Data Science Lab in Medicine: Big Data in Health Care

Federico Signoretta

May 12, 2020

Contents

1 Disegno dello studio e tassonomia degli studi nella ricerca biomedica

Come primo passo nella gestione dei Big Data in ambito Health Care, vi è la definizione dell'obiettivo di studio prima di tutto. In tale senso, la ricerca biomedica è l'insieme di studi con finalità mediche volte a stabilire una relazione tra una caratteristica o intervento (es. trattamento) ed una malattia o una condizione predisponente ad una malattia. Dunque, l'obiettivo è quello di osservare la relazione causa-effetto della malattia.

1.1 Caratteri distintivi

I caratteri distintivi di uno studio clinico (o biomedico) sono i seguenti:

- i ragionamenti,i metodi e le conclusioni devono essere basati sul confronto
- le conclusioni tratte dal campione devono essere estese alla popolazione (inferenza statistica), sulla base di un modello statistico-probabilistico
- le caratteristiche dell'esperimento devono essere ben dettagliate e documentate nella fase antecedente allo studio
- le conclusioni devono essere basate sul confronto di gruppi "omogenei"

Un buon studio clinico deve porre una domanda importante e condivisa e rispondervi in modo affidabile.
Dall'obiettivo primario - il quale costituisce l'elemento più importante dello studio - discendono il disegno, l'endpoint (punto di osservazione finale dello studio), la selezione dei malati, i trattamenti, il periodo di follow-up (periodo di osservazione), analisti statistica e interpretazione.

1.2 Tassonomia

Per definire una tassonomia negli studio della ricerca clinica, verrà utilizzato uno dei paper di maggior rilievo in questo ambito: *An overview of clinical research: the lay of the land - Lancet (2002)*. Secondo il testo appena citato, la ricerca clinica è suddivisa in due grandi regni: studi sperimentali e studi osservazionali.

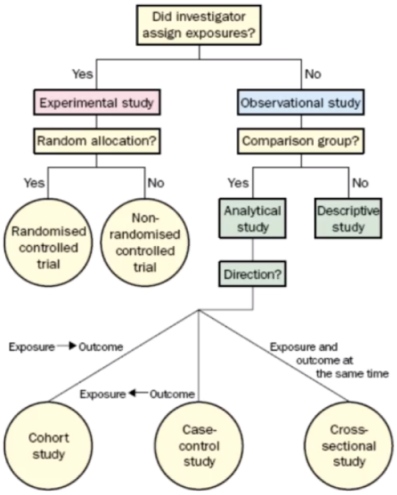

Figure 1: Studi osservazionali e sperimentali

Dalla figura 1 è possibile osservare le due diverse tipologie di studio: in particolare, la domanda iniziale è "il ricercatore ha assegnato l'esposizione?": tale domanda si riferisce al trattamento

terapeutico o più in generale un fattore che condiziona un effetto. Se la risposta è **sì**, allora lo studio è sperimentale e ci si chiede se l'esperimento è stato eseguito in modo casuale o meno. Se la risposta è **no**, allora lo studio è osservazionale e ci si chiede se esiste o meno la possibilità di confronto tra gruppi: se i gruppi sono confrontabili allora si parla di studio analitico in cui viene valutata l'associazione tra esposizione clinical outcome: può essere studio di coorte (si osserva dall'esposizione all'outcome), studio case-control (si osserva dall'outcome all'esposizione) oppure studio cross-sectional (si osserva l'outcome e l'esposizione sulla stessa timeline); viceversa, si parla di studio descrittivo.

Lo **studio sperimentale** implica la modifica (rispetto alla normale pratica clinica) del trattamento per studiarne l'effetto sull'esito. É condotto in condizioni controllate e più includere esperimenti casuali. Tale studio implica la manipolazione dell'esposizione (una o più variabili indipendenti) per studiarne l'effetto sull'esito (una o più variabili dipendenti).

Lo **studio osservazionale** implica che la decisione di prescrivere il farmaco al singolo paziente deve essere del tutto indipendente da quella di includere il paziente dello studio (AIFA). Quindi, in questo caso il ricercatore osserva l'effetto di un farmaco o un trattamento, senza modificarne la normale pratica clinica. In tale studio si studiano le relazioni tra le differenze in una o più variabili senza modificare l'esposizione.

1.3 Studi osservazionali

1.3.1 Studi descrittivi

Questo tipo di studio - di tipo qualitativo - fornisce le informazioni iniziali in nuove aree come:

- la frequenza di determinate caratteristiche/fattori
- distribuzione geografica, temporale e demografica
- possibili determinanti di una condizione
- sorveglianza sanitaria

Per quanto riguarda i risultati, si ottengono:

- la distribuzione e le caratteristiche di una patologia e/o dei soggetti affetti
- ipotesi di studio sull'ezologia (causa delle malattie)

1.3.2 Studi analitici

Questo tipo di studio fornisce la relazione tra esposizione e esito. Possono essere studi:

- **epidemiologici**: viene studiata la relazione tra intensità d esposizione a fattori di rischio/protettivi e frequenza della malattia
- **epidemiologici clinici**: viene valutato un impatto di un trattamento/procedura nella pratica clinica, fattori prognostici, effetti a lungo termine (farmacovigilanza)

La sostanziale differenza tra i due tipi di studio è il diverso legame dell'osservazione col tempo. Gli studi analitici possono essere:

- **cross-sectional**: viene osservata l'associazione di una esposizione con un esito misurati contemporaneamente. In questo tipo di studio non è facile interpretare la causa e l'effetto. Si basa sulla prevalenza, ossia prende in considerazione tutti i casi esistenti con determinate caratteristiche nel momento in cui sono osservati. In questo tipo di studi la relazione temporale non è chiara

- **caso-controllo**: si parte dall'esito e si procede in modo retrospettivo a misurare l'esposizione al fine di osservare una possibile associazione. Dunque, si basano sulla definizione di un gruppo con un outcome (detti casi) e di un gruppo senza outcome (detti controlli). Accertano la prevalenza d'esposizione ad un fattore di rischio di entrambi i gruppi. Se la prevalenza dell'esposizione è alta, allora l'esposizione è associata ad un rischio aumentato dell'outcome. Risultano essere studi utili quando i clinical outcome sono rari o hanno bisogno di tempo per essere sviluppati, richiedono meno tempo, sforzo e denaro rispetto a quelli di coorte. I punti di debolezza riguardano la scelta del gruppo di controllo appropriato, sono affetti da recall bias (errore sistematico dovuto alle differenze in accuratezza del richiamo alla memoria degli eventi passati e non è possibile calcolare i tassi di incidenza, i rischi relativi o i rischi attribuibili. L'associazione è misurata attraverso l'**odds ratio**
- **coorte**: si parte dalla misura dell'esposizione e viene valutato in modo prospettivo l'esito. Si basano sull'identificazione di un gruppo con una data esposizione. Vengono seguiti gruppi esposti e non esposti per determinare l'outcome. Se i gruppi esposti sviluppano un'incidenza maggiore rispetto ai gruppi non esposti, allora l'esposizione è associata ad un rischio incrementato dell'outcome (per incidenza si intendono i nuovi casi che presentano un evento di interesse in un determinato periodo di tempo). Gli aspetti positivi di questo tipo di studio riguardano la bassa recall bias, la possibilità di calcolare i tassi di incidenza, i rischi relativi o i rischi attribuibili e risulta essere il migliore per l'interpretazione causale. I punti di debolezza riguardano la necessità di tempi molto lunghi per l'ottenimento dei risultati di un evento e risultano essere molto costosi

1.4 Studi sperimentali

Sono gli studi nei quali l'esposizione è considerata come trattamento. L'effetto di un trattamento (sperimentale) è sempre quantificato relativamente ad un trattamento di controllo (standard o placebo). Quindi, sono studi nei quali in assenza di un gruppo di controllo non è possibile valutare se vi è un'associazione tra trattamento ed esito. Sono studi prospettici controllati in cui si definisce un gruppo di controllo che può essere randomizzato oppure non randomizzato: in quest'ultimo caso si ricorre a controlli concomitanti e controlli storici.

1.4.1 Studi non randomizzati

I **controlli storici** si basano sulla presenza di una base storica sui soggetti non esposti (sui quali è possibile fare un controllo storico). La base storica serve come base studio per i soggetti non esposti e guida il ricercatore per la costruzione dell'esposizione. In questo tipo di controllo i risultati vengono confrontati con pazienti simili osservati in precedenza nello stesso centro e per tale ragione risulta essere un controllo rapido e poco costoso. Inoltre, è possibile integrare i dati storici da altre fonti (Letteratura e Banche Dati). L'ipotesi sui cui si basa questo tipo di controllo è che i soggetti di controllo siano sovrapponibili ai soggetti esposti: è un'ipotesi molto forte e non sempre vera. Tali studi sono soggetti ad *errori sistematici* poiché si possono avere delle variazioni dell'esito nel tempo come:

- variazioni nelle caratteristiche dei soggetti
- variazioni nei criteri di selezione dei soggetti
- variazioni nelle modalità di diagnosi e di assistenza dei pazienti
- variazioni nei criteri diagnostici
- variazioni nella qualità dei dati

I **controlli paralleli** sono caratterizzati dall'assenza di una base storica. In questo caso il controllo avviene in parallelo tra soggetti esposti e soggetti non esposti. Tali studi non sono condizionati dal tempo, ma presentano ugualmente dei limiti come:

- bias di indicazione (sull'esposizione)
- preferenze del medico o del paziente
- differenti modalità di assistenza

Sia per gli studi a controllo storico che a controllo parallelo, il contrasto fra i gruppi confrontati stima l'effetto atteso dei trattamenti nei futuri malati.

1.4.2 Studi randomizzati

Gli studi randomizzati risultano essere i più accurati: si parte da una base di studio di pazienti non trattati ed il trattamento avviene in modo casuale sia sui pazienti trattati che quelli non trattati. Per questo tutti i pazienti hanno la stessa probabilità di ricevere uno dei trattamenti studiati (partendo dalla stessa base studio). I controlli sono per stesso disegno di studio concomitanti (paralleli).
La *randomizzazione* ha come effetto la ripartizione fra i gruppi i fattori prognostici (noti e ignoti) ed elimina gli errori sistematici nell'assegnamento dei trattamenti ai malati (consapevoli o inconsapevoli). Inoltre, risulta essere il modo più eticamente accettabile per assegnare ai malati i trattamenti confrontati ed i risultati sono più credibili. Infine, garantisce la validità dei test statistici.
Grazie a questo tipo di studio, è possibile evitare bias, effetti confondenti e riduce la informazioni derivanti dai bias.

1.5 Gerarchia dell'evidenza clinica

É possibile definire una gerarchia dell'evidenza clinica: partendo dal basso si hanno la **ricerca in vitro** e la **ricerca animale** che sono importanti per far nascere **opinioni e considerazioni scientifiche** sulla possibilità di far nascere nuovi farmaci e trattamenti. Proseguendo, si ha lo sviluppo e l'implementazione di **case report**, lo **studio caso controllo**, lo **studio di coorte** ed, infine, i test randomizzati controllati.

1.6 Protocollo

Sia gli studi sperimentali che quelli osservazionali necessitano di un protocollo. Il **protocollo** rappresenta il piano di studio. Un buon piano di studio deve portare a conclusioni affidabili e riproducibili e deve avere le seguenti caratteristiche:

1. **validità**: l'effetto osservato in corrispondenza di un trattamento (o esposizione) deve poter essere attribuito, senza ambiguità, al trattamento stesso. Viene ottenuta attraverso la randomizzazione
2. **precisione**: in uno studio valido, l'effetto osservato e quello vero differiscono a causa della variabilità casuale. Lo studio è tanto più preciso quanto più riesce a limitare tale variabilità. Viene ottenuta attraverso la numerosità del campione selezionato
3. **applicabilità**: lo studio deve garantire la generalizzazione dei risultati. É legata alla popolazione di riferimento degli studi e alle procedure (criteri di eleggibilità, modalità di applicazione del protocollo, etc.)

1.7 Processo ideale di un progetto clinico

Il processo ideale di un progetto clinico, che parte da buoni presupposti sperimentali in vitro e negli animali, è legato alla **farmacocinetica** e alla **farmacodinamica** e riguarda tre fasi:

1. **Studi di Fase I**: si pongono l'obiettivo di ottenre il profilo di sicurezza del prodotto nel volontario sano

2. **Studi di Fase II**: si pongono l'obiettivo di ottenere il profilo di sicurezza nel prodotto e nel paziente, valutare l'attività del prodotto nelle condizioni cliniche ideali (condizioni sperimentali controllate) e definire gli "ingredienti" di base per la fase III (popolazione target, dose, dimensione campionaria)
3. **Studi di Fase III**: si pongono l'obiettivo di confermare l'efficacia del prodotto nelle condizioni cliniche reali in cui l'ipotesi da sottoporre a verifica è espressa in termini quantitativi e le condizioni sperimentali definiranno le condizioni in cui il prodotto è efficace (saranno riportate nel foglietto illustrativo se il prodotto sarà accettato). Gli studi di fase III sono i più importanti e devono essere almeno due, indipendenti e riproducibili

2 Misure di effetto

2.1 Endpoint Binari - Variabili Dicotomiche

Le misure di effetto con endpoint binari hanno come obiettivo l'apprendimento degli indicatori fondamentali utilizzati della ricerca biomedica per descrivere l'effetto di un trattamento/esposizione (rispetto ad un gruppo di controllo) quando l'outcome dello studio sia rappresentato da un evento dicotomico. Ad esempio, siamo interessati a capire la mortalità (outcome dicotomica) attraverso la valutazione di un trattamento/esposizione A, rispetto ad un altro trattamento/esposizione B. Per farlo è necessario definire le misure ad effetto più comunemente utilizzate:

- misure basate sul **rischio**: differenza tra i rischi (riduzione assoluta di rischio), rischio relativo e il number needed to treat
- misure basate sul **odds ratio**

Definiamo il *rischio* come la probabilità che si verifichi un evento nel seguente modo (è una proporzione):

$$P(y_i|G) = p_G = \frac{casi\_favorevoli}{casi\_possibili}$$

dove y_i con $i \in \{0, 1\}$ e p_G è la probabilità dell'evento nel gruppo G (G è il trattamento considerato - in questo caso A o B). Nel nostro esempio, la probabilità di morte nel gruppo A sarà $P(y_0 = morto|A) = p_A$.
Definiamo l' **odds** come il rapporto tra la probabilità che si verifichi un evento nel seguente e la probabilità che non si verifichi:

$$ODDS = \frac{P(y_0|G)}{P(y_1|G)} = \frac{P(y_0|G)}{1 - P(y_0|G)}$$

Ora, è possibile andare a definire le misure d'effetto. La **differenza tra i rischi** è definita come la differenza di due probabilità:

$$\Delta = p_B - p_A$$

ossia, siamo interessati a vedere l'effetto di B rispetto a A. Nel nostro caso, se $\Delta \geq 0$ allora la probabilità di morte nel gruppo B è maggiore del $\Delta_{\%}$ rispetto ad A.
Da qui è possibile definire il **Number Needed to Treat**: è il numero di pazienti che devo trattare con A invece che con B per evitare un evento. In formule:

$$NNT = \frac{1}{p_B - p_A} = \frac{1}{\Delta}$$

L'idea è che se la differenza tra il trattamento A e il trattamento B è grande, allora il numero di pazienti da trattare con A (invece che con B) sarà più basso. Ogni NNT pazienti, risparmio un evento.
Ora, definiamo il **rischio relativo** nel seguente modo:

$$RR = \frac{p_A}{p_B}$$

ossia siamo interessati a capire il rischio di evento se vengo trattato con A, rispetto al rischio di evento ad essere trattato con B. Ricordando che p_A è la probabilità di morte col trattamento A e analogamente per p_B, se l'$RR > 0$ allora sarò favorito dal trattamento A. In modo complementare, viene definita la **Riduzione del Rischio Relativo** nel seguente modo:

$$RRR = 1 - RR$$

ossia di quanto si riduce il rischio di morire (in questo caso).
Infine viene definito l'**odds ratio** tra due trattamenti A e B nel seguente modo:

$$OR = \frac{\frac{p_A}{1-p_A}}{\frac{p_B}{1-p_B}}$$

dove se $OR < 1$, allora ho meno probabilità di morire col trattamento A rispetto al trattamento B.

2.2 Differenza tra misura assoluta e relativa

La **misura assoluta** ha un valore che assume rilevanza diversa a seconda dell'entità del fenomeno. Ossia, se prendiamo un Δ tra due trattamenti, esso avrà un rilevanza diversa a seconda dei casi (contesto). Per questo motivo vengono utilizzate le **misure relative**, le quali hanno la sensibilità di mostrare un valore rapportato al contesto. Per questo motivo le RR (o OR) sono preferibili per esprimere l'effetto del trattamento, mentre quelle assolute per esprimere il suo impatto clinico.
Per presentare ed interpretare i dati, non serve il solo valore relativo dell'effetto, ma anche il valore ottenuto nei singoli gruppi (o perlomeno nel gruppo di controllo). Banalmente, a parità di rischio relativo è necessario andare ad osservare anche le altre misure per poter trarre delle conclusioni.

É possibile notare che i valori di RR e OR tendono a coincidere (figura 2) solo se l'evento è **raro**, ossia per valori molto piccoli di p_A e p_B:

$$OR = \frac{\frac{p_A}{1-p_A}}{\frac{p_B}{1-p_B}} = \frac{p_A}{p_B} \cdot \frac{1-p_B}{1-p_A} = RR \cdot \frac{1-p_B}{1-p_A}$$

	p_A	p_B	Δ	RR	OR
1	0.01	0.02	0.01	**0.5**	0.495
2	0.02	0.04	0.02	**0.5**	0.490
3	0.10	0.20	0.10	**0.5**	0.444
4	0.20	0.40	0.20	**0.5**	0.375
5	0.40	0.80	0.40	**0.5**	0.167

Figure 2: Tabella con RR e OR

Quindi se $\frac{1-p_B}{1-p_A} \to 1$ allora $OR \cong RR$. L'OR è uno stimatore di rischio relativo distorto che tende ad enfatizzare l'effetto dell'esposizione. Come regola empirica, è possibile interpretare l'OR come se fosse un RR quando i rischi non sono superiori al 5% (figura 3).

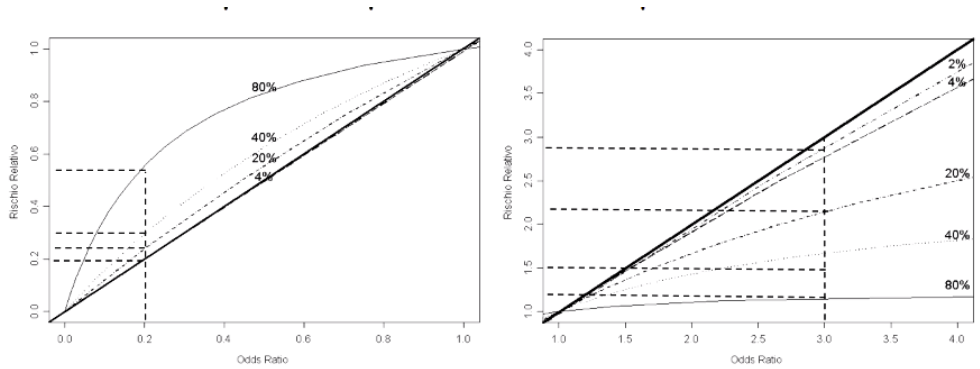

Figure 3: Rischio Relativo - Odds Ratio

Dati due trattamenti A e B è possibile affermare che:

- se A e B sono equivalenti, allora $OR \to 1$ e $RR \to 1$
- se A è peggiore di B, allora $OR > 1$ e $RR > 1$
- se A è migliore di B, allora $OR < 1$ e $RR < 1$

Quindi, OR e RR assumo risultati diversi, hanno significati diversi, ma procedono verso la stessa direzione. In generale, però, viene preferito l'OR siccome ha migliori proprietà matematico/statistiche, può essere modellato in funzione di variabili esplicative (come il modello logistico) e può essere utilizzato indipendentemente dalla natura dello studio (ad esempio l'RR non può essere utilizzato nel caso di studi caso-controllo, poiché i casi e i controlli sono fissati dallo sperimentatore e viene perciò valutata l'esposizione e non l'evento - studio retrospettivo). Come contro, l'OR risulta essere meno interpretabile dell'RR.

2.3 Precisione delle stime

Data la presenza di variabilità casuale e di quella campionaria, a tutte queste misure è associato un "indice di incertezza" chiamato **errore standard** (standard deviation - SD), il quale dipende dalla numerosità delle osservazioni. Da questo valore, vengono definiti gli **intervalli di confidenza** ad un livello di significatività α.

Nel caso delle **differenze tra rischi** Δ si avrà $\delta \pm Z_{1-\frac{\alpha}{2}} \cdot SD$, dove:

$$SD = \sqrt{\frac{p_B(1-p_B)}{n_B} + \frac{p_A(1-p_A)}{n_A}}$$

dove A e B sono i due trattamenti, p_i, $i \in \{A, B\}$ le rispettive probabilità e n_i le rispettive numerosità.

Nel caso del **Number Needed to Treat** gli intervalli di confidenza non forniscono delle informazioni vermente interessanti.

Nel caso del **Rischio Relativo**, poiché $log(RR) \sim N$, si ha $log(RR) \pm Z_{1-\frac{\alpha}{2}} \cdot SD_{log(RR)}$.

Equivalentemente, nel caso del **Odds Ration**, poiché $log(OR) \sim N$, si ha $log(OR) \pm Z_{1-\frac{\alpha}{2}} \cdot SD_{log(OR)}$.

Tali intervalli di confidenza permettono di valutare la differenza statisticamente significativa tra due trattamenti (ad esempio). Per verificare tale differenza vengono utilizzati dei test statistici quali:

- test sulle proporzioni
- testi *chi-quadro* (associazione)
- test esatto di *Fisher* (quando il valore atteso delle frequenze è ≤ 5)

2.4 Curva di sopravvivenza

Le **curve di sopravvivenza** rientrano nelle misure d'effetto nel tempo. La curva di sopravvivenza $S(t)$ descrive la probabilità di sopravvivere nel tempo a partire da una determinata origine, definita come $S(t) = P(T > t)$. É una misura di rischio nel tempo.

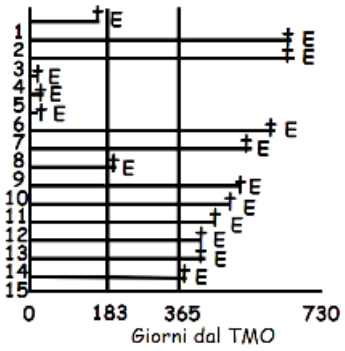

Figure 4: Grafico di sopravvivenza per ogni paziente ad ogni istante t_i

Come punto di partenza supponiamo che non ci siano dati censurati (tutti gli N soggetti hanno l'evento al tempo T): ad ogni tempo t, la probabilità di sopravvivenza è descritta dalla stima percentuale dei soggetti ancora vivi al tempo t. Dunque, osservati gli N soggetti con un determinato evento al tempo T (100% dei soggetti hanno l'evento), viene osservata la porzione di soggetti $n < N$ al tempo $t < T$ che non hanno ancora avuto l'evento. Allora, la sopravvivenza a t mesi sarà data da $S(t) = n/N$. Dal grafico in figura 4, vengono conteggiati i n pazienti che al tempo t non hanno ancora avuto l'evento. Ad esempio, prima del tempo $t = 183$, i pazienti sono $n = 11 (= 15 - 4)$. Quindi, la sopravvivenza a 183 giorni è $S(183) = 11/15$.
Come step successivo, supponiamo che vi siano dati censurati. Partendo dall'esempio precedente, supponiamo che il paziente 1, abbia il dato censurato (figura 5): è possibile ancora stimare che la sopravvivenza a 183 giorni sia pari a $S(183) = 11/15$? Se è vero che sono 11 i soggetti ancora vivi, non è vero però che i rimanenti 4 sono morti. Dunque, vi è una sottostima della sopravvivenza.

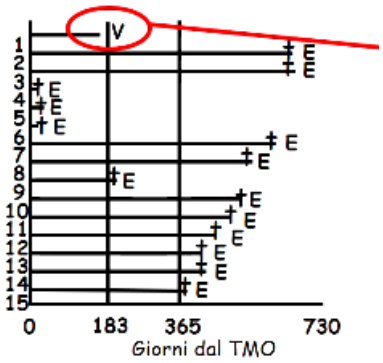

Figure 5: Grafico di sopravvivenza - dato censurato

Per stimare la probabilità di sopravvivenza si deve tenere conto:

- degli eventi (chi ha l'evento) e quando si verificano
- delle durate di osservazione dei soggetti diverse (poiché entrano nello studio in momenti differenti)
- dei dati censurati

Tenere conto dei dati censurati (che sono come se fossero dati mancanti) fornisce un'informazione aggiuntiva per la stima della curva di sopravvivenza. Ma in che modo? Attraverso lo **stimatore di Kaplan-Meier**.

2.4.1 Stimatore di Kaplan-Meier

Lo stimatore di Kaplan-Meier, noto anche come stimatore del prodotto limite, è uno stimatore che si usa per stimare la funzione di sopravvivenza di dati relativi alla durata di vita. Nella ricerca medica, si usa spesso per misurare la frazione di pazienti che vivono per una certa quantità di tempo dopo il trattamento.

Lo stimatore di Kaplan–Meier è la stima non parametrica della massima verosimiglianza di $S(t)$. É un prodotto con la forma:

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

dove n_i è il numero di pazienti vivi (o a rischio) al tempo t_i e d_i sono il numero di decessi (eventi) al tempo t_i. In altre parole, si stima la **probabilità condizionata di sopravvivenza**, in cui si stima la probabilità di sopravvivere ogni singolo giorno, dato che si è vissuti fino alla fine del giorno precedente. In questo modo, vengono tenuti in considerazione i dati censurati (per esempio se escono dallo studio per qualche ragione).

La stima di Kaplan–Meier delle probabilità di sopravvivenza viene rappresentata con un curva a gradini che ha come valore iniziale 1 (al tempo t_0 tutti i pazienti sono vivi per definizione), decresce nel tempo, cambia valore in corrispondenza di un evento e l'altezza dei gradini dipende sia dal numero di eventi che dal numero di soggetti a rischio (quindi anche dal numero dei dati censurati).

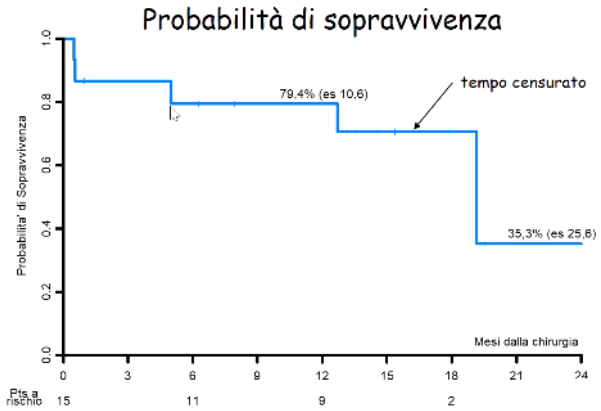

Figure 6: Curva di sopravvivenza

Sul diagramma in figura 6, piccoli segni di spunta verticali indicano le perdite, dove il tempo di sopravvivenza di un paziente è stato censurato a destra. É possibile notare che sotto l'asse

orizzontale, vengono conteggiati il numero di soggetti a rischio (ogni 6 mesi ad esempio), depurati dagli eventi e dati censurati.

Lo stimatore di Kaplan–Meier è un indicatore statistico, e vari stimatori sono utilizzati per approssimare la sua varianza. Uno dei più comuni di tali stimatori è la **formula di Greenwood**:

$$\hat{Var}(\hat{S(t)}) = \hat{S(t)}^2 \sum_{t_i<T} \frac{d_i}{n_i(n_i - d_i)}$$

La precisione della stima della probabilità di sopravvivenza è indicata dal suo **errore standard** $SD_{\hat{S(t)}} = \sqrt{\hat{Var}(\hat{S(t)})}$ che è inversamente proporzionale al numero di soggetti a rischio al tempo t_i. Pertanto, l'intervallo di confidenza al 95% per la stima della sopravvivenza al tempo t_i è pari a:

$$\hat{S(t_i)} \pm 1.96 \cdot SD_{\hat{S(t)}}$$

É buona norma riportare sulla curva i valori della stima a punti rilevanti (giorno, mese o anno), dell'errore standard o gli intervalli di confidenza ed il numero di soggetti a rischio.

Quando viene utilizzato lo stimatore di KM, vi sono degli aspetti molto importanti da considerare.

1. **Definizione dell'origine.** Influenza il numero di soggetti a rischio e degli eventi. É diversa a seconda della natura degli studi: studi randomizzati (randomizzazione) e o studi non randomizzati (diagnosi, inizio del trattamento, etc.)

2. **Definizione dell'evento di interesse (end-point).** Lo stimatore di Kaplan-Meier è in rado di considerare un solo evento su ogni soggetto. Se vi fossero più eventi di interesse, lo stimatore di K-M può essere usato solo per descrivere il primo degli eventi che si verifica. Se si fosse interessati a valutare il fallimento del trattamento sia in termini di ricaduta di malattia che morte in remissione completa (RC). Il metodo K-M potrebbe essere utilizzato per stimare la "sopravvivenza libera da malattia" (*DFS - Disease Free Survival*), considerando il tempo al primo dei due eventi che si manifesta (ricaduta o RC).

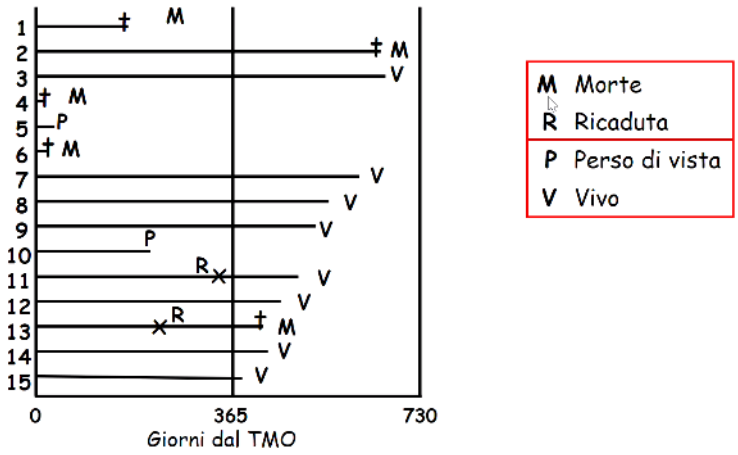

Figure 7: Grafico di sopravvivenza - Kaplan-Meier

In figura 7, è possibile osservare l'aggiunta di queste informazioni: in particolare, se è possibile conteggiare le ricadute con le morti (indipendentemente dall'evento finale) o analogamente escludere i vivi che hanno avuto una ricaduta.

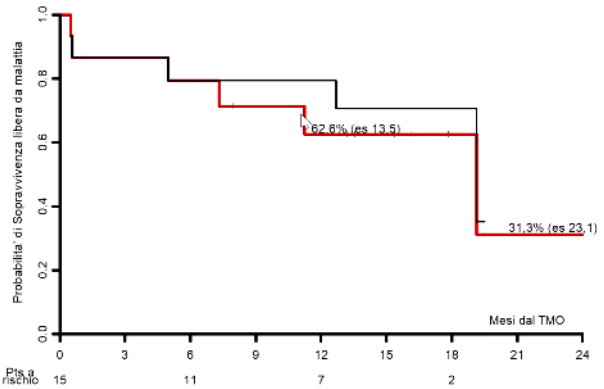

Figure 8: Curva di sopravvivenza con censura

In figura 8 si può osservare come cambia la curva di sopravvivenza (in rosso).

3. **Natura della censura.** Lo stimatore di K-M assume che la censura sia non-informativa: i soggetti censurati non devono "nascondere" un'esperienza diversa da quella degli altri. In altre parole, non si cerca di indagare la causa del dato censurato.

4. **Modalità di follow-up.** Indica una fase di controllo continuo o periodico e programmato (a seguito di un'azione o intervento). É importante che il follow-up sia sempre aggiornato. Può essere disponibile alla data limite (con annessa notifica di ricadute e morti) oppure può essere aggiornato in modo uniforme fino alla data limite. L'aggiornamento è importante al fine di non creare stime distorte.

Per una corretta lettura della curva di sopravvivenza è importante osservare la forma della curva più che i dettagli, è utile riferirsi ai valori della stima della probabilità di sopravvivenza a punti ciclicamente rilevanti ed è buona norma leggere la curva siano a quando vi sono almeno 10-20 soggetti a rischio (poiché chiaramente lo *standard deviation* sarà molto più alta con pochi soggetti).

Infine, viene anche utilizzata la stima del tempo di **sopravvivenza mediano**, ossia quel tempo t_{median} tale che vale la relazione $\hat{S}(t_{median}) = 0.5$. Può non esistere quando la curva di sopravvivenza non scende sotto lo 0.5.

Oltre alle stime, possono essere effettuati anche dei test d'ipotesi (parametrici e non-parametrici) basate sul confronto tra le curve.

3 Cox Regression Model

3.1 Time functions

A differenza del modello di Kaplman-Meier, in cui viene stimata la funzione di sopravvivenza $S(t) = P(T > t)$, nel **modello di regressione di Cox** viene stimato l'azzardo (*hazard function*). Partendo dalla funzione di sopravvivenza $S(t)$, la quale descrive la proporzione di individui liberi dall'evento (di sopravvivere) al tempo t, è possibile definire il suo complemento a 1 come **funzione di incidenza** $F(t) = 1 - S(t) = P(T \leq t)$: è la porzione di pazienti che hanno l'evento entro il tempo t. Queste due funzioni sono delle probabilità cumulative e ci descrivono una determinata situazione al tempo t. Dunque, viene definita la **funzione d'azzardo**, la quale può essere vista, a differenza delle precedenti, come una probabilità condizionata

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t | T > t)}{\Delta t}$$

In particolare, il numeratore ci dice qual è la probabilità di evento, dato che il paziente è vivo al tempo t (ossia non ha avuto l'evento), nella finestra immediatamente successiva al tempo t, quindi tra t e $t+\Delta t$. Dividendo il numeratore per Δt e applicando il limite per $\Delta t \to 0$, si ottiene il cosiddetto tasso istantaneo di evento (velocità istantanea). Per fare un esempio, tale funzione d'azzardo ci dice qual è la probabilità di evento dato che il paziente è ancora in vita alla fine del 5 anno, nel periodo immediatamente successivo; ossia, se $\Delta t = 1d$ allora sarà la probabilità di evento a 5 anni e un giorno. É possibile definirla anche come la probabilità di fallimento al tempo t, dato un rischio fino a t, ossia $\lambda(t)\Delta t$.

La funzione di sopravvivenza e quella d'azzardo sono legate tra loro. Definita la **funzione d'azzardo cumulato** come:

$$\Lambda(t) = \int_0^t \lambda(u)du$$

allora la relazione tra $S(t)$ e $\lambda(t)$ è data da:

$$S(t) = e^{-\Lambda(t)}$$

Per quanto riguarda gli andamenti delle curve nel tempo è possibile affermare che:

- la funzione di sopravvivenza è una funzione monotona non-crescente (o rimane costante o decresce)
- la funzione di incidenza è una funzione monotona non-decrescente (o rimane costante o cresce)
- la funzione d'azzardo può assumere andamenti differenti, come descritto dalla figura 9, a seconda del tipo di situazione in cui ci troviamo

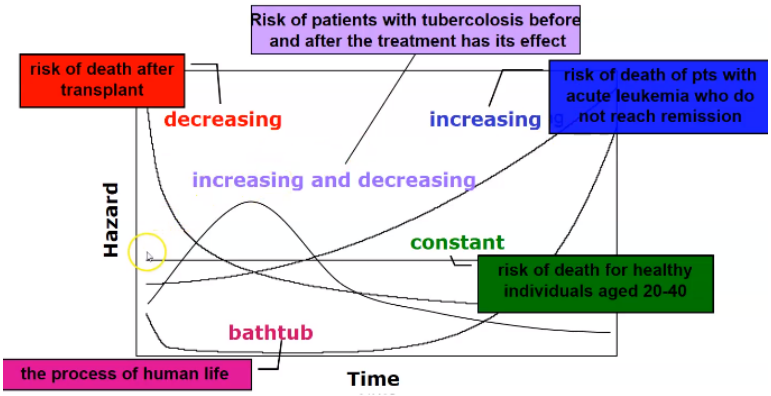

Figure 9: Hazard function

3.2 Stimatore non-parametrico dell'azzardo cumulato

Lo stimatore non-parametrico dell'azzardo cumulato è chiamato **stimatore di Aalen-Nelson** ed è definito nel seguente modo:

$$\hat{\Lambda}(t) = \sum_{j|t_j \leq t} \frac{d_j}{n_j}$$

dove d_j e n_j sono rispettivamente il numero di eventi ed il numero di soggetti a rischio al tempo j. Sarebbe possibile avere una stima puntuale in ogni istante dt attraverso $\lambda(t_j) = \frac{d_j}{t_j}$, che però è rappresentata da un andamento molto irregolare, per questo viene utilizzata la sua somma cumulata. La stima attraverso lo stimatore di Aalen-Nelson ci fornisce una funzione a gradini

monotona non-decrescente, dove il gradino rappresenta l'evento.
Per non cadere in errore, è necessario far notare che l'azzardo cumulato $\Lambda(t)$ (stimato attraverso A-N) non è equivalente alla probabilità cumulata di morte - incidenza - $F(t)$ (stimata attraverso K-M):

$$\Lambda(t) = -log(S(t)) = -log(1 - F(t)) \neq F(t)$$

Anche se, per $F(t)$ molto piccolo (0.2), i valori di $\Lambda(t)$ e $F(t)$ sono molto simili (figura 10).

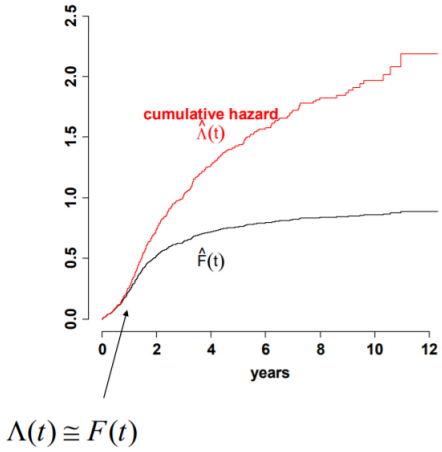

Figure 10: Confronto tra le curve $\hat{\Lambda}(t)$ e $\hat{F}(t)$

Da notare, però, che la funzione $F(t)$ varia su un intervallo $[0, 1]$, mentre $\Lambda(t)$, essendo una somma cumulata di probabilità varia su $[0, +\infty)$.

3.3 Modello di regressione di Cox

Il *modello di regressione di Cox* fa parte della famiglia dei cosiddetti *proportional hazard model* (PH), poiché ha come assunto di partenza molto forte che riguarda la proporzionalità degli azzardi. Per entrare nel più nel dettaglio, si prenda in considerazione un modello semplice con una sola variabile che descrive i pazienti malati di leucemia. La variabile si suppone essere booleana: assume valore 0 se i pazienti sono stati trattati col placebo, 1 se trattati col farmaco 6-MP. Pertanto, il modello sarà:

$$\lambda_X(t) = \lambda_0(t) \cdot e^{\beta X}$$

in cui viene descritto il paziente al tempo t, data una determinata caratteristica $X \in \{0, 1\}$. Il paziente avrà come base-line $\lambda_0(t)$, ossia con covariata $X = 0$ e dipende solo dal tempo t, moltiplicata per $e^{\beta X}$, il quale descrive l'azzardo di un soggetto che ha covariata pari a X e dipende solo da essa. Ovviamente, è possibile estendere il modello a più covariate.

Le pricipali caratteristiche del modello dei *PH models* sono:

- **Baseline hazard.** $\lambda_0(t)$ può essere non-parametrica (Cox) o parametrica ed è dipendente dal tempo t
- **Covariate effect.** $e^{\beta X}$ è parametrica, moltiplicativa ed è indipendente dal tempo t
- **Hazard ratio**. Il rapporto è costante nel tempo, infatti:

$$HR = \frac{\lambda_{X=x+m}(t)}{\lambda_{X=x}(t)} = \frac{\lambda_0(t) \cdot e^{\beta(x+m)}}{\lambda_0(t) \cdot e^{\beta(x)}} = e^{m\beta}$$

dove se $\beta > 0$ si ottiene $HR > 1$ (fattore di rischio), mentre se $\beta < 0$ si ottiene $0 < HR < 1$ (fattore protettivo).

La proporzionalità degli azzardi (PH), è difficile da valutare in ogni istante t su una scala esponenziale, ma è più se si prende il logaritmo del modello. Infatti:

$$\log \lambda_X(t) = \log \lambda_0(t) + \beta X$$

la cui differenza ad ogni istante t tra $\lambda_{x+m}(t)$ e $\lambda_x(t)$ è pari a $m\beta$. (aggiungere slide 10).

Dal modello $_X(t)$ è possibile ottenere la funzione di azzardo cumulato, $\Lambda_X(t)$:

$$\Lambda_X(t) = \Lambda_0(t) \cdot e^{\beta X}$$

e, analogamente, la funzione di sopravvivenza, $S_x(t)$:

$$S_X(t) = S_0(t)^{e^{\beta X}}$$

le quali sono legate tra loro.

3.4 Stima dei parametri nel modello di Cox

La stima del parametro β, viene eseguita attraverso la funzione di *massima verosimiglianza parziale* (PL - *Partial Likelihood*), in cui vengono sfruttate le caratteristiche della PH. In questo modo la LP dipende solamente da β e non dall'azzardo iniziale $\lambda_0(t)$ (baseline). Quindi, siano N il numero di osservazioni totali, $x_1, \ldots, x_N$ le covariate e $R(t_j) = R_j = \{i : t_i \geq t_j\}$ il set di soggetti a rischio al tempo t_j, allora è possibile definire la probabilità di un soggetto con covariata x_j di avere l'evento al tempo t_j data $R(t_j)$, nel seguente modo:

$$PL(\beta|data) = \prod_{j=1}^{J} \frac{\lambda_{x_j}(t_j)}{\sum\limits_{i \in R_j} \lambda_{x_i}(t_j)} = \cdots = \prod_{j=1}^{J} \frac{e^{\beta' x_j}}{\sum\limits_{i \in R_j} e^{\beta' x_i}}$$

Tale stima permette di non dover ipotizzare la forma della curva d'azzardo a priori, sotto l'ipotesi di proporzionalità degli azzardi. Passando alla forma logaritmica LLP, si avrà:

$$LLP(\beta|data) = \sum_{j=1}^{J} \beta' x_j - \log \left[\sum_{i \in R_j} e^{\beta' x_i} \right]$$

La cui derivata k-esima $U_k(\beta)$ (*score function*) sarà:

$$U_k(\beta) = \frac{\partial LLP(\beta)}{\partial \beta_k} = \sum_{j=1}^{J} \left[x_k - \frac{\sum\limits_{i \in R_j} x_k e^{\beta' x_i}}{\sum\limits_{i \in R_j} e^{\beta' x_i}} \right]$$

dove x_k è il valore della covariata del soggetto che fallisce al tempo t_j, mentre il rapporto tra le somme non è altro che la media pesata delle covariate x_k sul set di soggetti a rischio R_j. Quindi, è possibile stimare la matrice varianza-covarinza di $\hat{\beta}$:

$$U_k(\beta) = 0 \rightarrow \hat{\beta} = \hat{\beta}_1, \ldots, \hat{\beta}_k$$

da cui è possibile fare inferenza statistica. I test statistici utilizzati sono il Likelihood ratio Test, Score Test e il Wald test.

(esempio slide 15)

Nel caso in cui si voglia stimare la funzione di sopravvivenza $S(t)$ per un soggetto con covariata X, si ha:

$$S_{X_i}(t) = S_0(t)^{e^{\beta' x_i}}$$

dove si ha una stima di $\hat{\beta} = \hat{\beta}_1, \ldots, \hat{\beta}_k$. In questo caso, però, è necessario stimare $S_0(t)$. ' possibile ottenere tale stima secondo il metodo proposto da Norman Breslow, in cui viene effettuata una stima non parametrica di $S_0(t)$:

$$\hat{S}_0(t) = \prod_{t_j \leq t} 1 - \frac{d_j}{\sum\limits_{i \in R_j} e^{\hat{\beta}' x_i}}$$

in cui al denominatore si ha il numero totale di soggetti che hanno fallito al tempo t_j e al denominatore il contributo pesato in cui viene valutato lo scenario in cui tutti i soggetti hanno $x_i = 0$. Si potrebbe usare anche lo stimatore K-M, ma in quel caso non si avrebbe il vantaggio di utilizzare la stima di $\hat{\beta}$.

In figura (pag 17) è possibile osservare le curve costruite con Cox e con K-M. La differenza fondamentale tra i due modelli è l'assunzione aggiuntiva del modello di Cox. Per cui è necessario andare a verificare se l'ipotesi di azzardi proporzionali è vera (sensata). Vi sono vari metodi per farlo, ma verrà presentato il "metodo grafico": viene preso il logaritmo della dell'azzardo cumulato $\log S_X(t)$ (e quindi $\log[-\log S_X(t)]$) per tutte le covariate X, stimata con K-M, e viene rappresentato graficamente (funzione a gradini). Se la differenza tra le curve (in modo verticale) è, in modo approssimativo, costante per ogni t, allora è possibile affermare che siamo in presenza di PH, ed è perciò possibile applicare il modello di Cox.