penalizzazione applicato ai coefficienti, valori elevati di λ tendono a ridurre a zero un numero maggiore di coefficienti, favorendo modelli più parsimoniosi. La selezione ottimale di λ viene solitamente effettuata tramite convalida incrociata, una tecnica fondamentale per la valutazione delle prestazioni di un modello statistico, che permette di stimare la capacità di generalizzazione del modello su dati non utilizzati durante la fase di stima. Con questo metodo, il *dataset* viene suddiviso in n sottoinsiemi (*folds*) e il modello viene stimato su $n-1$ *folds*, mentre il *fold* rimanente viene utilizzato per la valutazione del modello. Ripetendo questo processo per ciascun *fold*, si ottiene una misura media dell'errore, riducendo la variabilità delle stime delle prestazioni del modello. Un aspetto rilevante nella convalida incrociata è l'utilizzo di un vettore di assegnazione dei *folds*.

Questo accorgimento consente di ottenere risultati riproducibili e consistenti, garantendo che la suddivisione dei dati tra i vari *folds* rimanga identica sia nella fase di stima del parametro di regolazione, sia nella fase di valutazione delle prestazioni del modello.

In sintesi, la penalizzazione lasso per il modello di Cox facilita l'identificazione delle variabili omiche rilevanti, riducendo i coefficienti delle variabili non rilevanti a zero, migliorando così l'interpretabilità del modello senza sacrificare la capacità predittiva.

Nei contesti clinici, l'inclusione di un insieme ristretto di fattori di rischio clinici consolidati è fondamentale, in quanto tali covariate possono rappresentare potenziali fattori di confondimento e sono ampiamente riconosciute per la loro rilevanza. Queste covariate possono essere integrate in modelli di regressione penalizzata senza essere soggette a penalizzazione, preservando così il loro contributo prognostico, mentre si applica la selezione di altre variabili meno conosciute, come quelle omiche.

Nel modello di Cox penalizzato, la funzione di *log*-verosimiglianza penalizzata, quando si desidera garantire che le variabili clinico-demografiche rimangano nel modello nonostante la selezione applicata alle caratteristiche omiche, assume la seguente forma:

$$\frac{2}{n}\log\prod_{i=1}^{n}\left(\frac{\exp(\mathbf{X}_{0i}\beta_0+\mathbf{X}_i\beta)}{\sum_{l\in\mathcal{R}_k}\exp(\mathbf{X}_{0l}\beta_0+\mathbf{X}_l\beta)}\right)^{\delta_i}-\text{pen}(\beta),$$

dove β_0 rappresenta i coefficienti associati alle covariate obbligatorie $\mathbf{X}_{0i}$ per l'individuo i-esimo e pen(β) rappresenta il termine di penalizzazione applicato ai coefficienti delle caratteristiche omiche.

Per implementare la regressione penalizzata lasso nel contesto del modello di Cox, è stata utilizzata la funzione `cv.glmnet` del pacchetto `glmnet`, la quale consente di eseguire la convalida incrociata per identificare il valore ottimale del parametro di regolazione λ, scelto su una griglia predefinita di valori variabile tra 10^3 e 10^{-3}, basandosi sulla *log*-verosimiglianza penalizzata. In questa analisi, sono stati utilizzati 5 *folds*, un vettore di assegnazione dei *folds* specificato per garantire la coerenza delle suddivisioni del *dataset* durante le fasi di stima e valutazione del modello, mentre le variabili cliniche sono state esentate dalla penalizzazione.

La regressione penalizzata lasso ha condotto alla selezione di un modello che includeva esclusivamente le covariate clinico-demografiche, tutte le variabili associate ai geni sono state ridotte a zero. Pertanto, il modello finale selezionato è risultato essere quello di base.

Tuttavia, in contesti in cui le covariate sono fortemente correlate tra loro, il lasso potrebbe non essere sufficiente per identificare le variabili più rilevanti. In queste situazioni, l'approccio *elastic net* può risultate maggiormente adeguato, poiché combina i vantaggi del lasso e della penalizzazione *ridge*, permettendo una gestione più efficace delle variabili correlate.

3.5.2 *Elastic net*

Il modello di Cox con penalizzazione *elastic net* (Simon, Friedman, Hastie et al., 2011) combina due approcci di penalizzazione: la penalizzazione lasso, che applica una norma ℓ_1 per consentire la selezione delle variabili, e la penalizzazione *ridge*, che utilizza una norma ℓ_2 per gestire l'effetto del raggruppamento di caratteristiche omiche correlate. La penalizzazione *ridge* riduce l'entità dei coefficienti senza annullarli, attenuando così il problema della collinearità tra le variabili. Questa combinazione consente all'*elastic net* di sfruttare la capacità del lasso di selezionare le variabili rilevanti e, allo stesso tempo, l'abilità del *ridge* di gestire le covariate correlate, migliorando spesso le prestazioni del modello rispetto al solo lasso.

L'approccio *elastic net* si basa su due parametri di regolazione: λ e α. Il parametro di penalizzazione λ, come nel lasso, regola l'intensità complessiva della penalizzazione applicata ai coefficienti, mentre il parametro di miscelazione $\alpha \in [0, 1]$ bilancia i due metodi: quando $\alpha = 1$, il modello applica una penalizzazione lasso pura, mentre con $\alpha = 0$ si ha una penalizzazione *ridge* pura. Per valori intermedi di α, il modello offre un compromesso tra i due, risultando particolarmente utile per mantenere l'informazione tra variabili correlate e ridurre la complessità del modello. Infatti, per ogni $\alpha < 1$ e $\lambda > 0$, il problema risulta strettamente convesso, garantendo una soluzione unica, indipendentemente dalla presenza di correlazioni o di coefficienti duplicati.

Il termine di penalizzazione nel caso dell'*elastic net* è definito dalla seguente espressione:

$$\lambda \left\{ \alpha ||\beta||_1 + \frac{1}{2}(1-\alpha)||\beta||_2^2 \right\}, \quad \text{dove} \quad ||\beta||_2^2 = \sum_{j=1}^{p} \beta_j^2,$$

dove $||\beta||_1$ e $||\beta||_2$ rappresentano rispettivamente le norme ℓ_1 e ℓ_2 dei coefficienti.

In sintesi, l'*elastic net* facilita l'identificazione di variabili omiche rilevanti in contesti ad alta dimensionalità, preservando l'informazione tra variabili correlate e migliorando la robustezza del modello.

Anche per implementare la regressione penalizzata *elastic net* nel contesto di un modello di Cox è stata utilizzata la funzione `cv.glmnet` del pacchetto `glmnet`, la quale

permette, appunto, di eseguire la convalida incrociata per determinare il valore ottimale di λ. Come nel caso del lasso, sono stati utilizzati 5 *folds*, un vettore di assegnazione dei *fold* specificato, una griglia predefinita di valori per λ e le variabili cliniche sono state esentate dalla penalizzazione.

La selezione del parametro di miscelazione α è stata realizzata esplorando una sequenza di valori compresi tra 0.1 e 1, suddivisi in dieci intervalli. Anche in questo caso, il modello ottimale è stato selezionato massimizzando la *log*-verosimiglianza parziale.

L'applicazione della regressione penalizzata *elastic net* ha evidenziato che tutti i coefficienti associati alle variabili geniche sono stati ridotti a zero, suggerendo l'assenza di contributi significativi da parte delle caratteristiche omiche nel miglioramento della capacità predittiva del modello. Di conseguenza, anche in questo caso, è stato confermato come modello migliore il modello di riferimento che include esclusivamente le covariate cliniche. Questo risultato sottolinea come, nonostante la ricchezza informativa dei dati omici, questi non apportino un contributo predittivo aggiuntivo rispetto ai fattori clinici consolidati. La mancanza di significatività delle variabili geniche potrebbe anche essere attribuibile ad altri fattori, tra cui la potenziale ridondanza delle informazioni omiche rispetto a quelle cliniche o la necessità di un campione più ampio per rilevare effetti genici sottili.

3.5.3 Interazioni tra geni

Nel contesto dell'analisi genomica per la predizione della sopravvivenza nei pazienti affetti da cancro, è ragionevole supporre che l'espressione di singoli geni possa non essere sufficiente a catturare completamente le dinamiche complesse che influenzano la prognosi. Oltre all'effetto individuale di ciascun gene, può risultare rilevante considerare interazioni tra più geni, poiché la coesistenza di variazioni nell'espressione genica potrebbe avere un impatto maggiore rispetto a ciascun gene considerato singolarmente. In particolare, le interazioni bivariate possono rivelare sinergie o antagonismi tra l'espressione di due geni che influenzano la sopravvivenza in modo non additivo. È importante notare, tuttavia, che l'effetto delle interazioni geniche può essere complesso e, talvolta, non lineare, motivo per cui queste relazioni potrebbero non essere completamente rappresentate da sole interazioni lineari.

Per esplorare questa ipotesi, sono stati applicati modelli di regressione penalizzata lasso e *elastic net* includendo tutte le possibili interazioni bivariate tra i geni. L'applicazione di tali penalizzazioni consente di ridurre la complessità del modello in contesti di alta dimensionalità, eliminando le interazioni meno informative e mantenendo solo

quelle più rilevanti per la predizione dell'*outcome*. In questo modo, è possibile identificare le coppie di geni che, congiuntamente, hanno un impatto significativo sulla prognosi dei pazienti, migliorando la capacità del modello di catturare le relazioni più sottili e complesse presenti nei dati genomici.

Questo approccio ha generato un *dataset* composto da 313 osservazioni e 2351 variabili, ampliando notevolmente lo spazio delle variabili rispetto al set originale. Per la selezione dei parametri di penalizzazione λ e α sono state applicate le medesime metodologie descritte in precedenza.

Tuttavia, con nessuna delle due tecniche è stato possibile identificare geni o interazioni con coefficienti diversi da zero, suggerendo che le interazioni tra i geni non contribuiscono in modo significativo alla predizione della sopravvivenza in questo contesto. Tale risultato potrebbe essere attribuibile a diversi fattori, tra cui la ridondanza informativa tra le interazioni e le variabili cliniche già incluse nel modello, la dimensione limitata del campione, che potrebbe non essere sufficiente per rilevare effetti sottili delle interazioni geniche, o la necessità di approcci alternativi in grado di rappresentare meglio interazioni non lineari presenti nei dati genomici.

3.6 Considerazioni e confronti

L'analisi condotta ha evidenziato che le caratteristiche omiche non hanno apportato un contributo significativo ai modelli finali. In particolare, i modelli di Cox con penalizzazione lasso ed *elastic net* non hanno selezionato geni che migliorassero significativamente la capacità predittiva rispetto al modello basato esclusivamente sulle covariate cliniche. Anche il modello sviluppato con selezione *forward* guidata dall'AIC, pur introducendo alcune variabili omiche, non ha evidenziato miglioramenti rilevanti nelle prestazioni sull'insieme di verifica.

La Tabella 3.3 confronta i modelli sviluppati, valutando le capacità di discriminazione e calibrazione nella predizione della sopravvivenza. Le misure riportate, come l'indice C_H di *Harrell*, l'indice C_U di *Uno*, l'AUC a 10 anni e l'indice di *Brier*, evidenziano che l'aggiunta di variabili omiche non ha portato a un miglioramento rispetto al modello basato esclusivamente sulle variabili cliniche. Anzi, gli indici di concordanza C mostrano un decremento delle prestazioni nei modelli che includono le covariate omiche e il *log-rank* test indica che questi modelli non sono in grado di discriminare efficacemente tra pazienti a basso e alto rischio, né tra pazienti a basso, medio e alto rischio.

TABELLA 3.3: *Confronto tra modelli analizzati nell'insieme di verifica.* **Colonna 1:** *Log-rank test* per i pazienti classificati ad alto e basso rischio. **Colonna 2:** *Log-rank test* per i pazienti classificati ad alto, medio e basso rischio. **Colonna 3:** AUC a 10 anni. **Colonna 4:** C_H di Harrell. **Colonna 5:** C_U di Uno. **Colonna 6:** Indice di *Brier*.

Modello	***Log-rank* - 2 gruppi**	***Log-rank* - 3 gruppi**	**AUC**	C_H	C_U	***Brier***
Variabili cliniche	$p = 0.0003$	$p = 0.0001$	0.67	0.75	0.66	0.20
Selezione forward AIC	$p = 0.14$	$p = 0.16$	0.75	0.63	0.57	0.20

Questi risultati suggeriscono che, nonostante la potenziale ricchezza informativa dei dati omici, fattori clinici consolidati come l'età alla diagnosi, il genere e lo stadio del tumore restano centrali nella predizione della sopravvivenza nei pazienti affetti da COAD. La mancanza di un contributo sostanziale da parte delle variabili omiche potrebbe riflettere una ridondanza delle informazioni rispetto ai dati clinici o l'esigenza di un campione più ampio per rilevare effetti più sottili.

Un aspetto rilevante nell'analisi di dati omici o ad alta dimensionalità è la stabilità della selezione delle caratteristiche. L'ottimizzazione dei parametri di penalizzazione tramite *cross-validation* (CV) introduce infatti una fonte di variabilità, che può influenzare la selezione finale delle variabili, come osservato da Kalousis et al. (2007). Diversi *folds* della CV potrebbero portare a selezioni variabili, ma poiché i modelli penalizzati non hanno selezionato alcuna variabile omica, questo aspetto non è stato esplorato nella presente analisi.

In conclusione, l'analisi ha evidenziato la predominanza delle caratteristiche cliniche nella predizione della sopravvivenza nei pazienti con COAD, mentre il contributo delle variabili omiche non è emerso in modo significativo. È possibile, tuttavia, che il potenziale delle variabili omiche non sia stato completamente esplorato a causa del numero limitato di geni considerati.

Nel prossimo capitolo, verranno analizzati ulteriori *pathway*, ampliando l'insieme di geni studiati. Questo approccio permetterà di sperimentare nuovi metodi di modellazione, tra cui la regressione *group lasso*, che sfrutta la struttura a gruppi dei dati, consentendo una selezione simultanea dei geni appartenenti a specifici *pathway* per migliorare

ulteriormente le capacità predittive del modello.

Capitolo 4

Analisi di sopravvivenza considerando più *pathway*

L'analisi di sopravvivenza applicata a dati genomici richiede metodi statistici capaci di gestire un numero elevato di covariate, spesso superiore al numero di pazienti. In questi contesti, il modello a rischi proporzionali di Cox, descritto nel Paragrafo 1.2, rappresenta uno strumento fondamentale per identificare le covariate che influenzano la prognosi del cancro. Quando il numero di covariate supera quello dei pazienti, si rende necessario l'uso di tecniche di penalizzazione, come le penalizzazioni lasso ed *elastic net* discusse nel Paragrafo 3.5. Queste penalizzazioni, mirate a produrre soluzioni sparse in cui alcuni coefficienti vengono ridotti a zero, migliorano l'efficienza predittiva dei modelli, ma presentano limiti di stabilità nella selezione delle variabili.

Inoltre, nonostante lasso ed *elastic net* siano strumenti utili per selezionare singoli geni rilevanti, il loro impiego risulta limitato nel cogliere appieno la complessità e le strutture intrinseche dei dati genomici. Infatti, i geni non agiscono isolatamente, ma spesso interagiscono all'interno di vie molecolari o *pathway* biologici, regolando insieme processi cellulari complessi. Le vie molecolari rappresentano insiemi di geni che agiscono in modo coordinato per regolare processi cellulari specifici e il loro coinvolgimento nei meccanismi di sviluppo e progressione tumorale può fornire informazioni preziose per la predizione della sopravvivenza. Considerare i geni singolarmente, senza tenere conto della struttura dei *pathway*, potrebbe limitare la capacità del modello di catturare queste interazioni e portare a una variabilità elevata nella selezione dei geni tra diversi campioni.

Nei capitoli precedenti è stato analizzato un singolo *pathway* molecolare, che però non ha mostrato un contributo significativo alla capacità predittiva del modello. Inoltre, è stata effettuata la medesima analisi descritta nel Capitolo 3, applicata però a tutti i geni a disposizione senza limitarsi a un singolo *pathway*. Tuttavia, anche in questo

caso, nessun gene è risultato significativo per la predizione della sopravvivenza. Questa evidenza, insieme alle considerazioni fatte in precedenza, ha spinto a riconsiderare l'approccio, estendendo l'analisi a più *pathway*, con l'obiettivo di cogliere sinergie tra *pathway* e interazioni più ampie tra gruppi di geni, che potrebbero avere un impatto congiunto più rilevante rispetto a quello di singoli geni o di un singolo *pathway*.

A questo proposito, il modello *group lasso* (Yuan e Lin, 2006) permette di incorporare informazioni strutturali attraverso la definizione di gruppi di covariate rappresentati dai *pathway*. Questo approccio si distingue dal lasso standard in quanto seleziona interi gruppi di covariate invece di singole variabili, questa caratteristica contribuisce a ridurre la variabilità nella selezione delle variabili; infatti, un problema comune quando si utilizzano tecniche come il lasso o l'*elastic net* è che anche piccoli cambiamenti nei dati, ad esempio dovuti alla suddivisione dei *fold* nella convalida incrociata, possono portare alla selezione di geni diversi.

In questo capitolo, si approfondirà, quindi, l'implementazione del *group lasso* applicato a gruppi di geni organizzati per *pathway* molecolari.

4.1 *Group lasso*

Il *group lasso* è particolarmente adatto all'analisi genomica per la sua capacità di selezionare interi gruppi di variabili, come i *pathway* molecolari, piuttosto che singoli geni. Questo approccio offre un importante vantaggio in contesti ad alta dimensionalità, caratterizzati da forte correlazione tra geni, poiché raggruppando le variabili in *pathway* predefiniti stabilizza la selezione, riducendo la variabilità e la sensibilità a variazioni minime nel *dataset*. Tale struttura migliora l'interpretabilità del modello, consentendo di concentrare l'analisi su processi biologici più ampi e biologicamente rilevanti, rendendo così il modello più robusto e coerente nella predizione della prognosi.

Nel contesto dei processi biologici, alcuni *pathway* possono condividere geni comuni, dunque gestire la sovrapposizione tra *pathway* è fondamentale per rappresentare accuratamente le reti molecolari. Per superare questa limitazione, seguendo l'approccio di Malenová et al. (2021), si applica una tecnica chiamata *overlapped group lasso* (Jacob et al., 2009; Obozinski et al., 2011), che prevede la duplicazione dei geni presenti in più *pathway*. Questo consente al modello di selezionare il gene solo nei *pathway* in cui risulta effettivamente rilevante per la predizione, permettendo di fornire una rappresentazione più accurata delle relazioni molecolari sottostanti.

Formalmente, il problema di ottimizzazione del *group lasso* nel contesto del modello di Cox (Kim et al., 2012) prevede la minimizzazione della somma tra la *log*-verosimiglianza negativa del modello di Cox e il termine di penalizzazione $R_\lambda(\beta)$. Denotando con g un gruppo di covariate, la penalizzazione impiegata è espressa come

$$R_\lambda(\beta) = \lambda \sum_g \|\beta_g\|_2,$$

dove $\|\beta_g\|_2$ è la norma ℓ_2 dei coefficienti associati al gruppo g e λ è il parametro che regola l'intensità della penalizzazione. La funzione obiettivo da minimizzare è quindi data da $-l(\beta) + R_\lambda(\beta)$, dove $-l(\beta)$ rappresenta la *log*-verosimiglianza parziale associata al modello di Cox.

Aggiungendo $R_\lambda(\beta)$ alla funzione di perdita $\ell(\beta)$, alcuni gruppi di coefficienti vengono ridotti a zero, producendo un modello sparso in cui solo alcuni gruppi di covariate hanno coefficienti diversi da zero. Questo approccio selettivo consente di identificare i *pathway* associati al rischio di evento, favorendo una maggiore interpretabilità biologica rispetto alla selezione di singoli geni.

I *pathway* molecolari utilizzati sono stati quelli definiti da Malenová et al. (2021), che hanno combinato vari *database* (SPEED, PROGENy e set curati dalla *Duke University* e dal *Curie Institute*) per identificare geni attivi o inibiti in risposta all'attivazione di *pathway* di segnalazione e con l'integrazione di queste fonti hanno generato 69 set unici di geni bersaglio. Malenová et al. (2021) hanno scelto di includere esclusivamente i geni che rappresentano bersagli dei *pathway* di segnalazione, escludendo altri set di geni disponibili come quelli di *Reactome* (Fabregat et al., 2018) o KEGG (Kanehisa e Goto, 2000) o processi biologici dal *Gene Ontology* (Ashburner et al., 2000). Questa scelta si basa sull'ipotesi che solo i geni bersaglio, essendo direttamente influenzati dai segnali molecolari, mostrino cambiamenti coordinati nell'espressione in risposta all'attivazione del *pathway*. In altre parole, quando un *pathway* viene attivato, i geni bersaglio a valle tendono a esprimersi in modo sincronizzato o modulato rispetto al segnale, riflettendo un effetto omogeneo e specifico. Questo comportamento rende tali geni particolarmente rilevanti per studiare i processi biologici connessi alla prognosi del cancro, poiché offrono una visione più diretta di come le reti di segnalazione possano influenzare l'esito clinico.

Per implementare il modello di Cox *group lasso* si è utilizzato il pacchetto R `grpreg` (Breheny e Huang, 2015; Breheny, Zeng et al., 2021), aggiungendo ai 69 *pathway* definiti da Malenová et al. (2021) un gruppo per le covariate clinico-demografiche. In particolare, la regressione è stata eseguita tramite convalida incrociata con `cv.grpsurv()`,

utilizzando 5 *folds*, un vettore di assegnazione dei *fold* specificato e una griglia di valori per λ.

Il parametro λ è stato selezionato minimizzando l'errore di convalida incrociata, questo approccio consente di ottenere un modello sufficientemente flessibile da adattarsi ai dati senza risultare eccessivamente complesso, mantenendo una buona capacità di generalizzazione su nuovi dati. Il valore ottimale di λ, 0.0248, il quale rappresenta il miglior compromesso tra adattamento e parsimonia, come illustrato nella Figura 4.1, seleziona un solo gruppo, corrispondente alle covariate clinico-demografiche. La selezione di un singolo gruppo di covariate clinico-demografiche implica che, in questo contesto, tali covariate forniscono una capacità predittiva sufficiente per il modello, senza la necessità di aggiungere complessità attraverso i *pathway* molecolari. Questo risultato riflette la robustezza dei dati clinici rispetto a quelli omici per la predizione della sopravvivenza in questo studio.

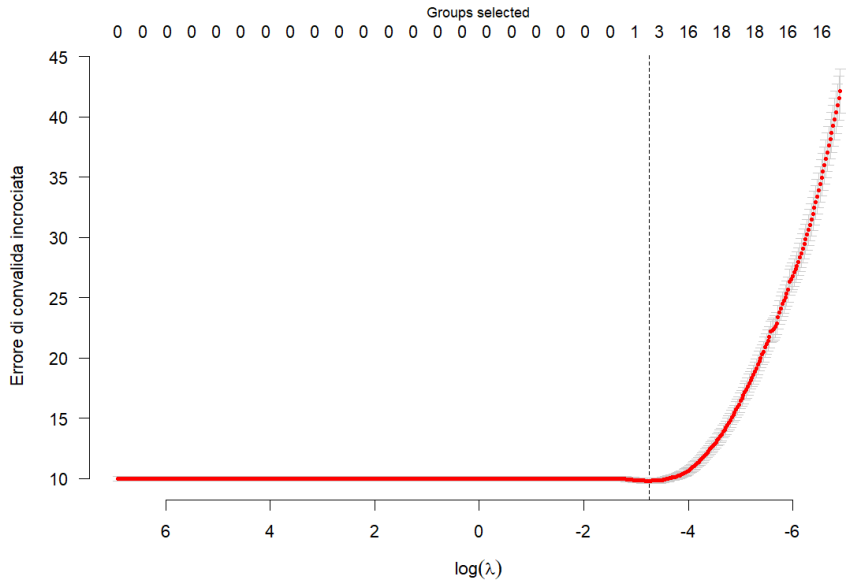

FIGURA 4.1: *Grafico dell'errore di convalida incrociata rispetto ai valori di* $log(\lambda)$ *per il modello group lasso.*
Il numero di gruppi selezionati per ciascun valore di λ è riportato nella parte superiore del grafico. Il punto in cui l'errore di convalida incrociata è minimo è indicato dalla linea tratteggiata verticale, corrispondente al valore ottimale di λ.

4.2 Considerazioni

I risultati ottenuti mostrano che, nonostante l'implementazione del modello *group lasso* permetta di esplorare l'influenza di diversi *pathway* molecolari sulla prognosi, il modello

finale ha selezionato solamente il gruppo delle variabili clinico-demografiche. Questo suggerisce che, nel contesto in esame, i *pathway* molecolari non hanno fornito un contributo significativo alla predizione della sopravvivenza. La selezione di sole covariate cliniche può indicare che le informazioni offerte dai *pathway* analizzati siano ridondanti rispetto ai dati clinici o che non aggiungano dettagli sufficientemente rilevanti per migliorare la capacità predittiva complessiva del modello.

Dal punto di vista dell'interpretabilità, il *group lasso* offre un vantaggio importante rispetto alla selezione di singoli geni, poiché concentra l'attenzione su interi *pathway* anziché su variabili individuali che potrebbero essere soggette a una variabilità elevata tra differenti campioni. Questo approccio riduce la sensibilità del modello a piccoli cambiamenti nei dati e garantisce una selezione delle covariate più stabile e biologicamente rilevante. Tuttavia, in questo contesto in cui il modello seleziona solo variabili cliniche, l'interpretazione specifica delle vie molecolari coinvolte diventa meno rilevante e il contributo dei *pathway* alla comprensione dei processi tumorali resta limitato.

Inoltre, se il modello avesse selezionato uno o più *pathway* molecolari, si sarebbe potuto proseguire con l'applicazione dello *sparse group lasso* (Simon, Friedman, Hastie e Tibshirani, 2013), una tecnica che combina la selezione dei gruppi con la selezione dei geni all'interno dei gruppi stessi. Questo metodo avrebbe consentito di identificare, all'interno dei *pathway* selezionati, i geni più rilevanti, aumentando ulteriormente la precisione e l'interpretabilità del modello.

Nel complesso, i risultati suggeriscono che, pur essendo biologicamente ricchi di informazioni, i *pathway* molecolari, come definiti in questo studio, non hanno fornito un valore aggiunto per la predizione della sopravvivenza nei pazienti con COAD. Sebbene questo risultato sia in parte deludente rispetto all'obiettivo di identificare *pathway* rilevanti, esso rimane coerente con quanto osservato nei capitoli precedenti.

A differenza di altri studi utilizzati come riferimento per le metodologie applicate in questa tesi, come il lavoro di Malenová et al. (2021) e Zhao et al. (2024), nel presente studio le variabili cliniche hanno dimostrato una rilevanza dominante. Tuttavia, nel lavoro di Malenová et al. (2021), ad esempio, le variabili cliniche non sono state considerate, mentre Zhao et al. (2024) non hanno applicato la standardizzazione dell'espressione genica, il che potrebbe aver favorito l'inclusione di geni nel modello unicamente per via di un'elevata variabilità. Questi approcci hanno permesso agli autori di individuare geni associati agli esiti di sopravvivenza, suggerendo un potenziale valore informativo dei dati omici in contesti specifici.

D'altra parte, l'utilizzo di sole informazioni clinico-demografiche, come l'età alla diagnosi, il genere e lo stadio del tumore, fornisce una base più immediatamente accessibile e stabile per la caratterizzazione dei pazienti. Le variabili cliniche, infatti, sono generalmente facili da raccogliere e interpretare e sono ben consolidate nei protocolli di valutazione prognostica in ambito oncologico. Questa disponibilità e affidabilità delle informazioni cliniche giustifica il loro impatto dominante, suggerendo che possano da sole rappresentare efficacemente il rischio di sopravvivenza nei pazienti con COAD, senza necessariamente integrare dati omici aggiuntivi.

Rimane comunque aperta la possibilità che l'integrazione di set di dati più ampio o specifico, o l'adozione di approcci più sofisticati per gestire la sovrapposizione tra i *pathway*, possano rivelare associazioni rilevanti che non sono emerse nell'attuale contesto analitico.

Capitolo 5

Evoluzioni future e conclusioni

La presente tesi ha esplorato la possibilità di identificare biomarcatori prognostici per la sopravvivenza nei pazienti con adenocarcinoma del colon (COAD) attraverso l'integrazione di dati clinici e omici. I risultati hanno evidenziato che, in questo contesto, le variabili cliniche offrono una capacità predittiva sufficiente da rendere superfluo l'utilizzo delle informazioni omiche. Nonostante l'impiego di metodologie avanzate, come le penalizzazioni lasso, *elastic net* e *group lasso* con l'inclusione di *pathway* multipli, le informazioni omiche non hanno migliorato significativamente la capacità predittiva rispetto all'utilizzo delle sole variabili cliniche. Tali risultati indicano che, per sfruttare appieno il potenziale delle informazioni omiche, potrebbero essere necessari campioni di dimensioni maggiori o strategie analitiche più sofisticate.

Tale risultato contrasta con alcuni studi precedenti menzionati in questa tesi, i quali attribuiscono ai dati omici un valore aggiunto rilevante nella stratificazione del rischio per i pazienti oncologici. Tuttavia, un'analisi approfondita di questi studi ha rivelato alcune limitazioni metodologiche che potrebbero spiegare tale discrepanza. In particolare, alcuni non affrontano in modo sistematico i potenziali fattori di confondimento, che possono influenzare significativamente i risultati. Mentre, in altre occasioni, i dati non vengono standardizzati prima dell'applicazione delle regressioni penalizzate, una pratica fondamentale per garantire la coerenza dei risultati. Questi aspetti metodologici possono quindi limitare l'interpretabilità e la generalizzabile delle conclusioni raggiunte, suggerendo che un approccio più rigoroso potrebbe ridimensionare il contributo delle informazioni omiche rispetto a quanto riportato in letteratura.

5.1 Evoluzioni future

Come menzionato nell'introduzione, nonostante i progressi compiuti nei campi dell'oncologia e della genomica, permangono sfide significative per una loro integrazione efficace. I *software* attualmente disponibili, infatti, non consentono una gestione integrata e agevole dei dati genomici e clinici, ostacolando il loro impiego diretto nei modelli di sopravvivenza. Una possibile area di miglioramento riguarda lo sviluppo di strumenti dedicati all'interno di *Bioconductor*, consentendo così di gestire questi dati in un formato unificato e standardizzato. Ciò contribuirebbe a ridurre i rischi di errori di sincronizzazione tra i diversi *dataset* e migliorerebbe la coerenza e la riproducibilità delle analisi, facilitando l'adozione di modelli predittivi più complessi e integrati.

Inoltre, come illustrato nella Figura 5.1, l'analisi del *Relative Log Expression* (RLE) mette in evidenza la necessità di ulteriori ottimizzazioni per garantire una normalizzazione adeguata dei dati. Una normalizzazione più accurata potrebbe migliorare la qualità dei dati omici, consentendo una valutazione più precisa del loro potenziale contributo. In particolare, si ipotizza che la gestione degli effetti di *batch*, i quali potrebbero introdurre variabilità indesiderata, possa contribuire a migliorare l'affidabilità dei risultati.

Sarebbe altresì importante estendere le analisi condotte in questa tesi a tutte le piattaforme e ai diversi tipi di tumore presenti nel TCGA, per ottenere un quadro più esaustivo delle implicazioni genomiche nella sopravvivenza. Questo permetterebbe di approfondire le differenze genetiche tra le varie forme tumorali e di analizzare come i profili genetici influiscano sui diversi livelli di rischio.

Sarebbe inoltre interessante esplorare altre misure di *outcome*, come il tempo libero da progressione, per verificare se forniscano indicazioni differenti rispetto alla sopravvivenza complessiva. Sebbene la durata del *follow-up* sia ritenuta adeguata da Liu et al., 2018, potrebbe tuttavia risultare troppo breve, influenzando i risultati ottenuti.

Infine, il programma TCGA ha tradizionalmente posto maggiore enfasi sull'analisi genomica rispetto al *follow-up* clinico. Tuttavia, i risultati di questa tesi suggeriscono che una più stretta integrazione dei dati genomici con le informazioni cliniche disponibili potrebbe migliorare la capacità predittiva dei modelli. Ad esempio, la raccolta di dati dettagliati sui trattamenti ricevuti dai pazienti potrebbe contribuire a ridurre l'impatto di potenziali fattori di confondimento associati alle terapie. La possibilità di identificare i pazienti che risponderanno positivamente a trattamenti specifici – come la chemioterapia – rappresenterebbe un significativo passo avanti, riducendo l'esposizione a terapie invasive per coloro che potrebbero non trarne beneficio.

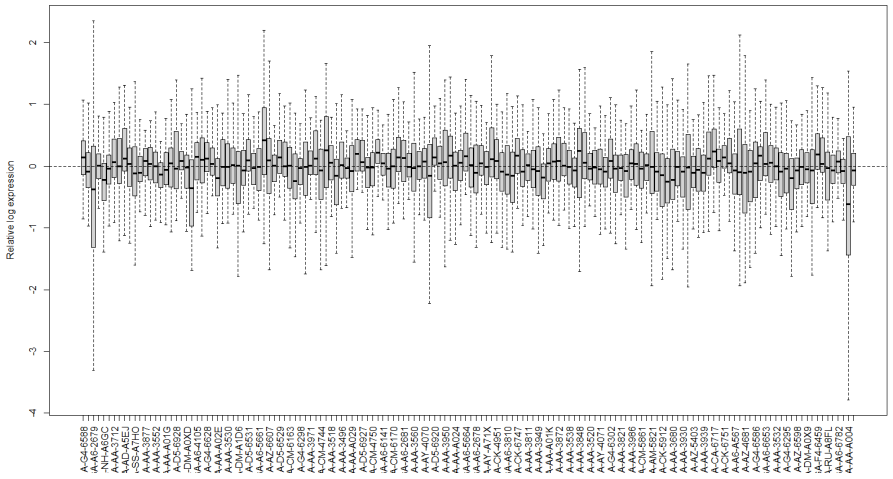

FIGURA 5.1: *RLE plot delle espressioni geniche normalizzate.*
Boxplot dei valori di *Relative Log Expression* (RLE) per un sotto-campione dei dati composto da 150 unità estratte casualmente. Ogni box rappresenta un campione, mentre l'asse verticale rappresenta i valori calcolati come:

$$z_{ij} = \log\left(\frac{y_{ij}}{\tilde{y}_{.j}}\right) = \log(y_{ij}) - \log(\tilde{y}_{.j}),$$

dove y_{ij} è il valore di espressione del gene j nel campione i e $\tilde{y}_{.j}$ è la mediana dell'espressione del gene j tra tutti i campioni. Un buon livello di normalizzazione si riflette in valori di RLE vicini allo zero. Questo tipo di grafico permette, quindi, di valutare la qualità della normalizzazione delle espressioni geniche tra i campioni.

Appendice

```
> sessionInfo()

R version 4.3.3 (2024-02-29 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 11 x64 (build 22631)

Matrix products: default

locale:
[1] LC_COLLATE=Italian_Italy.utf8  LC_CTYPE=Italian_Italy.utf8
[3] LC_MONETARY=Italian_Italy.utf8 LC_NUMERIC=C
[5] LC_TIME=Italian_Italy.utf8

time zone: Europe/Rome
tzcode source: internal

attached base packages:
[1] parallel stats4 stats graphics grDevices utils datasets
[8] methods base

other attached packages:
 [1] rms_6.8-1                Hmisc_5.1-3
 [3] survAUC_1.2-0            doParallel_1.0.17
 [5] iterators_1.0.14         foreach_1.5.2
 [7] risksetROC_1.0.4.1       MASS_7.3-60.0.1
 [9] glmnet_4.1-8             Matrix_1.6-5
[11] gridExtra_2.3            locfdr_1.1-8
```

```
[13] officer_0.6.6                org.Hs.eg.db_3.18.0
[15] GSEABase_1.64.0              annotate_1.80.0
[17] XML_3.99-0.17                AnnotationDbi_1.64.1
[19] curl_5.2.1                   EnrichmentBrowser_2.32.0
[21] graph_1.80.0                 survminer_0.4.9
[23] ggpubr_0.6.0                 survival_3.7-0
[25] lubridate_1.9.3              forcats_1.0.0
[27] stringr_1.5.1                purrr_1.0.2
[29] readr_2.1.5                  tidyr_1.3.1
[31] tibble_3.2.1                 ggplot2_3.5.1
[33] tidyverse_2.0.0              dplyr_1.1.4
[35] readxl_1.4.3                 SummarizedExperiment_1.32.0
[37] Biobase_2.62.0               GenomicRanges_1.54.1
[39] GenomeInfoDb_1.38.8          IRanges_2.36.0
[41] S4Vectors_0.40.2             BiocGenerics_0.48.1
[43] MatrixGenerics_1.14.0        matrixStats_1.3.0
[45] data.table_1.16.0

loaded via a namespace (and not attached):
  [1] splines_4.3.3               polspline_1.1.25
  [3] filelock_1.0.3              bitops_1.0-8
  [5] cellranger_1.1.0            rpart_4.1.23
  [7] lifecycle_1.0.4             rstatix_0.7.2
  [9] globals_0.16.3              lattice_0.22-5
 [11] regplot_1.1                 backports_1.5.0
 [13] magrittr_2.0.3              rmarkdown_2.28
 [15] plotrix_3.8-4               sm_2.2-6.0
 [17] zip_2.3.1                   askpass_1.2.1
 [19] plotmo_3.6.4                DBI_1.2.3
 [21] minqa_1.2.7                 multcomp_1.4-26
 [23] abind_1.4-8                 zlibbioc_1.48.2
 [25] RCurl_1.98-1.16             nnet_7.3-19
 [27] TH.data_1.1-2               sandwich_3.1-1
 [29] lava_1.8.0                  GenomeInfoDbData_1.2.11
 [31] KMsurv_0.1-5                listenv_0.9.1
 [33] MatrixModels_0.5-3          parallelly_1.38.0
```

[35] commonmark_1.9.1 codetools_0.2−19
[37] DelayedArray_0.28.0 ggtext_0.1.2
[39] xml2_1.3.6 tidyselect_1.2.1
[41] shape_1.4.6.1 farver_2.1.2
[43] lme4_1.1−35.5 beanplot_1.3.1
[45] BiocFileCache_2.10.2 base64enc_0.1−3
[47] vioplot_0.5.0 Formula_1.2−5
[49] systemfonts_1.1.0 tools_4.3.3
[51] ragg_1.3.3 cmprsk_2.2−12
[53] Rcpp_1.0.12 glue_1.7.0
[55] prodlim_2024.06.25 SparseArray_1.2.4
[57] mgcv_1.9−1 xfun_0.47
[59] withr_3.0.1 numDeriv_2016.8−1.1
[61] BiocManager_1.30.25 fastmap_1.2.0
[63] boot_1.3−29 fansi_1.0.6
[65] openssl_2.2.2 SparseM_1.84−2
[67] digest_0.6.36 timechange_0.3.0
[69] R6_2.5.1 textshaping_0.4.0
[71] colorspace_2.1−1 riskRegression_2023.12.21
[73] markdown_1.13 RSQLite_2.3.7
[75] utf8_1.2.4 generics_0.1.3
[77] httr_1.4.7 htmlwidgets_1.6.4
[79] S4Arrays_1.2.1 pkgconfig_2.0.3
[81] gtable_0.3.5 blob_1.2.4
[83] XVector_0.42.0 survMisc_0.5.6
[85] htmltools_0.5.8.1 carData_3.0−5
[87] scales_1.3.0 png_0.1−8
[89] knitr_1.48 km.ci_0.5−6
[91] rstudioapi_0.17.0 uuid_1.2−1
[93] tzdb_0.4.0 checkmate_2.3.2
[95] nlme_3.1−164 nloptr_2.1.1
[97] cachem_1.1.0 zoo_1.8−12
[99] foreign_0.8−86 pillar_1.9.0
[101] **grid**_4.3.3 vctrs_0.6.5
[103] car_3.1−3 dbplyr_2.5.0
[105] xtable_1.8−4 cluster_2.1.6

```
[107] htmlTable_2.4.3          Rgraphviz_2.46.0
[109] KEGGgraph_1.62.0         evaluate_1.0.1
[111] mvtnorm_1.2-5            cli_3.6.3
[113] compiler_4.3.3           rlang_1.1.4
[115] crayon_1.5.3             future.apply_1.11.2
[117] ggsignif_0.6.4           labeling_0.4.3
[119] timereg_2.0.5            grpreg_3.5.0
[121] stringi_1.8.4            Biostrings_2.70.3
[123] munsell_0.5.1            quantreg_5.98
[125] mets_1.3.4               hms_1.1.3
[127] bit64_4.0.5              future_1.34.0
[129] KEGGREST_1.42.0          gridtext_0.1.5
[131] memoise_2.0.1            broom_1.0.7
[133] bit_4.0.5
```

Listing A.1: *Session Info dell'ambiente di lavoro R*

TABELLA A.1: *z-value ottenuti con il modello di Cox univariato e con il modello di Cox che tiene conto dei fattori confondenti età, stadio del tumore e sesso.*
Colonna 1: *z*-value, per ciascun gene, ottenuti con il modello di Cox univariato. **Colonna 2:** *z*-value, per ciascun gene, ottenuti con il modello di Cox univariato da Smith e Sheltzer, 2022. **Colonna 3:** *z*-value, per ciascun gene, ottenuti con il modello di Cox con confondenti. **Colonna 4:** *z*-value, per ciascun gene, ottenuti con il modello di Cox con confondenti da Smith e Sheltzer, 2022.

Gene	*z*-value uni	*z*-value uni orig	*z*-value conf	*z*-value conf orig
AKT1	0.35	0.35	0.85	0.81
AKT2	0.47	0.47	0.13	0.06
AKT3	1.30	1.17	0.53	0.59
C12orf5	0.88	0.88	1.27	1.05
EGFR	2.21	2.21	1.08	0.97
ERBB2	0.75	0.75	0.26	0.27
FGFR1	0.49	0.49	-0.62	-0.25
FGFR2	-0.22	-0.22	-1.26	-0.82
FGFR3	-0.53	-0.54	-0.07	-0.52
FLT3	-1.14	-1.29	-0.26	-0.71
G6PD	0.59	0.59	0.62	0.66
GCK	-0.66	-0.59	-0.27	-0.28
GLS	0.33	0.33	-0.49	-0.58
GLS2	-1.00	-1.01	-0.27	-0.56
HIF1A	-0.39	-0.39	0.17	-0.03
HK1	-1.04	-1.04	-1.18	-1.03
HK2	0.08	0.09	0.66	0.48
HK3	0.66	0.67	0.91	1.01
HKDC1	-0.35	-0.31	-0.76	-0.77
HRAS	0.78	0.78	0.28	0.65
IDH1	-1.36	-1.36	-0.57	-0.75
IDH2	-1.42	-1.42	0.00	-0.21
KIT	0.40	0.40	0.38	0.24
KRAS	-0.30	-0.30	-0.55	-0.81
LDHA	-0.36	-0.36	-0.35	-0.63
LDHAL6A	1.61	2.56	1.01	1.81
LDHAL6B	0.05	-0.07	0.07	-0.12

Gene	z-value uni	z-value uni orig	z-value conf	z-value conf orig
LDHB	-0.35	-0.35	0.56	0.52
LDHC	0.41	0.83	0.14	0.39
MAP2K1	0.20	0.20	0.96	0.59
MAP2K2	-0.72	-0.72	-0.43	-0.62
MAPK1	-1.43	-1.43	-0.39	-0.94
MAPK3	0.31	0.31	-0.20	-0.21
MET	0.15	0.15	-0.36	-0.63
MTOR	0.29	0.29	0.60	0.46
MYC	-1.09	-1.09	-0.72	-0.83
NRAS	-1.42	-1.42	-0.94	-1.32
NTRK1	1.21	1.23	1.42	1.44
NTRK3	1.13	1.05	0.44	0.72
PDGFRA	-0.74	-0.73	-1.00	-0.84
PDGFRB	1.55	1.55	0.89	1.28
PDHA1	-1.25	-1.25	-0.39	-0.27
PDHA2	-0.66	NA	0.05	NA
PDHB	-2.55	-2.55	-2.01	-1.85
PDK1	0.71	0.71	1.22	0.73
PFKL	1.21	1.21	1.71	1.81
PFKM	0.00	0.00	1.08	0.85
PFKP	1.51	1.51	1.24	1.14
PGAM1	0.29	0.29	1.37	0.98
PGAM2	3.12	3.13	1.84	2.07
PGAM4	-0.29	-0.28	1.05	0.57
PIK3CA	1.58	1.58	0.57	0.67
PIK3CB	-0.22	-0.19	-0.95	-1.02
PIK3CD	1.31	1.31	0.84	1.02
PIK3R1	0.24	0.24	0.05	0.36
PIK3R2	0.35	0.35	1.53	1.41
PIK3R3	-0.39	-0.39	-0.05	0.19
PKM2	1.65	1.65	1.95	1.72
PTEN	1.03	1.04	0.87	0.88
RAF1	0.22	0.22	0.08	0.26
RET	0.24	0.03	-0.57	-0.65

Gene	z-value uni	z-value uni orig	z-value conf	z-value conf orig
SCO2	-0.42	-0.42	0.94	0.82
SIRT3	0.00	0.00	-0.50	-0.07
SIRT6	0.39	0.39	0.63	0.53
SLC16A3	0.35	0.35	-0.49	-0.74
SLC1A5	0.06	0.06	-1.36	-1.48
SLC2A1	0.60	0.60	-1.36	-1.66
SLC2A2	-1.22	-1.14	-0.81	-1.30
SLC7A5	0.57	0.57	0.20	-0.18
TP53	-1.69	-1.69	-0.82	-0.72

TABELLA A.2: *p-value ottenuti con il modello di Cox univariato.*
Colonna 1: *p*-value, per ciascun gene, ottenuti con il modello di Cox univariato. **Colonna 2:** *p*-value corretti per test multipli con il metodo FDR di Benjamini e Hochberg (Benjamini e Hochberg, 1995). **Colonna 3:** *p*-value corretti per (i) deviazioni dalla distribuzione nulla teorica sottostante e (ii) per test multipli attraverso la procedura lfdr di Efron (Efron, Turnbull et al., 2011).

Gene	***p*-value**	**BH-FDR**	**lfdr**
AKT1	0.73	0.92	1.00
AKT2	0.64	0.92	1.00
AKT3	0.19	0.84	1.00
C12orf5	0.38	0.92	1.00
EGFR	0.03	0.63	1.00
ERBB2	0.45	0.92	1.00
FGFR1	0.62	0.92	1.00
FGFR2	0.83	0.92	1.00
FGFR3	0.59	0.92	1.00
FLT3	0.26	0.87	1.00
G6PD	0.55	0.92	1.00
GCK	0.51	0.92	1.00
GLS	0.74	0.92	1.00
GLS2	0.32	0.88	1.00
HIF1A	0.69	0.92	1.00
HK1	0.30	0.88	1.00
HK2	0.93	0.98	1.00
HK3	0.51	0.92	1.00
HKDC1	0.73	0.92	1.00
HRAS	0.43	0.92	1.00
IDH1	0.17	0.84	1.00
IDH2	0.15	0.84	1.00
KIT	0.69	0.92	1.00
KRAS	0.76	0.92	1.00
LDHA	0.72	0.92	1.00
LDHAL6A	0.11	0.84	1.00
LDHAL6B	0.96	0.98	1.00

Gene	***p*-value**	**BH-FDR**	**lfdr**
LDHB	0.73	0.92	1.00
LDHC	0.68	0.92	1.00
MAP2K1	0.84	0.92	1.00
MAP2K2	0.47	0.92	1.00
MAPK1	0.15	0.84	1.00
MAPK3	0.76	0.92	1.00
MET	0.88	0.95	1.00
MTOR	0.78	0.92	1.00
MYC	0.27	0.87	1.00
NRAS	0.16	0.84	1.00
NTRK1	0.23	0.84	1.00
NTRK3	0.26	0.87	1.00
PDGFRA	0.46	0.92	1.00
PDGFRB	0.12	0.84	1.00
PDHA1	0.21	0.84	1.00
PDHA2	0.51	0.92	1.00
PDHB	0.01	0.38	1.00
PDK1	0.48	0.92	1.00
PFKL	0.23	0.84	1.00
PFKM	1.00	1.00	0.04
PFKP	0.13	0.84	1.00
PGAM1	0.77	0.92	1.00
PGAM2	0.00	0.13	0.03
PGAM4	0.77	0.92	1.00
PIK3CA	0.11	0.84	1.00
PIK3CB	0.82	0.92	1.00
PIK3CD	0.19	0.84	1.00
PIK3R1	0.81	0.92	1.00
PIK3R2	0.72	0.92	1.00
PIK3R3	0.70	0.92	1.00
PKM2	0.10	0.84	1.00
PTEN	0.30	0.88	1.00
RAF1	0.82	0.92	1.00
RET	0.81	0.92	1.00

Gene	***p*-value**	**BH-FDR**	**lfdr**
SCO2	0.68	0.92	1.00
SIRT3	1.00	1.00	0.04
SIRT6	0.70	0.92	1.00
SLC16A3	0.72	0.92	1.00
SLC1A5	0.95	0.98	1.00
SLC2A1	0.55	0.92	1.00
SLC2A2	0.22	0.84	1.00
SLC7A5	0.57	0.92	1.00
TP53	0.09	0.84	1.00

TABELLA A.3: *p-value ottenuti con il modello di Cox che tiene conto dei fattori confondenti età, stadio del tumore e sesso.*

Colonna 1: *p*-value, per ciascun gene, ottenuti con il modello di Cox univariato. **Colonna 2:** *p*-value corretti per test multipli con il metodo FDR di Benjamini e Hochberg (Benjamini e Hochberg, 1995). **Colonna 3:** *p*-value corretti per (i) deviazioni dalla distribuzione nulla teorica sottostante e (ii) per test multipli attraverso la procedura lfdr di Efron (Efron, Turnbull et al., 2011).

Gene	***p*-value**	**BH-FDR**	**lfdr**
AKT1	0.40	0.95	1.00
AKT2	0.90	0.97	0.92
AKT3	0.60	0.95	1.00
C12orf5	0.20	0.95	1.00
EGFR	0.28	0.92	1.00
ERBB2	0.44	0.92	1.00
FGFR1	0.80	0.95	1.00
FGFR2	0.73	0.92	1.00
FGFR3	0.60	0.95	1.00
FLT3	0.67	0.92	1.00
G6PD	0.90	0.95	1.00
GCK	0.11	0.92	1.00
GLS	0.80	0.95	1.00
GLS2	0.51	0.92	1.00
HIF1A	0.38	0.92	1.00
HK1	0.67	0.92	1.00
HK2	0.31	0.92	1.00
HK3	0.90	0.95	1.00
HKDC1	0.67	0.92	1.00
HRAS	0.28	0.92	1.00
IDH1	0.45	0.92	1.00
IDH2	0.90	0.95	1.00
KIT	0.73	0.92	1.00
KRAS	0.10	0.95	1.00
LDHA	0.90	0.95	1.00
LDHAL6A	0.53	0.92	1.00
LDHAL6B	0.39	0.92	1.00

Gene	***p*-value**	**BH-FDR**	**lfdr**
LDHB	0.80	0.95	1.00
LDHC	0.62	0.92	1.00
MAP2K1	0.80	0.95	1.00
MAP2K2	0.23	0.92	1.00
MAPK1	0.55	0.92	1.00
MAPK3	0.70	0.92	1.00
MET	0.44	0.92	1.00
MTOR	0.90	0.95	1.00
MYC	0.45	0.92	1.00
NRAS	0.33	0.92	1.00
NTRK1	0.91	0.92	1.00
NTRK3	0.10	0.95	1.00
PDGFRA	0.90	0.95	1.00
PDGFRB	0.23	0.92	1.00
PDHA1	0.30	0.92	1.00
PDHA2	0.62	0.92	1.00
PDHB	0.44	0.92	1.00
PDK1	0.90	0.95	1.00
PFKL	0.20	0.92	1.00
PFKM	0.14	0.92	1.00
PFKP	0.78	0.92	1.00
PGAM1	0.80	0.92	1.00
PGAM2	0.30	0.92	1.00
PGAM4	0.40	0.92	1.00
PIK3CA	0.80	0.95	1.00
PIK3CB	0.10	0.95	1.00
PIK3CD	0.20	0.92	1.00
PIK3R1	0.30	0.92	1.00
PIK3R2	0.60	0.92	1.00
PIK3R3	0.50	0.95	1.00
PKM2	0.30	0.92	1.00
PTEN	0.10	0.92	1.00
RAF1	0.30	0.92	1.00
RET	0.22	0.92	1.00

Gene	***p*-value**	**BH-FDR**	**lfdr**
SCO2	0.10	0.95	1.00
SIRT3	0.20	0.92	1.00
SIRT6	0.70	0.92	1.00
SLC16A3	0.80	0.95	1.00
SLC1A5	0.10	0.95	1.00
SLC2A1	0.12	0.92	1.00
SLC2A2	0.30	0.92	1.00
SLC7A5	0.20	0.95	1.00
TP53	0.80	0.95	1.00

Bibliografia

Akaike, Hirotsugu (1974). A new look at the statistical model identification. In: *IEEE Transactions on Automatic Control*, **19**.6, 716–723.

Altman, Douglas G, Lisa M McShane, Willi Sauerbrei e Sheila E Taube (2012). Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): Explanation and Elaboration. In: *PLoS Med*, **9**.5, e1001216.

Amin, Mahul B, Frederick L Greene, Stephen B Edge, Carolyn C Compton, Jeffrey E Gershenwald, Richard K Brookland, Laura Meyer, David M Gress, David R Byrd e David P Winchester (2017). The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. In: *CA: A Cancer Journal for Clinicians*, **67**.2, 93–99.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry et al. (2000). Gene Ontology: Tool for the Unification of Biology. In: *Nature Genetics*, **25**, 25–29.

Azzalini, A. e B. Scarpa (2004). *Analisi dei dati e data mining*. Springer-Verlag Italia. ISBN: 88-470-0272-9.

Bailey, Matthew H., Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C. Wendl, Jaegil Kim, Brendan Reardon, Kwok-Shing Ng, Kang Jin Jeong, Song Cao, Zixing Wang, JianJiong Gao, Qingsong Gao, Fang Wang, Eric Minwei Liu, Loris Mularoni, Carlota Rubio-Perez, Niranjan Nagarajan, Isidro Cortés-Ciriano, Daniel Cui Zhou, Wen-Wei Liang, Julian M. Hess, Venkata D. Yellapantula, David Tamborero, Abel Gonzalez-Perez, Chayaporn Suphavilai, Jia Yu Ko, Ekta Khurana, Peter J. Park, Eliezer Van Allen, Han Liang, The MC3 Working Group, The Cancer Genome Atlas Research Network, Michael Lawrence, Adam Godzik, Nuria Lopez-Bigas, Josh Stuart, David Wheeler, Gad Getz, Ken Chen, Alexander J. Lazar, Gordon B. Mills, Rachel Karchin e Li Ding

(2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. In: *Cell*, **173**.2, 371–385.e18.

Benjamini, Yoav e Daniel Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. In: *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**.1, 289–300.

Breheny, Patrick e Jian Huang (2015). Group Descent Algorithms for Nonconvex Penalized Linear and Logistic Regression Models with Grouped Predictors. In: *Statistics and Computing*, **25**, 173–187.

Breheny, Patrick, Yaohui Zeng e Ryan Kurth (2021). *Grpreg: Regularization Paths for Regression Models with Grouped Covariates*. URL: `https://CRAN.R-project.org/package=grpreg`.

Breslow, N.E. (1972). Multiple Imputation of Incomplete Data: A Bayesian Approach. In: *Journal of the American Statistical Association*, **67**.338, 728–733.

Brier, Glenn W. (1950). Verification of forecasts expressed in terms of probability. In: *Monthly Weather Review*, **78**, 1–3.

Broström, Göran (2022). *Event History Analysis with R, 2nd ed.* Boca Raton, FL: Chapman & Hall/CRC Press, xxxv + 304. ISBN: 978-1-138-58771-7.

Cairns, R. A., I. Harris, S. McCracken e T. W. Mak (2011). Cancer cell metabolism. In: *Cold Spring Harbor Symposia on Quantitative Biology*, **76**. Epub 2011 Dec 12, 299–311.

Corsello, Steven M., Rohith T. Nagari, Ryan D. Spangler, Jordan Rossen, Mustafa Kocak, Jordan G. Bryan, Ranad Humeidi, David Peck, Xiaoyun Wu, Andrew A. Tang, Vickie M. Wang, Samantha A. Bender, Evan Lemire, Rajiv Narayan, Philip Montgomery, Uri Ben-David, Colin W. Garvie, Yejia Chen, Matthew G. Rees, Nicholas J. Lyons, James M. McFarland, Bang T. Wong, Li Wang, Nancy Dumont, Patrick J. O'Hearn, Eric Stefan, John G. Doench, Caitlin N. Harrington, Heidi Greulich, Matthew Meyerson, Francisca Vazquez, Aravind Subramanian, Jennifer A. Roth, Joshua A. Bittker, Jesse S. Boehm, Christopher C. Mader, Aviad Tsherniak e Todd R. Golub (2020). Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. In: *Nature Cancer*, **1**.2, 235–248.

Cox, D. R. (1972). Regression Models and Life-Tables. In: *Journal of the Royal Statistical Society B,* **34**.2, 187–202.

Cox, D. R. e D. Oakes (1987). Analysis of Survival Data. In: *Biometrical Journal*, **29**.1, 114.

Efron, B., R. Tibshirani, J. D. Storey e V. Tusher (2001). Empirical Bayes Analysis of a Microarray Experiment. In: *Journal of the American Statistical Association*, **96**, 1151–1160.

Efron, B., B. B. Turnbull e B. Narasimhan (2011). *R Package*. Ver. 1.1. URL: `http://CRAN.R-project.org/package=locfdr`.

Fabregat, Antonio, Sarah Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Pradeep Garapati et al. (2018). The Reactome Pathway Knowledgebase. In: *Nucleic Acids Research*, **46**, D649–D655.

Friedman, Jerome, Trevor Hastie e Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. In: *Journal of Statistical Software*, **33**.1, 1–22.

Gatza, Michelle L., Joseph E. Lucas, William T. Barry, Joo Won Kim, Q. Wang, Melissa D. Crawford et al. (2010). A Pathway-Based Classification of Human Breast Cancer. In: *Proceedings of the National Academy of Sciences*, **107**, 6994–6999.

Gilis, Jeroen, Steff Taelman, Lucas Davey, Lennart Martens e Lieven Clement (2020). Pitfalls in re-analysis of observational omics studies: a post-mortem of the human pathology atlas. In: *bioRxiv*,

Hanley, J.A. e B.J. McNeil (1982). The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. In: *Radiology*, **143**.1, 29–36.

Harrell, Frank E., Kerry L. Lee e Daniel B. Mark (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. In: *Statistics in Medicine*, **15**.4, 361–387.

Harrell Frank E., Jr., Robert M. Califf, David B. Pryor, Kerry L. Lee e Robert A. Rosati (1982). Evaluating the Yield of Medical Tests. In: *JAMA*, **247**.18, 2543–2546.

Hastie, Trevor, Robert Tibshirani e Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. New York: Springer. ISBN: 978-0-387-84857-0.

Heagerty, Patrick J. e Yingye Zheng (2005). Survival Model Predictive Accuracy and ROC Curves. In: *Biometrics*, **61**.1, 92–105.

Jacob, Laurent, Guillaume Obozinski e Jean-Philippe Vert (2009). Group Lasso with Overlap and Graph Lasso. In: *Proceedings of the 26th Annual International Conference on Machine Learning*, 433–440.

Kalousis, A., J. Prados e M. Hilario (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. In: *Knowledge and Information Systems*, **12**, 95–116.

Kanehisa, Minoru e Susumu Goto (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. In: *Nucleic Acids Research*, **28**, 27–30.

Kaplan, Edward L. e Paul Meier (1958). Nonparametric estimation from incomplete observations. In: *Journal of the American Statistical Association*, **53**.282, 457–481.

Kattan, M. W., T. M. Wheeler e P. T. Scardino (1999). Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. In: *Journal of Clinical Oncology*, **17**.5, 1499–1507.

Kim, Jong, Inhwa Sohn, Sin-Ho Jung, Seung Kim e Cheolwoo Park (2012). Analysis of Survival Data with Group Lasso. In: *Communications in Statistics - Simulation and Computation*, **41**, 1593–1605.

Kleinbaum, David G. e Mitchel Klein (2012). *Survival Analysis: A Self-Learning Text*. Springer.

Li, Bo e M. Celeste Simon (2013). Molecular Pathways: Targeting MYC-induced Metabolic Reprogramming and Oncogenic Stress in Cancer. In: *Clinical Cancer Research*, **19**.21, 5835–5841.

Liu, Jing, Tara Lichtenberg, Katherine A Hoadley, Lonesha M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Annette V Lee et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. In: *Cell*, **173**.2, 400–416.e11.

Lockhart, Richard, Jonathan Taylor, Ryan J. Tibshirani e Robert Tibshirani (2014). A significance test for the lasso. In: *The Annals of Statistics*, **42**.2, 413.

Malenová, Gabriela, Daniel Rowson e Valentina Boeva (2021). Exploring Pathway-Based Group Lasso for Cancer Survival Analysis: A Special Case of Multi-Task Learning. In: *Frontiers in Genetics*, **12**, 771301.

Martignetti, Laure, Laurence Calzone, Emeric Bonnet, Emmanuel Barillot e Andrei Zinovyev (2016). Roma: Representation and Quantification of Module Activity from Target Expression Data. In: *Frontiers in Genetics*, **7**, 18.

Mohr, Alex E., Carmen P. Ortega-Santos, Corrie M. Whisner, Judith Klein-Seetharaman e Paniz Jasbi (2024). Navigating Challenges and Opportunities in Multi-Omics Integration for Personalized Healthcare. In: *Biomedicines*, **12**.7, 1496.

Obozinski, Guillaume, Laurent Jacob e Jean-Philippe Vert (2011). Group Lasso with Overlaps: The Latent Group Lasso Approach. In: *arXiv preprint arXiv:1110.0413*,

Parikh, Jigar R., Björn Klinger, Yu Xia, Jarrod A. Marto e Nils Blüthgen (2010). Discovering Causal Signaling Pathways through Gene-Expression Patterns. In: *Nucleic Acids Research*, **38**, W109–W117.

Press, William H, Saul A Teukolsky, William T Vetterling e Brian P Flannery (2007). *Numerical Recipes: The Art of Scientific Computing*. 3rd. Cambridge University Press.

Rahman, M. Shafiqur, Gareth Ambler, Babak Choodari-Oskooei e Rumana Z. Omar (2017). Review and evaluation of performance measures for survival prediction models in external validation settings. In: *BMC Medical Research Methodology*, **17**, 60.

Ramos, M., L. Geistlinger, S. Oh, L. Schiffer, R. Azhar, H. Kodali, I. de Bruijn, J. Gao, V. Carey, M. Morgan e L. Waldron (2020). Multiomic Integration of Public Oncology Databases in Bioconductor. In: *JCO Clinical Cancer Informatics*, **1**.4, 958–971.

Royston, P. e D. G. Altman (2013). External validation of a Cox prognostic model: principles and methods. In: *BMC Medical Research Methodology*, **13**, 33.

Rydenfelt, Malin, Björn Klinger, Martin Klünemann e Nils Blüthgen (2020). SPEED2: Inferring Upstream Pathway Activity from Differential Gene Expression. In: *Nucleic Acids Research*, **48**, W307–W312.

Schubert, Michael, Björn Klinger, Martin Klünemann, Anne Sieber, Florian Uhlitz, Sascha Sauer et al. (2018). Perturbation-response Genes Reveal Signaling Footprints in Cancer Gene Expression. In: *Nature Communications*, **9**, 20–11.

Simon, N., J. Friedman, T. Hastie et al. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. In: *Journal of Statistical Software*, **39**, 1–13.

Simon, Noah, Jerome Friedman, Trevor Hastie e Robert Tibshirani (2013). A sparse-group lasso. In: *Journal of Computational and Graphical Statistics*, **22**, 231–245.

Smith, Joan C. e Jason M. Sheltzer (2022). Genome-wide identification and analysis of prognostic features in human cancers. In: *Cell Reports*, **38**.13, 110569.

The Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander e Joshua M Stuart (2013). The Cancer Genome Atlas Pan-Cancer analysis project. In: *Nature Genetics*, **45**, 1113–1120.

Tibshirani, Robert (1997). The lasso method for variable selection in the Cox model. In: *Statistics in Medicine*, **16**, 385–395.

Tomczak, Kinga, Patrycja Czerwińska e Maciej Wiznerowicz (2015). The Cancer Genome Atlas (TCGA): an Immeasurable Source of Knowledge. In: *Contemporary Oncology (Poznań)*, **19**, A68–A77.

Uhlen, M., C. Zhang, S. Lee, E. Sjöstedt, L. Fagerberg, G. Bidkhori, R. Benfeitas, M. Arif, Z. T. Liu, F. Edfors, K. Sanli, K. von Feilitzen, P. Oksvold, E. Lundberg, S. Hober, P. Nilsson, J. Mattsson, J. M. Schwenk, H. Brunnstrom, B. Glimelius, T. Sjöblom, P. H. Edqvist, D. Djureinovic, P. Micke, C. Lindskog, A. Mardinoglu e F. Pontén (2017). A pathology atlas of the human cancer transcriptome. In: *Science*, **357**.6352, 660.

Uno, Hajime, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino e Lee-Jen Wei (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. In: *Statistical Medicine*, **30**, 1105–1117.

Ventura, L. e W. Racugno (2017). *Biostatistica. Casi di Studio in R*. Italiano. Milano, Italia: Egea, 1–326.

Warburg, Otto (1956). On the Origin of Cancer Cells. In: *Science*, **123**, 309–314.

Wishart, David S., Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox e Michael Wilson (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. In: *Nucleic Acids Research*, **46**.D1, D1074–D1082.

Yuan, Ming e Yi Lin (2006). Model selection and estimation in regression with grouped variables. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 49–67.

Yue, Zongliang (2024). *Central Carbon Metabolism in Cancer - Homo sapiens (human)*. URL: `https://discovery.informatics.uab.edu/PAGER/index.php/geneset/view/WAG001958`.

Zhao, Zhi, John Zobolas, Manuela Zucknick e Tero Aittokallio (2024). Tutorial on survival modeling with applications to omics data. In: *Bioinformatics*, **40**.3, btae132.