POLITECNICO DI TORINO

DIPARTIMENTO DI MATEMATICA GIUSEPPE LUIGI LAGRANGE

LM-44 - MODELLISTICA MATEMATICO-FISICA PER L' INGEGNERIA

Corso di Laurea Magistrale in Ingegneria Matematica

Tesi di Laurea Magistrale

TEST D'IPOTESI PER L'ANALISI DELLA SOPRAVVIVENZA IN PRESENZA DI RISCHI COMPETITIVI

Relatore: Prof. Roberto Fontana

Candidato: Dante Futia

ANNO ACCADEMICO 2017-2018

Indice

Ringraziamenti

Quando si affronta un percorso faticoso è importante guardarsi intorno e continuare a vedere fiducia nei propri confronti. Ringrazio mamma, papà e Giuseppe perché nei momenti di difficoltà, d'incertezza, di smarrimento, hanno comunque pensato che la cosa migliore fosse continuare a darmi fiducia.
Ringrazio immensamente il professor Roberto Fontana, perché il suo sostegno e il suo conforto sono stati cruciali per convertire le mie paure e le mie insicurezze in determinazione e grinta. Un ringraziamento speciale va al professor Preziosi, perché fin dal nostro primo incontro mi ha dimostrato (e ha sempre dimostrato ai miei colleghi studenti) una vicinanza e una presenza uniche, che vanno senz'altro oltre i doveri di un coordinatore di un corso di laurea magistrale. Ringrazio la professoressa Roberta Sirovich, per aver seguito con molta dedidizione la parte finale del mio primo percorso universitario e la professoressa Isabella Lami per avermi affidato un ruolo di responsabilità all'interno di un progetto accademico e per le sue parole di apprezzamento.
Ringrazio **SAS®** Italia, nella persona di Cinzia Gianfiori e l'Università degli Studi di Padova per il supporto tecnico che mi hanno fornito e Metis Ricerche srl nelle persone di Flavio Bonifacio e Veronica Baldisserri, per avermi offerto la prima vera opportunità professionale.
Ringrazio Cinzia, Roberto, Mauro, Manuela, Andrea, Rosa, Marisa e Gianni per avermi semplicemente trattato come un figlio, un nipote, un cugino, gioendo e soffrendo insieme a me.
Ringrazio Manuele perché nel corso di tre anni passati insieme mi ha sempre ricordato con la sua curiosità e la sua capacità di stimolarmi perché 8 anni fa io abbia deciso d'iscrivermi a matematica.
Ringrazio Luigi per aver rappresentato la mia àncora quando sono approdato

in un porto sconosciuto e per l'insegnamento più importante di tutti:
"STAI SENZA PENSIERI BIG D ".
Ringrazio Francesca, Eleonora, Pietro, Ilaria e Monica per avermi fatto vedere l'università e lo studio da una diversa prospettiva.
Ringrazio Alberto per aver vigilato su di me, mettendo il mio studio sempre davanti al suo.
Ringrazio Roberto per la sua presenza fraterna e Daniele, Marco ed Elias per avermi regalato nottate di risate e spensieratezza.
Ringrazio Davide per il suo aiuto concreto con la programmazione e Simone per avermi riportato a un amore che nei primi anni di università avevo trascurato.
Ringrazio Leonardo, Enrico e Irene per avermi accompagnato in un momento delicato come la fine di questo lungo percorso.
Infine ringrazio con tutto il mio amore Giulia Aurora per aver saputo amplificare tutti i sentimenti che ho vissuto nella mia vita, l'amore, la rabbia, l'egoismo, l'empatia, la gioia e la tristezza. Alla fine non sono diventato il futuro della matematica italiana come le avevo promesso, mi è bastato ritrovarmi una persona più forte e matura al suo fianco.

Introduzione

L'analisi della sopravvivenza costituisce un ramo molto vasto della statistica applicata, storicamente legato ad applicazioni in ambito biomedico. L'analisi "classica "fa riferimento allo studio del tempo che intercorre fino al verificarsi di un unico e specifico evento. Tuttavia nelle applicazioni legate all'analisi della sopravvivenza, ricorre spesso la necessità di studiare l'incidenza di un evento tenendo conto del verificarsi di altri possibili scenari che potrebbero influenzarne l'accadimento. È qui che avviene il passaggio dall'analisi tradizionale alla cornice dei rischi competitivi (l'idea è appunto che diversi eventi possono accadere e sono perciò tra loro in competizione). Come spesso però capita in tutti gli ambiti della modellistica matematica, quando si vuole passare a una descrizione più realistica del fenomeno sottoposto a studio, risulta necessario di pari passo variare la complessità del modello matematico sotteso a tale descrizione. All'interno di questo lavoro si è inteso osservare questa variazione di complessità, nel passaggio da analisi classica a rischi competitivi, rispetto a uno dei diversi strumenti statistici che vengono utilizzati in questo contesto: il test d'ipotesi. Esso infatti viene abitualmente adoperato in analisi della sopravvivenza. L'esempio più rappresentativo è costituito dal log-rank test, utilizzato per il confronto tra le curve di sopravvivenza di due o più gruppi sperimentali. All'interno di questa tesi si è cercato invece di studiare quali sono e che caratteristiche presentano i principali test d'ipotesi in presenza di rischi competitivi, e si è tentata l'implementazione di una nuova procedura statistica, basata sulla metodologia *Nonparametric combination* (NPC), in modo da progettare un nuovo test d'ipotesi confrontabile con quelli già presenti in letteratura.

Date queste premesse, la trattazione è organizzata nel seguente modo: nel

primo capitolo vi è un richiamo agli elementi di base dell'analisi della sopravvivenza classica, come punto di partenza per introdurre invece quali sono gli aspetti matematico-statistici che caratterizzano il passaggio alla situazione dei rischi competitivi. Vengono per questo motivo definiti e analizzati i due tipi di *hazard rate* utilizzati in questo contesto, e viene poi introdotta la grandezza principale corrispondente alla curva di sopravvivenza, ovvero la funzione d'incidenza cumulativa.
Il secondo capitolo è dedicato a una rapida rassegna dei principali test d'ipotesi impiegati in presenza di rischi competitivi, con un focus sui due metodi più diffusi: il test di Gray e il test di Pepe-Mori. Di questi due test viene altresì discussa l'implementazione in ambiente **SAS®**, software con il quale tutte le analisi di questo lavoro sono state svolte.
Nel tesrzo capitolo vengono presentati gli aspetti principali della metodologia NPC e contestualmente viene descritta e discussa l'implementazione di un algoritmo in grado di generare un nuovo test d'ipotesi. Tale algoritmo è basato sull'estensione di una procedura NPC già utilizzata recentemente nell'ambito dell'analisi della sopravvivenza tradizionale.
Il quarto capitolo riporta le caratteristiche e i risultati di uno studio di simulazione utilizzato per mettere a confronto i metodi descritti nei capitoli precedenti.
In appendice infine sono trascritte alcune parti di codice **SAS®**, utilizzato nell'implementazione della procedura NPC e dello studio di simulazione.

Capitolo 1

I rischi competitivi nell'analisi della sopravvivenza

Lo studio dei rischi competitivi è intrinsecamente legato al contesto teorico e metodologico dell'analisi della sopravvivenza in quanto, come verrà tra breve descritto, rappresenta un raffinamento dei modelli di base che caratterizzano questo settore dell'inferenza statistica. Per tale motivo la prima sezione di questo capitolo è dedicata a un rapido richiamo degli elementi costitutivi dell'analisi della sopravvivenza, poiché essi rappresentano il punto di partenza per introdurre gli aspetti probabilistico-statistici che caratterizzano i rischi competitivi. Per una trattazione completa dei metodi e modelli più ulizzati in analisi della sopravvivenza si rimanda ad esempio a Collett [1994] e Kalbfleisch and Prentice [2002] oppure a Kleinbaum and Klein [2010] per un approccio alla materia più applicativo e meno formale dal punto di vista matematico. Nella sezione 1.2 verranno in questo modo introdotte le ragioni pratiche che portano alla definizione dei rischi competitivi e quelle che sono le ripercussioni a livello probabilitistico e statistico che tale cambio di prospettiva porta con sè.

1.1 Elementi di base dell'analisi della sopravvivenza

1.1.1 Survivor function e hazard function

L'analisi della sopravvivenza rappresenta un insieme di modelli e metodologie che hanno come oggetto di studio il tempo che intercorre fino al verificarsi di un determinato evento; nel contesto biomedico si tratta spesso di decessi, ricadute, contrazione di malattie, ma numerosi casi di studio sono discussi in altre situazioni (esistono anche ambiti di studio legati al contesto industriale, in cui però piuttosto che di sopravvivenza si parla di teoria dell'affidabilità e l'impostazione metodologica risulta leggermente differente). Dato che come detto ci si focalizza sul tempo fino al verificarsi di un evento, s'indicherà con T la variabile aleatoria che rappresenta tale tempo e con t l'istante temporale preciso in cui il fenomeno studiato si verifica. La funzione di ripartizione associata a T (che all'interno del nostro contesto di lavoro chiameremo funzione d'incidenza cumulativa e che in presenza di rischi competitivi svolgerà un ruolo cruciale) sarà come di consueto definita da

$$F(t) = \mathbb{P}(T < t) = \int_0^t f(u)\, du,$$

con f densità di probabilità. La funzione di sopravvivenza $S(t)$ (che invece svolge un ruolo di primo piano nell'analisi classica) è definita come nient'altro che la probabilità che T sia maggiore o uguale a uno specifico t:

$$S(t) = \mathbb{P}(T \geq t) = 1 - F(t). \tag{1.1}$$

$S(t)$ rappresenta perciò la probabilità che un soggetto sperimenti l'evento in un tempo maggiore rispetto a t. Essa è inoltre una funzione non crescente e $S(0) = 1$. Un'altra quantità fondamentale è la *hazard function* $h(t)$:

$$h(t) = \lim_{\delta t -> 0} \left\{ \frac{\mathbb{P}(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}. \tag{1.2}$$

Da un punto di vista intuitivo, tale funzione si può interpretare come il potenziale istantaneo per unità di tempo affinché un individuo della popolazione

sperimenti l'evento sottoposto a studio, posto che tale individuo sia restato immune all'evento fino al tempo t [Kleinbaum and Klein, 2010]. $h(t)$ è inoltre una funzione non negativa e superiormente illimitata.
Per come sono definite, $S(t)$ e $h(t)$ consentono una descrizione completa di un modello di analisi della sopravvivenza, in quanto mentre $S(t)$ fornisce una misura diretta della sopravvivenza all'interno della popolazione studiata, la *hazard function* è in grado di identificare uno specifico modello parametrico (come ad esempio quello esponenziale o di Weibull). Inoltre i modelli matematici di sopravvivenza sono solitamente espressi in termini di $h(t)$ [Kleinbaum and Klein, 2010].
Dalla formula (1.2) è possibile ricavare una relazione che lega la *hazard function* $h(t)$ con la funzione di sopravvivenza. Facendo riferimento alla definizione di probabilità condizionata il numeratore della (1.2) si può riscrivere come

$$\frac{\mathbb{P}(t \leq T < t + \delta t)}{\mathbb{P}(T \geq t)} = \frac{F(t + \delta t) - F(t)}{S(t)}.$$

A questo punto la (1.2) si può riscrivere nel seguente modo:

$$h(t) = \lim_{\delta t -> 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)} = \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)},$$

da cui consegue

$$h(t) = -\frac{d}{dt}\{\log S(t)\}$$

che consente infine di ricavare la relazione

$$H(t) = -\log\, S(t), \tag{1.3}$$

dove

$$H(t) = \int_0^t h(u)\, du. \tag{1.4}$$

$H(t)$ prende il nome di *cumulative hazard*. Un'osservazione che tornerà utile successivamente è che dalle relazioni (1.3) e (1.1) risulta una corrispondenza 1 a 1 tra $h(t)$ e la funzione d'incidenza cumulativa $F(t)$.

1.1.2 Il censoring

Nell'ambito degli studi di analisi della sopravvivenza ci si trova spesso nella situazione per cui non è possibile stabilire con certezza, per alcuni soggetti,

l'istante preciso in cui essi hanno sperimentato l'evento: in questo caso si parla di *censoring* o dati censurati. Nello specifico ci si riferisce a una situazione di *censoring* destro quando il tempo di sopravvivenza reale è maggiore o uguale rispetto a quello osservato. Questo scenario si profila ad esempio se all'interno di uno studio clinico un paziente si trova a dover abbandonare prima di aver sperimentato l'evento in esame oppure se al termine dello studio l' evento non si è ancora verificato (situazione anche identificata come *censoring* amministrativo). In letteratura si fa riferimento anche al *censoring* sinistro e più in generale al *interval censoring* ma all'interno degli esempi che verranno presentati e rispetto alla teoria che si intende sviluppare, questo genere di situazioni non verrà menzionato; per alcuni esempi si rimanda a Kleinbaum and Klein [2010] e Collett [1994].
Il *censoring* rappresenta, come vederemo, un importantissimo fattore di cui tener conto nella stima delle grandezze principali e nel confronto tra due curve di sopravvivenza, anche e soprattutto nel contesto dei rischi competitivi. Un'assunzione che comporta notevoli conseguenze rispetto alla validità di molti modelli utilizzati in letteratura è quella di *indipendent censoring*: si tratta dell'ipotesi per cui il tempo di sopravvivenza effettivo di un soggetto risulta indipendente dalla causa che genera il *censoring* e quindi i soggetti che lo sperimentano in un determinato istante temporale, sono comunque da considerarsi rappresentativi degli altri individui sopravvissuti fino a quel momento [Collett, 1994]. La *hazard function* al tempo t per i soggetti non censurati deve perciò coincidere con quella degli individui censurati [Kleinbaum and Klein, 2010].
Un altro assunto molto importante è quello di *non-informative censoring* la cui ipotesi sottostante è che la distribuzione di probabilità di T non fornisce alcuna informazione sulla distribuzione della variabile che genera il *censoring*; tale assunzione è spesso giustificabile in presenza di *indipendent censoring* [Kleinbaum and Klein, 2010].

1.1.3 Stimatore di Kaplan-Meier

Uno dei metodi non parametrici di stima di $S(t)$ più utilizzati in assoluto è sicuramente lo stimatore di Kaplan and Meier [1958]. Per ricavarlo si

ottiene anzitutto dai dati il numero r di istanti in cui uno o più soggetti hanno sperimentato l'evento d'interesse e succcessivamente si suddivide l'asse temporale in r istanti ordinati $t_{(1)}, \ldots, t_{(r)}$ (d'ora in poi per comodità useremo l'espressione *failure time* al fine di indicare questi istanti di tempo). È da notare che in corrispondenza di un generico $t_{(j)}$ si possa avere anche più di un *failure*. Si ottengono in questa maniera intervalli del tipo $[t_0, t_{(1)}[, [t_{(1)}, t_{(2)}[, [t_{(2)}, t_{(3)}[, \ldots$, in cui l'estremo di sinistra è sempre un *failure time* (tranne nel primo intervallo, in cui t_0 è l'istante iniziale). A ciascuno degli n individui che costituiscono la popolazione iniziale è associato un *survival time* da t_1 a t_n, tenendo presente che alcune osservazioni però potrebbero essere censurate a destra (tipicamente nel dataset è presente una variabile binaria che consente di capire se il tempo osservato è un *failure time* oppure un'osservazione censurata). Per convenzione, se contemporaneamente a un *failure time* sopraggiunge una censura, si assume che l'osservazione sia stata censurata un istante dopo il verificarsi del *failure*. Il numero di individui che sono vivi (non hanno sperimentato l'evento) un attimo prima dell'istante $t_{(j)}$ (e che comprende anche quelli che in $t_{(j)}$ falliranno) è indicato con $n(t_{(j)})$ (con $j = 1, \ldots, r$), mentre il numero di soggetti che in tale istante di tempo falliranno si indica con $d(t_{(j)})$. Se si denota con δ un tempo infinitesimale, si otterrà che la probabilità di fallimento nell'intervallo $]t_{(j)} - \delta, t_{(j)}[$ sarà pari a $d(t_{(j)})/n(t_{(j)})$ e quindi la probabilità di sopravvivenza stimata sarà $(n(t_{(j)}) - d(t_{(j)}))/n(t_{(j)})$. Dato che nell'intervallo di tempo da $t_{(j)}$ a $t_{(j+1)} - \delta$ non avvengono *failure*, per δ che tende a 0, $(n(t_{(j)}) - d(t_{(j)}))/n(t_{(j)})$ diventa la stima della sopravvivenza nell'intervallo da $t_{(j)}$ a $t_{(j+1)}$. Assumendo inoltre che i *failure* avvengano indipendentemente l'uno dagli altri, si perviene alla stima di Kaplan-Meier per la funzione di sopravvivenza:

$$\widehat{S}(t) = \prod_{j=1}^{l} \left(\frac{n(t_{(j)}) - d(t_{(j)})}{n(t_{(j)})} \right), \tag{1.5}$$

per $t_{(l)} \leq t < t_{(l+1)}$, $l = 1, 2, \ldots r$, e $\widehat{S}(t) = 1$ per $t < t_{(1)}$. La stima di Kaplan-Meier è anche chiamata stima prodotto-limite per la funzione di sopravvivenza. Sfruttando il fatto che il numero di persone che sopravvivono durante l'intervallo di tempo che comincia in $t_{(j)}$ si distribuisce come una variabile aleatoria binomiale di parametri $n(t_{(j)})$ e $(n(t_{(j)}) - d(t_{(j)}))/n(t_{(j)})$ e

la formula per il calcolo della varianza della funzione di una variabile aleatoria [Collett, 1994], si giunge alla formula di Greenwood per la varianza della stima di Kaplan-Meier:

$$var\{\hat{S}(t)\} \approx [\hat{S}(t)]^2 \sum_{j=1}^{k} \frac{d(t_{(j)})}{n(t_{(j)})(n(t_{(j)}) - d(t_{(j)}))} \,. \tag{1.6}$$

Per tutti i dettagli relativi a come si ricava la (1.6) si rimanda a Collett [1994]. Assumendo che la *hazard function* sia costante tra due *failure times consecutivi* è anche possibile ottenere una stima di $h(t)$ di tipo Kaplan-Meier:

$$\widehat{h}(t) = \frac{d(t_{(j)})}{n(t_{(j)})(t_{(j+1)} - t_{(j)})}, \tag{1.7}$$

per $t_{(j)} \leq t < t_{(j+1)}$.

1.1.4 Confronto tra due curve di sopravvivenza

Dal momento che questa trattazione è focalizzata sui test d'ipotesi (nel contesto dei rischi competitivi) è importante comprendere quali siano i modelli più utilizzati per confrontare curve di sopravvivenza. Nella pratica è molto spesso importante mettere a confronto le curve di sopravvivenza di due popolazioni, per testare ad esempio l'efficacia di una nuova cura, rispetto a un medicinale precedentemente utilizzato oppure a un placebo. Il metodo piû semplice e immediato per visualizzare qualitativamente le differenze tra le stime di due *survivor function* consiste nell'esaminare un plot congiunto delle due curve di sopravvivenza stimate. Il passo successivo, come avviene classicamente in statistica, è domandarsi se le differenze riscontrate nel grafico, siano frutto del caso oppure se effettivamente è stata sperimentata una sopravvivenza significativamente diversa all'interno dei due gruppi.
Il test d'ipotesi maggiormente utilizzato all'interno di questo contesto è il log-rank test. Immaginando di considerare r *failure times* congiuntamente tra i due gruppi (che indicheremo con A e B) e mantenendo una notazione coerente con quanto presentato precedentemente, possiamo indicare con $d^{(A)}(t_{(j)})$ e $d^{(B)}(t_{(j)})$ gli individui che fallirranno nell'istante $t_{(j)}$, rispettivamente nel primo e nel secondo gruppo. Allo stesso modo, gli individui a rischio tra i due gruppi, nel medesimo istante, saranno $n^{(A)}(t_{(j)})$ e $n^{(B)}(t_{(j)})$. Complessivamente perciò,

all'istante $t_{(j)}$ ci saranno $d(t_{(j)}) = d^{(A)}(t_{(j)}) + d^{(B)}(t_{(j)})$ individui che falliranno e $n(t_{(j)}) = n^{(A)}(t_{(j)}) + n^{(B)}(t_{(j)})$ soggetti a rischio. A questo punto è possibile procedere con la stessa logica sottesa al test del chi-quadrato, immaginando cioè di mettere a confronto il numero di *failure* osservati tra i due gruppi con quello dei *failure* attesi nell'ipotesi nulla di ugual distribuzione tra le due curve di sopravvivenza. Focalizzando l'attenzione solo su $d^{(A)}(t_{(j)})$ (ciò è possibile perché la tabella di contingenza 2×2 che ha per colonne il numero di *failure* e quello di sopravvissuti al tempo $t_{(j)}$ e sulle righe i due gruppi presenta un solo grado di libertà fissando i marginali di riga e colonna) si può osservare che tale quantità può essere vista come una variabile aleatoria ipergeometrica che assume valori da 0 a $\min(d(t_{(j)}), n(t_{(j)}))$ e perciò la probabilità che all'istante $t_{(j)}$ il numero di *failure* nel gruppo 1 sia proprio $d^{(A)}(t_{(j)})$ risulta:

$$\frac{\binom{d(t_{(j)})}{d^{(A)}(t_{(j)})}\binom{n(t_{(j)})-d(t_{(j)})}{n^{(A)}(t_{(j)})-d^{(A)}(t_{(j)})}}{\binom{n(t_{(j)})}{d(t_{(j)})}}.$$

In questo modo si possono ottenere valore atteso e varianza di d_{1j}:

$$e_{Aj} = n^{(A)}(t_{(j)})d(t_{(j)})/n(t_{(j)}), \tag{1.8}$$

$$v_{Aj} = \frac{n^{(A)}(t_{(j)})n^{(B)}(t_{(j)})d(t_{(j)})(n(t_{(j)}) - d(t_{(j)}))}{n(t_{(j)})^2(n(t_{(j)}) - 1)}. \tag{1.9}$$

A questo punto, per avere una misura complessiva del discostamento tra il numero di *failure* osservati e quelli attesi sotto ipotesi nulla, sarà sufficiente effettuare una sommatoria su tutti i *failure time*:

$$U_L = \sum_{j=1}^{r}(d^{(A)}(t_{(j)}) - e_{Aj}) \tag{1.10}$$

e poiché i singoli *failure* sono indipendenti, la varianza di U_L sarà semplicemente

$$V_L = var(U_L) = \sum_{j=1}^{r} v_{Aj}. \tag{1.11}$$

Dato che per r sufficientemente grande U_L si distribuisce approssimativamente come una normale, si ottiene la statistica del test:

$$W_L = \frac{U_L^2}{V_L} \sim \chi_1^2. \tag{1.12}$$

Si può anche notare che rimaneggiando la statistica U_L si ottiene

$$\begin{aligned} U_L &= \sum_{j=1}^{r} \left(d^{(A)}(t_{(j)}) - n^{(A)}(t_{(j)}) \frac{d(t_{(j)})}{n(t_{(j)})} \right) \\ &= \sum_{j=1}^{r} n^{(A)}(t_{(j)}) \left(\frac{d^{(A)}(t_{(j)})}{n^{(A)}(t_{(j)})} - \frac{d(t_{(j)})}{n(t_{(j)})} \right) \\ &= \sum_{j=1}^{r} \frac{n^{(A)}(t_{(j)})\, n^{(B)}(t_{(j)})}{n^{(A)}(t_{(j)}) + n^{(B)}(t_{(j)})} \left(\frac{d^{(A)}(t_{(j)})}{n^{(A)}(t_{(j)})} - \frac{d^{(B)}(t_{(j)})}{n^{(B)}(t_{(j)})} \right), \end{aligned} \tag{1.13}$$

perciò U_L equivale a una differenza pesata tra le stime degli *hazard* dei due gruppi messi a confronto (quest'osservazione sarà molto utile quando, all'inizio del secondo capitolo, verrà discusso l'utilizzo del log-rank in presenza di rischi competitivi).
Esistono diverse varianti del log-rank, tutte basate sull'assegnazione di un peso alla differenza $d^{(A)}(t_{(j)}) - e_{Aj}$; tra le altre vale la pena ricordare quella che origina il test di Wilcoxon, basati sulla statistica:

$$U_W = \sum_{j=1}^{r} n(t_{(j)})(d^{(A)}(t_{(j)}) - e_{Aj}), \tag{1.14}$$

in cui la differenza tra *failure* osservati e attesi risulta pesata dal numero totale di soggetti a rischio in quell'istante. Ciò comporta che rispetto al log-rank test, la variante di Wilcoxon enfatizzi maggiormente le differenze nei primi istanti del *follow up*, quando ancora il campione complessivo risulta numeroso. Il log-rank appare più adatto rispetto al test di Wilcoxon quando l'ipotesi alternativa all'uguaglianza tra due popolazioni è di *hazard* proporzionali, mentre per altri tipi di discostamenti il test di Wilcoxon si comporta meglio.

1.1.5 Il modello di Cox per hazard proporzionali

Anche se all'interno di questo lavoro non faremo riferimento a modelli alle covariate, per la centralità che riveste all'interno dell'analisi della sopravvivenza, è bene fare un breve accenno al modello statistico non parametrico più utilizzato per l'inclusione di variabili esplicative al fine di spiegare la sopravvivenza sperimentata da una popolazione. L'obiettivo di molti studi clinici è infatti quello di individuare quali possano essere le caratteristiche

del soggetto (come età, sesso, livello di emoglobina, stili di vita, tipo di dieta, ecc...) che hanno un impatto notevole sulla sopravvivenza sperimentata. Dato che l'interesse principale sta nel monitorare costantemente il rischio che si verifichi l'evento studiato, è proprio un modello della *hazard function* che si vorrebbe ricavare. Quello a cui qui si desidera fare accenno è il cosiddetto modello di regressione di Cox per *hazard* proporzionali. Si indichi con $\mathbf{x}$ un vettore di p variabili $x_1, \ldots, x_p$. L'idea è che per ogni individuo i della popolazione, con $i = 1, 2, \ldots, n$, il proprio *hazard* può essere visto come il prodotto tra un termine che non coinvolge le covariate e che dipende dal tempo e un termine che invece dipende dai valori di $\mathbf{x}$ per il soggetto i:

$$h_i(t) = h_0(t)\psi(\mathbf{x}_i). \tag{1.15}$$

Il termine $h_0(t)$ prende il nome di *baseline hazard function* e su di esso non occorre fare particolari assunzioni[Collett, 1994]. La parte che coinvolge le variabili esplicative non prevede un legame di tipo temporale col rischio, perciò i valori di ciscuna variabile vengono raccolti all'inizio dello studio (il tema delle covariate dipendenti dal tempo costituisce un raffinamento del modello di partenza, ampiamente trattato in letteratura). ψ può essere altresì interpretato come il rischio al tempo t di fallire per un soggetto con valori delle variabili esplicative pari a $\mathbf{x}_i$, relativamente al rischio per un individuo tale che $\mathbf{x} = \mathbf{0}$[Collett, 1994]. Una scelta conveniente per la funzione $\psi(\mathbf{x}_i)$ può essere $\psi(\mathbf{x}_i) = \exp(\eta_i)$, con η_i combinazione lineare dei predittori e che prende appunto il nome di componente lineare del modello:

$$\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{1p}.$$

In questo modo la (1.15) si può esprimere come

$$h_i(t) = h_0(t)\exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{1p}), \tag{1.16}$$

da cui si deriva

$$\log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{1p}. \tag{1.17}$$

Il fitting di questo genere di modello avviene attraverso la stima dei coefficienti $\beta_1, \beta_2, \ldots, \beta_p$, massimizzando la seguente funzione di verosimiglianza:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^t \boldsymbol{x}_{(j)})}{\sum_{l\,\epsilon\, n_{(t_{(j)})}} \exp(\boldsymbol{\beta}^t \boldsymbol{x}_l)}, \tag{1.18}$$

in cui $\boldsymbol{\beta}$ è il vettore dei coefficienti, $\boldsymbol{x}_{(j)}$ è il vettore di variabili per il soggetto che fallisce in $t_{(j)}$ e la sommatoria a denominatore è eseguita su tutti i soggetti l a rischio in $t_{(j)}$. Per approfondimenti sulla stima, sugli intervalli di confidenza e i test di ipotesi per i coefficienti e per ulteriori precisazioni si rimanda a Collett [1994].

1.2 Definizione dei rischi competitivi

Come è stato ricordato all'inizio del capitolo, l'analisi della sopravvivenza si fonda sullo studio statistico del tempo che intercorre fino al verificarsi di un evento d'interesse. Volendo formalizzare con dei diagrammi di stato, ciò che si verifica è un passaggio da uno stato 0, caratterizzato dall'assenza di eventi, a uno stato in cui si è verificato l'evento sottoposto a studio. Lo schema in figura 1.1 esemplica bene la situazione. In molti contesti reali tuttavia

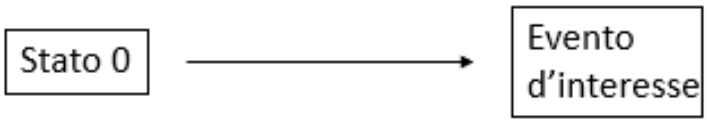

Figura 1.1: Struttura dell'analisi della sopravvivenza tradizionale

si ottiene un modello più funzionale all'interpretazione del fenomeno che si sta studiando se si tiene conto che a partire da uno stato iniziale si possano verificare 2 o più eventi diversi tra loro; è proprio all'interno di tale scenario che si definisce la presenza dei rischi competitivi. La situazione che rispecchia il framework appena introdotto è rappresentata dallo schema in figura 1.2. Nel corso dei capitoli successivi faremo per lo più riferimento a situazioni in cui gli eventi possibili saranno due, l'evento principale e un evento competitivo, sia perché si può sempre effettuare una generalizzazione a posteriori dei modelli statistici impiegati, sia perché se l'attenzione è rivolta ad un solo evento, tutti gli altri eventi possono essere collassati in un unico rischio competitivo. Riguardo allo schema appena introdotto, si può anche notare che in taluni scenari i possibili avvenimenti che possono accadere durante il *follow-up* non arrivino a verificarsi in maniera improvvisa o inaspettata, bensì che vengano

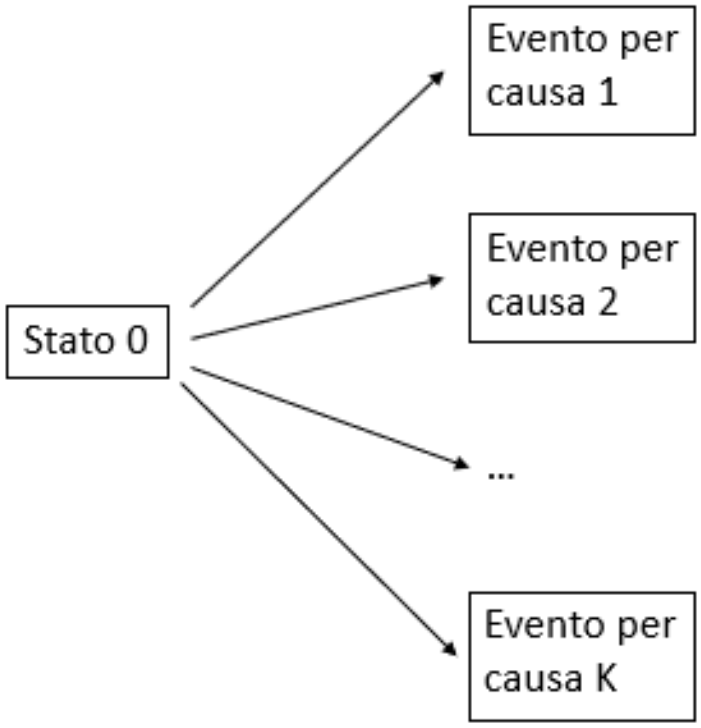

Figura 1.2: Struttura dell'analisi della sopravvivenza in presenza di rischi competitivi

anticipati da specifici stati intermedi: si parla in quel caso precisamente di modelli multistato, che non verranno qui presi in considerazione ma che sono ampiamente trattati in letteratura.
Uno degli obiettivi dell'analisi in presenza di rischi competitivi è quello di mantenere, come nel contesto classico, il focus su un unico evento d'interesse, tenendo tuttavia presente la possibilità che anche altri eventi si possano verificare interferendo con l'incidenza del fenomeno principale. Va altresì precisato che all'interno di tale configurazione solo uno dei diversi possibili eventi si può verificare e può farlo una volta soltanto [Kleinbaum and Klein, 2010], perché altrimenti ci si va a riferire alla cornice dei cosiddetti eventi ricorrenti, situazione in cui alcuni eventi possono verificarsi più volte durante il *follow-up* (si tratta di una situazione che capita diverse volte in ambito applicativo ma che non viene qui approfondita). Per chiarire meglio quanto

appena detto si pensi intuitivamente alla situazione in cui si sta effettuando uno studio di sopravvivenza rispetto all'insorgenza di malattie cardiache in una popolazione di persone anziane [Pintilie, 2006]: data l'età avanzata dei soggetti sottoposti a studio, alcuni di essi potrebbero morire per altre cause prima di sperimentare l'evento d'interesse, che in questo senso è come se risultasse ¨ostacolato˝dal verificarsi di eventi di altro tipo.
Ci si può domandare a questo punto se gli strumenti quantitativi introdotti nella precedente sezione possano essere utilizzati per descrivere efficacemente il framework di lavoro che viene adesso a delinearsi. Lo stimatore di Kaplan-Meier utilizzato in presenza di rischi competitivi, non dà più la possibilità di intepretare le stime come delle probabilità. In [Pintilie, 2006, p. 4] è riportato un esempio specifico: il dataset è costituito da 20 pazienti che possono sperimentare un infarto del miocardio oppure una morte per altre cause. Al termine del *follow up* (16 mesi), metà della popolazione ha sperimentato l'infarto. La stima di Kaplan-Meier della sopravvivenza all'infarto (considerando la morte per altre cause come censura) risuta pari a circa 0.16. Questo significa, per la definizione di sopravvivenza, che la probabilità che si verifichi l'infarto in un tempo minore o uguale ai 16 mesi risulta $1 - 0.16 = 0.84$. Eseguendo lo stesso calcolo in modo che la morte per altre cause sia l'evento d'interesse, si ottiene che $1 - \widehat{S}^{(KM)} = 1$. La somma tra 0.84 e 1 dovrebbe perciò rappresentare la probabilità che al sedicesimo mese uno qualunque tra i due eventi si sia verificato ma come si osserva immediatamente tale somma è superiore a 1, rendendo un'interpretazione probabilistica priva di senso.
Per quanto riguarda invece i test d'ipotesi, il log-rank test viene adoperato tenendo presente che il tipo di interpretazione che se ne ricava è molto specifico e non sempre utile ai fini pratici; questo aspetto, insieme ai principali metodi alternativi di confronto presenti in letteratura, sarà oggetto del prossimo capitolo. Nei prossimi paragrafi invece si introdurranno prima alcuni esempi a cui poi si farà riferimento nelle implementazioni successive e dopo il formalismo matematico che fa da cornice ai rischi competitivi, con un focus sugli aspetti che costituiscono il punto di partenza per introdurre il tema dei test d'ipotesi.

1.2.1 Alcuni esempi

Verranno ora presentati alcuni esempi tratti da situazioni reali, che verranno utilizzati a partire dal prossimo capitolo, per discutere e confrontare le varie tipologie di test d'ipotesi. L'obiettivo adesso è solo quello di cercare di capire quali possano essere le situazioni applicative principali in cui si ricorre a un modello che prevede la presenza di rischi competivi. Per quanto riguarda i dati, si tratta per lo più di sottoinsiemi dei dataset utilizzati negli studi originari.

Linfoma follicolare

Questo studio, trattato e analizzato in Pintilie [2006], fa riferimento a un sottoinsieme di 541 pazienti affetti da linfoma di tipo follicolare. Essi sono stati sottoposti a due trattamenti (sole radiazioni o radiazioni più chemioterapia) al fine di verificare gli effetti sul lungo termine. L'evento d'interesse in questo caso è il *failure* provocato dalla malattia (situazione che comprende sia i soggetti che hanno manifestato una ricaduta che gli individui i quali non hanno dato risposta ai trattamenti) mentre il rischio competitivo è rappresentato da cause diverse dal tumore (figura 1.3).

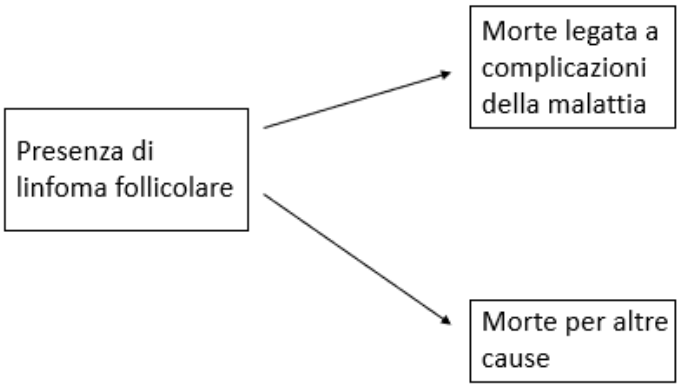

Figura 1.3: Schema relativo allo studio sul linfoma follicolare

Trapianto di midollo osseo

Questo dataset è citato e studiato in diverse fonti, ad esempio in Klein and Moeschberger [2005]. I dati fanno riferimento a 137 pazienti affetti da diversi tipi di leucemia e sottoposti a trapianto del midollo osseo. Un parametro per verificare i benifici di un trapianto di midollo è studiare il tempo che intercorre tra l'operazione e la ricaduta, l'evento di interesse di questo studio. Il rischio competitivo è rappresentato dalla morte nel periodo di latenza della malattia, in seguito al trapianto.

Malattie cardiovascolari

I dati [Hosmer et al., 2008] fanno riferimento a 453 pazienti con età media pari a 70 anni. L'evento principale è un qualunque episodio cardiovascolare (CVD), mentre il rischio competitivo è nuovamente rappresentato dalla morte per altre cause

1.2.2 Due tipi di hazard function

Un primo aspetto da prendere in considerazione nel passaggio dall'analisi della sopravvivenza classica all'introduzione dei rischi competitivi è che nel contesto tradizionale esiste una relazione 1 a 1 tra l'*hazard function* e l'incidenza cumulativa dell'evento studiato [Geskus, 2015]. Tale legame è espresso dalla formula (1.3) la quale è ovviamente invertibile e consente di trovare la funzione di sopravvivenza (e conseguentemente l'incidenza) in termini di $h(t)$. In presenza di rischi competitivi invece è possibile definire due tipi di *hazard function*: uno dei due non è in corrispondenza 1 a 1 con la funzione di incidenza cumulativa della tipologia di evento a cui fa riferimento, l'altro sì. Questi due modi di definire il rischio, conducono perciò a due tipi diversi d'interpretazione dei modelli matematici utilizzati.

Il cause-specific hazard

Indicando con (T, E) la coppia costituita dal tempo fino all'occorrenza di un evento e il tipo di evento verificatosi, il *cause-specific hazard*, indicato con

$\lambda_k(t)$ è così definito:

$$\lambda_k(t) = \lim_{\delta t->0} \frac{\mathbb{P}(t \leq T < t + \delta t, E = k | T \geq t)}{\delta t}, \quad k = 1, \ldots, K, \tag{1.19}$$

dove K è il numero di eventi possibili. Tale definizione è pressoché analoga a quella standard fornita da (1.2). Rappresenta di fatto la frazione di individui che sviluppano l'evento di interesse a un certo istante di tempo tra tutti i soggetti che a quell'istante non hanno ancora sperimentato alcun evento.
È possibile poi ottenere il rischio complessivo al tempo t sommando tutti i *cause-specific hazard*:

$$\begin{aligned} \Lambda(t) = \sum_{k=1}^{K} \lambda_k(t) &= \lim_{\delta t->0} \frac{\sum_{k=1}^{K} \mathbb{P}(t \leq T < t + \delta t, E = k | T \geq t)}{\delta t} \\ &= \lim_{\delta t->0} \left\{ \frac{\mathbb{P}(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}. \end{aligned} \tag{1.20}$$

In maniera analoga alla definizione (1.4) è possibile introdurre il *cumulative hazard* per l'evento di tipo k:

$$\Lambda_k(t) = \int_0^t \lambda_k(s)\, ds. \tag{1.21}$$

Dall'ultima relazione e invertendo la (1.3) è così possibile ricavare la funzione di sopravvivenza da un qualunque tipo di evento, detta anche *event-free survival*:

$$S(t) = \mathbb{P}(T > t) = \exp\left(- \sum_{k=1}^{K} \Lambda_k(t) \right) = \exp\left(- \sum_{k=1}^{K} \int_0^t \lambda_k(s)\, ds \right). \tag{1.22}$$

È da notare che si può definire la quantità $S_k(t) = \mathbb{P}(T > t, E = k) = \exp(-\Lambda_k(t))$, ma che quest'ultima non è interpretabile come una *survivor function* marginale [Geskus, 2015].
Presentate le quantità che gravitano intorno al *cause-specific hazard* è adesso possibile definire la funzione di incidenza cumulativa per la causa k:

$$\begin{aligned} F_k(t) = \mathbb{P}(T \leq t, E = k) &= \int_0^t \mathbb{P}(T > t) \lambda_k(s) ds \\ &= \int_0^t S(s) \lambda_k(s) ds. \end{aligned} \tag{1.23}$$

Tale relazione si può interpretare intuitivamente poiché la probabilità che l'evento k si verifichi in un determinato istante di tempo s, può essere vista come il prodotto tra la probabilità di non aver sperimentato eventi di alcun genere prima di s e la probabilità che all'istante s avvenga l'evento k posto che non si sia ancora verificato, quest'ultima rappresentata proprio da $\lambda_k(s)$ [Geskus, 2015]. Va osservato che $F_k(t)$ non è una distribuzione di probabilità vera e propria in quanto $\lim_{t->\infty} F_k(t) = \mathbb{P}(K = k) \leq 1$; prende infatti anche il nome di *subdistribution*. Vale inoltre che $F_k(t) + S_k(t) = \mathbb{P}(K = k)$ [Pintilie, 2006].
Dalla relazione (1.23) si evince che $F_k(t)$ risulta determinata da tutti i K *cause-specific hazard* (tramite $S(s)$ attraverso la formula (1.22)) e perciò come accennato precedentemente, non esiste una corrispondenza 1 a 1 tra il *cause-specific hazard* e la funzione d'incidenza cumulativa per la k-esima causa. Di conseguenza potrebbe capitare ad esempio che in un intervallo di tempo, per un sottogruppo, il *cause specific hazard* risulti più basso rispetto agli altri, per tutto l'intervallo, mentre la funzione di incidenza cumulativa risulti maggiore rispetto agli altri sottogruppi in almeno una parte dell'intervallo [Geskus, 2015].
Per quando riguarda la stima del *cause-specific hazard* bisogna tener presente anzitutto che un individuo il quale sperimenta un evento competitivo rispetto a quello in analisi, viene rimosso dall'insieme degli individui a rischio. Mantenendo perciò la notazione introdotta precedentemente (con l'unica differenza che d'ora in poi comprarirà il pedice k riferito alla tipologia di evento presa in considerazione), la stima di λ_k al j-esimo *failure time* risulterà:

$$\widehat{\lambda}_k(t_{(j)}) = \frac{d_k(t_{(j)})}{n(t_{(j)})} \tag{1.24}$$

Nell'interpretazione di questa stima si assume quindi che i soggetti che sperimentano un evento competitivo non siano da considerare come rappresentativi della popolazione che rimane all'interno del *follow-up* [Geskus, 2015].
Dalla (1.24) è possibile ricavare lo stimatore per il *cause-specific hazard* cumulativo al tempo t: $\widehat{\Lambda_k}(t) = \sum_{t_{(j)} \leq t} \widehat{\lambda}_k(t_{(j)})$, che prende il nome di stimatore di Nelson-Aalen.

Il subdistribution hazard

L'*hazard function* che risulta in corrispondenza 1 a 1 rispetto a $F_k(t)$ è il cosiddetto *subdistribution hazard* $h_k(t)$ introdotto da Gray [1988] (all'interno del medesimo lavoro è anche presentato un test d'ipotesi basato sul confronto tra le medie pesate delle stime di $h_k(t)$ tra 2 gruppi, il quale sarà ampiamente approfondito nel prossimo capitolo). Il *subdistribution hazard* risulta così definito:

$$h_k(t) = \lim_{\delta t->0} \frac{\mathbb{P}(t \leq T < t + \delta t, E = k | T \geq T \,\mathrm{o}\, (T \leq t \,\mathrm{e}\, E \neq k))}{\delta t}. \tag{1.25}$$

La relazione che lega $F_k(t)$ e $h_k(t)$ è

$$1 - F_k(t) = \exp\left\{ - \int_0^t h_k(s) ds \right\} \tag{1.26}$$

che si può leggere alternativamente come

$$h_k(t) = -\frac{d \log(1 - F_k(t))}{dt}, \tag{1.27}$$

da cui si ricava anche

$$h_k(t) = \frac{f_k(t)}{1 - F_k(t)}. \tag{1.28}$$

$h_k(t)$ è interpretabile come l'*hazard* relativo a una variabile aleatoria T' tale che

$$T' = \begin{cases} T & E = k \\ \infty & E \neq k \end{cases}$$

Come si può notare, la differenza principale rispetto al *cause-specific hazard* sta nel fatto che i soggetti che sperimentano un evento competitivo non vengono rimossi dal risk set fintantoché non subiscano *censoring*. A livello di stima ciò implica che rispetto alla (1.24) si modifichi il denominatore:

$$\widehat{h}_k(t_{(j)}) = \frac{d_k(t_{(j)})}{n^*(t_{(j)})}, \tag{1.29}$$

dove $n^*(t_{(j)})$ è un risk set esteso che si può esprimere nel seguente modo:

$$n^*(t_{(j)}) = n(t_{(j)}) \, \frac{1 - F_k(t_{(j)}-)}{\widehat{S}^{(KM)}(t_{(j)}-)}, \tag{1.30}$$

in cui $\widehat{S}^{(KM)}(t)$ rappresenta lo stimatore di Kaplan-Meier per la sopravvivenza da un qualunque tipo di evento.
Il *cause-specific* e il *subdistribution hazard* sono legati dalla seguente relazione [Geskus, 2015]:

$$\lambda_k(t) \times S(t-) = h_k(t) \times (1 - F_k(t-)). \tag{1.31}$$

1.3 Stima della funzione d'incidenza cumulativa e relazione con lo stimatore di Kaplan-Meier

La stima di massima verosimiglianza della funzione d'incidenza cumulativa per l'evento k conduce alla seguente relazione:

$$\widehat{F_k}(t) = \sum_{t_{(j)} \leq t} \widehat{S}^{(KM)}(t_{(j)}-) \times \widehat{\lambda}_k(t_{(j)})\,. \tag{1.32}$$

La varianza di $\widehat{F_k}(t)$ risulta[Pintilie, 2006]:

$$\begin{aligned} var\{\widehat{F_k}(t)\} &= \sum_{t_{(j)} \leq t} var\{[d_k(t_{(j)})/n(t_{(j)})]\widehat{S}^{(KM)}(t_{(j)})\} \\ &+ 2 \sum_{t_{(j)}<t} \sum_{t_{(j)}<t_{(\nu)} \leq t} cov([d_k(t_{(j)})/n(t_{(j)})]\widehat{S}^{(KM)}(t_{(j)}), [d_k(t_{(\nu)})/n(t_{(\nu)})]\widehat{S}^{(KM)}(t_{(\nu)})). \end{aligned} \tag{1.33}$$

Lo stimatore di tale varianza, nella versione di Aalen [1978](in Pintilie [2006]) è rappresentato da:

$$\begin{aligned} \widehat{var}\{\widehat{F_k}(t)\} = &\sum_{t_{(j)} \leq t} \left\{ [\widehat{F_k}(t) - \widehat{F_k}(t_{(j)})]^2 \frac{d(t_{(j)})}{(n(t_{(j)}) - 1)(n(t_{(j)}) - d(t_{(j)}))} \right\} \\ &+ \sum_{t_{(j)} \leq t} \widehat{S}(t_{(j-1)})^2 \frac{d_k(t_{(j)})(n(t_{(j)}) - d_k(t_{(j)}))}{n(t_{(j)})^2(n(t_{(j)}) - 1)} \\ &- 2 \sum_{t_{(j)} \leq t} [\widehat{F_k}(t) - \widehat{F_k}(t_{(j)})]\widehat{S}^{(KM)}(t_{(j-1)}) * \\ &* \frac{d_k(t_{(j)})(n(t_{(j)}) - d_k((t_{(j)}))}{n(t_{(j)})(n(t_{(j)}) - d(t_{(j)}))(n(t_{(j)}) - 1)}. \end{aligned} \tag{1.34}$$

Dalla (1.34) è possibile ricavare intervalli di confidenza approssimati per la funzione d'incidenza cumulativa:

$$\widehat{F_k}(t) \pm z_{1-\alpha/2}\sqrt{\widehat{var}\widehat{F_k}(t)}, \tag{1.35}$$

dove z_α è il quantile di ordine α della normale standard. Tale intervallo di confidenza pone tuttavia il problema di avere limiti inferiori e superiori negativi oppure maggiori di 1. Individuando un intervallo di confidenza per $\log(-\log(\widehat{F_k}(t)))$ e riportando i limiti alla scala originale si ottiene un intervallo di confidenza della forma:

$$\widehat{F_k}(t)^{\exp\left\{\pm\frac{z_{1-\alpha/2}\sqrt{\widehat{var}(\widehat{F_k}(t))}}{\widehat{F_k}(t)\log(\widehat{F_k}(t))}\right\}}. \tag{1.36}$$

È interessante, nella prospettiva dello studio dei test d'ipotesi, capire quale sia la relazione che lega la stima della funzione d'incidenza cumulativa con la stima di Kaplan-Meier. Nello specifico si può dimostrare che in un generico *failure time* $t_{(j)}$, lo stimatore di Kaplan-Meier sovrastima la funzione d'incidenza cumulativa, ovvero $1 - \widehat{S}^{(KM)}$ è maggiore o uguale a $\widehat{F_k}$ [Pintilie, 2006]. Si supponga di avere solo due eventi, quello d'interesse (indicato con 1 a pedice) e un evento competitivo. Lo stimatore di Kaplan-Meier dell'evento d'interesse si può esprimere come:

$$\begin{aligned}
\widehat{S}_1^{(KM)}(t_{(j)}) &= \prod_{l=1}^{j} \frac{n(t_{(l)}) - d_1(t_{(l)})}{n(t_{(l)})} \\
&= \frac{n(t_{(j)}) - d_1(t_{(j)})}{n(t_{(j)})} \widehat{S}_1^{(KM)}(t_{(j-1)}) \\
&= \widehat{S}_1^{(KM)}(t_{(j-1)}) - \frac{d_1(t_{(j)})}{n(t_{(j)})} \widehat{S}_1^{(KM)}(t_{(j-1)}),
\end{aligned}$$

da cui, cambiando segno e sommando 1 a entrambi i membri si ottiene

$$\begin{aligned}
1 - \widehat{S}_1^{(KM)}(t_{(j)}) &= 1 - \widehat{S}_1^{(KM)}(t_{(j-1)}) + \frac{d_1(t_{(j)})}{n(t_{(j)})} \widehat{S}_1^{(KM)}(t_{(j-1)}) \\
&= \sum_{l=1}^{j} \frac{d_1(t_{(l)})}{n(t_{(l)})} \widehat{S}_1^{(KM)}(t_{(l-1)}).
\end{aligned}$$

L'ultima quantità risulta confrontabile con la stima della funzione d'incidenza cumulativa per l'evento 1:

$$\widehat{F_1}(t_{(j)}) = \sum_{l=1}^{j} \widehat{\lambda}_1(t_{(l)}) \widehat{S}^{(KM)}(t_{(l-1)}) = \sum_{l=1}^{j} \frac{d_1(t_{(l)})}{n(t_{(l)})} \widehat{S}^{(KM)}(t_{(l-1)}).$$

Si può infatti osservare che $\widehat{S}_1^{(KM)}$ fa riferimento al solo evento d'interesse, mentre $\widehat{S}^{(KM)}$ è la sopravvivenza dall'evento 1 e dal rischio competitivo, quindi $\widehat{S}^{(KM)}(t) \leq \widehat{S}_1^{(KM)}(t)$ in ogni istante, da cui conseguentemente si ha che $\widehat{F_1}(t_{(k)}) \leq 1 - \widehat{S}_1^{(KM)}(t_{(k)})$.

In questo primo capitolo sono stati richiamati gli elementi costitutivi dell'analisi della sopravvivenza, specificando poi sia informalmente sia dal punto di vista matematico, cosa comporti assumere la presenza di rischi competitvi. I vari metodi di confronto tra gruppi che verranno presentati nel capitolo successivo, si differenziano per la grandezza, tra quelle appena elencate, che scelgono di stimare e mettere a confronto (il *cause-specific hazard*, il *subdistribution hazard*, la stessa funzione d'incidenza cumulativa).

Capitolo 2

Test delle ipotesi statistiche in presenza di rischi competitivi

Introdotti nel primo capitolo gli elementi che caratterizzano l'analisi della sopravvivenza in presenza di rischi competitivi, passiamo adesso a presentare la metodologia statistica al centro di questa trattazione ovvero il test di ipotesi. Nella sezione 1.1.4 ci siamo già riferiti al test di ipotesi maggiormente utilizzato nell'analisi della sopravvivenza classica, il log-rank test. È lecito perciò a questo punto domandarsi quali siano l'applicabilità e l'interpretazione dei risultati di questo test quando si fa riferimento alla presenza di rischi competitivi. Poiché come già visto, il log-rank test effettua di fatto un confronto pesato tra gli *hazard rate* di due gruppi (si veda la formula (1.13)), la prima osservazione che in tal senso è opportuno segnalare prende spunto da quanto ampiamente descritto precedentemente: la presenza cioé di due tipi di *hazard function*, il *cause-specific* e il *subdistribution*. L'estensione più naturale del log-rank in questo contesto appare essere quella di sostituire nel confronto tra due gruppi l'*hazard function* classico con il *cause-specific hazard rate* dell'evento di interesse per ciascun gruppo. A questo proposito va ricordato tuttavia come questa tipologia di funzione di rischio non sia in corrispondenza 1 a 1 con la funzione di incidenza cumulativa. Ciò significa che il log-rank test potrebbe fornire risultati contrastanti con la visualizzazione grafica del confronto tra le curve di incidenza. Pintilie [2006] mostra tale evidenza con un esempio numerico: ci s'immagina di confrontare due gruppi

che presentino sia l'evento di interesse che il rischio competitivo con una distribuzione esponenziale. In particolare s'ipotizza per il gruppo A i *rate* $\lambda_1^{(A)} = 0.1$ e $\lambda_2^{(A)} = 0.1$ mentre per il gruppo B $\lambda_1^{(B)} = 0.2$ $\lambda_2^{(B)} = 0.4$. Quello che si osserva è che nonostante il rischio dell'evento principale sia maggiore nel gruppo B rispetto al gruppo A ($\lambda_1^{(B)} > \lambda_1^{(A)}$) la curva di incidenza cumulativa del gruppo B a un certo istante, va a finire al di sotto dell'incidenza cumulativa relativa al gruppo A. Tale inconsistenza è dovuta al fatto che la funzione di icidenza cumulativa non dipende esclusivamente dal *cause-specific rate* dell'evento principale ma anche da quello dell'evento competitivo (si riveda la definizione (1.23)). Quanto detto sinora non significa che l'utilizzo del log-rank test sia da considerarsi scorretto in presenza di rischi significativi, va però precisata l'interpretazione che se ne da all'interno di questa specifica situazione. A questo proposito Pintilie [2006] riporta nuovamente un esempio, tratto da un caso reale: l'autrice fa riferimento a uno studio riguardante due gruppi di pazienti (con età rispettivamente minore o uguale e maggiore di 30 anni) che sono stati affetti dalla malattia di Hodgkin (cancro che solitamente si manifesta in giovane età) per confrontare l'incidenza di un secondo episodio maligno nelle due fasce di età. L'evento competitivo è costituito dalla morte per altre cause. Il log-rank test applicato per l'evento di interesse risulta significativo (p-value = 0.0022) mentre il test di Gray (che verrà ampiamente illustrato nella sezione 2.1), risulta non significativo (p-value = 0.5196), segnalando pertanto che il secondo episodio maligno risulta incidere più o meno allo stesso modo nei due gruppi. Il punto in questione è che il confronto tra *cause-specific hazard* tramite log-rank viene eseguito come se non esistessero altri tipi di eventi (che come abbiamo già detto vengono trattati alla stregua di osservazioni censurate) e per tale motivo quella che si confronta in realtà è esclusivamente la biologia dei due gruppi di pazienti nella predisposizione a contrarre il secondo episodio maligno. Il test di Gray invece non risulta significativo perchè l'incidenza delle morti per altre cause risulta molto maggiore nel gruppo degli over 30 (com'è intuitivo supporre) al punto di influenzare l'incidenza dell'evento di interesse rendendola numericamente simile tra i due gruppi. Quest'esempio mostra perciò come il confronto tra *cause-specific hazard* può risultare interessante da un punto di vista biologico mentre la comparazione tra le funzioni di incidenza cumulativa è in grado di

fornire maggiori informazioni sul numero effettivo di pazienti che subiscono un *failure*.
Di seguito verranno presentati i principali test di ipotesi in letteratura specificamente utilizzati in presenza di rischi competitivi.

2.1 Test di Gray

Il metodo di Gray [1988] è sicuramente quello più diffuso nell'ambito dei test di ipotesi in presenza di rischi competitivi. Come evidenziato nell'introduzione di questo capitolo, il confronto tra *cause-specific hazard* all'interno di tale contesto pone alcuni problemi di interpretazione per il fatto di non essere in corrispondenza con la funzione di incidenza cumulativa. Nella sezione 1.2.2 è stato tuttavia fatto presente come invece il *subdistribution hazard* sia in corrispondenza 1 a 1 con l'incidenza cumulativa: ciò che ha fatto Gray è stato proprio utilizzare questa funzione di rischio per effettuare un confronto tra gruppi, andando in questo modo a configurare una variante del log-rank test. Indicati i due gruppi sperimentali con A e B, l'ipotesi nulla che si vuole adesso testare è del tipo

$$H_0 : F_1^{(A)}(t) = F_1^{(B)}(t) = F_1^0(t), \tag{2.1}$$

dove il pedice $_1$ fa riferimento all'evento principale e $F_1^0(t)$ rappresenta una funzione d'incidenza cumulativa non specificata. La statistica del test per i due gruppi riferita all'evento d'interesse si basa su uno score della forma:

$$Z_1 = \sum_{j=1}^{r} K(t_{(j)}) \left[\frac{d_1^{(A)}(t_{(j)})}{n^{*(A)}(t_{(j)})} - \frac{d_1(t_{(j)})}{n^*(t_{(j)})} \right] \tag{2.2}$$

dove r è il numero totale di *failure time* ottenuto dall'unione dei due gruppi e $K(t)$ è una funzione di peso che si può esprimere come prodotto $K(t) = W(t)\, n^{*(A)}(t)$ tra una generica $W(t)$ e il risk set esteso del gruppo A, definito dalla formula (1.30) (è da ricordare che laddove non compare l'apice relativo al gruppo ci si riferisce, in conformità alla sezione 1.1.4, al campione ottenuto considerando l'unione tra i due gruppi). Sfruttando la scomposizione di $K(t)$

appena introdotta, la (2.2) si può riscrivere come [Geskus, 2015]

$$Z_1 = \sum_{j=1}^{r} W(t_{(j)}) \frac{n^{*(A)}(t_{(j)})\, n^{*(B)}(t_{(j)})}{n^{*(A)}(t_{(j)}) + n^{*(B)}(t_{(j)})} \left\{ \widehat{h}_1^{(A)}(t_{(j)}) - \widehat{h}_1(B)(t_{(j)}) \right\}. \quad (2.3)$$

L'interpretazione di questo test come variante del log-rank è giustificata dalla somiglianza tra la (1.13) e la (2.3) in quanto quest'ultima esprime proprio la media pesata dei *subdistribution hazard* così come la (1.13) rappresenta una media pesata degli *hazard* semplici dei due gruppi.
Lo score calcolato nelle relazioni (2.2) e (2.3) costituisce solo una parte della statistica del test di Gray, è infatti necessario dividere Z_1 per la propria varianza. Quest'ultima tuttavia presenta un'espressione molto complicata, Pintilie [2006] ha sintetizzato il calcolo delle parti principali che la compongono. Viene anzitutto esplicitata la forma di F^0 nel caso di due gruppi:

$$F^0(t_{(j)}) = \sum_{l=1}^{j} \frac{d_1^{(A)}(t_{(j)}) + d_1^{(B)(t_{(j)})}}{m^A(t_{(j)}) + m^{(B)}(t_{(j)})} \quad (2.4)$$

dove

$$m^{(G)}(t_{(j)}) = \frac{n^{(G)}(t_{(j)})}{\widehat{S}^{(KM)(G)}(t_{(j-1)})}, \quad G \in \{A, B\}.$$

Occorre ricordare nuovamente che il pedice numerico fa riferimento all'evento che si considera, nella (2.4) ci si riferisce perciò al solo evento principale.
Nel presentare le varie componenti che saranno utilizzate per costruire la varianza verrà omesso l'argomento $t_{(j)}\,, j = 1 \dots r$ (dato che tutte le quantità che seguiranno sono calcolate in quell'istante) e per riferirsi all'istante $t_{(j-1)}$ si utilizzerà un underscore(_):

$$D = \frac{m^{(A)}\, m^{(B)}}{m^{(A)} + m^{(B)}},$$

$$C = \sum_{\forall\, t_{(j)} < t} D \frac{d_1^{(A)} + d_1^{(B)}}{(m^{(A)} + m^{(B)})(1 - F_{\_}^0)},$$

$$\tilde{C} = \sum_{\forall t_{(j)}} D \frac{d_1^{(A)} + d_1^{(B)}}{(m^{(A)} + m^{(B)})(1 - F_{\_}^0)}.$$

Per ciascuno dei due gruppi $G \in \{A, B\}$ e per ogni *failure time* si calcolano le seguenti quantità:

$$tev4^{(G)} = 1 - \frac{1 - F^0}{S^{(KM)(G)}},$$

$$tev3^{(G)} = (d_1^{(A)} + d_1^{(B)}) \frac{S_-^{(KM)(G)}}{(m^{(A)} + m^{(B)}) n^{(G)}} \left(1 - \frac{d_1^{(A)} + d_1^{(B)} - 1}{S_-^{(KM)(G)}(m^{(A)} + m^{(B)}) - 1} \right)$$

$$tev1^{(G)} = \begin{cases} (D - tev4^{(G)}\, C) & \text{G=A} \\ -(D - tev4^{(G)}\, C) & \text{G=B} \end{cases}.$$

Per ogni gruppo e per ciascun istante in cui si è invece verificato un rischio competitivo si calcolano le seguenti quantità:

$$tcr4^{(G)} = \frac{1 - F^0}{S^{(KM)(G)}},$$

$$tcr3^{(G)} = d_2^{(G)} \frac{(S_-^{(KM)(G)})^2}{(n^{(G)})^2} \left(1 - \frac{d_1^{(G)} - 1}{n^{(G)} - 1} \right),$$

$$tcr1^{(G)} = \begin{cases} tcr4^{(G)}\, C & \text{G=A} \\ -tcr4^{(G)}\, C & \text{G=B} \end{cases},$$

$$v3^{(G)} = \sum_{\forall\, t_{(j)}} \{(tev4^{(G)})^2\, tev3^{(G)} + (tcr4^{(G)})^2\, tcr3^{(G)}\},$$

$$v2^{(G)} = \sum_{\forall\, t_{(j)}} \{tev1^{(G)}\, tev4^{(G)}\, tev3^{(G)} + tcr1^{(G)}\, tcr4^{(G)}\, tcr3^{(G)}\}$$

$$vpart = \sum_{G \in \{A,B\}} \sum_{\forall\, t_{(j)}} \{(tev1^{(G)})^2\, tev3^{(G)} + (tcr1^{(G)})^2\, tcr3^{(G)}\}.$$

La varianza si esprime infine attraverso la relazione

$$V = vpart + \tilde{C}^2 v3^{(A)} + 2\tilde{C} v2^{(A)} + \tilde{C}^2 v3^{(B)} + 2\tilde{C} v2^{(B)}. \tag{2.5}$$

La statistica test si distribuisce asintoticamente come una χ^2.
Quando si esegue il test di Gray in un contesto reale è necessario eseguirlo prima per il confronto delle *subdistribution* relative all'evento d'interesse ed eseguirlo nuovamente scambiando di ruolo l'evento principale e quello competitivo (in realtà è opportuno eseguire due volte un qualunque test d'ipotesi che confronti le funzioni d'incidenza cumulativa oppure loro derivati). Il motivo di questo modo di procedere è strettamente legato a quanto abbiamo detto

all'inizio del capitolo relativamente all'interpretazione del test: il confronto tra le *subdistribution* dell'evento principale tiene conto dei rischi competivi e perciò una differenza non significativa può essere causata da un'ampia differenza tra gli eventi competitivi (come nell'esempio della malattia di Hodgkin). Tale aspetto verrà ulteriormente ripreso negli esempi successivi.

2.1.1 Il test di Gray in ambiente SAS

Nell'ultima versione di **SAS®**, la **9.4**, la procedura `proc lifetest` implementa il test di Gray e calcola le stime della funzione d'incidenza cumulativa per ciascun gruppo [Ward and Weber, 2016]. Il codice per eseguire il test è del tipo

```
proc lifetest data=dataset plots=cif(test);
    time t*status(0) / eventcode=1;
    strata group;
run;
```

in cui `t` è la variabile tempo, mentre `status` è l'etichetta che assume i valori che servono per stabilire se il soggetto in questione abbia sperimentato l'evento d'interesse, il rischio competitivo oppure sia stato censurato. Nello `strata` statement è da inserire la variabile che discrimina i gruppi sperimentali. Nel `time` statement il valore tra parentesi rappresenta quali sono le osservazioni da considerare come censurate, mentre l'opzione `eventcode` consente di specificare il valore dello `status` corrispondente all'evento principale (è possibile inserire più valori della variabile `status` qualora si voglia eseguire il test più volte andando a variare l'evento d'interesse). Precisiamo che nello svolgimento dello studio di simulazione che verrà discusso nel quarto capitolo, ci si è avvalsi di un server dell'Università degli Sudi di Padova, in cui non è presente l'ultima versione di **SAS®**, per cui non è stato possibile utilizzare la `proc lifetest`. Si è fatto ricorso in quel caso alla funzione macro `%CIF` rilasciata dalla **SAS®** foundation per versioni precedenti alla **9.4**, che però restituisce i medesimi risultati.
Vediamo adesso qualche esempio prendendo spunto dai dataset descritti in sezione 1.2.1.

Linfoma follicolare

Pintilie [2006] riporta i risultati dell'esecuzione del test di Gray su questo dataset. I due gruppi sperimentali sono individuati suddividendo i pazienti tra chi ha un'età superiore e chi una inferiore ai 65 anni al tempo dello studio. In figura 2.1 è mostrata una porzione della stima della funzione d'incidenza

Stime della funzione di incidenza cumulativa				
Strato 1: gruppo = Over 65				
time	Incidenza cumulativa	Errore standard	Intervallo di confidenza al 95%	
0	0	0	.	.
0.002738	0.0881	0.0225	0.0504	0.1386
0.238193	0.0943	0.0233	0.0552	0.1460
0.303901	0.1006	0.0239	0.0600	0.1534
0.33128	0.1069	0.0246	0.0649	0.1608
0.350445	0.1132	0.0252	0.0698	0.1681
0.391513	0.1195	0.0258	0.0748	0.1754
0.457221	0.1258	0.0264	0.0798	0.1826
0.481862	0.1321	0.0269	0.0849	0.1898
0.531143	0.1384	0.0275	0.0900	0.1970
0.558522	0.1447	0.0280	0.0951	0.2041
0.689938	0.1509	0.0285	0.1003	0.2112
0.711841	0.1572	0.0290	0.1055	0.2183
0.766598	0.1635	0.0294	0.1108	0.2253
0.854209	0.1698	0.0299	0.1160	0.2323
0.960986	0.1761	0.0303	0.1214	0.2393
0.988364	0.1824	0.0307	0.1267	0.2462
1.043121	0.1887	0.0311	0.1320	0.2532

Figura 2.1: Porzione della stima della *subdistribution* dell'evento principale in riferimento agli over 65 nel dataset follic

cumulativa per l'evento principale in riferimento al gruppo degli over 65.In figura 2.2 è invece riportato il confronto tra le curve d'incidenza stimata nei due gruppi per la morte legata al linfoma, l'evento d'interesse di questo studio. È inoltre riportato il p-value del test di Gray, pari a 0.1047, il che significa che l'evento principale non incide con una differenza significativa nei due

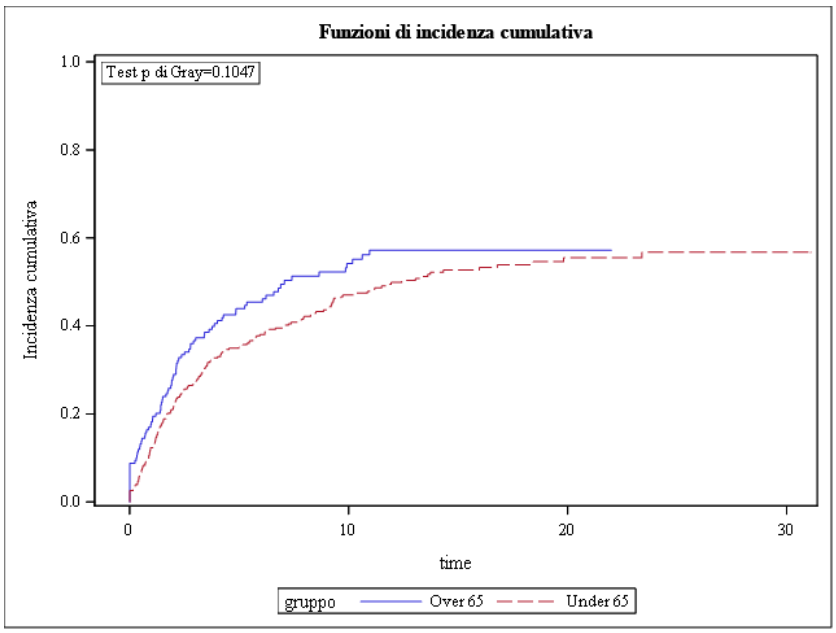

Figura 2.2: Confronto tra le *subdistribution* dell'evento d'interese e p-value del test di Gray per il dataset follic

gruppi sperimentali. Il log-rank test (eseguito considerando la morte per altre cause come censura) risulta invece significativo (figura 2.3), evidenziando una sopravvivenza maggiore per il gruppo dei pazienti sotto i 65 anni. Con un ragionamento di fatto analogo a quello fatto per la malattia di Hodgin si può concludere che l'incidenza della morte del linfoma risulta più o meno simile nei gruppi per via di un'influenza da parte del rischio competitivo. Se infatti si esegue il test di Gray confrontando le *subdistribution* della morte per altre cause si osserva una differenza significativa con una più alta incidenza per i pazienti con età superiore ai 65 anni (figura 2.4).

Trapianto di midollo osseo

All'interno di questo dataset i gruppi sperimentali sono determinati da tre tipi di leucemia. Il confronto tra le funzioni d'incidenza cumulativa è stato eseguito tra il gruppo di pazienti con i tipi di leucemia identificati come *AML-Low Risk* e *AML-High Risk*. Il test di Gray eseguito sull'evento principale ricaduta

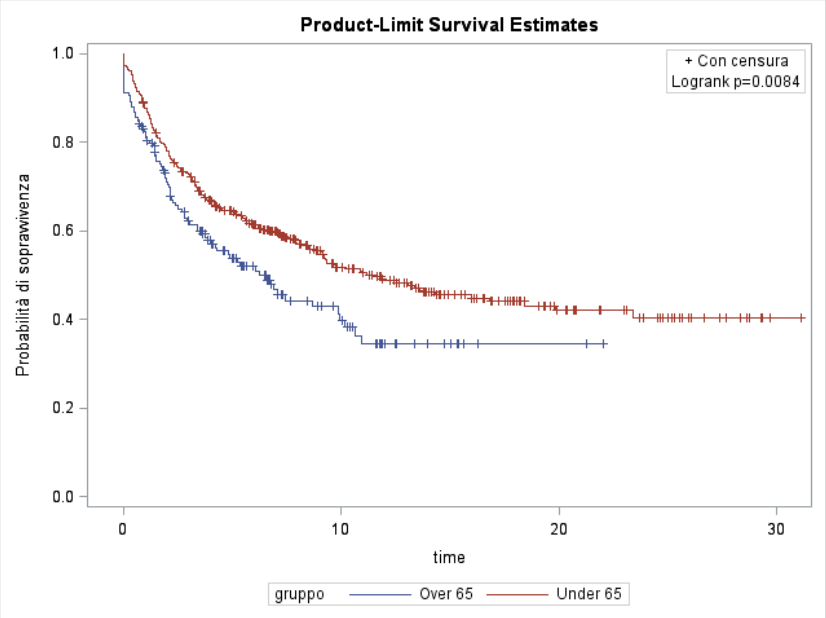

Figura 2.3: Stima di Kaplan-Meier e p-value del log-rank test dell'evento d'interesse per il dataset follic

risulta significativo evidenziando una maggiore incidenza nella tipologia *AML-High Risk* (figura 2.5). Il confronto tra i due gruppi relativamente alla morte nel periodo di latenza non risulta invece significativo (figura 2.6) mostrando dunque che il rischio competitivo non esercita alcun tipo d'influenza e che l'incidenza dell'evento ricaduta è effettivamente diversa nei due gruppi (inoltre anche il log-rank test risulta significativo). Questo tipo d'interpretazione del risultato dell'analisi è riportato ad esempio da Lin [1997].

2.2 Il test di Pepe e Mori

Qualche anno dopo la pubblicazione di Gray viene proposto un secondo test d'ipotesi utilizzabile in presenza di rischi competitivi. Il test sviluppato da Pepe and Mori [1993] si caratterizza per il fatto che effettua un confronto diretto tra le *subdistribution* dei due gruppi, nello specifico calcola l'area pesata tra le due funzioni d'incidenza cumulative stimate. La statistica test

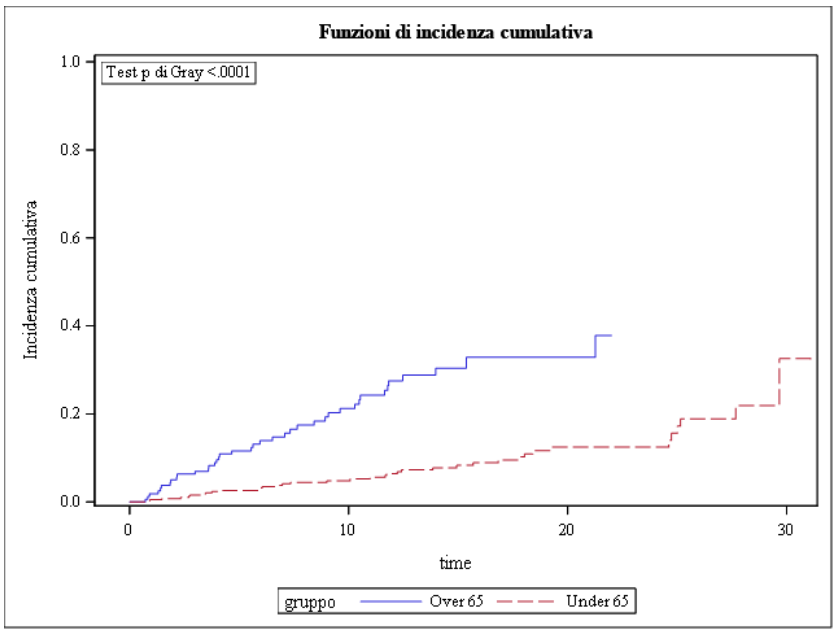

Figura 2.4: Confronto tra le *subdistribution* dell'evento competitivo e p-value del test di Gray per il dataset follic

si basa sulla quantità

$$s = \sqrt{\frac{N^{(A)}\,N^{(B)}}{N^{(A)}+N^{(B)}}}\int_0^\tau W(t)\{\widehat{F}_1^{(A)}(t)-\widehat{F}_1^{(B)}(t)\}dt \tag{2.6}$$

che Pepe [1991] ha dimostrato convergere asintoticamnete a una normale di media nulla e varianza σ^2. $N^{(A)}$ e $N^{(B)}$ rappresentano rispettivamente il numero totale di individui in ciascun gruppo. Calcolata partendo dai *failure time* ordinati la (2.6) diventa

$$s = \sqrt{\frac{N^{(A)}\,N^{(B)}}{N^{(A)}+N^{(B)}}}\sum_{\forall t_{(j)}}\{W(t_{(j)})[\widehat{F}_1^{(A)}-\widehat{F}_1^{(B)}](t_{(j+1)}-t_{(j)})\}. \tag{2.7}$$

L'espressione della funzione di peso $W(t)$ risulta [Pintilie, 2006]

$$W(t_{(j)}) = \frac{(N^{(A)}+N^{(B)})\widehat{C}^{(A)}(t_{(j-1)})\,\widehat{C}^{(B)}(t_{(j-1)})}{N^{(A)}\widehat{C}^{(A)}(t_{(j-1)})+N^{(B)}\widehat{C}^{(B)}(t_{(j-1)})}, \tag{2.8}$$

dove $1-\widehat{C}(t)$ è la stima di Kaplan-Meier relativa alla distribuzione del *censoring*, ovvero calcolata considerando l'evento principale e il rischio competitivo

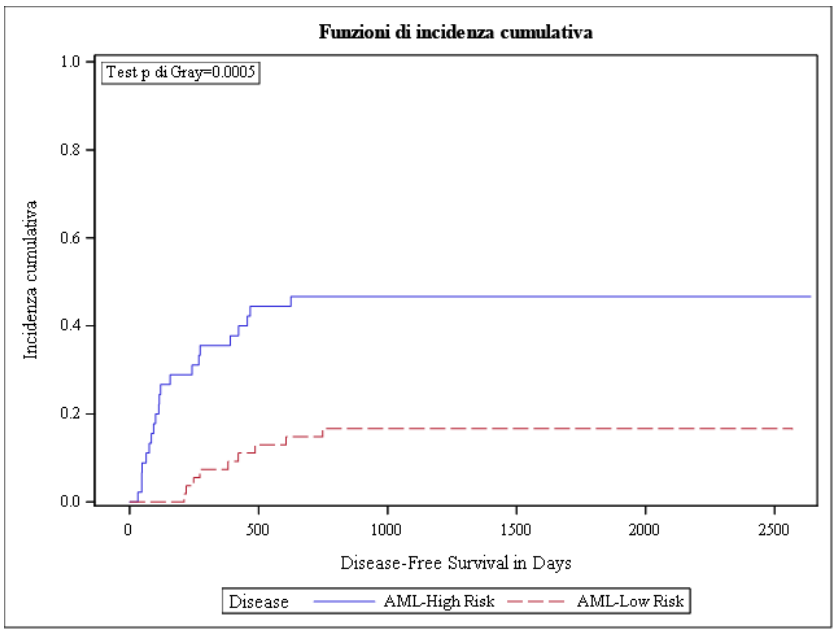

Figura 2.5: Confronto tra le *subdistribution* dell'evento principale e p-value del test di Gray per i gruppi *AML-Low Risk* e *AML-High Risk* del dataset bmt

come osservazioni censurate e il *censoring* come *failure*. Come fa notare Pintilie [2006] $W(t)$ è una funzione decrescente e perciò il suo effetto tende a diminuire durante il follow up, dando meno peso alla differenza tra le *subdistribution* man mano che il tempo aumenta.
Per ottenere la formula della varianza della statistiica test è necessario calcolare prima le varianze relative ai due gruppi (per facilitare la lettura omettiamo l'apice riferito al gruppo perché questa quantità è da intendersi calcolata per entrambi i gruppi sperimentali):

$$\widehat{\sigma}^{2(G)} = \sum_{\forall\, t_{(j)}} \frac{[v_I(t_{(j)}) - \widehat{F}_2(t_{(j)})\, v_{II}(t_{(j)})]^2\; d_1(t_{(j)}) + v_{II}^2(t_{(j)})\,[d(t_{(j)}) - d_1(t_{(j)})]}{n(t_{(j)})(1 - n(t_{(j)}))} \tag{2.9}$$

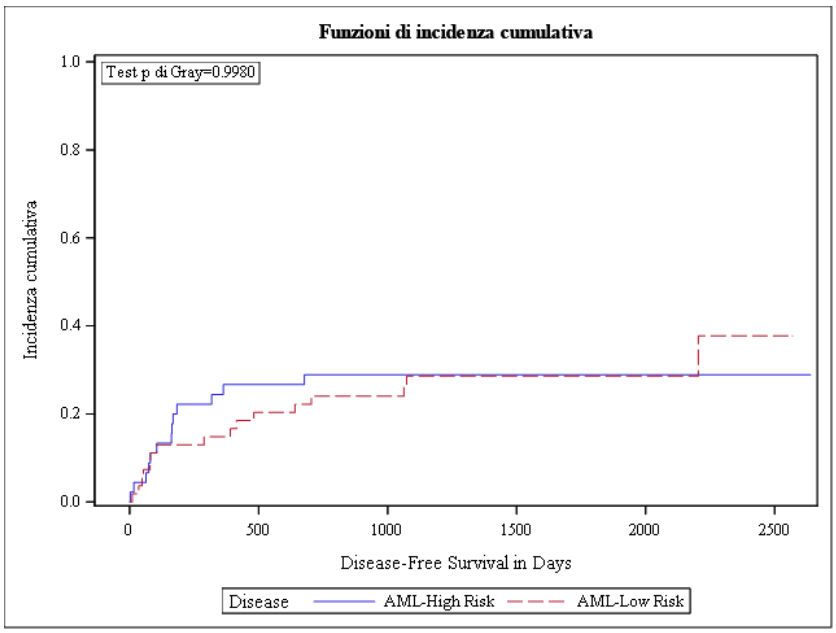

Figura 2.6: Confronto tra le *subdistribution* dell'evento competitivo e p-value del test di Gray per i gruppi *AML-Low Risk* e *AML-High Risk* del dataset bmt

dove $d(t_{(j)})$ comprende sia l'evento d'interesse che il rischio competitivo e

$$v_I(t_{(j)}) = \sum_{t_{(l)} \geq t_{(j)}} W(t_{(l)})\,(t_{(l+1)} - t_{(l)})\,(1 - \widehat{F}_1(t_{(l)})),$$
$$v_{II}(t_{(j)}) = \sum_{t_{(l)} \geq t_{(j)}} W(t_{(l)})\,(t_{(l+1)} - t_{(l)}).$$

L'espressione della varianza di s infine risulterà

$$\widehat{\sigma}^2 = \frac{N^{(A)} N^{(B)}\,(\widehat{\sigma}^{2(A)} + \widehat{\sigma}^{2(B)})}{N^{(A)} + N^{(B)}}. \tag{2.10}$$

2.2.1 Il test di Pepe-Mori in ambiente SAS

All'interno di questa trattazione faremo riferimento all'implementazione del test di Pepe-Mori fornita da Pintilie [2006], che all'interno della sua opera ha inserito in appendice il codice della funzione macro `%compcif` da lei scritta. Tale macro calcola prima le stime per i due gruppi della funzione d'incidenza

cumulativa (che come già detto viene anche fatto nell'ultima versione di **SAS®** nella `proc lifetest`), da cui ricava successivamente statistica e p-value del test. La sintassi richiesta per eseguire il test d'ipotesi è del tipo

```
%compcif(ds=, time=, cens=, group=,val1=,val2=);
```

i cui parametri corrispondono rispettivamente al dataset su cui viene eseguita l'analisi, la variabile tempo, il flag che determina la tipologia di evento che è accaduto (si da per implicito che il *censoring* è rappresentato dallo 0), la variabile che specifica gruppi sperimentali e infine i due valori da tale variabile, da considerare per effettuare il confronto.

Il test applicato al dataset follic restituisce un risultato praticamente identico

Pepe and Mori Test

N Group 1	N Group 2	Score	Chi-square	p-value
159	382	7.3016	2.6174	0.10570

Figura 2.7: Statistica e p-value del test di Pepe-Mori sull'evento principale per il dataset follic

Pepe and Mori Test

N Group 1	N Group 2	Score	Chi-square	p-value
159	382	7.0284	17.7670	0.00002

Figura 2.8: Statistica e p-value del test di Pepe-Mori sull'evento competitivo per il dataset follic

a quello fornito dal test di Gray, sia per l'evento principale (output in figura 2.7), sia per quanto riguarda il rischio competitivo (figura 2.8).

Malattie cardiovascolari

In questo dataset i due gruppi sperimentali sono individuati dal genere. Fissando un livello di significatività al 5% il test sull'evento principale risulta di poco non significativo (figura 2.9) mentre l'assenza di significatività è più marcata se si considera la morte per altre cause (figura 2.10).

Pepe and Mori Test

N Group 0	N Group 1	Score	Chi-square	p-value
267	186	-834.7972	3.2348	0.07209

Figura 2.9: Statistica e p-value del test di Pepe-Mori sull'evento principale per il dataset cvd

Pepe and Mori Test

N Group 0	N Group 1	Score	Chi-square	p-value
267	186	337.6636	1.1667	0.28008

Figura 2.10: Statistica e p-value del test di Pepe-Mori sull'evento competitivo per il dataset cvd

2.3 Altri test d'ipotesi

In questa sezione faremo cenno ad altre due tipologie di test d'ipotesi meno diffuse ma pur sempre adottate in presenza di rischi competitivi. Per questi due metodi non verranno presentate le implementazioni in ambiente statistico.

2.3.1 Test di tipo Kolmogorov-Smirnov

Lin [1997] ha proposto un test di tipo Kolmogorov-Smirnov utilizzando una tecnica di ricampionamento per la stima degli scarti tra le funzioni d'incidenza cumulativa dei due gruppi sperimentali. La statistica test (riferita all'evento principale) risulta

$$Q = \sup_t |\widehat{F}_1^{(A)}(t) - \widehat{F}_1^{(B)}(t)|, \tag{2.11}$$

con $K(f)$ funzione di peso. Un ampio numero di repliche della statistica test tramite la tecnica di ricampionamento consente di ricavare un p-value empirico. Questo test risulta meno sensibile degli altri a specifiche violazioni dell'ipotesi nulla[Lin, 1997].

2.3.2 Un approccio di tipo Renyi

Bajorunaite and Klein [2008] hanno proposto un metodo di confronto che risulta maggiormente indicato rispetto ai precedenti nella situazione in cui si abbiano i *subdistribution hazard* dei due gruppi sperimentali che s'incrociano. Si tratta di un test analogo alla statistica di Renyi già proposta in analisi della sopravvivenza classica [Klein and Moeschberger, 2005]. L'idea è quella anzitutto di calcolare la statistica di Z_1 Gray definita nella (2.2) su tutti gli r failure time ottenuti considerando congiuntamente i due gruppi sperimentali. Definito τ come il massimo $t_{(j)}$ tale che $n^{(A)}(t)$,$n^{(B)}(t) > 0$, la quantità

$$O = \sup\{|Z_1(t)|,\, t \leq \tau\}/V \tag{2.12}$$

rappresenta la statistica di questo test. V è la varianza dello score di Gray calcolata nella (2.5). Sotto l'ipotesi nulla la statistica O converge debolmente nell'intervallo $[0, \tau]$ a un processo gaussiano a media nulla che può essere rappresentato da un moto Browniano. Billingsley [2013] ha fornito la seguente relazione riguardo un moto Browniano standard $B(t)$, che può essere molto utile per ricavare il p-value del test:

$$\mathbb{P}\{\sup\, |B(t)| > y\} = 1 - \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\left\{-\pi^2\, \frac{(2k+1)^2}{8y^2}\right\}.$$

Capitolo 3

La metodologia NPC in analisi della sopravvivenza

Nel capitolo precedente sono stati presentati i test d'ipotesi più diffusi nel contesto dei rischi competitivi. Oggetto di questo capitolo sarà una breve introduzione di una metodologia sviluppata di recente, che nasce come un test d'ipotesi non parametrico generico per l'analisi multivariata ma che può essere adoperato anche nel contesto dell'analisi della sopravvivenza. Si tratta della procedura *Non Parametric Combination* (NPC). In un recente lavoro [Arboretti et al., 2017] è stato mostrato come la metodologia NPC possa rappresentare, nell'ambito dell'analisi della sopravvivenza classica, un'ottima alternativa al log-rank test in presenza di *informative censoring*, performando molto bene anche in diversi scenari di *non-informative censoring*. Cogliendo proprio l'auspicio con cui tale pubblicazione si conclude, riporteremo in questo capitolo il tentativo di estendere l'algoritmo sotteso alla procedura NPC per un suo utilizzo in presenza di rischi competitivi. Per questo motivo verrà prima presentata una breve descrizione degli aspetti salienti di questa metodologia statistica, per poi proporre un'implementazione in grado di produrre un test d'ipotesi che possa essere confrontato con quelli introdotti nel precedente capitolo.

3.1 Aspetti principali della procedura NPC

La procedura NPC si fonda a partire dall'assunto che in molti problemi reali, le ipotesi da studiare attraverso un test, possono essere complesse nel senso che comprendono numerose variabili risposta e vari aspetti d'interesse (si pensi proprio al confronto di curve nell'analisi della sopravvivenza in presenza di rischi competitivi: è vero che il confronto che si effettua è tra le sole funzioni d'incidenza cumulativa, tuttavia i processi che entrano in gioco nel determinare l'esito di tale confronto sono numerosi, poiché potremmo immaginare di considerare come si distribuiscono singolarmente tutti i singoli eventi possibili più la censura). Per certi versi può risultare perciò naturale pensare di scomporre il problema iniziale in sottoproblemi, nei quali i dati si possono processare attraverso l'utilizzo di un numero $k > 1$ di test d'ipotesi parziali, che da un lato consentirebbero di effettuare procedure inferenziali separate (attraverso metodi che tengano conto di un aggiustamento di confronti multipli) e dall'altro potrebbero essere considerati congiuntamente per fornire un'informazione generale in grado di rispondere al quesito iniziale. Dal punto di vista operativo tutto ciò si traduce con la possibilità di suddividere l'ipotesi nulla H_0 del test che si vuole eseguire in k ipotesi $H_{0i},\ i = 1, \ldots, k$, per ciascuno degli aspetti parziali d'interesse. Questo significa che H_0 risulta vera se le H_{0i} sono congiuntamente vere, perciò l'ipotesi nulla si può riscrivere come $\left\{\bigcap_{i=1}^{k} H_{0i}\right\}$. Analogo ragionamento si può fare in riferimento all'ipotesi alternativa H_1, che risulta anch'essa scomponibile in k ipotesi $H_{1i},\ i = 1, \ldots, k$. Questa volta però H_1 risulterà vera se almeno una delle H_{1i} è vera, per questo motivo l'ipotesi alternativa potrà essere scritta come $\left\{\bigcup_{i=1}^{k} H_{1i}\right\}$. Una volta definiti i due set d'ipotesi nulle e alternative, all'interno della metodologia NPC vengono eseguiti dei test di permutazione: un'ipotesi nulla globale di equidistribuzione di una determinata quantità tra due gruppi sperimentali, dovrebbe infatti implicare la scambiabilità dei soggetti tra i due gruppi. Per essere più precisi ci si affida al principio di *permutation multivariate testing* (per maggiori dettagli su quest'aspetto e più in generale sulla teoria matematica dei test di permutazione e procedura NPC si rimanda a Pesarin and Salmaso [2010]).

Su ciascuna permutazione dei dati rispetto all'appartenenza dei soggetti ai 2

gruppi sperimentali, viene solitamente applicata una procedura parametrica di test d'ipotesi, la quale produce una statistica test che viene opportunamente confrontata con quelle ottenute da altre permutazioni (nella sezione 3.2, in cui sarà descritto dettagliatamente l'algoritmo per sviluppare un test d'ipotesi in presenza di rischi competitivi, questi passi che ora vengono sommariamente citati, saranno esplicitati in maniera diretta). L'insieme di statistiche test viene calcolato per ciascuno dei test d'ipotesi parziali che si decide di eseguire, con il risultato di avere k p-value diversi. Il punto cruciale risulta quindi trovare il modo per sintetizzare l'informazione contenuta in ciacuno di questi p-value, in modo da otterne uno solo con cui decidere se rifiutare o meno l'ipotesi nulla globale H_0. Per questo scopo si utlizzano delle funzioni di combinazione, le quali presentano opportune proprietà matematiche [Pesarin and Salmaso, 2010] in grado di garantire la possibilità di pesare opportunamente ciascun p-value parziale. In letteratura vi sono diversi tipi di funzione di combinazione ψ, ma noi facendo riferimento a Arboretti et al. [2017] utilizzeremo nella sezione successiva la funzione di Tippet (ψ_T) e quella di Fisher (ψ_F), definite nel seguente modo:

$$\psi_F(\lambda_1, \ldots, \lambda_k) = \min_{i=1,\ldots k} \{\lambda_i\} \tag{3.1}$$

$$\psi_T(\lambda_1, \ldots, \lambda_k) = -\sum_{i=1}^{k} \log \lambda_i, \tag{3.2}$$

dove i λ_i sono i p-value dei test parziali. In realtà come vedremo nella prossima sezione i risultati dell'applicazione delle funzioni di combinazione non forniscono i p-value definitivi del test globale ma delle quantità intermedie che vanno ulteriormente processate. Alla fine dell'intero processo si ottiene un p-value combinato che viene utilizzato per rifiutare o meno H_0.

3.2 Un algoritmo basato su metodologia NPC per l'esecuzione di un test d'ipotesi in presenza di rischi competitivi

In Arboretti et al. [2017] viene proposto un algoritmo basato su NPC per effettuare un test d'ipotesi che possa essere eseguito in alternativa al log-rank test. L'obiettivo di questa sezione è quello di estendere e modificare opportunamente tale procedura al fine di costruire un test d'ipotesi che possa operare in presenza di rischi competitivi e in alternativa ai metodi presentati nel capitolo 2.
Un primo aspetto importante da considerare è quello di individuare le ipotesi nulle dei test parziali in cui si può suddividere il test globale: l'idea è quella di valutare separatamente l'incidenza cumulativa, tra i due gruppi sperimentali, dei tre eventi che possono effettivamente accadere: la censura, l'evento principale e il rischio competitivo. Per far ciò verrà calcolata la statistica di Gray tre volte, variando di volta in volta l'evento da considerarsi d'interesse (nella descrizione dei vari passi dell'algoritmo si utilizzerà una notazione consistente a quella di Arboretti et al. [2017], indicando con T_G la statistica del test di Gray, e nello specifico con T_{G0}, T_{G1} e T_{G2} i valori di tale statistica riferiti rispettivamente ai tre tipi di eventi). Quanto detto rappresenta la parte iniziale della procedura che viene qui di seguito descritta. S'indicherà con B il numero di permutazioni dell'appartenenza dei soggetti ai gruppi sperimentali che verranno eseguite.

1. Partendo dai dati osservati si calcolano le statistiche di Gray riferite ai tre eventi che, per garantire una coerenza con i passi successivi, verranno indicate con $T_{G0}^{(0)}$, $T_{G1}^{(0)}$ e $T_{G2}^{(0)}$.

2. Si eseguono B permutazioni del dataset osservato. Per far ciò viene semplicemente scambiata in maniera casuale l'appartenenza dei soggetti a ciascuno dei due gruppi sperimentali. Per tutti i dataset permutati vengono calcolate le statistiche test $T_{G0}^{(b)}$, $T_{G1}^{(b)}$ e $T_{G2}^{(b)}$, con $b = 1, \ldots, B$.

3. Una volta calcolate tutte le statistiche test, vengono calcolati i p-value

relativi ai tre eventi sottoposti a studio nel seguente modo:

$$p_{G0} = \frac{\frac{1}{2} + \sum_{b=1}^{B} I(|T_{G0}^{(b)}| \geq |T_{G0}^{(0)}|)}{B+1} \tag{3.3}$$

$$p_{G1} = \frac{\frac{1}{2} + \sum_{b=1}^{B} I(|T_{G1}^{(b)}| \geq |T_{G1}^{(0)}|)}{B+1} \tag{3.4}$$

$$p_{G2} = \frac{\frac{1}{2} + \sum_{b=1}^{B} I(|T_{G2}^{(b)}| \geq |T_{G2}^{(0)}|)}{B+1}, \tag{3.5}$$

dove $I()$ è la funzione indicatrice. In pratica, per ciascuno dei tre eventi, si tratta di contare quante volte la statistica test calcolata sui dati permutati supera in valore assoluto la statistica ottenuta dai dati iniziali, correggendo tale conteggio come mostrato sopra, aggiungendovi $1/2$ e dividendo tutto per $B+1$.

4. Per ogni permutazione si calcolano delle statistiche p-value like:

$$\lambda_{G0}^{(b)} = \frac{\frac{1}{2} + \sum_{j \in \{1,\ldots,B\}, j \neq b} I(|T_{G0}^{(j)}| \geq |T_{G0}^{(b)}|)}{B+1} \quad b = 1, \ldots, B \tag{3.6}$$

$$\lambda_{G1}^{(b)} = \frac{\frac{1}{2} + \sum_{j \in \{1,\ldots,B\}, j \neq b} I(|T_{G1}^{(j)}| \geq |T_{G1}^{(b)}|)}{B+1} \quad b = 1, \ldots, B \tag{3.7}$$

$$\lambda_{G2}^{(b)} = \frac{\frac{1}{2} + \sum_{j \in \{1,\ldots,B\}, j \neq b} I(|T_{G2}^{(j)}| \geq |T_{G2}^{(b)}|)}{B+1} \quad b = 1, \ldots, B. \tag{3.8}$$

Si tratta qui di ripetere calcoli analoghi al punto precedente, considerando questa volta come statistica test osservata quella ottenuta dalla b-esima permutazione, ottenendo in questo modo B valori differenti.

5. Sfruttando le funzioni di combinazione (3.2) e (3.1), si calcolano i definitivi p-value combinando quelli ottenuti in precedenza per i tre eventi:

$$p_{GT} = \frac{\frac{1}{2} + \sum_{b=1}^{B} I(\min\{\lambda_{G0}^{(b)}, \lambda_{G1}^{(b)}, \lambda_{G2}^{(b)}\} \leq \min\{p_{G0}, p_{G1}, p_{G2}\})}{B+1} \tag{3.9}$$

$$p_{GF} = \frac{\frac{1}{2} + \sum_{b=1}^{B} I(-\log \lambda_{G0}^{(b)} - \log \lambda_{G1}^{(b)} - \log \lambda_{G2}^{(b)} \geq -\log p_{G0} - \log p_{G1} - \log p_{G2})}{B+1}, \tag{3.10}$$

con le lettere T e F che si riferiscono rispettivamente a Tippet e Fisher. È eventualmente anche possibile escludere dalla combinazione i p-value riferiti al *censoring*:

$$p_{GT12} = \frac{\frac{1}{2} + \sum_{b=1}^{B} I(\min\{\lambda_{G1}^{(b)}, \lambda_{G2}^{(b)}\} \leq \min\{p_{G1}, p_{G2}\})}{B+1} \tag{3.11}$$

$$p_{GF12} = \frac{\frac{1}{2} + \sum_{b=1}^{B} I(-\log \lambda_{G1}^{(b)} - \log \lambda_{G2}^{(b)} \geq -\log p_{G1} - \log p_{G2})}{B+1}. \tag{3.12}$$

Come si può notare, la procedura descritta fornisce alla fine 2 p-value (più altri due se si vuole esculdere dalla combinazione il *censoring*).
Un'osservazione importante è che la scelta di utilizzare la statistica di Gray, è stata arbitraria, avremmo potuto riferirci anche a un'altra statistica parametrica, come quella di Pepe-Mori, ma si è preferito il test di Gray in quanto è quello più utilizzato in letteratura. Infatti nonostante NPC sia una procedura non parametrica si appoggia comunque a una statistica test parametrica (in Arboretti et al. [2017] la statistica di riferimento è quella del log-rank).

3.3 La procedura NPC in ambiente SAS

L'algoritmo descritto nella precedente sezione è stato implementato in ambiente **SAS**® in modo da creare un test d'ipotesi utilizzabile e confrontabile con quelli descritti in precedenza. All'interno del file `NPC.sas` è stata implementata l'intera procedura NPC, attraverso la definizione di due funzioni macro: `%npc`, in cui si effettua il calcolo delle statistiche di Gray e si eseguono le permutazioni, e la funzione `%comp_pval`, utilizzata per il computo dei p-value e delle statistiche p-value like (il codice è inserito in appendice).
In fugura 3.1 sono riportati i p-value del test d'ipotesi applicato al dataset sul linfoma follicolare, fissando un numero di permutazioni $B = 1000$. PG1 e PG2 risultano perfettamente consistenti con i risultati evidenziati dal test di Gray eseguito su questi stessi dati in sezione 2.1.1. Entrambi i pvalue relativi alla procedura NPC mostrano una differenza significativa tra i due gruppi sperimentali. Considerazioni analoghe possono esser fatte riferendosi

Oss	PG0	PG1	PG2	pGF	pGT	pGF12	pGT12
1	0.048452	0.11239	.000499500	.001498501	.002497502	.001498501	.001498501

Figura 3.1: Risultati del test NPC sul dataset follic con $B = 1000$

ai dati del linfoma follicolare (figura 3.2) rispetto ai gruppi *AML-Low Risk* e *AML-High Risk* (fissando sempre $B = 1000$).

Oss	PG0	PG1	PG2	pGF	pGT	pGF12	pGT12
1	.005494505	.000499500	0.99750	.002497502	.002497502	.008491508	.001498501

Figura 3.2: Risultati del test NPC sul dataset bmt per i gruppi *AML-Low Risk* e *AML-High Risk* con $B = 1000$

Capitolo 4

Studio di simulazione per il confronto tra i vari metodi

Dopo aver introdotto i metodi principali presenti in letteratura per effettuare test d'ipotesi in presenza di rischi competitivi e dopo aver tentato di proporre una nuova procedura da affiancare a quelle esistenti, si propone adesso uno studio di simulazione per mettere a confronto le tre principali metodologie: test di Gray, test di Pepe-Mori e l'algoritmo NPC. Prima però di entrare nel dettaglio dello studio da cui siamo partiti e descrivere le simulazioni effettivamente svolte, abbiamo ritenuto necessario fare una piccola premessa sulla tecnica di simulazione adoperata per generare i dataset a cui applicare i vari test d'ipotesi. La modalità che si decide di adottare per simulare i dati, influenza infatti la possibilità d'interpretare correttamente un modello, che altrimenti potrebbe poi essere inadeguato per l'applicazione in contesti reali.

4.1 Simulazione di dati in presenza di rischi competitivi

Esistono diversi modi per simulare dati in presenza di rischi competitivi, come ad esempio quello dei *failure time* latenti, in cui s'immagina che i tempi in cui si verificano i vari eventi siano delle variabili aleatorie ciascuna con la propria distribuzione. In fase di simulazione, per ogni soggetto si prende il

minimo di questi tempi e si identifica in tal modo il tipo di evento che si è verificato. In tale contesto la soppravvivenza sarà rappresentata dalla distribuzione congiunta dei vari *failure time* latenti. Questo modello però è stato fortemente criticato perché pone un problema di non identificabilità [Tsiatis, 1975]: per un set di marginali specificato esistono in realtà infinite possibili scelte di distribuzione congiunta, ottenibili andando a variare arbitrariamente il grado di dipendenza tra le varie distribuzioni marginali.
Il modello a cui qui ci riferiamo è basato invece sulla simulazione dei *cause-specific hazard* [Beyersmann et al., 2009]. Si tratta in questa prospettiva, di modellare il verificarsi di un evento attraverso un diagramma multistato, in cui un soggetto, partendo da uno stato iniziale potrà passare a trovarsi a uno stato 1 (l'evento d'interesse) oppure uno stao 2 (il rischio competitivo). La propensione a raggiungere uno stato piuttosto che l'altro è appunto determinata dai *cause-specific hazard* dei due eventi. Dal momento che le possibili cause di *failure* sono modellate facendo riferimento a una sola variabile, non si pone il problema di dover applicare il concetto d'indipendena tra variabili aleatorie. Riferendosi inoltre ai *cause-specific hazard*, come ampiamente discusso nel primo capitolo, si ha la certezza di poter determinare completamente il comportamento stocastico sotteso al processo in cui intervengono i rischi competitivi (la *event-free survival* (1.22) è completamente determinata dai $\lambda_k(t)$).
Per decidere, in fase di simulazione, quale sia la probabilità con la quale si verifica ad esempio l'evento principale, è possibile partendo dalla definizione (1.19) di cause-specific hazard fare riferimento alla seguente relazione:

$$\mathbb{P}(E = 1|T \in \delta t,\, T \geq t) = \frac{\mathbb{P}(T \in \delta t,\, K = 1|T \geq t)}{\mathbb{P}(T \in \delta t|T \geq t)} = \frac{\lambda_1(t)}{\lambda_1(t) + \lambda_2(t)}. \quad (4.1)$$

L'interpretazione di questo risultato appare piuttosto evidente: assumendo che un soggetto subisca un evento al tempo t, la probabilità che quest'ultimo sia quello principale è pari alla proporzione rappresentata da $\lambda_1(t)$ rispetto al rischio complessivo di un qualunque evento.
Fatta questa premessa, Beyersmann et al. [2009] ricavano un algoritmo per la simulazione di dati in presenza di rischi competitivi:

1. Si stabilisce la forma funzionale di $\lambda_1(t)$ e $\lambda_2(t)$.

2. Si simula il *survival time* T con un rischio complessivo $\Lambda(t) = \lambda_1(t) + \lambda_2(t)$.

3. Per il tempo simulato t si genera una *Bernoulli* di parametro $\lambda_1(t)/(\lambda_1(t)+\lambda_2(t))$ in riferimento all'evento 1 (e di conseguenza viene anche generato l'evento 2).

4. Si genera a posteriori la variabile aleatoria C che definisce la distribuzione del *censoring*.

4.2 Lo studio di simulazione

La situazione sperimentale a cui è stato fatto riferimento per proporre un nuovo studio di simulazione, è tratta dal lavoro di Freidlin and Korn [2005], all'interno del quale però è stato effettuato un confronto tra il test di Gray e il log-rank. Gli autori hanno cercato di mostrare come il test di Gray risulti maggiormente in grado di non rifiutare l'ipotesi nulla di uguale sopravvivenza quando l'evento principale agisce in maniera non significativamente diversa tra i due gruppi sperimentali, e allo stesso tempo si riscontra una differenza significativa per quel che riguarda l'evento competitivo. Alla luce di quanto ampiamente discusso nel secondo capitolo, sappiamo che non è il log-rank test che si comporta meglio del test di Gray, è che quest'ultimo tiene conto del rischio competitivo nel valutare l'incidenza dell'evento d'interesse, il log-rank invece tratta come censure gli individui che sperimentano l'evento 2 (determinando come più volte sottolineato un'interpretazione dei fenomeni coinvolti di natura differente).
Al di là di queste considerazioni, ci interessa soffermarci sul piano sperimentale impostato da Freidlin and Korn [2005], poiché è da qui che siamo partiti per sviluppare il nostro studio di simulazione. I due autori hanno utilizzato un setting da 5 scenari, in cui si è sempre fatto uso di *cause-specific hazard* esponenziali sia per l'evento d'interesse, sia per il rischio competitivo.

Sc.1: $\lambda_1^A = 1$, $\lambda_1^B = 1$, $\lambda_2^A = 1$, $\lambda_2^B = 1.2$

Sc.2: $\lambda_1^A = 1$, $\lambda_1^B = 1$, $\lambda_2^A = 1$, $\lambda_2^B = 1.6$

Sc.3: $\lambda_1^A = 1$, $\lambda_1^B = 1.6$, $\lambda_2^A = 1$, $\lambda_2^B = 1$

Sc.4: $\lambda_1^A = 1$, $\lambda_1^B = 1.6$, $\lambda_2^A = 1$, $\lambda_2^B = 1.2$

Sc.5: $\lambda_1^A = 1$, $\lambda_1^B = 1.6$, $\lambda_2^A = 1$, $\lambda_2^B = 1.6$

Nello studio sono state effettuate 10000 repliche con dataset di taglia pari a 200 (per ciascuno scenario è stata indotta anche una correlazione di volta in volta differente, aspetto di cui noi non teniamo conto perché il nostro impianto di simulazione non è basato sul modello dei *failure time* latenti). Da questo setting sperimentale sono stati estrapolati gli scenari 1, 2 e 3, in più è stato simulato uno scenario 0 in cui tutti i *cause-specific hazard* sono stati posti uguali 1. I test sono stati eseguiti mettendo a confronto, tra i due gruppi sperimentali, solo l'incidenza cumulativa dell'evento principale.
All'interno del nostro studio di simulazione, in tutti gli scenari è stato indotto unicamente del *censoring* distribuito come una variabile aleatoria uniforme nell'intervallo $]0, 5[$ e quindi non informativo (a differenza di quanto fatto da Arboretti et al. [2017], in cui l'utilizzo di scenari che presentassero anche *informative censoring* si è rivelato importante per mostrare l'ottimo comportamento della metodologia NPC). È stato inoltre fissato l'istante di tempo pari a 5 come termine per il *censoring* amministrativo: gli individui che hanno sperimentato l'evento interesse oppure il rischio competitivo in un istante di tempo maggiore a 5 sono stati trattati come osservazioni censurate. Per tutti e tre gli scenari sono state eseguite 350 repliche di dataset con 25 soggetti per il gruppo A e 20 per il gruppo B. Nell'applicare la metodologia NPC il numero di permutazioni è stato sempre fissato a 1000 (le cifre appena illustrate sono fortemente influenzate dagli strumenti di calcolo impiegati, nell'effettuare lo studio abbiamo avuto l'opportunità di sfruttare un server dell'Università degli Studi di Padova, ma per ragioni di carattere temporale e computazionale non è stato possibile poter usufruire di numeri più grandi).

4.3 Struttura computazionale dello studio

In questa sezione s'intende delineare brevemente l'infrastruttura software progettata per eseguire l'intero studio di simulazione. È stato anzitutto

realizzato un primo file, `settings.sas`, all'interno del quale è stato possibile diversificare alcune impostazioni riguardanti la simulazione, in base al tipo di utente che esegue l'intera routine (se lo studio di simulazione viene eseguito su server si possono considerare numeri di repliche e permutazioni più grandi rispetto all'esecuzione in locale e si possono simulare più scenari in una sola chiamata. Inoltre sul server che abbiamo utilizzato era disponibile solo la versione **9.3** di **SAS®**, perciò è stato necessario differenziare una porzione di codice rispetto all'esecuzione in locale, su cui è stata resa invece disponibile la versione **9.4**). Nel file `simulation.sas` viene anzitutto definito il dataset contenente tutti gli scenari precedentemente descritti. Compare poi la funzione macro `simulation.sas` che contiene l'esecuzione effettiva dello studio di simulazione: per tutte le repliche che si desiderano eseguire, viene generato il dataset con le caratteristiche specificate nella sezione 4.2 (riguardanti la simulazione del tempo in cui si verifica uno dei due eventi, l'inserimento del *censoring*, ecc.). Successivamente partendo da tali dati vengono eseguite e collezionate le statistiche test relative ai metodi di Gray, Pepe-Mori e procedura NPC (facendo utilizzo del già citato file `NPC.sas`). Infine nel file `stime_potenza.sas` è stata inserita la funzione macro `%compare_test` che esegue i confronti tra i tre test per uno specifico scenario.
Tutti gli script prodotti sono stati trascritti in appendice.

4.4 Risultati principali della simulazione

Dopo aver effettuato la simulazione dei quattro scenari citati nella sezione 4.2, è stata eseguita un'analisi qualitativa per confrontare complessivamente le performance dei tre test d'ipotesi sottoposti a studio. Tale confronto è stato possibile facendo riferimento al grafico in cui la probabilità di rifiuto empirica viene plottata per diversi valori di α nominale (tale probabilità viene semplicemente calcolata riferendosi ai p-value relativi a ciascuna replica di tutti i test, e dividendo il numero di volte in cui tali p-value risultano minori dell'α nominale per il numero totale di repliche).
Lo scenario 1 (figura 4.1), in cui differisce lievemente tra i due gruppi solo l'incidenza del rischio competitivo, tutti i test si assestano intorno alla bisettrice, con i due metodi NPC che presentano una probabilità di rifiuto

leggermente più alta. Nel secondo scenario (figura 4.2), in cui la deviazione

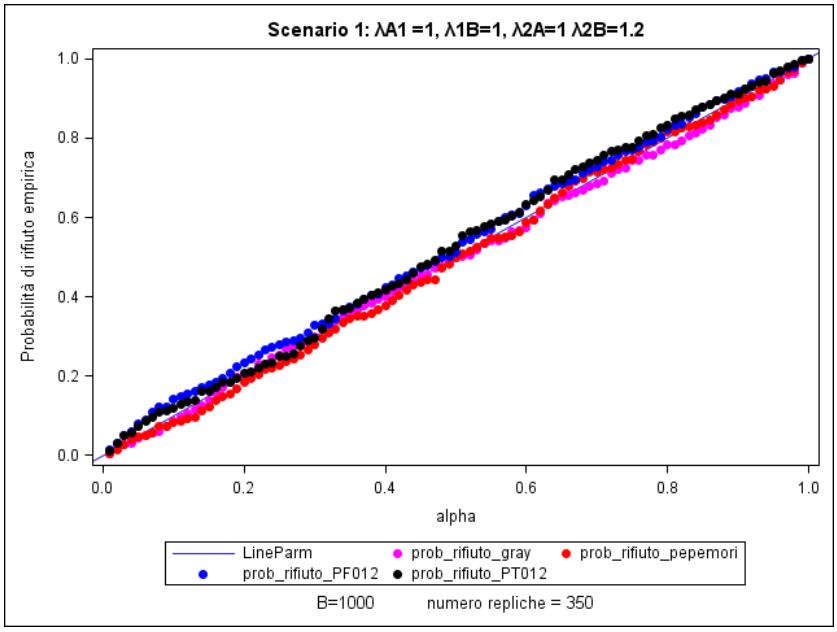

Figura 4.1: Confronto tra test di Gray, Pepe-Mori e NPC per lo scenario di simulazione 1

dall'ipotesi nulla per il rischio competitivo risulta più marcata (ricordiamo che i test sono eseguiti solamente sulle funzioni d'incidenza cumulativa relative all'evento 1), il test di Gray e il Pepe-Mori si assestano nuovamente intorno al α nominale, mentre i due test NPC esibiscono palesemente una probabilità di rifiuto più alta (con una prima fase in cui il metodo di Fisher sembra rifiutare maggiormente). Nello scenario 3 (figura 4.3), di marcato rifiuto dell'ipotesi nulla rispetto all'evento 1, il test di Pepe-Mori appare quello con minor capacità di rifiuto, assestandosi al di sotto degli altri. Eccetto una prima parte, in cui il metodo Fisher, sembra assestarsi al di sopra degli altri, le due metodologie NPC e il test di Gray appaiono comportarsi più o meno nello stesso modo.

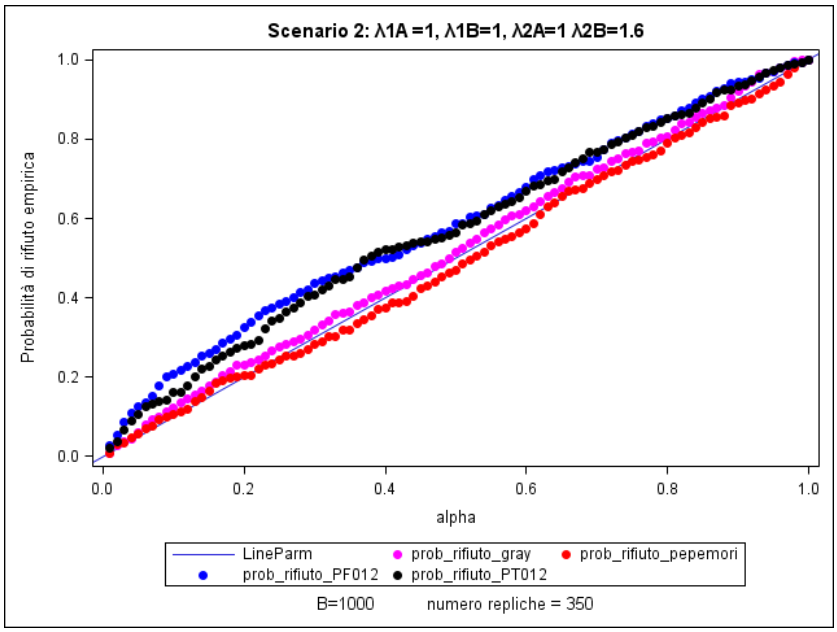

Figura 4.2: Confronto tra test di Gray, Pepe-Mori e NPC per lo scenario di simulazione 2

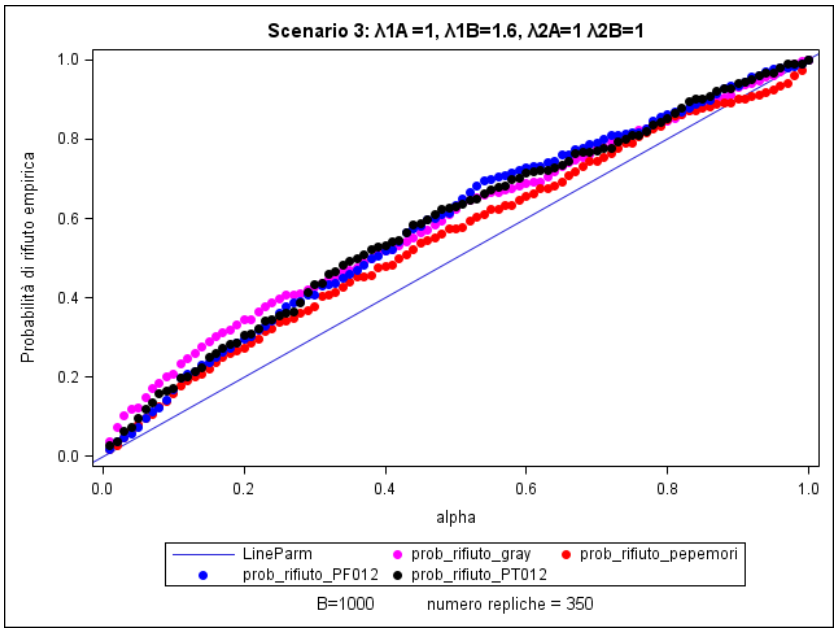

Figura 4.3: Confronto tra test di Gray, Pepe-Mori e NPC per lo scenario di simulazione 3

Conclusioni

All'interno di questa trattazione si è inteso far luce sugli aspetti che riguardano un ambito ristretto dell'analisi della sopravvivenza: l'utilizzo dei test d'ipotesi, non nel setting classico dell'analisi bensì nel contesto dei rischi competitivi. Dopo aver passato in rassegna i metodi principali presenti in letteratura, si è tentato di presentare un metodo innovativo basato su NPC, come estensione di un lavoro precedente [Arboretti et al., 2017], svolto in un contesto di analisi classica. Il confronto tra i vari test ha messo in luce che le tecniche basate su procedura NPC, per il fatto di separare l'analisi per ciascuno dei possibili eventi che possono verificarsi, sono più adatte per riconoscere una differenza complessiva nel comportamento dei due gruppi sperimentali. C'è da dire tuttavia che si tratta di un primo risultato esplorativo, il quale andrebbe approfondito attraverso ulteriori studi di simulazione. Sarebbe anzitutto necessario fare riferimento a numeri più grandi sia per quanto riguarda le repliche, sia per ciò che concerne il numero di permutazioni delle procedure NPC ma per ragioni di carettere temporale e computazionale ciò non è stato possibile all'interno di questo lavoro. Un'altra questione da prendere in considerazione riguarda gli scenari simulati: si è fatto utilizzo esclusivo di *cause-specific hazard* costanti, che rappresentano sicuramente la situazione più semplice e compatta da descrivere, ma non quella più realistica. Sarebbe perciò necessario riflettere sull'utilizzo di λ dipendenti dal tempo. Inoltre non sono stati studiati setting con la presenza di *censoring* informativo, il che sarebbe stato utile per arricchire la valutazione delle performance dei vari metodi.
Sulla procedura NPC ci sono alcune riflessioni che vale la pena riportare. La prima riguarda il fatto che per certi versi tale metodologia non è del tutto

standardizzata ma si presta ad adattarsi ad opportune scelte metodologiche del ricercatore. Si pensi in primo luogo alla scelta della statistica di Gray, all'interno di questo lavoro è stata ritenuta la più adatta per realizzare un primo test d'ipotesi NPC. Tuttavia il confronto con esperti di analisi della sopravvivenza potrebbe portare alla scelta di statistiche diverse in specifiche situazioni. Anche la scelta delle funzioni di combinazione di Tippet e Fisher potrebbe essere modificata; in Arboretti et al. [2017] sono state individuate come le migliori da utilizzare in quel contesto sperimentale tuttavia in presenza di rischi competitivi potrebbe essere opportuno riflettere sulla possibilità di vagliare delle alternative.
Nonostante tutti questi spunti di miglioramento e approfondimento, si è comunque mostrato all'interno di questa tesi la potenzialità di un metodo innovativo che potrebbe essere utilizzato (per la sua portabilità e facilità d'implementazione) in analisi della sopravvivenza quando la prospettiva risulta di carattere multivariato.

Appendice A - Lo studio di simulazione: simulation.sas

Si riporta di seguito il codice **SAS®** utilizzato per implementare lo studio d'implementazione. Vi sono alcuni riferimenti a variabili macro che sono state definite e settate in uno script di impostazioni ma la cui interpretazione è abbastanza ovvia.

```
/*Il path in cui salvare i risultati della simulazione
 dipende dallo utente che esegue la simulazione tramite la
 macro %settings*/

libname results "&path_risultati";

/* Nel seguente dataset sono presenti i parametri relativi agli
scenari possibili che si possono simulare (tale formato va bene
soprattutto per cause-specific hazard costanti). La scelta di quali
e quanti scenari si vogliono mandare in esecuzione vengono impostate
nella macro settings tramite le variabili &lista_scenari e &nscenari.
t_admin rappresenta il tempo oltre il quale scatta il censoring
amministrativo. Il censoring viene sempre simulato come una variabile
uniforme di estremi 0 e theta_c */

data scenari;
input id_scen t_admin n_a n_b lambda_A1 lambda_A2 lambda_B1
```

```
lambda_B2 theta_c eol $;
datalines;
1 5 25 20 1 1 1 1.2 5 x
2 5 25 20 1 1 1 1.6 5 x
3 5 25 20 1 1 1.6 1 5 x
4 5 25 20 1 1 1.6 1.2 5 x
5 5 25 20 1 1 1.6 1.6 5 x
6 5 25 20 1 1 1 1 5 x
;
run;
data scenari;
set scenari;
drop eol;
run;

/* La macro esegue lo studio di simulazione per tutti gli scenari
selezionati dal dataset di partenza*/

%macro simulation;
%do iiii=1 %to &nscenari;
%let id_scen=%scan(&lista_scenari,&iiii);
data _NULL_;
set scenari;
if id_scen=&id_scen then do;
call symput("n_a",n_a);
call symput("n_b",n_b);
call symput("lambda_A1",lambda_A1);
call symput("lambda_A2",lambda_A2);
call symput("lambda_B1",lambda_B1);
call symput("lambda_B2",lambda_B2);
call symput("t_admin",t_admin);
```

```
call symput("theta_c",theta_c);
end;
run;

/* Inizializzazione di tre dataset che conterranno i p-values
per i tre test eseguiti*/
%let scen=&id_scen;
proc datasets library = results nolist;
delete final_report&scen NPC_pvalues_scen&scen
Pepemori_pvalues_scen&scen Gray_pvalues_scen&scen;
run;

%do rep=1 %to &n_sim;

/*Risulta necessario cancellare il ds con i p-value di Gray dato
che richiamato molte volte al di fuori di una procedura*/
proc datasets library = work nolist;
delete output_Gray;
run;
DM "clear log";

data simul_data_scen&scen.rep&rep( keep= id_paz gruppo tempo
tempo_cens evento );
/*call streaminit(4321 + &rep)*/
mu_A1 = 1 / &lambda_A1;
mu_A2 = 1 / &lambda_A2;
mu_B1 = 1 / &lambda_B1;
mu_B2 = 1 / &lambda_B2;
pA = %sysevalf( &lambda_A1 /(&lambda_A1+ &lambda_A2) );
pB = %sysevalf( &lambda_B1 /(&lambda_B1+ &lambda_B2) );
do id_paz = 1 to %eval(&n_A+&n_B);
```

```
if id_paz le &n_A then do;
gruppo = "A";
tempo = (mu_A1+mu_A2) * rand("Exponential");
evento = rand("Binomial", pA , 1);
if evento = 0 then evento = 2;
tempo_cens = 0 + (&theta_c-0)*rand("Uniform");
if tempo > &t_admin or tempo_cens < tempo then evento = 0;
output;
end;
else do;
gruppo = "B";
tempo = (mu_B1+mu_B2) * rand("Exponential");
evento = rand("Binomial", pB , 1);
if evento = 0 then evento = 2;
tempo_cens = 0 + (&theta_c-0)*rand("Uniform");
if tempo > &t_admin or tempo_cens < tempo then evento = 0;
output;
end;
end;
run;

/* Esecuzione del test di Gray per i due eventi col salvataggio
della statistica test e del p-value (ci sono due possibili
esecuzioni a seconda della versione di SAS) */

%if &versione_sas = _9_4 %then %do;
proc lifetest data = simul_data_scen&scen.rep&rep plots= &grafici;
time tempo*evento(0)/ eventcode = 1 2;
strata gruppo;
ods output GrayTest = output_Gray;
run;
```

```
%end;
%else %if &versione_sas = _9_3 %then %do;
%CIF(data=simul_data_scen&scen.rep&rep, time=tempo, status=evento,
group=gruppo, event=1, censored=0, alpha=.05, options=NOPLOT)
data testresult; set testresult; Failcode = 1; run;
proc append base = output_Gray data = testresult( rename =
(test_stat = ChiSq pval = ProbChiSq df = DF) ); run;
%CIF(data=simul_data_scen&scen.rep&rep, time=tempo, status=evento,
group=gruppo, event=2, censored=0, alpha=.05, options=NOPLOT)
data testresult; set testresult; Failcode = 2; run;
proc append base = output_Gray data = testresult( rename =
(test_stat = ChiSq pval = ProbChiSq df = DF) ); run;
%end;
proc append base = results.Gray_pvalues_scen&scen data = output_Gray;
run;

/* Esecuzione del test di Pepe-Mori per i due eventi col salvataggio
della statistica test e del p-value*/

%compcif(ds= simul_data_scen&scen.rep&rep, time= tempo, cens= evento,
group= gruppo,val1= "A",val2= "B")
quit;

data output_Pepemori;
keep failcode pvalue;
set result;
failcode=1;
run;
proc append base = results.Pepemori_pvalues_scen&scen
data = output_Pepemori; run;
```

```
data tmp;
drop evento;
set simul_data_scen&scen.rep&rep;
nevento=evento;
if evento=1 then nevento=2;
if evento=2 then nevento=1;
run;

data tmp1;
set tmp;
rename nevento=evento;
label nevento=evento;
run;

%compcif(ds= tmp1, time= tempo, cens= evento, group= gruppo,
val1= "A",val2= "B")
quit;
data output_Pepemori;
keep failcode pvalue;
set result;
failcode=2;
run;
proc append base = results.Pepemori_pvalues_scen&scen
data = output_Pepemori; run;
DM "clear log";

/* Esecuzione della procedura NPC col salvataggio dei
vari p-value */

%NPC(ds=simul_data_scen&scen.rep&rep,tempo=tempo,evento=evento,
gruppo=gruppo)
```

```
proc append base = results.NPC_pvalues_scen&scen
data = final_result_NPC; run;
%end;

/* Inseriamo in due variabili distinte i p-value del
test di Gray */

data Gray_pvalues_scen&scen._1;
keep gray_pval1;
set results.Gray_pvalues_scen&scen;
if Failcode=1 then do; gray_pval1=ProbChiSq; output; end;
run;

data Gray_pvalues_scen&scen._2;
keep gray_pval2;
set results.Gray_pvalues_scen&scen;
if Failcode=2 then do; gray_pval2=ProbChiSq; output ; end;
run;

data results.Gray_pvalues_scen&scen;
merge Gray_pvalues_scen&scen._1 Gray_pvalues_scen&scen._2;
run;

DM "clear log";
/* Inseriamo in due variabili distinte i p-value del
test di Pepe-Mori */

data Pepemori_pvalues_scen&scen._1;
keep Pepemori_pval1;
set results.Pepemori_pvalues_scen&scen;
if Failcode = 1 then do; Pepemori_pval1 = pvalue;
```

```
output; end;
run;

data Pepemori_pvalues_scen&scen._2;
keep Pepemori_pval2;
set results.Pepemori_pvalues_scen&scen;
if Failcode = 2 then do; Pepemori_pval2 = pvalue; output; end;
run;

data results.Pepemori_pvalues_scen&scen;
merge Pepemori_pvalues_scen&scen._1
Pepemori_pvalues_scen&scen._2;
run;

DM "clear log";

/*Creiamo il dataset col report definitivo*/
data assembla;
merge results.Gray_pvalues_scen&scen
results.NPC_pvalues_scen&scen
results.Pepemori_pvalues_scen&scen;
run;
proc append base = results.final_report&scen data = assembla;
run;
%end;

%mend simulation;
```

Appendice B - La procedura npc: NPC.sas

Segue adesso la porzione di codice usata per implementare una metodologia NPC basata sul test di Gray.

```
/* Implementiamo prima una routine che consente di effettuare
il calcolo dei p-value e dei p-value-like, da utilizzare
nella routine NPC (tipo = pval_like o pval) */

%macro comp_pval(type,ds,var);

proc datasets nolist;
delete result_conteggio;
run;

%if &type = pval_like %then %let eoc = (&B + 1);
%if &type = pval %then %let eoc = 1;

%do scanner=1 %to &eoc;
/* A ogni ripetizione del ciclo si punta alla osservazione
i-esima e la si fa diventare una MV soglia */
data _NULL_;
ii = &scanner;
set &ds point=ii;
```

```
call symput('soglia',&var);
stop;
run;

/* Fissata la soglia si scorre tutto il ds e si contano i superi.
Alla ultima iterazione del datastep il conteggio, che contiene
il numero tot di superi, viene trasformato in MV */
data _NULL_;
set &ds;
if _N_^ = &scanner and &var >= &soglia then count+1;
call symput('n_superi',count);
run;

data tmp;
scanner=&scanner;
soglia = &soglia;
n_superi  = &n_superi;
pval=(1/2+&n_superi)/(&B+1);
run;
proc append data=tmp base=result_conteggio; run;
%end;
%mend;

/* La procedura NPC viene implementata usando la statistica
test di Gray come riferimento */

%macro NPC(ds,tempo=,evento=,gruppo=);

/* Eliminiamo tre dataset che conterranno le statistiche
test di Gray per ciascuno dei tre eventi */
proc datasets library=work nolist;
```

```
delete chi_0 chi_1 chi_2;
run;

/* Calcolo della statistica test di Gray per la censura.
In questo caso il censoring viene settato a 3 (valore fittizio)
mentre lo evento principale e il competitivo sono trattati
come rischi competitivi */

%if &versione_sas = _9_4 %then %do;
proc lifetest data = &ds plots = &grafici;
time &tempo*&evento(3) / eventcode = 0;
strata &gruppo;
ods output GrayTest = output_G_NPC(keep= failcode chisq);
run;
proc append base = chi_0 data = output_G_NPC; run;
%end;

%else %if &versione_sas = _9_3 %then %do;
%CIF(data=&ds, time=&tempo, status=&evento,
group=&gruppo, event=0, censored=3, alpha=.05, options=NOPLOT)
data testresult; set testresult(keep= test_stat);
Failcode = 0; run;
proc append base = chi_0 data = testresult
( rename = (test_stat = ChiSq) ); run;
%end;

/* Calcolo della statistica Gray per lo evento principale e il
rischio competitivo. In questo caso il censoring
viene settato ovviamente in corrispondenza dello 0 */

%if &versione_sas = _9_4 %then %do;
```

```
proc lifetest data = &ds plots = &grafici;
time &tempo*&evento(0) / eventcode = 1 2;
strata &gruppo;
ods output GrayTest = output_G_NPC(keep= failcode chisq);
run;
proc append base = chi_1 data = output_G_NPC
( where = (failcode=1) ); run;
proc append base = chi_2 data = output_G_NPC
( where = (failcode=2) ); run;
%end;

%else %if &versione_sas = _9_3 %then %do;
%CIF(data=&ds, time=&tempo, status=&evento, group=&gruppo,
event=1, censored=0, alpha=.05, options=NOPLOT)
data testresult; set testresult(keep= test_stat); Failcode = 1; run;
proc append base = chi_1 data = testresult
( rename = (test_stat = ChiSq) ); run;
%CIF(data=&ds, time=&tempo, status=&evento, group=&gruppo,
event=2, censored=0, alpha=.05, options=NOPLOT)
data testresult; set testresult(keep= test_stat); Failcode = 2; run;
proc append base = chi_2 data = testresult( rename =
(test_stat = ChiSq) ); run;
%end;

/* Adesso questa routine viene ripetuta permutando il
dataset B volte */
%do P = 1 %to &B;
DM 'clear log';

/* Per permutare la appartenenza ai gruppi delle varie
osservazioni, alla configurazione iniziale dei gruppi
```

```
affianchiamo una variabile con osservazioni simulate
da una uniforme e ordiniamo il dataset ottenuto
rispetto ai numeri distribuiti uniformemente */

data da_permutare;
set &ds( keep = &gruppo );
rename &gruppo = &gruppo._perm&P;
/*call streaminit(123+&P);*/
ordine = rand("uniform");
run;
proc sort data = da_permutare out =
gruppi_permutati( drop = ordine );
by ordine;
run;

data dati_permutati( drop = &gruppo );
set &ds;
merge gruppi_permutati &ds;
run;

%if &versione_sas = _9_4 %then %do;
proc lifetest data = dati_permutati plots = &grafici;
time &tempo*&evento(3) / eventcode = 0;
strata &gruppo._perm&P;
ods output GrayTest = output_G_NPC(keep= failcode chisq);
run;
proc append base = chi_0 data = output_G_NPC; run;
proc lifetest data = dati_permutati plots = &grafici;
time &tempo*&evento(0) / eventcode = 1 2;
strata &gruppo._perm&P;
ods output GrayTest = output_G_NPC(keep= failcode chisq);
```

```
run;
proc append base = chi_1 data = output_G_NPC
( where = (failcode=1) ); run;
proc append base = chi_2 data = output_G_NPC
( where = (failcode=2) ); run;
%end;

%else %if &versione_sas = _9_3 %then %do;
%CIF(data=dati_permutati, time=&tempo, status=&evento,
group=&gruppo._perm&P, event=0, censored=3, alpha=.05,
options=NOPLOT)
data testresult; set testresult(keep= test_stat);
Failcode = 0; run;
proc append base = chi_0 data = testresult( rename =
(test_stat = ChiSq) ); run;
%CIF(data=dati_permutati, time=&tempo, status=&evento,
group=&gruppo._perm&P, event=1, censored=0, alpha=.05,
options=NOPLOT)
data testresult; set testresult(keep= test_stat);
Failcode = 1; run;
proc append base = chi_1 data = testresult( rename =
(test_stat = ChiSq) ); run;
%CIF(data=dati_permutati, time=&tempo, status=&evento,
group=&gruppo._perm&P, event=2, censored=0, alpha=.05,
options=NOPLOT)
data testresult; set testresult(keep= test_stat);
Failcode = 2; run;
proc append base = chi_2 data = testresult( rename =
(test_stat = ChiSq) ); run;

%end;
```

```
%end;

/* Calcoliamo i p-value e i p-value-like per i tre eventi */
%comp_pval(pval_like,chi_0,chisq)
data chi_0; merge chi_0 result_conteggio; run;

%comp_pval(pval_like,chi_1,chisq)
data chi_1; merge chi_1 result_conteggio; run;

%comp_pval(pval_like,chi_2,chisq)
data chi_2; merge chi_2 result_conteggio; run;

/* Uniamo i p-value-like per i tre eventi per poter
calcolare le funzioni di combinazione di Tippet
e Fisher */
data pval_0(keep = PG0);
set chi_0(rename = (pval = PG0));
run;
data pval_1(keep = PG1);
set chi_1(rename = (pval = PG1));
run;
data pval_2(keep = PG2);
set chi_2(rename = (pval = PG2));
run;
data output_NPC;
merge pval_0 pval_1 pval_2;
Tippet_012 = min(PG0,PG1,PG2);
Fisher_012 = - log(PG0) - log(PG1) - log(PG2);
Tippet_12 = min(PG1,PG2);
Fisher_12 = - log(PG1) - log(PG2);
run;
```

```
/* Combiniamo i Fisher e i Tippet sia considerando tutti
gli eventi
che prendendo in considerazione solo 1 e 2 */

%comp_pval(pval,output_npc,Fisher_012)
data f012;
keep PF_012;
set result_conteggio;
PF_012=pval;
run;

%comp_pval(pval,output_npc,Tippet_012)
data t012;
keep PT_012;
set result_conteggio;
PT_012=1 - pval; /* per i Tippet i valori devono essere
minori o uguali della soglia */
run;

%comp_pval(pval,output_npc,Fisher_12)
data f12;
keep PF_12;
set result_conteggio;
PF_12=pval;
run;

%comp_pval(pval,output_npc,Tippet_12)
data t12;
keep PT_12;
set result_conteggio;
PT_12=1 - pval; /* per i Tippet i valori devono essere
```

```
minori o uguali della soglia */
run;

data pgs;
keep pg0 pg1 pg2;
set output_npc;
if _n_=1 then output;
run;

data final_result_NPC;
merge pgs f012 t012 f12 t12;
run;

%mend NPC;
```

Appendice C - Le impostazioni utente: settings.sas

Di seguito lo script relativo alle impostazioni utente.

```
/* Definiamo le impostazioni da utilizzare nelle
simulazioni a seconda dell'utente che effettua
le prove. La settings consente di inserire
un nome utente, a cui si riferisce un set d'impostazioni
per effettuare le simulazioni */

%global path_risultati;
%global inserimento; * inserimento utente;
%global n_sim; * numero repliche simulazione;
%global lista_scenari nscenari; * lista e numero
scenari da simulare (presi dal simulation study;)
%global B; * numero permutazioni per NPC;
%global grafici; /* nel server di Padova non si possono
eseguire plot */
%global versione_sas; /* nelle versioni prima
della 9_4 la proc lifetest non implementa il test
di Gray (macro %CIF) */
%macro settings(name);
%let inserimento = &name;
```

```
%if &inserimento = Dante %then %do;
%let path_risultati = C:/Users/Dante/Desktop/thesis/results;
%let n_sim = 6;
%let lista_scenari =5;
%let nscenari = 1;
%let B = 4;
%let grafici = cif(test);
%let versione_sas = _9_4;
%end;

%else %if &inserimento = Roberto %then %do;
%let path_risultati = C:/Users/roberto/Documents/2018 tesi
futia/results;
%let n_sim = 2;
%let lista_scenari=1 3;
%let nscenari=2;
%let B = 3;
%let grafici = none;
%let versione_sas = _9_4;
%end;

%else %if &inserimento = Padova %then %do;
%let path_risultati = /home/dottgest/rischi_comp/results;
%let n_sim = 100;
%let lista_scenari =2;
%let nscenari = 1;
%let B = 1000;
%let grafici = none;
%let versione_sas = _9_3; %end;
%mend settings;
```

Appendice D - Il main dell'applicazione: main.sas

Di seguito il main della routine di simulazione.

```
*options symbolgen mcompilenote=all mprint;

/* FUTIA */
%let utente = Dante;
%let path_programmi_sas = C:/Users/Dante/Desktop/thesis;


/* FONTANA */
*%let utente=Roberto;
/*%let path_programmi_sas = C:/Users/roberto/Documents
/2018 tesi futia/programmi sas;*/

/* Padova */
*%let utente=Padova;
*%let path_programmi_sas = /home/dottgest/rischi_comp;


/* Assicurarsi che nel &path_programmi_sas ci siano
i seguenti file .sas:
gray_macro.sas
```

```
CompCif_corretta.sas
settings_finale.sas
simulation_study_finale.sas
npc_finale.sas    */

%include "&path_programmi_sas/gray_macro.sas";
%include "&path_programmi_sas/compcif_corretta.sas";
%include "&path_programmi_sas/settings_finale.sas";

/*Impostare l'utente che svolge la simulazione. I valori
possibili sono
Dante   Roberto   Padova */
%settings(&utente)

%include "&path_programmi_sas/npc_finale.sas";
%include "&path_programmi_sas/simulation_study_finale.sas";

ods html close;
ods html;

*Mandare in esecuzione lo studio di simulazione;
%simulation
```

Appendice E - Il confronto grafico tra i test: compare_test.sas

Di seguito il codice utilizzato per effettuare il confronto tra i vari test d'ipotesi.

```
ibname results "C:/Users/Dante/Desktop/thesis/results";

ods escapechar='~'; /* serve per usare le lettere greche */

%macro compare_test(ds);
proc datasets nolist;
delete rifiuto_estim_prob;
run;

proc sql noprint;
SELECT COUNT(*) INTO :size_ds
FROM &ds;
quit;

%do alpha = 1 %to 100 %by 1;
data _NULL_;
set &ds;
if gray_pval1 < &alpha/100 then count_gray + 1;
```

```
if pepemori_pval1 < &alpha/100 then count_pepemori + 1;
if PT_012 < &alpha/100 then count_PT012 + 1;
if PF_012 < &alpha/100 then count_PF012 + 1;
call symput('lessthenalpha_gray',count_gray);
call symput('lessthenalpha_pepemori',count_pepemori);
call symput('lessthenalpha_PT012',count_PT012);
call symput('lessthenalpha_PF012',count_PF012);
run;

data current_estim_prob;
alpha = &alpha/100;
prob_rifiuto_gray = &lessthenalpha_gray / &size_ds;
prob_rifiuto_pepemori = &lessthenalpha_pepemori / &size_ds;
prob_rifiuto_PT012 = &lessthenalpha_PT012 / &size_ds;
prob_rifiuto_PF012 = &lessthenalpha_PF012 / &size_ds;
run;

proc append data =  current_estim_prob base = rifiuto_estim_prob;
run;
%end;

%mend compare_test;

ods graphics on;
ods listing gpath = 'C:/Users/Dante/Desktop/thesis/monografia';

%compare_test(results.final_report1_350rep_1000b)
proc sgplot data=rifiuto_estim_prob;
title "Scenario 1: ~{unicode lambda}A1 =1, ~{unicode lambda}1B=1,
~{unicode lambda}2A=1 ~{unicode lambda}2B=1.2";
lineparm x=0 y=0 slope=1; /** intercept, slope **/
```

```
yaxis label="ProbabilitÃ  di rifiuto empirica";
scatter x=alpha y=prob_rifiuto_gray / markerattrs=
graphdata2(symbol = circlefilled size=8 color = fuchsia);
scatter x=alpha y=prob_rifiuto_pepemori / markerattrs=
graphdata2(symbol = circlefilled size=8 color = red);
scatter x=alpha y=prob_rifiuto_PF012 / markerattrs=
graphdata2(symbol = circlefilled size=8 color = blue);
scatter x=alpha y=prob_rifiuto_PT012 / markerattrs=
graphdata2(symbol = circlefilled size=8 color = black);
footnote "B=1000          numero repliche = 350";
run;

%compare_test(results.final_report2_350rep_1000b)
proc sgplot data=rifiuto_estim_prob;
title "Scenario 2: ~{unicode lambda}1A =1, ~{unicode lambda}1B=1,
~{unicode lambda}2A=1 ~{unicode lambda}2B=1.6";
lineparm x=0 y=0 slope=1; /** intercept, slope **/
yaxis label="ProbabilitÃ  di rifiuto empirica";
scatter x=alpha y=prob_rifiuto_gray / markerattrs=
graphdata2(symbol = circlefilled size=8 color = fuchsia);
scatter x=alpha y=prob_rifiuto_pepemori / markerattrs=
graphdata2(symbol = circlefilled size=8 color = red);
scatter x=alpha y=prob_rifiuto_PF012 / markerattrs=
graphdata2(symbol = circlefilled size=8 color = blue);
scatter x=alpha y=prob_rifiuto_PT012 / markerattrs=
graphdata2(symbol = circlefilled size=8 color = black);
footnote "B=1000          numero repliche = 350";
run;

%compare_test(results.final_report3_350rep_1000b)
proc sgplot data=rifiuto_estim_prob;
```

```
title "Scenario 3: ~{unicode lambda}1A =1,
~{unicode lambda}1B=1.6, ~{unicode lambda}2A=1
~{unicode lambda}2B=1";
lineparm x=0 y=0 slope=1; /** intercept, slope **/
yaxis label="ProbabilitÃ  di rifiuto empirica";
scatter x=alpha y=prob_rifiuto_gray / markerattrs=
graphdata2(symbol = circlefilled size=8 color = fuchsia);
scatter x=alpha y=prob_rifiuto_pepemori / markerattrs=
graphdata2(symbol = circlefilled size=8 color = red);
scatter x=alpha y=prob_rifiuto_PF012 / markerattrs=
graphdata2(symbol = circlefilled size=8 color = blue);
scatter x=alpha y=prob_rifiuto_PT012 / markerattrs=
graphdata2(symbol = circlefilled size=8 color = black);
footnote "B=1000          numero repliche = 350";
run;

%compare_test(results.final_report6_350rep_1000b)
proc sgplot data=rifiuto_estim_prob;
title "Scenario 0: ~{unicode lambda}1A =1,
~{unicode lambda}1B=1, ~{unicode lambda}2A=1
~{unicode lambda}2B=1";
lineparm x=0 y=0 slope=1; /** intercept, slope **/
yaxis label="ProbabilitÃ  di rifiuto empirica";
scatter x=alpha y=prob_rifiuto_gray / markerattrs=
graphdata2(symbol = circlefilled size=8 color = fuchsia);
scatter x=alpha y=prob_rifiuto_pepemori / markerattrs=
graphdata2(symbol = circlefilled size=8 color = red);
scatter x=alpha y=prob_rifiuto_PF012 / markerattrs=
graphdata2(symbol = circlefilled size=8 color = blue);
scatter x=alpha y=prob_rifiuto_PT012 / markerattrs=
graphdata2(symbol = circlefilled size=8 color = black);
```

```
footnote "B=1000          numero repliche = 350";
run;

ods graphics off;
```

Bibliografia

Odd Aalen. Nonparametric estimation of partial transition probabilities in multiple decrement models. *The Annals of Statistics*, pages 534–545, 1978.

Rosa Arboretti, Roberto Fontana, Fortunato Pesarin, and Luigi Salmaso. Nonparametric combination tests for comparing two survival curves with informative and non-informative censoring. *Statistical methods in medical research*, page 0962280217710836, 2017.

Ruta Bajorunaite and John P Klein. Comparison of failure probabilities in the presence of competing risks. *Journal of Statistical Computation and Simulation*, 78(10):951–966, 2008.

Jan Beyersmann, Aurelien Latouche, Anika Buchholz, and Martin Schumacher. Simulating competing risks data in survival analysis. *Statistics in medicine*, 28(6):956–971, 2009.

Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

David Collett. *Modelling survival data in medical research*. Chapman & Hall, 1994.

Boris Freidlin and Edward L Korn. Testing treatment effects in the presence of competing risks. *Statistics in medicine*, 24(11):1703–1712, 2005.

Ronald B Geskus. *Data analysis with competing risks and intermediate states*. CRC Press, 2015.

Robert J Gray. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics*, pages 1141–1154, 1988.

David W Hosmer, Stanley Lemeshow, and Susanne May. *Applied survival analysis*. John Wiley & Sons, 2008.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2002.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282): 457–481, 1958.

John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

David G Kleinbaum and Mitchel Klein. *Survival analysis*. Springer, 2010.

DY Lin. Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in medicine*, 16(8):901–910, 1997.

Margaret Sullivan Pepe. Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association*, 86(415): 770–778, 1991.

Margaret Sullivan Pepe and Motomi Mori. Kaplan-meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in medicine*, 12(8):737–751, 1993.

Fortunato Pesarin and Luigi Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.

Melania Pintilie. *Competing risks: a practical perspective*. John Wiley & Sons, 2006.

Anastasios Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22, 1975.

Tyler Ward and Zachary Weber. Two shades of gray: Implementing gray's test for equivalence of cif in sas 9.4. In *MWSUG 2016 Conference Proceedings*, October 2016.