CAIMOD Tutorial
Analisi della sopravvivenza mediante dati omici: dalla teoria alla pratica.

Claudia Angelini, Francesca Calanca, Andrea Raiconi

Istituto per le Applicazioni del Calcolo "M. Picone"
Consiglio Nazionale delle Ricerche

11/12/2025

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
Consiglio Nazionale delle Ricerche
IAC

Section 1

Introduzione e notazioni

Introduzione

Molto spesso all'interno di studi clinici di follow-up si misura il tempo T fino al verificarsi di un determinato evento:

- il tempo fino alla morte di un paziente.
- il tempo fino all'insorgenza di una determinata patologia.
- il tempo fino ad una nuova recidiva del cancro dopo il trattamento/operazione.

I dati che riguardano tali studi sono spesso chiamati dati **time-to-event** (*tempo all'evento*) o anche detti *tempo di fallimento* (mutuando il nome quando l'evento riguarda la durata di un dispositivo).

Più in generale, dati di questo tipo si incontrano in diversi settori

- **Medicina**: Valutare l'efficacia di nuovi farmaci, prevedere l'esito di una terapia, identificare i fattori che influenzano la sopravvivenza o la recidiva in malattie come il cancro.
- **Ingegneria**: Stimare la durata di vita dei componenti meccanici o elettronici (modelli di durata).
- **Scienze sociali**: Analizzare la durata di eventi come la disoccupazione.

La branca della statistica che si occupa di analizzare questo tipo di dati è chiamata **Analisi della Sopravvivenza**.

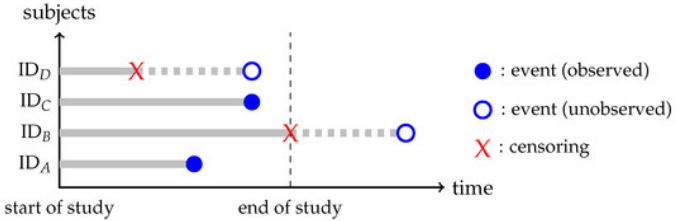
Introduzione

I dati **time-to-event** sono un tipo di dati diversi dai dati categorici o continui classici e richiedono approcci specifici per la loro analisi statistica.

A prima vista, potrebbe sembrare che il tempo T fino al verificarsi di un evento sia una variabile continua, tuttavia, esiste una caratteristica fondamentale dei dati tempo-evento:

- **l'evento non si verifica sempre** o più in generale non sempre si verifica all'interno dell'intervallo di tempo dello studio (tempo di follow-up).

In questo caso, il tempo all'evento è incognito (non sappiamo se e quando l'evento si verificherà). Tuttavia si sà che fino ad un certo istante, l'evento non si è verificato. $\Rightarrow$ Questo tipo di dati vengono detti **censurati** o censored.

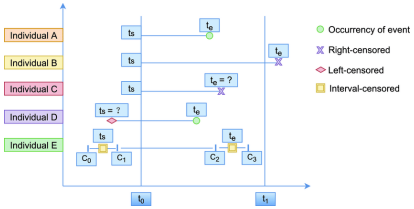


Introduzione

La censura (**censoring**) è un fenomeno piuttosto comune; può essere causata da molti fattori:

- la fine dello studio di follow-up,
- il paziente abbandona lo studio e si perdono i contatti con lui/lei,
- il paziente muore per motivi indipendenti dalla patologia che si sta studiando

Ci possono essere diverse forme di censoring, sulla base dei quali le analisi possono essere effettuate in modo diverso.



Di seguito assumiamo di avere dati di **independent (non-informative) right censoring**, che è la tipologia più comune.

Notazioni e Concetti Chiave

Sia $T \geq 0$ una casuale variabile non negativa che rappresenta il tempo fino al verificarsi di un determinato evento.

- **Funzione di Survival** è la probabilità che un individuo sopravviva più del generico tempo t,

$$S(t) = P(T > t), \ t > 0$$

.

- **Funzione di Hazard** denota il tasso istantaneo di fallimento/morte al tempo t, dato che un individuo è sopravvissuto almeno fino al tempo t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}.$$

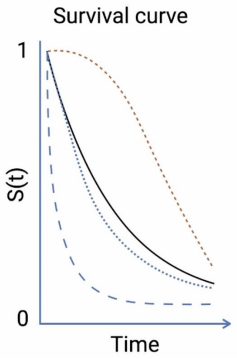
Queste due quantità sono funzioni non negative e legate tra loro dalla seguente relazione:

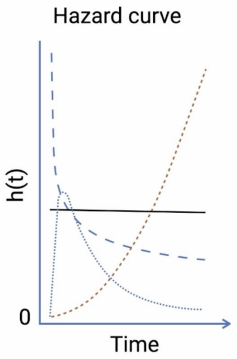
$$h(t) = -\frac{d \log S(t)}{dt} \quad \text{and} \quad S(t) = exp\left(-\int_0^t h(s)ds\right).$$

Esempio

La funzione *S(t)* di sopravvivenza è non negativa e non crescente. Tanto più rapida è la discesa della funzione di sopravvivenza verso lo zero, tanto più breve sarà il tempo di sopravvivenza.

Nota la funzione ***S(t)*** per un dato individuo/patologia è possibile fornire una stima della probabilità di "sopravvivere" almeno fino ad un certo tempo.





La funzione di sopravvivenza ***S(t)*** riguarda l'intera popolazione di interesse. Se sono disponibili delle covariate **X** osservate sul singolo campione, si può ottenere ***S(t|X)*** che è una curva di sopravvivenza "***personalizzata***". Più in generale, una popolazione può essere suddivisa (i.e., **stratificata**) in sottogruppi con curve di sopravvivenza simili all'interno di ciascun gruppo.

La funzione *h(t)* di hazard è non negativa.

h(t) **crescente:** situazione tipica quando l'evento di interesse è legato all'usura/invecchiamento per cui più passa il tempo maggiore è il rischio che si verifichi l'evento (es. rottura di un dispositivo).

h(t) **decrescente:** situazione tipica quando più passa il tempo più il rischio che si verifichi l'evento diminuisce (es. l'evento è una complicanza post intervento chirurgico).

Obiettivi dell'Analisi di Sopravvivenza

Gli obiettivi di un'analisi di sopravvivenza sono

- Stimare il tempo ad un evento (i.e., la funzione di sopravvivenza o la funzione di Hazard) per un gruppo di individui di interesse. E.g., essere in grado di calcolare la probabilità che $T > t^*$, per valori di t^* che dipendono dalla patologia oggetto di studio.
- Confrontare il tempo di sopravvivenza per due o più gruppi di individui. E.g., vedere se un determinato trattamento è in grado di migliore l'aspettativa di vita oppure se una determinata variabile clinica o genetica è in grado di stratificare la popolazione in gruppi con classi di rischio diverse.
- Studiare la relazione tra la funzione di sopravvivenza $S(t|\mathbf{X})$ o la funzione di Hazard $h(t|\mathbf{X})$ ed una o più variabili esplicative $\mathbf{X} = (X_1, \ldots, X_p)^T$. E.g., individuare le variabili che possono agire come fattori di rischo o fattori di protezione per una determinata patologia e costruire curve di sopravvivenza paziente specifiche.

Pertanto, se si vuole analizzare il tempo ad un evento di interesse per una popolazione di individui, occorre raccogliere un insieme di dati osservati su un campione casuale di taglia n della popolazione, ovvero i tempi T_i degli eventi, le informazioni sul censoring C_i, e eventuali variabili di interesse. Questo si ottiene mendiante uno studio di follow-up in cui gli n individui vengono monitorati nel tempo.

Notazioni e Concetti chiave

Consideriamo uno studio di follow-up con n pazienti.

Per $i = 1, \ldots, n$:

- T_i denota il tempo di sopravvivenza.
- C_i denota il tempo di censura e $\delta_i = \mathrm{I}(T_i \leq C_i)$ è l'indicatore della censura, dove $\mathrm{I}()$ denota la funzione indicatrice, i.e.,
  - $\delta_i = 1$ se il tempo dell'evento è osservato
  - $\delta_i = 0$ se il tempo dell'evento è censurato
- Pertanto, $\xi_i = min\{T_i, C_i\}$ denotano i dati di sopravvivenza osservati.
- Sia $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ un vettore di p-variabili da usare come predittore o variabili esplicative.

Pertanto $(T_i, \delta_i, \mathbf{X}_i) \quad i = 1, ..., n$, rappresenta l'insieme dei dati osservati durante lo studio.

- **Metodi non parametrici**: si calcola una stima di $S(t)$ senza fare assunzioni sulla densità di probabilità di T. E.g., Il metodo di Kaplan-Meier
- **Modelli parametrici**: si assume che T sia una variabile aleatoria avente una determinata distribuzione di probabilità (tipicamente esponenziale, di Weibull o lognormale) e si stima $S(t)$ di conseguenza.

Section 2

Stimatore di Kaplan-Meier e log-rank test

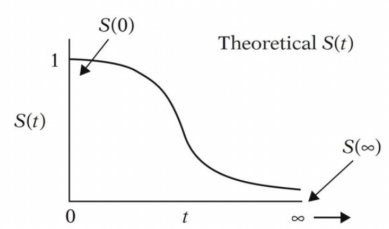
Lo stimatore di Kaplan-Meier

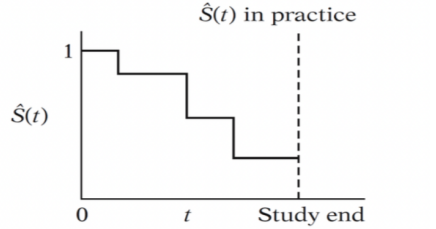
Il metodo di **Kaplan–Meier** è un approccio **non parametrico** per stimare la funzione di sopravvivenza $S(t)$.

Siano $t_1 < t_2 < t_3 < \cdots < t_K$ i K tempi di sopravvivenza distinti ordinati in senso crescente, lo stimatore di KM è dato da

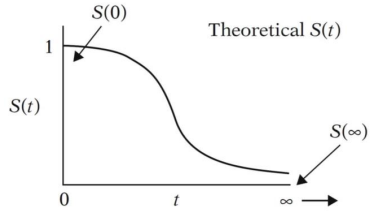
$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

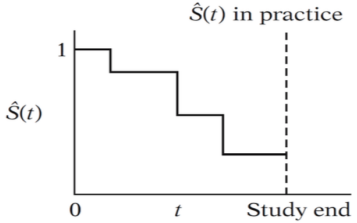
dove n_i denota il **numero di individui a rischio evento all'istante** t_i (i.e., che non hanno ancora avuto un evento o che sono stati censurati) e d_i indica il **numero di eventi osservati all'istante** t_i.

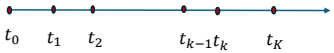




Lo stimatore di Kaplan-Meier







Al tempo $t = 0$ sono presenti n individui a rischio
La diminuzione degli individui n_k a rischio di evento è dovuta

1) Verificarsi di d_k eventi nell'intervallo di tempo considerato
2) Censoring di alcuni pazienti/tempi all'interno dellintervallo di tempo considerato

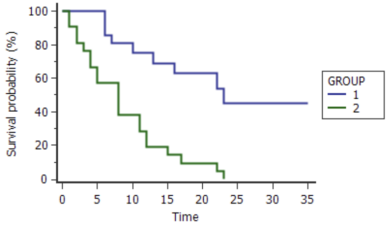
Intervallo	Individui a rischio	Numero Eventi	Sopravvivenza intervallo di tempo
$t < t_1$	n	0	1
$t_1 \leq t < t_2$	n_1	d_1	$\frac{n_1 - d_1}{n_1}$
$t_2 \leq t < t_3$	n_2	d_2	$\frac{n_2 - d_2}{n_2}$
...	...	....	...
$t_{K-1} \leq t < t_K$	n_K	d_K	$\frac{n_K - d_K}{n_K}$

Il log-rank test

Il log-rank test è un test statistico non parametrico per confrontare le curve di sopravvivenza di due o più gruppi di individui e stabilire se sono statisticamente diverse oppure se le differenze osservate nelle curve sono legate al caso.

I gruppi possono rappresentare stratificazioni di una popolazione in classi di rischio sulla base di parametri clinici o molecolari.

Nel caso di due gruppi: si vuole testare $H_0 : S_1 = S_2$ contro l'alternativa $H_1 : S_1 \neq S_2$. Come nel caso di KM, si considerano i $t_1 \leq t_2 \leq t_3 \leq, \cdots \leq t_K$ i K tempi di sopravvivenza ordinati in senso crescente. Per ciascun intervallo di tempo, si calcolano le frequenze di occorrenza osservate nei due gruppi.



Nel generico intervallo $[t_{k-1}, t_k]$ si hanno le seguenti **frequenze osservate**

	Gruppo 1	Gruppo 2	Totale
Eventi	$d_{1,k}$	$d_{2,k}$	$d_k = d_{1,k} + d_{2,k}$
Non Eventi	$n_{1,k} - d_{1,k}$	$n_{2,k} - d_{2,k}$	$n_k - d_k = n_{1,k} + n_{2,k} - (d_{1,k} + d_{2,k})$
Totale	$n_{1,k}$	$n_{2,k}$	$n_k = n_{1,k} + n_{2,k}$

Il log-rank test

Sotto l'ipotesi nulla, la frequenza attesa di $D_{l,k}$ per $l = 1, 2$ segue una distribuzione hypergeometrica con media

$$e_{l,k} = n_{l,k} \frac{d_k}{n_k}$$

e varianza

$$v_{l,k} = n_{l,k} \frac{d_k}{n_k} \frac{n_k - d_k}{n_k} \frac{n_k - n_{l,k}}{n_k - 1}$$

Lo stimatore è dato dalla statistica test

$$\frac{W}{\sqrt{V}} \sim N(0, 1)$$

dove $W = \sum_k (D_{i,k} - e_{1,k})$ e $V = \sum_k v_{1,k}$.

In alternativa, si ha che (sotto l'ipotesi nulla)

$$\frac{W^2}{V} \sim \chi(1).$$

Il log-rank test si generalizza al caso di più gruppi.

Section 3

Regressione di Cox

L'analisi di sopravvivenza come modello di regressione

L'obiettivo principale di un modello di regressione per l'analisi di sopravvivenza è:

- Studiare la funzione di sopravvivenza $S(t|\mathbf{X})$ o la funzione di hazard $h(t|\mathbf{X})$ come funzione delle variabili osservate $\mathbf{X} = (X_1, \ldots, X_p)^T$ attraverso dei coefficienti incogniti che devono essere stimati dai dati.

Ulteriori obiettivi riguardano

- Identificare un piccolo sottoinsime di variabili $X_i \quad i \in S$ che contribuiscono in modo significativo alla predizione. Tali variabili possono agire come signature/marcatori ed aiutare nella comprensione dei meccanismi molecolari.
- Calcolare per ciascun paziente uno score o prognostic index $PI_i(\mathbf{X}_i)$ che può essere utilizzato per stratificare i pazienti in classi di rischio.

Il modello di Cox

Il modello di Cox (Cox 1972) è un modello di tipo proportional hazard **semi-parametrico** e costituisce una tecnica statistica multivariata usata per analizzare il tempo che intercorre prima di un evento (come una malattia o la morte) in relazione ad una o più variabili esplicative (come l'età, il sesso o un trattamento, ed più in generale anche informazioni molecolari).

Assumiamo di osservare $(t_i, \delta_i, \mathbf{X}_i^T) \quad i = 1, ..., n$, e di modellizzare la **hazard function** $h(t)$ come

$$h(t|\mathbf{X}_i) = h_0(t)\exp\Big(\sum_{j=1}^{p} x_{i,j}\beta_j\Big) = h_0(t)\exp\Big(\mathbf{X}_i^T\beta\Big)$$

$h_0(t)$: è la **baseline hazard function** che descrive il rischio quando $\mathbf{X}_j = 0 \quad j = 1, ..., p$,

$\exp\Big(\sum_{j=1}^{p} x_{i,j}\beta_j\Big)$: **relative risk**, i.e., proporzionale all'aumento o diminuzione di x_{ij}

$\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$: è il vettore dei coefficienti (da stimare dai dati).

Cox-regression

- $h_0(t)$ rappresenta la parte non parametrica dove non si fanno assunzioni sulla forma della baseline hazard function.
- $\exp\left(\mathbf{X}_i^T \beta\right)$ rappresenta la parte parametrica funzione delle variabili indipendenti.

Il metodo della massima verosimiglianza parziale (**partial maximum likelihood**) consente di stimare i coefficienti $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, senza conoscere $h_0(t)$.

Stima dei parametri del modello: i parametri $\boldsymbol{\beta}$ della regressione sono ottenuti massimizzando la **Cox's partial likelihood**, o equivalentemente, minimizzando la **negative Cox's log-partial likelihood** $-l(\beta)$

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}}\big[- l(\boldsymbol{\beta})\big] = \text{argmin}_{\boldsymbol{\beta}}\left[- \frac{1}{n} \sum_{i=1}^{n} \delta_i \left\{ \mathbf{X}_i^T \boldsymbol{\beta} - \log \left[\sum_{j \in R(t_i)} \exp(\mathbf{X}_j^T \boldsymbol{\beta}) \right] \right\} \right],$$

dove $R(t_i)$ denota l'insieme di individui a rischio al tempo t_i.

Una volta ottenuti $\hat{\boldsymbol{\beta}}$, si può stimare $h_0(t)$ con il metodo di Breslow.

Interpretazione dei coefficienti del modello di Cox

Hazard ratio: $e^{\hat{\beta}_j}$ indica di quanto un aumento di una unità della variabile X_j, tenendo costanti tutte le altre variabili indipendenti, amplifica o attenua $h(t)$

- Se $\hat{\beta}_j > 0$ allora $e^{\hat{\beta}_j} > 1$, pertanto se X_j aumenta, anche $h(t)$ aumenta, ovvero X_j è un fattore di rischio (date le altre variabili).
- Se $\hat{\beta}_j < 0$ allora $e^{\hat{\beta}_j} < 1$, pertanto se X_j aumenta, $h(t)$ diminuisce, ovvero X_j è un fattore di protezione (date le altre variabili).
- Se $\hat{\beta}_j = 0$ allora $e^{\hat{\beta}_j} = 1$,la variabile X_j non ha impatto su $h(t)$ (date le altre variabili).

Il **Wald test** serve per determinare se i coefficienti β_j sono significativamente diversi da 0: $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$.

Si noti che se $\beta_j \neq 0$ allora la corrispondente variabile X_j può essere considerata come un marker dell'evento.

Predittore lineare ed score prognostico

Un **indice prognostico** (PI) derivato da un modello di regressione è un punteggio (*score*) che rappresenta il rischio di un paziente ottenuto combinando i valori dei suoi predittori X_j con i corrispondenti coefficienti $\hat{\beta}_j$ di regressione stimati dal modello.

In particolare per il modello di Cox, si ha

$PI_i = \mathbf{X}_i^T \hat{\beta}$ (predittore lineare).

Punteggi (score) di rischio più elevati derivanti dall'PI indicano un rischio maggiore per l'evento di interesse.

Questo score può essere utilizzato per suddividere i pazienti in **classi di rischio**. Ad esempio, è possibile individuare una soglia th e suddividere i pazienti $i = 1, \ldots, n$ nelle classi di rischio alto e basso,

$$\begin{cases} PI_i > th\ , & \textbf{Alto rischio} \\ PI_i \leq th\ , & \textbf{Basso rischio}. \end{cases} \tag{1}$$

Una volta ottenuta la stratificazione, si può vedere se la differenza di classi di rischio è effettivamente significativa.

Section 4

Regressione e metodi di penalzzazione

Modello di regressione di Cox e metodi di penalizzazione

Quando il numero di predittori aumenta, la stima di massima verosimiglianza non può essere ottenuta. Analogamente al caso della regressione lineare si possono utilizzare **metodi di penalizzazione** al fine di regolarizzare il problema ed imporre una struttura desiderata alla soluzione.

Pertanto, si determina un **path** di soluzioni $\hat{\beta}_{\lambda}^{Pen}$ come

$$\hat{\beta}_{\lambda}^{Pen} = \text{argmin}_{\beta}\big[l_{pen}(\beta)\big] = \text{argmin}_{\beta}\big[-l(\beta) + P_{\lambda}(\beta)\big]$$

dove nel caso della regressione di Cox $l(\beta)$ è la funzione di log-likelihood parziale e $P_{\lambda}(\beta)$ è una funzione di penalizzazione che dipende da un parametro λ.

Per ciascun valore di λ si ottiene un corrispondente vettore di parametri $\hat{\beta}_{\lambda}^{Pen}$.

Il parametro $\lambda \geq 0$ è detto **parametro di regolarizzazione** e in pratica si deve determinare separatamente. Esso agisce come **tuning paramenter**.

- Se $\lambda = 0 \quad \Rightarrow \hat{\beta}^{Pen} = \hat{\beta}^{MLE}$
- Se $\lambda \to \infty \quad \Rightarrow \hat{\beta}^{Pen} \to \mathbf{0}$

.

Modello di regressione di Cox e metodi di penalizzazione

Esistono tanti modelli di regressione penalizzata che differiscono sulla base del termine di penalizzazione $P_\lambda(\beta)$ scelto

Alcuni esempi famosi:

- **Ridge Regression:** $P_\lambda(\beta) = \lambda||\beta||_2$
- **Lasso Regression:** $P_\lambda(\beta) = \lambda||\beta||_1$
- **Elastic-net Regression:** $P_\lambda(\beta) = \lambda_1||\beta||_1 + \lambda_2||\beta||_2$.

La regressione Ridge effettua una "riduzione" (**shrinkage**) dei coefficienti $\hat{\beta}_\lambda^{Pen}$ che aumenta al crescere del parametro λ.

La regressione Lasso, oltre ad effettuare una riduzione dei coefficienti, effettua anche una **selezione delle variabili** più importanti (i.e., mette a zero i coefficienti non importanti). Tuttavia, presenta dei limiti quando le variabili sono tra loro correlate.

La regression Elastic-net è un compromesso tra i due approacci precedenti.

Esistono molti altri approcci basati su tecniche di penalizzazione più sofisticate (vedere Poster Session).

La scelta del parametro di regolarizzazione

Le performance dei metodi di regolarizzazione dipendono fortemente dalla scelta del parametro di regolarizzazione λ (che tuttavia non è noto). Idealmente, il parametro di regolarizzazione λ dovrebbe essere scelto in modo da minimizzare l'errore di predizione (i.e., MSE, devianza, C-index), a seconda del tipo di regressione.

Tuttavia non è possibile calcolare esattamente tali errori, poichè dipendono dalla funzione di regressione incognita. Di conseguenza, l'idea è quello effettuare una stima di questo errore e di conseguenza, scegliere λ minimizzando questa stima. In questo modo

$$\hat{\lambda} = argmin_{\lambda>0} \quad C(\lambda, \mathbf{X}, \mathbf{Y})$$

è uno stimatore che dipende dai dati (i.e., scelta **data driven**) e $C(.)$ è un opportuno criterio.

Tra i criteri più utilizzati ricordiamo la **Cross-Validation**.

1. Leave-one-out cross validation.
2. K-fold cross-validation.

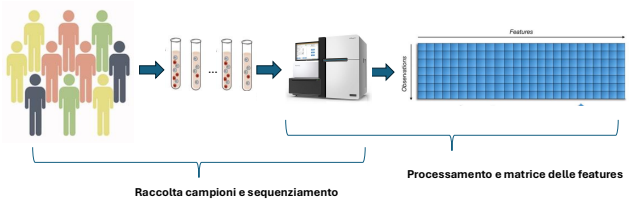
Esistono comunque altre possibilità, ed una parte significativa della messa a punto delle strategie di apprendimento riguarda i criteri per il tuning dei parametri.

Section 5

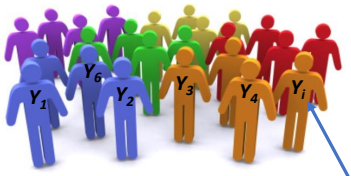
Dati omici e Analisi di Sopravvivenza

Dati Omici

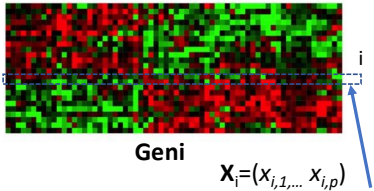
- I moderni sequenziatori consentono di acquisire **informazioni molecolari** dettagliate a livello di intero genoma per ciascun campione/individuo, in tempi rapidi ed ad un costo contenuto.
- Dopo opportune fasi di **pre-processing**, queste informazioni possono essere convertite in un vettore $(x_1, \ldots, x_p)^T$ di p variabili, dove $p \approx 10.000 - 50.000$ a seconda del tipo di dato molecolare.
- Uno studio che coinvolge n individui determina una **matrice di dati** $\mathbf{X} \in R^{n \times p}$ dove $p >> n$.
- I dati sono pertanto ad alta dimensione, estremamente rumorosi e correlati.
- La moderna ricerca in campo biomedico ha come obiettivo l'utilizzo di questo tipo di dati per la **medicina di precisione** $\Rightarrow$ Necessità di sviluppare **modelli ed algoritmi** adeguati.



Dati Omici ed Analisi della Sopravvivenza

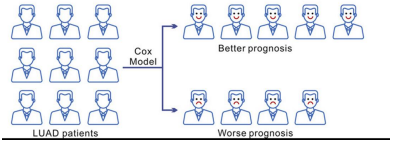


Pazienti presenti nello studio di follow-up per cancro



I dati osservati consistono in

$$(t_i, \delta_i, \boldsymbol{X}_i)_{i=1,\ldots,n} \qquad t_i = \min(T_i, C_i)$$



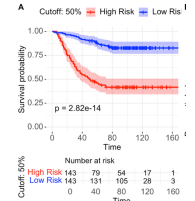
Studiare la **relazione** tra la **sopravvivenza** e le variabili **X.**

Fornire una predizione per **ciascun individuo** della sopravvivenza $S(X_i)$ a un dato tempo di interesse.

Fornire uno **score di rischio** PI_i a **ciascun paziente**.

Stratificare gli individui in classi di rischio sulla base del loro profilo di espressione.

Identificare una **gene-signature** (i.e., potenziali ***biomakers***) che può spiegare/predire il responso fornendo fattori di rischio e/o di protezione.



Section 6

Parte pratica: Un Caso Studio su Breast Cancer (in R)

Link al tutorial pratico e Ringraziamenti

La parte pratica:https://github.com/FraCalanca/TutorialBari

SCAN ME

⇒ Per ulteriori approfondimenti siete invitati alla Poster Session: **Bridging Multi-omics Data and Survival Analysis with Omics2Surv.**

This tutorial is part of the dissemination activities supported by the P2022BLN38 project *Computational approaches for the integration of multi-omics data* – funded by European Union – Next Generation EU within the PRIN 2022 PNRR program (D.D. 1409 del 14-09-2022 Ministero dell'Università e della Ricerca) CUP B53D23027810001.