Universitá della Calabria
Dipartimento di Ingegneria Meccanica, Energetica e Gestionale

Corso di Dottorato in Ricerca Operativa

Tesi di Dottorato di Ricerca

Modelli Integrati di Analisi di Sopravvivenza applicati alla Prognosi del Trapianto Renale

Candidato:
Ing. Danilo Lofaro

Relatore:
Prof. Domenico Conforti

Correlatore:
Ing. Rosita Guido

Anno Accademico 2012–2013

Dedicato a Dori.

INDICE

ELENCO DELLE FIGURE

ELENCO DELLE TABELLE

INTRODUZIONE

Il trapianto di rene è il trattamento di scelta per i pazienti con una insufficienza renale terminale (ESRD) che necessita di un trattamento dialitico sostitutivo (RRT) [RABBAT *et al.*, 2000; Schnuelle *et al.*, 1998; Wolfe *et al.*, 1999].

Nonostante un costante miglioramento della sopravvivenza dei trapianti renali negli ultimi due decenni [Gondos *et al.*, 2013; Hariharan *et al.*, 2002], la durata a lungo termine del trapianto è ancora una delle problematiche non risolte per la comunità trapiantologica [Colvin, 2003; Meier-Kriesche *et al.*, 2004; Pascual *et al.*, 2002] principali e circa il 20% dei pazienti sottoposti a dialisi ha ricevuto almeno un trapianto in precedenza [Rao, Schaubel, Jia *et al.*, 2007; Rao, Schaubel e Saran, 2005].

il problema della long-term survival

Per avere sia un utilizzo ottimale degli organi disponibili, che il più basso rischio di perdita del trapianto, diverse decisioni importanti devono essere prese rispetto a fattori *(pre-trapianto)* quali il tipo di donatore (vivente o cadavere), i criteri di allocazione (compatibilità, durata della dialisi, età del donatore) e la tempistica del trapianto stesso (età del paziente). L'Identificazione prima del trapianto di pazienti potenzialmente a maggior rischio di perdita dell'organo potrebbe, infatti, aiutare sia i medici che le famiglie nel fare queste scelte potenzialmente decisive [Ponticelli e Graziani, 2012]. Per queste ragioni, la valutazione dei fattori pre-trapianto, come età, compatibilità HLA, tipo donatore ed età dialitica assume sempre maggior importanza [Fellström *et al.*, 2005; Rao, Schaubel, Guidinger *et al.*, 2009; Terasaki e Ozawa, 2004]. Inoltre, è noto che l'andamento nel primo periodo post-trapianto, in termini di recupero precoce della funzionalità renale e normalizzazione dei parametri clinici, è un indicatore molto importante della prognosi a lungo termine del graft. Anche questo potrebbe essere di aiuto per i pazienti e le loro famiglie nella valutazione e programmazione del follow-up [First, 2003; Hariharan *et al.*, 2002; Kasiske *et al.*, 2005; Quiroga *et al.*, 2006; Terasaki e Ozawa, 2004].

fattori pre-trapianto

fattori clinici post-trapianto

Sia i fattori pre-trapianto che i dati sull'andamento clinico nell'immediato post-trapianto [Kasiske *et al.*, 2005; Quiroga *et al.*, 2006; Toma *et al.*, 2001], sono stati precedentemente valutati come predittori dell'outcome del trapianto renale a lungo terminie. Tuttavia, gli studi fin'ora si limitano alla valutazione dei singoli fattori e poco si sa circa il valore predittivo delle loro interazioni sulla prognosi. Inoltre, sono pochi i lavori che hanno indagato se queste interazioni possano identificare specifici sottogruppi di pazienti con un rischio particolarmente alto o basso di fallimento del trapianto [Goldfarb-Rumyantzev *et al.*, 2003; Krikov *et al.*, 2007].

Il modello di Cox

La stragrande maggioranza dei lavori che studiano le variabili di sopravvivenza con dati *censurati* utilizzano il modello di regressione di Cox e le sue estensioni per l'analisi delle relazioni funzionali. Queste tecniche di statistica parametrica (e semi-parametrica) sono molto utili, perché consentono interpretazioni semplici degli effetti di variabili e possono facilmente essere utilizzate per l'inferenza (test di ipotesi e valutazione del rischio).Tuttavia, tali modelli sono limitati dal fatto che presuppongono forzatamente un legame specifico tra le variabili prese in esame e l'evento oggetto di studio. Anche se è possibile incorporare all'interno dei modelli le interazioni tra variabili, queste devono essere specificate a priori nella definizione del modello. Inoltre, in pratica, le conclusioni di carattere inferenziale sono fatte solo dopo la prova di molti modelli diversi e le proprietà statistiche di tale inferenza dopo la selezione del modello sono ancora in gran parte sconosciute.

limiti dei modelli di statistica classica

Nel caso in cui non si voglia imporre aprioristicamente una funzione che metta in relazione dati e eventi, sono disponibili approcci quantitativi sicuramente più flessibili.

Alberi di sopravvivenza

Gli alberi di sopravvivenza sono modelli non parametrici e rappresentano una valida alternativa ai modelli (semi-) parametrici. Lo sviluppo di alberi di sopravvivenza è cominciato dalla metà degli anni 80 fino alla metà della decade successiva, e ha avuto il principale obiettivo di estendere gli algoritmi per la creazione di alberi già esistenti al caso dell'analisi di sopravvivenza e dei dati censurati. Questi modelli offrono una grande flessibilità, in quanto sono in grado di rivelare automaticamente l'esistenza di alcuni tipi di interazioni tra variabili senza doverle specificare in anticipo. Inoltre, un singolo albero può raggruppare naturalmente i soggetti a secondo dell'andamento della sopravvivenza in base ai loro dati. Diventa, quindi, piuttosto semplice derivare dai modelli ad albero dei sottogruppi di individui omogenei per prognosi. Un'altra caratteristica che rende gli approcci basati su alberi maggiormente flessibili rispetto ai modelli di regressione è la possibilità di trattare i dati mancanti senza la necessità di doverli imputare o di escludere le osservazioni che presentino missing values.

L'obiettivo del presente lavoro è stato, quindi, quello di creare modelli basati sugli alberi di sopravvivenza di alberi di sopravvivenza per l'analisi della prognosi del trapianto renale sia nei pazienti pediatrici che adulti. In particolare

IL PRIMO CAPITOLO offre una visione d'insieme dell'analisi di sopravvivenza e dei principali strumenti statistici, con particolare attenzione al popolare modello a rischi proporzionali (regressione di Cox).

IL SECONDO CAPITOLO presenta il modello degli alberi di sopravvivenza, partendo da una breve revisione degli algoritmi svilup-

pati fino ad oggi e illustrando le tecniche più popolari per la crescita degli alberi e per il pruning.

IL TERZO CAPITOLO presenta l'applicazione degli alberi di sopravvivenza nel dominio del trapianto renale, partendo da una comparazione di due regole di split su un ampio dataset di trapianti renali pediatrici e illustrando l'applicazione di una strategia incrementale per l'inclusione di dati longitudinali su un campione di pazienti seguiti nel tempo.

1 ANALISI DI SOPRAVVIVENZA

1.1 INTRODUZIONE

Il termine *analisi di sopravvivenza* si riferisce ad approcci statistici di tipo non-parametrico o semi-parametrico progettati per prendere in considerazione la variabile tempo negli studi clinici di follow-up. In particolare si riferisce all'analisi del periodo tra l'inizio dell'osservazione e l'occorrenza di un determinato *evento* (da qui il terimne *time-to-event analysis*). Originariamente l'evento di interesse era la morte del soggetto, mentre ormai le applicazioni di questa tecnica considerano gli eventi più diversi: dallo sviluppo o la diagnosi di malattia, fino ai terremoti o ai crolli del mercato azionario. Questo tipo di modelli statistici sono infatti applicabili a tutti gli scenari in cui un evento sia rappresentabile come una transizione da uno *stato* discreto ad un altro in un dato istante di tempo. Questo tipo di modelli sono particolarmente efficaci per l'analisi di trial clinici per la capacità di gestire dati incompleti a causa di precoci drop-out o deceduti per cause indipendenti dallo studio o provenienti da soggetti con diversi tempi di inizio dell'osservazione.

time-to-event-analysis

La Figura figura 1 nella pagina seguente mostra alcune situazioni tipiche che si incontrano nell'analisi di dati di sopravvivenza:

- non tutti gli individui entrano nello studio allo stesso tempo (*staggered entry*);
- alla fine dello studio non tutti i soggetti hanno subito un evento;
- alcuni individui escono dallo studio o vengono persi al follow-up (*Drop-out*) durante la conduzione dello studio, e tutto ciò che sappiamo è che fino ad un certo tempo erano liberi da eventi.

Le ultime due situazioni si riferiscono al concetto di *censoring*: cioè quando si ha qualche informazione circa la sopravvivenza di un certo individuo, ma non si conosce esattamente il suo *tempo-all'evento*.

Formalmente ogni individuo $i, i = 1, \dots, n$ è associato con un tempo-all'evento T_i e un tempo di censoring C_i. L'outcome osservato per ciascun individuo è dato da $(T_i^*; \delta_i)$ dove

$$T_i^* = \min(T_i; C_i)$$

$$\delta_i = \begin{cases} 1, & T_i \leqslant 0, \\ 0, & T_i > 0. \end{cases}$$

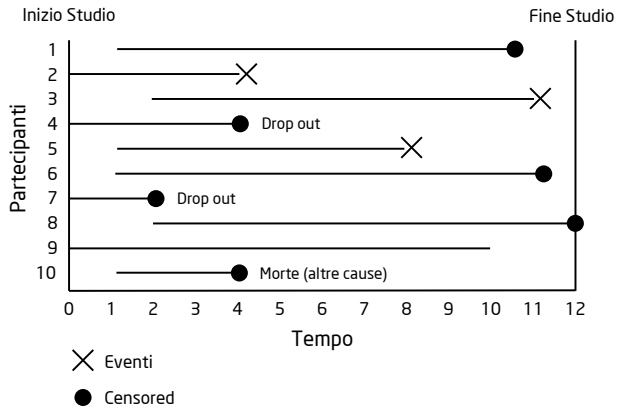


Figura 1: Un esempio di andamento del follow-up di uno studio di sopravvivenza.

Individui a rischio e failed

Ad ogni possibile tempo-all'evento discreto osservato $t_j, j = 1, \ldots, J$, si definiscono $\mathcal{R}_j$ come sottoinsieme degli individui R_j a rischio prima del tempo t_j e $\mathcal{D}_j$ come sottoinsieme degli individui D_j che hanno avuto un evento prima del tempo t_j

1.2 FUNZIONE DI SOPRAVVIVENZA

La funzione $S(t)$

L'obbiettivo fondamentale dell'analisi di sopravvivenza è quello di modellare il tempo-all'evento rappresentato dalla variabile random T. T può essere modellata attraverso la funzione di sopravvivenza

$$S(t) = Pr(T > t); \tag{1.1}$$

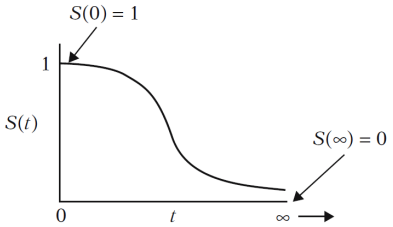
che rappresenta la probabilità che un evento T non si sia ancora verificato al tempo t. Questa *distribuzione di sopravvivenza* è strettamente correlata alle funzioni di distribuzione della probabilità $f(x)$ e di distribuzione cumulativa $F(x)$ dalla relazione [Moeschberger e Klein, 2003]:

$$S(t) = 1 - F(t) = \int_t^\infty f(x)\, dx, \tag{1.2}$$

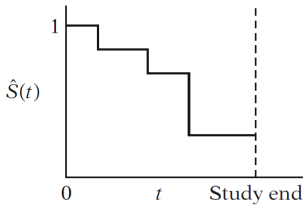
Una descrizione alternativa della distribuzione di T è rappresentata dalla *funzione di rischio*, che modella il tasso *istantaneo di fallimento*. La funzione di rischio è definita come

$$\lambda(t) \simeq \lim_{\Delta t \to \infty} \frac{P[T \leqslant t + \Delta t | T \geqslant t]}{\Delta t}, \tag{1.3}$$

dove $\lambda t \Delta t$ è un'approssimazione della probabilità di evento nell'intervallo $t + \Delta t$, data la sopravvivenza al tempo t. Per la funzione di rischio (1.3), vale la relazione



(a) Andamento teorico di S(t).



(b) Andamento reale di $\hat{S}(t)$.

Figura 2: Andamento teorico e reale della funzione di sopravvivenza.

$$\lambda t = \frac{f(t)}{s(t)} \tag{1.4}$$

dove $f(t)$ è la funzione di distribuzione della probabilità e $S(t)$ la funzione di sopravvivenza (1.1). La funzione di rischio *cumulativo* può essere quindi derivata dalla distribuzione di sopravvivenza [Peterson Jr, 1977]:

$$\Lambda(t) = -\ln[S(t)] \tag{1.5}$$

$S(t)$ è una funzione continua monotona decrescente tale che

$$t = 0, S(t) = S(0) = 1;$$
$$t = \infty, S(t) = S(\infty) = 0.$$

Funzione di Kaplan-Meier

Per studiare il comportamento di T in caso di censoring è conveniente usare un approccio non parametrico per stimare la funzione $S(t)$. Lo stimatore più popolare per la funzione di sopravvivenza è lo stimatore di Kaplan-Meier (o del *prodotto-limite*) [Kaplan e Meier, 1958]

$$\hat{S}(t) = \begin{cases} 1, & t < t_1, \\ \prod_{t_j \leqslant t} \left[1 - \frac{D_j}{R_j}\right], & t_1 \leqslant t. \end{cases} \tag{1.6}$$

Dalla (1.6) è possibile quindi derivare una stima del rischio cumulativo al tempo t tramite la (1.5).

Stimatore di Nelsoon-Aalen

Una modellazione alternativa per la funzione di rischio cumulativo (1.5) si può ottenere attraverso lo stimatore di *Nelson-Aalen*, definito come

$$\hat{\Lambda}(t) = \begin{cases} 1, & t < t_1, \\ \sum_{t_j \leqslant t} \left[\frac{D_j}{R_j}\right], & t_1 \leqslant t. \end{cases} \tag{1.7}$$

Sia lo stimatore di Nelson-Aalen (1.7) che per quello di Kaplan-Meier (1.6) sono basati sul concetto di censoring *non informativo*: cioè

la conoscenza del tempo al quale un individuo sia *censurato* non fornisce alcuna informazione ulteriore sulla probabilità di sopravvivenza se l'osservazione fosse continuata oltre il tempo t [Moeschberger e Klein, 2003].

1.3 IL MODELLO DEI RISCHI PROPORZIONALI

1.3.1 Variabili tempo-invarianti

Sebbene possa essere già molto informativa la stima del rischio o della sopravvivenza di un dato gruppo di individui, spesso risulta molto più rilevante, dal punto di vista clinico, modellare queste grandezze attraverso la loro relazione con uno specifico set di variabili.

La regressione di Cox

Il modello più utilizzato per la regressione tra variabili indipendenti e tempo-all'evento è il modello dei rischi proporzionali di Cox [Cox, 1972, 1975].

Formalmente, un evento al tempo t_j, condizionato da un insieme di covariate, è espresso come

$$\lambda(t_j|X) = \lambda_0(t_j)e^{(X\beta)} \tag{1.8}$$

dove $\beta = [\beta_1, \ldots, \beta_p]'$ rappresenta il vettore dei coefficienti della regressione e $X = x_{ik}$, dove $i = 1, \ldots, n$ sono gli individui e $k = 1, \ldots, p$ i valori degli attributi. In questo modello, $\lambda_0(t_j)$ è una funzione non nota che definisce il tasso di rischio quando $X = 0$, ed è normalmente chiamata *tasso di rischio basale* Un vantaggio del modello di Cox (1.8) è il fatto che sia *semi-parametrico*, cioè per stimare il valore dei parametri β, non è necessario fare alcuna assunzione sulla forma della distribuzione $\lambda_0(t)$. L'unica assunzione necessaria è che, poiché X non è funzione del tempo, confrontando diversi gruppi di individui, le relative funzioni di rischio siano proporzionali nel tempo (da qui il nome *modello dei rischi proporzionali*) e che tutti gli individui condividano presentino lo stesso tasso di rischio basale $\lambda_0(tj)$.

Partial likelihood

Per fare inferenza sui parametri della regressione, Cox sviluppò un metodo chiamato *partial likelihood* che utilizza solo i parametri di β e X (non λ_0), basato sulla distribuzione marginale dell'ordine dei tempi-all'evento osservati, in caso di presenza di censoring [Cox, 1972].

Per prendere in considerazione tempi discreti e includere individui con lo stesso tempo-all'evento, si utilizza la sommatoria delle variabili dei soggetti con eventi al tempo $t_j, s_j = \sum_{i \in \mathcal{D}_j} X_i$ [Moeschberger e Klein, 2003].

Il partial likelihood di Cox è espresso come

$$L(\beta) = \prod_{j=1}^{J} \frac{e^{(s_j \beta)}}{\left[\sum_{i \in \mathcal{R}_j} e^{(x_i \beta)}\right]^{\delta_j}} \tag{1.9}$$

Il numeratore dipende solamente da informazioni relative agli individui con tempo-all'evento t_j, e il denominatore, invece, su tutti i soggetti *liberi da eventi* per tutti i $t < t_j$. Una stima del *maximum likelihood* per i p coefficienti di regressione $\beta = [\beta_1, \ldots, \beta_p]'$,può essere ottenuta massimizzando il partial likelihood di Cox rispetto ai parametri di interesse.

Prognostic Index

Ottenuta la stima del maximum likelihood attraverso la (1.9), è possibile effettuare una predizione della probabilità di sopravvivenza di un singolo individuo i, utilizzando il cosiddetto *pregnostic index*,

$$PI_i = x_{i1}\hat{\beta}_1 + \cdots + x_{ip}\hat{\beta}_p = X_i\hat{\beta}, \tag{1.10}$$

definita come la somma dei valori delle variabile dell'individuo i, pesata per le stime corrispondenti dei coefficienti di regressione.

1.3.2 Variabili tempo-dipendenti

Il modello di Cox può essere esteso per includere variabili con misure ripetute modificando l'espressione (1.8):

$$\lambda(t_j | W(t_j)) = \lambda_0(t_j) e^{\left(W(t_j)\beta\right)}, \tag{1.11}$$

dove $W(t_j) = \{w_{ik}(t_j)\}$, con $i = 1, \ldots, n$ individui e $k = 1, \ldots, p$ variabili. Le stime del maximum likelihood per questo modello si ottengono riscrivendo il partial likelihood in (1.9):

$$L(\beta) = \prod_{j=1}^{J} \frac{e^{\left(s(t_j)\beta\right)}}{\left[\sum_{i \in \mathcal{R}(t_j)} e^{\left(W_i(t_j)\beta\right)}\right]}, \tag{1.12}$$

dove $s(t_j) = \sum_{i \in \mathcal{D}_j} W_i(tj)$ e $W_i(t_j) = [w_{i1}(t_j), \ldots, w_{ip}(t_j)]$ è il vettore delle $k = 1, \ldots, p$ osservazioni di attributi per ogni individuo i al tempo t_j. Nella (1.12), il valore delle variabili presente sia al numeratore che al denominatore può essere diverso per ogni tempo-all'evento t_j, seguendo il cambiamento delle variabili tempo-dipendenti. Vista la complessità dell'equazione, le stime del maximum likelihood peri coefficienti di regressioni nel modello esteso sono molto più costose in termini computazionali, perciò vengono spesso utilizzati l'algoritmo di Newton-Raphson o altre tecniche iterative [Moeschberger e Klein, 2003].

Nel caso del modello esteso di Cox, la predizione dell'outcome attraverso il prognostic index (1.10) non è di facile interpretazione. In particolare, usando variabili tempo-dipendenti, la capacità di predizione viene a perdersi poiché il modello è basato su valori degli

attributi che cambiano e valoti futuri potrebbero essere sconosciuti . Quando esistono valori a tempi maggiori, questo implica una conoscenza dello status di sopravvivenza del soggetto, introducendo quindi un possibile bias. Viste tali complicazioni, usare un modello di Cox con misure ripetute per predire un evento deve essere fatto con molta cautela [Fisher e Lin, 1999].

1.3.3 Classificazione con il modello di Cox

Il Cox proportional hazards model è sicuramente molto utile per l'analisi di tempo-all'evento. Purtroppo, se l'obiettivo dell'analisi è di creare gruppi di soggetti secondo una classificazione basata sulla prognosi, questo modello può non essere la scelta migliore per diversi motivi:

1. Il modello di Cox richiede il calcolo del rischio individuale per ogni soggetto, che potrebbe essere difficoltoso nel caso di individui *nuovi*.
2. Per studiare tutte le possibili interazioni tra variabili sono necessarie analisi molto costose sia in termini di tempo che di calcolo.
3. Per formare gruppi di calssificazione, i punti di cut-off per ogni singola variabile devono essere scelti ad hoc. Questo perché il modello di Cox fornisce una stima del rischio associato con un set di valori di variabili, ma nessuna informazione circa i valori che meglio differenziano il rischio individuale.

Viste queste limitazioni, l'analisi con *Alberi di Sopravvivenza* può essere un'alternativa più appropriata se l'obiettivo è quello creare un modello che possa classificare nuovi individui secondo gruppi basati sulla prognosi [Graf *et al.*, 1999].

1.4 DIAGNOSTICA DELL'ANALISI DI SOPRAVVIVENZA

misure di accuratezza

Valutare l'accuratezza delle predizioni ottenute attraverso un'analisi di sopravvivenza è molto utile a dare una misura della affidabilità del modello. Due diverse misure utilizzate per valutare l'accuratezza di un modello prognostico sono il *Brier score* o l'errore quadratico medio integrato (MISE) [BRIER, 1950; Graf *et al.*, 1999; Hothorn *et al.*, 2004]. Uno dei tanti vantaggi di queste statistiche è che le probabilità stimate sono direttamente usate per predire lo status dell'evento ad un dato tempo-all'evento t_j. In pratica, le predizioni sono espresse in termini di probabilità che un individuo sia classificato all'interno di

una data categoria prognostica, invece che semplicemente classificare ogni soggetto come avente o no un dato evento.

Brier score

Il Brier score per dati censurati ad un dato tempo-all'evento $t_j, j = 1, \ldots, J$, è dato da

$$BS(t_j) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1(T_i^* \leqslant t_j)\hat{S}(t_j^*|X_i)^2\delta_i}{\hat{G}(T_i^*)} + \frac{1(T_i^* > t_j)\{1-\hat{S}(t_j^*|X_i)\}^2}{\hat{G}(t_j)}\right], \tag{1.13}$$

dove $\hat{G}(T_i^*)$ è la stima della distribuzione dei valori di censoring $G(T_i^*)$, e $\hat{S}(T_i^*)$ è la stima della distribuzione della sopravvivenza $S(T_i^*)$. I contributi allo score possono essere suddivisi in tre categorie:

1. $T_i^* \leqslant t_j$ e $\delta_i = 1$;
2. $T_i^* > t_j$ e $\{\delta_i = 1$ o $\delta_i = 0\}$;
3. $T_i^* \leqslant t_j$ e $\delta_i = 0$.

La prima categoria contiene i casi i che non presentano un evento prima del tempo t_j. Il loro contributo al Brier score è pari a $\frac{\hat{S}(t_j|X_i)^2}{\hat{G}(T_i^*)}$. Nella categoria 2 sono presenti gli individui i che hanno avuto l'evento o sono censurati dopo l'istante t_j. Il contributo di questi individui allo score è pari a $\frac{1-\hat{S}(t_j^*|X_i)^2}{\hat{G}(t_j)}$. Per coloro con dati censurati prima di t_j (categoria 3), invece, lo status relativo all'evento è sconosciuto e quindi il loro contributo al Brier non può essere calcolato, anche se naturalmente entrano nel calcolo di $1/N$ [Graf *et al.*, 1999].

Ogni contributo individuale deve essere pesato per compensare la perdita di informazioni dovuta al censoring. Per questo, i soggetti nella categoria 1 che sopravvivono solamente fino a $T_i^* < t_j$ sono pesati da $\hat{G}(T_i^*)^{-1}$, così come coloro nella categoria 2 che sopravvivono al meno fino a t_j sono pesati da $\hat{G}(t_j)^{-1}$. Sebbene gli individui nella 3ª categoria non contribuiscano alle probabilità di sopravvivenza $\hat{S}(T_j^*|X_i)$, comunque esiste un loro contributo ai pesi attraverso $\{N\hat{G}(T_i^*)^{-1}\}$ e $\{N\hat{G}(t_j)^{-1}\}$.

Brier score tempo-variante

Nel caso di variabili tempo-varianti è possibile modificare l'equazione (1.13) dello score di Brier [Schoop *et al.*, 2008]. Al tempo t_j il Brier score con variabili a misure ripetute è calcolato come

$$BS(t_j) = 1N\sum_{i=1}^{N}\left[\frac{1(T_i^*\leqslant t_j)\hat{S}(t_j;W_i^{t_j})\delta_i}{\hat{G}(T_i^*)} + \frac{1(T_i^*>t_j)\{1-\hat{S}(t_j;W_i^{t_j})\}}{\hat{G}(t_j)}\right], \tag{1.14}$$

dove $W_i^{t_j} = \{W_i(t_{ij}), t_{i1} \leqslant t_{ij} \leqslant tj\}$ indica le variabili dei soggetti i fino al tempo t_j, $\hat{S}(t_j; W_i^{t_j})$ è una stima della distribuzione di sopravvivenza fino al tempo t_j, e $\hat{G}$ rappresenta una stima della distribuzione del censoring.

Lo score di Brier per dati censurati (1.13) può essere calcolato ad ogni singolo istante di tempo $t_j, j = 1, \ldots, J$, separatamente, ma può essere anche integrato rispetto ad una funzione di peso w_j,

$$\text{IBS} = \int_{j=1}^{J} \text{BS}(t_j)\, dw_j \tag{1.15}$$

Scelte naturali per la funzione di peso sono $w_j = t_j/t_J$ o $w_j = \frac{\{1-\hat{S}(t_j)\}}{\{1-\hat{S}(t_J)\}}$, dove $\hat{S}(t_j)$ è il tasso osservato di pazienti liberi da eventi al tempo t_j. Nel caso di tempi-all'evento discreto, il Brier score integrato può essere scritto come

$$\text{IBS} = \sum_{j=1}^{J} \frac{1}{w_j} \text{BS}(t_j), \tag{1.16}$$

dove J rappresenta il numero di tempi-all'evento unici.

In entrambi i casi di variabili tempo-indipendenti o dipendenti, l'eventualità teorica di previsione perfetta risulterebbe in uno score uguale a 0, che indica nessuna differenza fra la predizione del modello è gli status di fallimento dei casi analizzati. Uno score di 0.25 indica che il modello non predice meglio di una assegnazione del 50% della probabilità di sopravvivenza, e quindi potremmo dire essere un limite superiore per una predizione accettabile. Un valore di 1, corrispondente al massimo valore possibile per lo score, equivale ad una perfetta predizione *inversa*.

2 ALBERI DI SOPRAVVIVENZA

2.1 INTRODUZIONE

Definizione 2.1. Un albero binario consiste di un insieme finito non vuoto H di interi positivi $1, 2, \ldots, q$ e due funzioni $\text{left}(\cdot), \text{right}(\cdot) : T \to T \cup \{0\}$, per cui

(1) $\big[\text{left}(h) > h \to \text{right}(h) = \text{left}(h)\big] \vee \big[\text{right}(h) = \text{left}(h) = 0\big]$;

(2) $\forall h \in H, \exists$al più unu $\in H : h = \text{right}(u) \vee h = \text{left}(u)$.

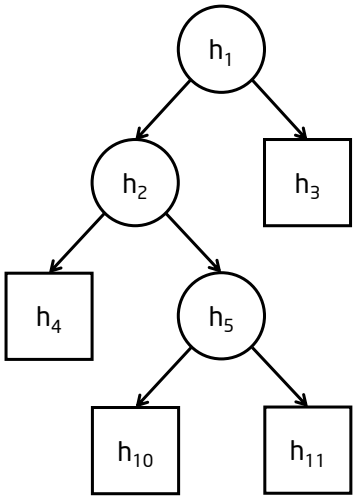


Figura 3: Un esempio di albero binario.

L'algoritmo per la creazione di un albero binario, come l'esempio in Figura 3, consiste nel partizionare ricorsivamente lo spazio delle variabili in sottoinsiemi, detti *nodi* h, omogenei per l'outcome di interesse. Questo processo è spesso compiuto utilizzando un criterio di *split* basato sulla minimizzazione di una misura di *impurità* dei nodi. Per gli output di tipo categorico i criteri più comunemente usati sono basati sul *Gini index* e sulla funzione di *entropia*, mentre la somma degli scarti quadratici dalla media è tipicamente usata per outcome continui.

impurità

L'approccio di base per la creazione dell'albero si basa su split binari usando una variabile per volta. La forma degli split è [Breiman *et al.*, 1984; LeBlanc e Crowley, 1992]:

$$\begin{aligned} &X \leqslant c, && X \text{ variabile continua o ordinale e } c \text{ costante;} \\ &X \in \{c_1, \ldots, c_k\}, && X \text{ variabile nominale con valori } c_1, \ldots, c_k. \end{aligned} \tag{2.1}$$

l'algoritmo

L'algoritmo tipico inizia a partire dall'intero spazio dei predittori χ_{h_1}, che equivale al *nodo radice* h_1, e i passi successivi sono:

RICERCA esaustiva delle possibili suddivisioni binarie considerando tutti i possibile valori di tutte le variabili;

SELEZIONE della suddivisione *migliore* in base al criterio di split in uso;

RIPETIZIONE del processo ricorsivamente sui nodi figli fino al raggiungimento di un dato criterio di stop (tipicamente quando un nodo contiene il numero minimo consentito di record).

Il risultato di questo processo è un albero completo di grandi dimensioni che normalmente produce overfitting sui dati: a questo punto si applica un algoritmo di *pruning* (potatura) e di selezione, al fine di avere un modello meno complesso e più generalizzabile. Alla fine del processo si avrà un insieme di nodi H e, all'ultimo livello dell'albero, i nodi foglia $\tilde{H}$, che rappresentano i gruppi di classificazione selezionati.

2.2 STORIA

La prima modellazione e implementazione degli algoritmi basati sugli alberi risale a Morgan e Sonquist [1963], che proposero l'algoritmo *AID* (Automatic Interaction Detection) sviluppato, appunto, per rilevare facilmente le interazioni tra variabili. Questo approccio divenne, però, largamente conosciuto, solo a partire dagli anni '80, dopo l'introduzione dell'algoritmo CART *(Classification and Regression Trees)* a cura di Breiman *et al.* [1984]. In particolare CART usa una misura di varianza *intra-nodo* basata sull'*indice di Gini*come criterio di *split*.L'innovazione più importante di CART alla programmazione dei modelli ad albero, fu l'introduzione di un algoritmo efficiente ed automatizzato di cost-complexity pruning per l'ottmizzazione delle dimensioni dell'albero completo.

CART

alberi di sopravvivenza

La prima applicazione dei modelli basati sugli alberi all'analisi di sopravvivenza è a cura di A. Ciampi *et al.* [1981] e Marubini *et al.* [1983] nei primi anni '80. Quello di Gordon e Olshen [1985] fu il primo lavoro a combinare gli alberi di sopravvivenza con CART e il suo algoritmo di pruning. L'algoritmo proposto da Gordon e Olshen [1985] usava una stima della funzione di Kaplan-Meier per trovare i valori di split per massimizzare la distanza tra le funzioni di sopravvivenza stimate. Anche se il criterio di split proposto nel loro lavoro non ebbe molta popolarità, Gordon e Olshen [1985] indicavano la possibilità di usare un *log-rank* o un *likelihood-ratio* test per misurare la *distanza* tra i duenodi figli, e questa idea fu largamente usata nei

lavori che seguirono. Davis e Anderson [1989], assumendo una funzione di sopravvivenza esponenziale con funzione di rischio costante, proposero la perdita di log-likelihood esponenziale come criterio di split. Anche in LeBlanc e Crowley [1992] viene usata il likelihood test, impiegando una misura di devianza tra un modello log–likelihood saturato ed uno massimizzato. Ahn e Loh [1994] usano invece un approccio diverso sviluppando un test per lo split ottimo basato sullo studio dei residui di Cox lungo gli assi delle singole variabili. Mark Robert Segal [1988] fu il primo ad applicare un criterio basato su una misura di separazione *inter-nodo* invece che di omogeneità intranodo all'analisi di dati censurati, proponendo il test di Tarone-Ware TARONE e WARE, 1977 come misura di split. Lo svantaggio era che l'algoritmo di pruning di CART non era direttamente applicabile. Mark Robert Segal [1988] propone un algoritmo di pruning non automatico, mentre in seguito LeBlanc e Crowley [1993] svilupparono un algoritmo automatizzato basato su approcci training/test set o bootstrapping in caso di piccoli campioni.

Sebbene CART non permetta l'analisi di outcome del tipo tempo-all'evento, esistono alcune implementazioni software degli algoritmi esistenti. REPCAM è un software creato da Antonio Ciampi *et al.* [1988] per l'analisi di dati censurati. Il programma permette la selezione di criteri di split sia intra- che inter-nodo A Ciampi, 1995, anche se non tutti gli algoritmi di pruning sono automatizzati. RPART di Therneau *et al.* [2013] è un'implementazione di CART per R, che permette l'analisi di dati di sopravvivenza usando il criterio di splitting proposto da LeBlanc e Crowley [1992]. RPART si contraddistingue per l'efficienza del suo algoritmo automatico di pruning basato su cross-validation, anche se non permette lo split basato su misure di distanza tra nodi.

REPCAM

2.3 CREAZIONE DELL'ALBERO

Si definisce un insieme di regole per creare una selezione di sottogruppi. I possibili split dello spazio dei predittori χ_h al nodo h sono indotti da tutti i quesiti del tipo «è $X \in S$?», dove $S \subset \chi_h$ e sono nella forma in 2.1. Ad ogni nodo h, i possibili split, s_h, che dividono lo spazio delle variabili χ_h in due sottoinsiemi disgiunti: S_h è l'insieme di ogni possibile split al nodo h.

valori di split

Per valutare la qualità di ogni potenziale divisione $s_h \in S_h$, il criterio di split $G(s_h)$ valuta la il miglioramento predittivo risultante dalla suddivisione dello spazio χ_h. Il criterio di split $G(s_h)$ è calcolato per ogni possibile punto si split $s_h \in S_h$ e il miglior split per il nodo h, è lo split (s_h^*) tale per cui

criterio di split

$$G(s_h^*) = \max_{s_h \in S_h} G(s_h). \tag{2.2}$$

Il miglior split s_h^* divide il nodo h in due nodi figli, h_L e h_R,che possono essere sottoposti allo stesso procedimento. Questo processo di splitting continua fino alla creazione del massimo albero possibile, H_{MAX}, dove i nodi non possono essere ulteriormente suddivisi per il raggiungimento di una dimensione minima pre-specificata o perché tutti gli elementi contenuti all'interno appartengono alla stessa classe. Questi nodi che non possono essere suddivisi ulteriormente vengono detti nodi terminali o *foglia* ed indicati con $\tilde{H}$ [Breiman *et al.*, 1984; LeBlanc e Crowley, 1992].

La selezione di un criterio di split $G(s_h)$, dipende largamente dalla natura dell'outcome di interesse. L'algoritmo CART di Breiman *et al.* usa un criterio di split intra-nodo del tipo

$$G(s_h) = G(h) - \big(G(h_L) + G(h_R)\big), \tag{2.3}$$

albero di regressione

Nel caso di outcome continuo y, CART crea un albero di regressione dove il criterio di split (2.3)è basato sulla riduzione del quadrato della devianza:

$$G(h) = \frac{1}{N_h} \sum_{i \in \mathcal{L}_h} \left(y_i - \frac{1}{N_h} \sum_{i \in \mathcal{L}_h} y_i \right)^2. \tag{2.4}$$

dove y_i è un outcome continuo per l'individuo i e $\mathcal{L}_h$ è l'insieme degli N_h individui al nodo h. La divisione migliore s_h^* risulta dalla massimizzazione del decremento del quadrato della devianza come definito nella 2.2.

Quando un modello ad albero viene sviluppato su dati di sopravvivenza, bisogna selezionare una statistica che possa essere usate come criterio di split in caso di dati censurati. Possono essere scelte statistiche parametriche basate sul likelihood test, semi-parametriche basate sul modello di Cox o non-parametriche basate su test sul rango di due campioni.

2.3.1 Test del rango di due campioni

Mark Robert Segal per primo propose l'uso di un test del rango su due campioni come possibile criterio di split in un albero di sopravvivenza. In particolare, i vantaggi di questo approccio sono diversi: le variabili di split e i valori di cut-off sono invarianti a trasformazioni monotone, i modelli ottenuti sono poco sensibili ad *outliers* per la natura non-parametrica, gli algoritmi di splitting sono così poco complessi ed efficienti, incorporando facilmente dati censurati.

Mark Robert Segal e LeBlanc e Crowley mostrano come qualunque test delle famiglie Tarone-Ware o Harrington-Fleming [HARRINGTON e FLEMING, 1982; TARONE e WARE, 1977] possono essere usate per impementare un criterio di split. Questi test valutano la

qualità di uno split s_h, quantificando la differenza tra le distribuzioni di sopravvivenza dei due nodi figli.

un criterio inter-nodo

In Bacchetti e M. R. Segal [1995] il test della classe Tarone-Ware usato per lo split è ottenuto costruendo una tabella 2 x 2, ognuna delle quali rappresenta un tempo-all'evento t_j.Per il $j-esimo$ tempo-all'evento, viene costruita la seguente tabella

	Evento	no Evento	
Nodo sinistro	x_j		m_{j1}
Nodo destro			
	n_{j1}		n_j

La statistica Tarone-Ware usa le informazioni contenute in ogni tabella costruita ad ogni tempo-all'evento discreto $t_j, j = 1, \ldots, J$, ed è calcolata come segue:

$$G(s_h) = \frac{\sum_{j=1}^{J} w_j \left[x_j - E_0(X_j)\right]}{\sqrt{\sum_{j=1}^{J} w_j^2 Var_0(X_j)}}, \tag{2.5}$$

dove w_j è il peso per il tempo t_j. Sotto l'ipotesi nulla di uguale tasso di sopravvivenza in entrambi i nodi figli, la variabile che rappresenta il numero di eventi nel nodo sinistro X_j al tempo t_j, segue la distribuzione ipergeometrica

$$E_0(X_j) = \frac{m_{j1} n_{j1}}{n_j} \tag{2.6}$$

e

$$Var_0(X_j) = \frac{m_{j1} n_{j1} (n_j - m_{j1})}{(n_j - 1) n_j^2} \tag{2.7}$$

Il cut-off selezionato per suddividere il nodo padre risulta dalla separazione inter-nodo maggiore, come definita dalla 2.2.

Modificando il peso w_j si ottengono diverse statistiche della famiglia Tarone-Ware: se $w_j = 1$ al tempo t_j, per ogni $j = 1, \ldots, J$, la statistica diventa il *log-rank test* [Cox, 1972; Mantel, 1966]. Ponendo $w_j = n_j$ al tempo t_j, per ogni $j = 1, \ldots, J$ si avrà il test di Wilcoxon [BRESLOW, 1970; GEHAN, 1965]. Le statistiche della classe Tarone-Ware seguono una distribuzione asintotica χ^2 con $r - 1$ gradi di libertà, con r pari al numero dei gruppi da comparare. Nel caso degli alberi binari, $r = 2$ nodi figli da comparare, e quindi la statistica è distribuita come una χ_1^2 [TARONE e WARE, 1977].

2.3.2 Devianza del modello esponenziale

LeBlanc e Crowley [1992] propongono un criterio di split basato sul calcolo della devianza di un full-likelihood esponenziale. Il metodo usa un modello a rischi proporzionali,

$$\lambda_h(t) = \theta_h \lambda_0(t), \tag{2.8}$$

dove θh è un parametro strettamente positivo specifico del nodo h e $\lambda_0(t)$ è la funzione di rischio basale. il modello di rischi proporzionale è normalmente basato sul partial likelihood, se, però, la funzione di rischio basale $\Delta_0(t)$ è nota, è preferibile l'uso del full likelihood per le stima e la selezione del modello. Per un albero H il full likelihood è espresso come

$$L = \prod_{h \in H} \prod_{i \in \mathcal{L}_h} (\lambda_0(T_i^*)\theta_h)^{\delta_i} e^{\Lambda_0(T_i^*)\theta_h}, \tag{2.9}$$

dove $\mathcal{L}_h$ è l'insieme degli individui al nodo h e $\Lambda_0(t)$ la funziona di rischio basale [LeBlanc e Crowley, 1992].

Per calcolare il likelihood, bisogna stimare il rischio cumulativo basale $\Lambda_0(T_i^*)$ e il parametro θ_h. Quando il rischio basale è noto, la stima del massimo likelihood (MLE) per θ_h è

$$\tilde{\theta}_h = \frac{\sum_{i \in \mathcal{L}_h} \delta_i}{\sum_{i \in \mathcal{L}_h} \Lambda_0(T_i^*)}. \tag{2.10}$$

Non essendo noto Λ_0, LeBlanc e Crowley [1992] utilizzano lo stimatore di Breslow [Breslow, 1972]:

$$\hat{\Lambda}_0(t) = \sum_{i:T_i^* \leqslant t} \frac{\delta_i}{\sum_{h \in H} \sum_{i:T_i^* \geqslant t, i \in \mathcal{L}_h} \hat{\theta}_h}. \tag{2.11}$$

É possibile quindi stimare la MLE in maniera iterativa. All'iterazione j, il rischio cumulativo è pari a:

$$\hat{\Lambda}_0^{(j)}(t) = \sum_{i:T_i^* \leqslant t} \frac{\delta_i}{\sum_{h \in H} \sum_{i:T_i^* \geqslant t, i \in \mathcal{L}_h} \hat{\theta}_h^{j-1}}, \tag{2.12}$$

da cui

$$\hat{\theta}_h^{(j)} = \frac{\sum_{i \in \mathcal{L}_h} \delta_i}{\sum_{i \in \mathcal{L}_h} \hat{\Lambda}_0^{(j)}(T_i^*)}. \tag{2.13}$$

Per efficienza l'algoritmo proposto da LeBlanc e Crowley [1992] ferma le iterazioni al primo passo. Avendo quindi $j = 1$, essendo $\theta_h^{(0)} = 1$, per stimare $\hat{\Lambda}_h^{(1)}$ avremo quindi

$$\hat{\theta}_h^{(1)} = \frac{\sum_{i \in \mathcal{L}_h} \delta_i}{\sum_{i \in \mathcal{L}_h} \hat{\Lambda}_0^{(1)}(T_i^*)}, \tag{2.14}$$

che rappresenta il rapporto fra il numero osservato e quello aspettato di fallimenti al nodo h.

La *devianza* al nodo h è definita come

$$G(h) = 2\{L_h(\textit{saturato}) - L_h(\tilde{\theta}_h)\}, \tag{2.15}$$

dove $L_h(\textit{saturato}$ è il log-likelihood di un modello *saturato* avente una variabile per ogni record, e $L_h(\tilde{\theta}_h)$ è il log-likelihood massimizzato quando $\Delta_0(t)$ è noto. Per una osservazione i la devianza è definita

$$d_i = 2\left[\delta_i \log\Big(\frac{\delta_i}{\Lambda_0(T_i^*)\hat{\theta}_h}\Big) - \big(\delta_i - \Lambda_0(T_i^*)\hat{\theta}_h\big)\right]. \tag{2.16}$$

La misura usata per il criterio di split nella 2.3 è quindi

$$G(h) = \frac{1}{N}\sum_{i\in\mathcal{L}_h}\left[\delta_i \log\Big(\frac{\delta_i}{\hat{\Lambda}_0^{(1)}(T_i^*)\hat{\theta}_h^{(1)}}\Big) - \big(\delta_i - \hat{\Lambda}_0^{(1)}(T_i^*)\hat{\theta}_h^{(1)}\big)\right]. \tag{2.17}$$

Il valore per la creazione della divisione binaria è quello che massimizza il decremento della devianza intra-nodo nei nodi figli, come definito nella 2.2

2.4 PRUNING

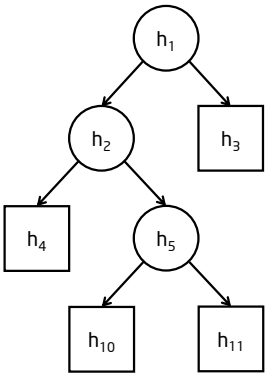
Definizione 2.2. Un albero H_h è detto *sottoalbero* di H se è un albero con lo stesso nodo radice di H e, $h \in H, \forall h \in H_h$

Definizione 2.3. Un albero H^h è detto *ramo* di H se è un albero con nodo radice $h \in H$ e ogni discendente di h in H sono discendenti di h in H^h

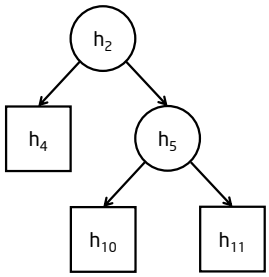
rami e sottoalberi

La complessità dell'albero definito alla fine della fase di creazione H_{MAX} (Figura 4a) è maggiore quanto maggiore è il numero di variabili nel modello o i record del data-set. L'albero totale H_{MAX} è costituito da *sottoalberi* creati eliminando successivamente i rami dall'albero totale. Un ramo H^h (Figura 4b) è costituito, quindi, da un nodo h e da tutti i nodi discendenti in H . Potando un ramo H^h da H, si rimuove tutto il ramo H^h eccetto il nodo h. Questo sottoalbero potato è indicato con $H_h = H - H^h$ (Figura 4c), per indicare il ramo mancante H^h, o $H_h \prec H$ enfatizzando che H^h è un sottoalbero di H [Breiman *et al.*, 1984; LeBlanc e Crowley, 1993].

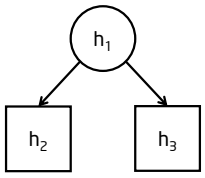
Come nei modelli di regressione tradizionali, un albero eccessivamente esteso contenente tutte le variabili possibili potrebbe essere di difficile interpretazione clinica. É necessario quindi definire una misura di accuratezza ed efficienza del modello al fine di selezionare il miglior sottoalbero fra i possibili $H_h \prec H_{MAX}$. Il metodo usato può variare in base al criterio di split usato.



(a) Albero: H



(b) Ramo: H^{h_2}



(c) Sottoalbero: $H - H^{h_2}$

Figura 4: Esempio di pruning.

LeBlanc e Crowley [1992] propongono un algoritmo molto simile a quello descritto per CART [Breiman *et al.*, 1984] per il loro criterio di split basato sulla devianza intra-nodo. Per valutare il valore predittivo di un sottoalbero H_h, viene calcolata una funzione di perdita di accuratezza derivante dalla potatura, detta *cost-complexity* di H_h

cost-complexity

$$G_\alpha(H_h) = \sum_{h \in \tilde{H}_h} G(h) + \alpha|\tilde{H}_h|, \tag{2.18}$$

Dove $G(h)$ è la devianza del nodo h, $|\tilde{H}_h|$ il numero di nodi terminali di H e α è un parametro di complessità, positivo, che può essere impostato per penalizzare più o meno la dimensione dell'albero.

Nel caso di criteri inter-nodo LeBlanc e Crowley [1993] propongono una misura di *split-complexity* per selezionare il miglior sottoalbero. Per ogni sottoalbero $H_h \prec H_{MAX}$,$I = H_h - \tilde{H}_h$ è l'insieme dei nodi interni. Si definisce la complessità $|I|$ come il numero di insiemi I. Dato un parametro di complessità $\alpha \geqslant 0$, la complessità di split $G_\alpha(H_h)$ è:

split-complexity

$$G_\alpha(H_h) = G(H_h) - \alpha|I|, \tag{2.19}$$

dove

$$G(H_h) = \sum_{h \in I} G(s_h^*). \tag{2.20}$$

Sia la cost- (2.18) che la split-complexity (2.19) di un sottoalbero $H_h \prec H_{MAX}$ valutano il bilanciamento tra le capacità predittive del modello e la sua complessità. Inizialmente viene calcolata l'accuratezza prognostica usando la sommatoria della statistica di split sui nodi h ($\in I$ nel caso della split-complexity), per poi sottrarre un indice di penalità per la dimensione dell'albero. L'effetto di questa penalizzazione è controllato dal parametro di complessità $\alpha \geqslant 0$. Tanto più α è vicino a 0, tanto più il costo di un elevato numero di nodi (interni) è basso e il sottoalbero che massimizza la misura di complessità sarà dimensionato. All'aumentare di α, viene contestualmente aumentata la penalizzazione sulle dimensione del modello, e il numero di nodi del sottoalbero ottimo saranno inferiori. Il sottoalbero formato dal solo nodo radice è associato al valore massimo di α [LeBlanc e Crowley, 1993]

il parametro di complessità α

Definizione 2.4. Un sottoalbero H_h è detto *potato ottimamente* per un dato parametro di complessità α se:

$$G_\alpha(H_h) = \min_{H_h \preceq H_{MAX}} G_\alpha(H_h) \quad \text{(cost-complexity)} \tag{2.21}$$
$$G_\alpha(H_h) = \max_{H_h \preceq H_{MAX}} G_\alpha(H_h) \quad \text{(split-complexity)}$$

H_h^* è il sottoalbero ottimamente potato *minimo* se $H_h^* \preceq H_h$ per ogni sottoalbero ottimamente potato $H_h \preceq H_{MAX}$. Sia $H^*(\alpha)$ il minimo sottoalbero ottimamente potato di H per il parametro α.

Si può dimostrare sia per la cost- che per la split-complexity [Breiman *et al.*, 1984; LeBlanc e Crowley, 1992, 1993] che

$$\forall \alpha, H \qquad \exists! H_h^*(\alpha),$$

e che all'aumentare di α, la sequenza ottima di sottoalberi è una sequenza innestata di alberi e che il *weakest link cutting* è un algoritmo efficiente per calcolarla.

Taglio del Legame più Debole

2.4.1 Weakest Link Cutting

Questo metodo individua i successivi parametri α relativi ai sottoalberi ottimi della sequenza, calcolando, inoltre, i sottoalberi stessi.
Detto $H_1 \equiv H^*(0)$, è possibile definire:

$$\forall \text{ nodo } h \in H_1, \qquad G_\alpha(\{h\}) = G(h) + \alpha \tag{2.22}$$
$$\forall \text{ ramo } H_h, \qquad G_\alpha(H_h) = G(h) + \alpha|(\tilde{H}_h)| \tag{2.23}$$

Sappiamo inoltre che

$$\alpha = 0, G_0(H_h) < G_0(\{h\}) \tag{2.24}$$

e questo vale sempre per valori sufficientemente piccoli di α. Aumentando il valore di α, troveremo un valore critico di α per cui $G_\alpha(H_h) = G_\alpha(\{h\})$ e oltre il quale la 2.24 cambia verso. Risolvendo la 2.24 si ottiene

$$\alpha < \frac{G(h) - G(H_h)}{|\tilde{H}_h| - 1} \geqslant 0$$

Si può definire una funzione $r_1(h), h \in H_1$

$$g_1(h) = \begin{cases} \frac{G(h)-G(H_h)}{|\tilde{H}_h|-1}, & t \notin \tilde{H}_1 \\ +\infty, & t \in \tilde{H}_1 \end{cases}$$

Nodi con legame più debole

Definizione 2.5. Si definisce *weakest link* il nodo $\bar{h}_1$ per il quale

$$g_1(\bar{h}_1) = \min_{h \in H_1} g_1(h)$$

All'aumentare di $\alpha, \bar{h}_1$ è il primo nodo che diventa *preferibile* del ramo $T_{\bar{h}}$ di cui è radice. Posto $\alpha_2 = g_1(\bar{h}_1)$, questo è il primo valore dopo $\alpha_1 = 0$ che permette di ottenere un sottoalbero ottimo $\prec H_1$:

$$\forall \alpha_1 \leqslant \alpha < \alpha_2, H^*_\alpha = H_1.$$

Il sottoalbero ottimo corrispondente ad α_2 deriva dalla rimozione del ramo discendente da $\bar{t}_1$

$$H_2 = H_1 - H_{\bar{t}_1}.$$

Il processo precedente viene ripetuto ricorsivamente:

$$g_k(h) = \begin{cases} \frac{G(h)-G(H_h)}{|\tilde{H}_h|-1}, & t \notin \tilde{H}_k \\ +\infty, & t \in \tilde{H}_k \end{cases} \tag{2.25}$$

$$g_k(\bar{h}_k) = \min_{h \in H_k} g_k(h) \tag{2.26}$$

$$\alpha_{k+1} = g_k(\bar{h}_k)$$

$$H_{k+1} = H_k - H_{\bar{t}_k}.$$

Nel caso al passo k esistano più nodi che minimizzano la 2.26, ovvero $g_k(\bar{h}_k) = g_k(\bar{h}'_k)$, si rimuoveranno tutti i rami discendenti da quei nodi:

$$H_{k+1} = H_k - H_{\bar{t}_k} - H_{\bar{t}'_k}.$$

Alla fine della procedura viene prodotta una sequenza di sottoalberi innestati:

$$H_1 \succ H_2 \succ H_3 \succ \ldots \{h_1\} \tag{2.27}$$

Teorema 1. *Le $\{\alpha_k\}$ rappresentano una sequenza crescente: $\alpha_k < \alpha_{k+1}, \forall k \geqslant 1, \alpha_1 = 0$*

$$\forall k \geqslant 1, \alpha_k \leqslant \alpha < \alpha_{k+1}, H(\alpha) = H(\alpha_k) = H_k.$$

Il teorema 1 implica che il sottoalbero ottimamente potato H_k rimane ottimo per tutti gli α, partendo da k fino a raggiungere α_{k+1}: sebbene si abbia una sequenza finita di sottoalberi, quindi, esistono valori ottimi per il parametro continuo α.

L'algoritmo basato sul weakest link cutting inizialmente presentato in Breiman *et al.* [1984], è la base per la fase di pruning negli alberi di sopravvivenza presentati in LeBlanc e Crowley [1992] e ulteriormente adattato ad un criterio inter-nodo in LeBlanc e Crowley [1993].

Ottenuta la sequenza di sottoalberi innestati è necessario selezionarne uno, di seguito saranno descritte alcune metodologie applicabili per effettuare una selezione efficiente.

2.4.2 Selezione del sottoalbero ottimo attraverso Training e Test Set

Nel caso di un campione abbastanza grande, è possibile dividere il dataset $\mathcal{L}$ in un training set $\mathcal{L}_T$ e un test set $\mathcal{L}_{(T)} = \mathcal{L} - \mathcal{L}_T$. Il training set, $\mathcal{L}_T$ è usato per la fase di *accrescimento* del modello e per creare la 2.27. Il test set $\mathcal{L}_{(T)}$ è invece utilizzato su ognuno dei sottoalberi della 2.27 per calcolare il valore della cost- (split-) complexity $G_\alpha(H)$. Il sottoalbero che ottimizza $G_\alpha(H)$ è scelto come albero migliore. Alternativamente la scelta può essere effettuata attraverso la regola *1 Standard Error* [Breiman *et al.*, 1984], per aumentare la *generalizzabilità* deel modello. L'idea è quella di selezionare l'albero più semplice tra i sottoalberi potati che hanno valori di $G_\alpha(H)$ sul test-set non significativamente più elevati del sottoalbero ottimo, vicini al limite superiore dell'intervallo di confidenza di $G_\alpha(H)$ dell'albero ottimo.

La regola 1 SE

2.4.3 Selezione del sottoalbero ottimo attraverso Cross-validation

Nel caso del criterio di split presentato in LeBlanc e Crowley [1992], la devianza per i sottoalberi potati viene stimata da una V-fold cross-validation. I dati $\mathcal{L}$ sono divisi in V insiemi $\mathcal{L}_v, v = 1, \ldots, V$ e $\mathcal{L}^{(v)} = \mathcal{L} - \mathcal{L}_v$ di dimensioni simili. Detto H_{MAX} l'albero sviluppato sull'intero set di dati, l'operazione di accrescimento è ripetuta per ogni fold v. Si otterranno così V alberi addizionali, H_v Gli alberi, $H^{(v)}_{MAX}$ sono sviluppati a partire dai sottoinsiemi $\mathcal{L}^{(v)}$. Per ogni parametro di complessità α siano $H(\alpha)$ e $H^v(\alpha), v = 1, \ldots, V$ i sottoalberi ottimamente potati rispettivamente per $H_{MAX}, H^{(v)}_{MAX}$. Per ogni parametro di complessità α è quindi possibile calcolare e il sottoalbero ottimo corrispondente $H^v(\alpha)$ e le stime della 2.14 $\hat{\theta}^1_h(v) : h \in \tilde{H}^v(\alpha)$. Per ogni albero H^v approssimativamente $1/V$ dei dati è usato per la crescita dell'albero. Le performance dei modelli generati sul campione $\mathcal{L}^{(v)}$ con i dati contenuti in $\mathcal{L}_v$. La devianza residua per ogni record $i \in \mathcal{L}_v$ è infatti pari a

devianza residua cross-validata

$$d_i(\delta_i, \hat{\theta}_h^1(v)) = 2\left[\delta_i \log\left(\frac{\delta_i}{\hat{\Lambda}_0^1(t_i)\hat{\theta}_h^1(v)}\right) - (\delta_i - \hat{\Lambda}_0^1(t_i)\hat{\theta}_h^1(v))\right], \quad (2.28)$$

dove $\hat{\Lambda}_0^1(t_i)$ è basato sull'insieme originale $\mathcal{L}$. Sia α^* il parametro di complessità che minimizza la media della devianza residua cross-validata per gli alberi $H^v(\alpha)$ sule V fold.

Si noti che per ogni sottoinsieme v l'intero processo di crecita del modello viene ripetuto: se la struttura dell'albero rimanesse fissa e venissero ricalcolati solo le stime sui nodi, la procedura porterebbe ad una sottostima della devianza degli alberi.

devianza infinita

Possono sorgere problemi nel calcolare le stime cross-validate per dati censurati. Se tutte le osservazioni di un nodo h sono censurate nel sottoinsieme usato per la crescita dell'albero, la stima $\theta_h \hat{=} 0$ Se nell'insieme di validazione corrispondente a quel nodo esistono osservazioni non censurate, la stima della devianza aspettata sarà infinito. L'uso di stimatori contratti nel setting parametrico può rapprsentare un possibile aggiustamento. Una possibile alternativa è quella di rimpiazzare i nodi con zero fallimenti con 0.5, come suggerito in Davis e Anderson [1989] per il modello esponenziale. La stima di θ_k diventa per la cross-validation è

$$\hat{\theta}_h = \frac{1}{2\sum_{i \in S_h} \hat{\Lambda}_0^1(t_i)}$$

per un nodo h senza fallimenti osservati.

Scelto un albero $H(\alpha^*)$ che minimizza la stima cross-validata della devianza aspettata, è possibile ottenere la stima del maximum likelihood iterando le equazioni 2.12 e 2.13. Anche in questo caso è possibile applicare tecniche per aumentare la stabilità del modello come la regola 1 SE [Negassa *et al.*, 2005].

2.4.4 Selezione del sottoalbero ottimo attraverso Bootstrapping

L'approccio training/test set per la selezione del sottoalbero migliore è sicuramente efficiente dal punto di vista computazionale, ma richiede una dimensione campionaria elevata per essere anche efficace. LeBlanc e Crowley [1993] adattano una tecnica per la correzione di bias nei problemi di prognosi [Efron, 1983] basata sul bootstrapping al loro criterio inter-nodo. Non necessitando di un N elevato, è preferibile nel caso di piccoli campioni. Sia $G(\mathbf{X}_1; \mathbf{X}_2, H) \equiv G(H)$, dove $\mathbf{X}_2$ rappresenta l'insieme usato per sviluppare l'albero iniziale H, e $\mathbf{X}_1$ è testato su H per calcolare la statistica. Si definiscono

bootstrapping

$$
\begin{aligned}
G^* &= E_F G(\mathbf{X}^*; \mathbf{X}, H);\\
G &= G(\mathbf{X}; \mathbf{X}, H);\\
o &= G^* - G;\\
\omega &= E_F\{G^* - G\}.
\end{aligned}
$$

Se l'effettiva distribuzione dei dati F fosse nota, allora $G = G(\mathbf{X}; \mathbf{X}, H) + \omega$ potrebbe essere usata come $G(T)$ bias-corretta, dove ω rappresenta *l'ottimismo* dovuto all'ottimizzazione del punto di split. Per calcolare questa quantità in pratica, si rimpiazza F con la distribuzione del training-set, $\mathbf{X}$, utilizzano tecniche *Monte-Carlo* per la stima di ω.

Il metodo prevede di crescere (e potare) l'albero sull'intero insieme di apprendimento e ottenere la sequenza di sottoalberi ottimamente potati H_k con i relativi parametri di complessità α_k. Posto $\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}}$. Si costituiscono quindi B campioni bootstrap da $\mathcal{L}$. Per ogni insieme $\mathcal{L}^{(b)}, b = 1, \ldots, B$ si costruisce un albero, il sottoalbero ottimo per ogni α'_k e si calcola

$$o_{b_k} = G\big(\mathbf{X}; \mathbf{X}_b, H_b(\alpha'_k)\big) - G\big(\mathbf{X}_b; \mathbf{X}_b, H_b(\alpha'_k)\big). \tag{2.29}$$

Quindi si può trovare la media sui B insiemi di bootstrap

$$\hat{\omega}_k = \frac{1}{B} \sum_{b=1}^{B} o_{b_k}.$$

Si sceglie l'albero che massimizza $\hat{G}_{\alpha_c}(H(\alpha'_k))$, dove

$$\hat{G}_{\alpha_c}(H(\alpha'_k)) = G\big(\mathbf{X}; \mathbf{X}, H(\alpha'_k)\big) + \hat{\omega}_k$$

e α_c è il parametro di complessità selezionato in precedenza.

2.5 CLASSIFICAZIONE DI NUOVI INDIVIDUI

Dopo che il modello ad albero è stato disegnato e potato, un nuovo caso può essere classificato semplicemente rispondendo alle domande corrispondenti ad ogni split binario sull'albero fino ad arrivare ad un nodo terminale. Poiché ad ogni nodo foglia è associata una misura di sopravvivenza, come uno stimatore di Kaplan-Meier o un hazard rate, questa rappresenterà la sopravvivenza predetta per tutti gli individui classificati in quel nodo [Mark Robert Segal, 1997].

2.6 ALBERI DI SOPRAVVIVENZA TEMPO-DIPENDENTI

Gli algoritmi per la costruzione di modelli ad alberi utilizzano tradizionalmente variabili tempo-invarianti per la predizione. Per esem-

pio, caratteristiche cliniche misurate all'inizio dell'osservazione, o informazioni descrittive come l'età e il genere sono tutte variabili *baseline* che possono essere usate per predire il gruppo di rischio di un particolare individuo. Esistono delle situazioni, però, nelle quali i valori di alcune delle variabili prese in considerazione cambiano nel corso del tempo e possono essere misurate ripetutamente per ottenere informazioni più complete. Questi dati acquisiti longitudinalmente potrebbero essere dei marker dello stato di salute (esami del sangue) o degli indicatori specifici del grado di malattia (SOFA score), ma anche degli eventi intercorsi durante l'osservazione come ospedalizzazioni, interventi chirurgici o cambi nella terapia. Al di là dello specifico tipo di variabile tempo-variante, l'acquisizione ripetuta del dato può fornire informazioni aggiuntive al fine di costruire i gruppi prognostici per la predizione dell'outcome.

2.6.1 Alberi Tempo-Dipendenti: modello Esponenziale a Tratti

Nel 1998 Huang *et al.* [1998] propongono un metodo per l'utilizzo di variabili misurate ripetutamente nei modelli ad albero. Il metodo proposto suddivide i nodi in base all'interazione tra il valore delle variabili e il tempo, definendo una misura di miglioramento delle prestazioni basata sulla distribuzione di sopravvivenza esponenziale a tratti.

L'algoritmo è basato su una modellazione alternativa della funzione di rischio. Invece di usare una assunzione di rischi proporzionali, il rischio è modellato in una forma generale $\lambda(t) = \lambda\big(\boldsymbol{W}(t), t\big)$, semplificata a $\lambda(t) = \lambda\big(\boldsymbol{W}, t\big)$ nel caso tutte le variabili analizzate siano tempo-dipendenti. Se vengono introdotte misure ripetute, l'algoritmo stima il rischio approssimando che la sopravvivenza di ogni individuo segua una distribuzione esponenziale a tratti, dove i punti di discontinuità sono selezionati direttamente dal metodo di accrescimento dell'albero. Nel caso le variabili siano tutte tempo-indipendenti l'algoritmo è identico a quello proposto in Davis e Anderson [1989]

distribuzione del rischio esponenziale a tratti

Questo metodo approssima la distribuzione dei tempi-all'evento, T_i, con una distribuzione esponenziale definita a tratti di k tratti, con funzione di densità:

$$f_i(t) = \begin{cases} \lambda_{i_1} e^{(-\lambda_{i_1} t)}, & 0 = t_{i_0} < t \leqslant t_{i_1} \\ \lambda_{i_2} e^{\left(t_{i_1}(\lambda_{i_2} - \lambda_{i_1}) - \lambda_{i_2} t\right)}, & t_{i_1} < t \leqslant t_{i_2} \\ \vdots & \vdots \\ \lambda_{i_k} e^{\left[\sum_{j=1}^{k-1} t_{i_j}(\lambda_{i_{j+1}} - \lambda_{i_j} - \lambda_{i_k} t)\right]}, & t_{i_{k-1}} < t \leqslant t_{i_k} \end{cases} \tag{2.30}$$

dove $0 = t_{i_0} < t_{i_1} < \cdots < t_{i_k} = \infty$ e $\lambda_{i_1}, \lambda_{i_2}, \ldots, \lambda_{i_k}$ sono positivi.

Per esempio, nel caso di variabili con misure ripetute non decrescenti nel tempo, se il nodo radice, h_0, si divide su un valore c_0

sulla covariata $W(t)$, un soggetto i, con osservazioni tempo-variate $W_i(t_{ij}) = [w_i(t_{i1}), \ldots, w_i(t_{iJ_i})]$, può appartenere ad una delle seguenti categorie:

1. $W_i(t_{ij}) \leqslant c, \quad \forall t \in T_i^* \rightarrow i \in \text{left}(h_0)$;
2. $W_i(t_{ij}) > c, \quad \forall t \in T_i^* \rightarrow i \in \text{right}(h_0)$;
3. $\begin{cases} W_i(t_{ij}) \leqslant c, & 0 < t \leqslant t_i^* \rightarrow i \in \text{left}(h_0), \\ W_i(t_{ij}) > c, & t_i^* < t \leqslant T_i^* \rightarrow i \in \text{right}(h_0). \end{cases}$

L'algoritmo è gestito analogamente per variabili strettamente decrescenti e può essere esteso per includere dati con andamenti non monotoni.

test sull'algoritmo di Huang et al.

L'algoritmo in Huang *et al.* [1998] è stato testato per confrontarne le performance contro un modello classico di regressione di Cox. Questo confronto era di particolare interessante per capire se il metodo proposto, che crea delle funzioni a gradino per modellare il cambiamento del rischio nel tempo, potesse funzionare con funzioni di rischio continue. Tre differenti setting sono stati usati per testare l'algoritmo: (1) tempi-all'evento dipendenti da un'unica variabile tempo-invariante; (2) tempi-all'evento con una sola variabile tempo-variante; (3) tempi-all'evento in relazione a variabili dipendenti e indipendenti dal tempo. Per misurare le performance dei due modelli veniva confrontato il rischio relativo con la media del rischio relativo stimato usando l'errore quadratico medio e la sua deviazione standard. Mentre la capacità di predizione del modello ad alberi era leggermente più precisa nel caso di sole variabili tempo-dipendenti o indipendenti, il modello a rischi proporzionali aveva performance migliori nel caso di modelli con entrambi i tipi di variabili.

2.6.2 Alberi Tempo-Dipendenti: test sul Rango di due Campioni

Un algoritmo alternativo per lo sviluppo di alberi di sopravvivenza tempo-dipendenti è presentato in Bacchetti e M. R. Segal [1995]. Questo metodo utilizza un test sul rango di due campioni per gestire variabili tempo-varianti. Come per l'algoritmo di Huang *et al.*,i soggetti sono divisi in *pseudo-soggetti* quando il valore di una delle variabili tempo-dipendenti risulta inferiore al cut-off selezionato per lo split di uno dei nodi dell'albero per alcuni istanti di tempo e superiore per altri.

pseudo-soggetti

Si consideri un individuo $i, i = 1, \ldots, N$, con un tempo di inizio dell'osservazione τ_i, un tempo all'evento T_i^* e una variabile tempo-variante $W_i(t_{ij}) = [w_i(t_{i1}), \ldots, w_i(t_{iJ_i})]$. Per un determinato split s_h, i soggetti i con $W_i(t_{ij}) > s_h$ ad ogni istante di tempo sarà mandato al nodo figlio destro, mentre gli i per cui $W_i(t_{ij}) \leqslant s_h$ per tutto il tempo sono mandati al figlio sinistro. Può succedere, però, che

per alcuni soggetti la variabile $W_i(t_{ij})$ può essere maggiore di s_h in alcuni istanti di tempo, e minore o uguale per altri. In qualche maniera questi soggetti dovrebbero contribuire sia al figlio destro che al sinistro.

Prendendo una $W_i(t_{ij})$ non decrescente nel tempo, e ponendo t_i^* come l'ultimo istante di tempo in cui $W_i(t_{ij}) \leqslant s_h$, con $\tau_i < t_{ij}^* \leqslant T_i^*$. Per testare in maniera corretta lo split c, il soggetto i deve essere considerato parte del figlio sinistro per i tempi t_{ij} tali che $\tau_i < t_{ij}^* \leqslant t_i^*$, e parte del nodo destro per tempi di fallimento $t_i^* < t_{ij} \leqslant T_i^*$. Uno scenario del genere è facilmente modellato se l'algoritmo è in grado di incorporare dati troncati a sinistra o *left-censored* cioè dati che siano disponibili a partire da un tempo successivo al reale inizio dell'osservazione.

left censoring

Consideriamo la sopravvivenza dell'$i-esimo$ soggetto come composta da due sopravvivenze non sovrapposte di due pseudo-soggetti, $i_1 e i_2$. Lo pseudo-soggetto i_1 è a rischio solo fino al tempo t_i^*, dopo del quale viene ad essere censurato. Al contrario lo pseudo-soggetto i_2 è troncato a sinistra a t_i^* ed è a rischio fino a tempo del suo evento T_i^*. Gli pseudo-soggetti i_1, i_2 sono assegnati al figlio destro e sinistro rispettivamente.

Il test di due campioni per il potenziale split s_h è calclato come

$$G(s_h) = \frac{\sum_{j=1}^{J} w_j \left[x_j - E_0(X_j)\right]}{\sqrt{\left[\sum_{j=1}^{J} w_j^2 \mathrm{Var}_0(X_j)\right]}} \tag{2.31}$$

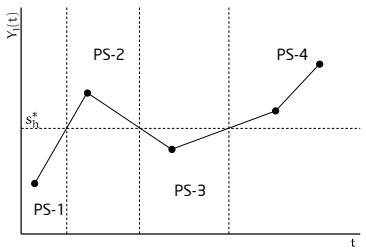
dove w_j è il peso per il tempo t_j e

$$x_j = \sum_{i \in \mathcal{R}_j} I\big(\delta_i(t_j) = 1 \vee z_i(t_j) \leqslant s_h\big), \tag{2.32}$$

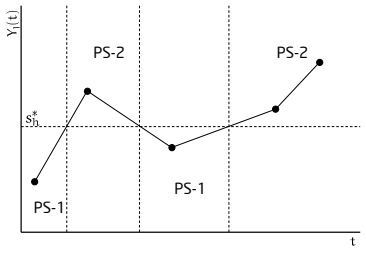
che è interpretato come il numero di individui nell'insieme a rischio al tempo t_j, che presentano un evento esattamente a t_j e un valore della variabile tempo-dipendente al tempo t_j che sia minore o uguale del potenziale split s_h. Quando $w_j = 1$ nella 2.31, la statistica è pari a un log-rank test su un modello univariato a rischi proporzionali di Cox con una sola variabile tempo-dipendente.

Bacchetti e M. R. Segal propongono inoltre un metodo per generalizzare l'algoritmo a variabili con andamenti non monotoni, come descritto in Figura 5. Questi metodi non differiscono nel calcolo del criterio di split, ma nel modo in cui vengono assegnati i diversi pseudo-soggetti dopo la selezione del cut-off per lo split. Il primo metodo (Figura 5a) divide un individuo i in due o più pseudo-soggetti, ognuno dei quali rappresenta uno degli intervalli di tempo durante i quali $W_i(t_{ij}) \leqslant s_h^*$ o $W_i(t_{ij}) > s_h^*$. Se, per esempio $W_i(t_{ij}) \leqslant s_h^*$ per $t_{ij} \leqslant t_i^{*1}$ e $t_{ij} > t_i^{2*}$, ma $W_i(t_{ij}) > s_h^*$ per $t_i^{*1} < t_{ij} \leqslant t_i^{*2}$. Per dividere questa osservazione in questo caso assegneremo tre insiemi di pseudo-osservazioni:

metodo a pseudo-soggetti multipli



(a) Metodo a pseudo-soggetti multipli.



(b) Metodo a 2 pseudo-soggetti

Figura 5: Due metodi per la creazione di Pseudo-Soggetti.

1. $(\tau_{i1}, T^*_{i1}, \delta_{i1}) = (\tau_i, t_i^{*1}, 0)$;
2. $(\tau_{i2}, T^*_{i2}, \delta_{i2}) = (t_i^{*2}, t_i^{*1}, 0)$;
3. $(\tau_{i3}, T^*_{i3}, \delta_{i3}) = (t_i^{*2}, T^*_i, \delta_i)$.

metodo a 2 pseudo-soggetti

Poiché la creazione pseudo-soggetti multipli può rivelarsi particolarmente complessa e ridurre la comprensibilità del modello, Bacchetti e M. R. Segal propongono un secondo metodo per la creazione degli pseudo-soggetti nel caso di variabili tempo-varianti non monotone. Questa tecnica restringe un individuo i ad essere suddiviso in soli due pseudo-soggetti, creando una variabile binaria valorizzata a 1 nel caso in cui i sia a rischio al $j-esimo$ tempo-all'evento, 0 altrimenti. Lo pseudo-soggetto i_1 è assegnato a tutti gli intervalli di tempo dove per i quali la variabile binaria sia $= 1$, e i_2 a quelli per cui è $= 0$.

3 MODELLI DI PROGNOSI DI TRAPIANTO RENALE

3.1 TRAPIANTO RENALE IN ETÀ PEDIATRICA

Per i pazienti in età pediatrica con Insufficienza Renale Terminale (End Stage Renal Disease, ESRD), il trapianto renale è sicuramente la prima opzione terapeutica. Per questa particolare classe di pazienti, candidati a ricevere fino a quattro trapianti durante la loro vita [Groothoff *et al.*, 2004], l'identificazione precoce di soggetti potenzialmente più ad alto rischio di perdita del trapianto potrebbe risultare di grande importanza. Per questa ragione, la valutazione dei fattori non-immunologici pre-trapianto, quali l'età, la malattia di base o la durata della dialisi prima del trapianto è divenuta ormai sempre più decisiva. Questi fattori, insieme all'andamento del follow-up nel primo periodo post-trapianto, possono aiutare, tra le altre cose, a fornire ai clinici informazioni dettagliate per un counselling ottimale per le famiglie in merito alla possibile prognosi del trapianto.

3.2 ALBERI DI SOPRAVVIVENZA SUL TRAPIANTO PEDIATRICO

Il registro della European Society of Pediatric Nephrology/European Renal Association - European Dialysis and Transplant Association (ESPN/ERA-EDTA) è un database europeo che raccoglie i dati dei pazienti in terapia renale sostitutiva provenienti da 31 paesi europei.

il registro ESPN/ERA-EDTA

Per il presente lavoro sono stati selezionati le informazioni riguardanti tutti i trapianti renali presenti all'interno del registro, eseguiti fra il 1990 e il 2009. Solo i pazienti per i quali era presente un follow-up completo e le informazioni sull'outcome del trapianto sono stati selezionati.

3.2.1 Analisi dei Dati

Gli obbiettivi dell'analisi dei dati del registro ESPN sono stati due:

scopo dello studio

1. Confronto di due diversi criteri di split;
2. Confronto degli alberi con la regressione di Cox.

In particolare per questa analisi si è scelto di utilizzare insieme i modelli ad albero e il modello dei rischi proporzionali di Cox, proceden-

do prima con la selezione dell'albero migliore anche con il confronto con gli esperti di dominio, e inserendo in seguito i nodi terminali come variabile dummy all'interno di un modello di regressione, per verificare se l'integrazione di tecniche differenti possa offrire prestazioni migliori e più facilmente interpretabili.

Due analisi differenti sono state condotte:

LA PRIMA prendendo in considerazione solo fattori noti prima del trapianto quali: età del ricevente, sesso, età all'inizio della terapia sostitutiva, durata della dialisi pre-trapianto, tipo di donatore, trapianto preemptive e rischio di recidiva della malattia di base.

LA SECONDA integrando anche alcuni parametri clinici misurati nel primo anno post-trapianto, in particolare: velocità di filtrazione glomerulare (eGFR), emoglobina sierica (Hb), pressione arteriosa sistolica e diastolica e altezza del ricevente.

La sopravvivenza del trapianto è stata calcolata partendo dalla data del trapianto fino alla data del ritorno in dialisi. I dati sono stati censurati nel caso di sopravvivenza oltre i 5 anni, per i pazienti con trapianto funzionante al 31 Dicembre 2009 o in caso di trasferimento ad un centro per adulti.

3.3 FATTORI PRE-TRAPIANTO

Dei 7839 trapianti registrati sul database ESPN ne sono stati selezionati 5275 che non presentavano informazioni sui fattori pre-trapianto mancanti. Nella tabella 1 nella pagina successiva sono presentati i dati del campione analizzato:

3.3.1 Comparazione criteri di split

Per l'analisi dei fattori pre-trapianto nei riceventi pediatrici di trapianto renale si sono confrontati alberi di sopravvivenza costruiti con due criteri di split differenti; uno basato su un criterio intra-nodo, la devianza del modello esponenziale (paragrafo 2.3.2), e un modello basato su uno split inter-nodo, il log-rank test (paragrafo 2.3.1). Le performance in termini di predizione della prognosi sono state misurate tramite lo score di Brier adattato per dati censurati (sezione 1.4 a pagina 6).

misura della performance

Nella figura 6 a pagina 30 è illustrato l'albero per la sopravvivenza del trapianto a 5 anni costruito con il criterio di split della devianza del modello esponenziale. Nei nodi terminali sono rappresentate le curve di sopravvivenza relativi ai pazienti relativi a quel nodo foglia e il numero complessivo dei soggetti classificati in quel sottogruppo.

Il modello è stato sviluppato usando il package di R *rpart* [Therneau *et al.*, 2013], ponendo un limite minimo di soggetti nei nodi fo-

Tabella 1: Dati relativi ai fattori pre-trapianto dei pazienti pediatrici riceventi di trapianto renale

Variabili	n=5275
Genere	
Femmine	2160(41.0)
Età all'inizio della RRT (anni)	10.5(5.4 – 14.3)
Età al trapianto (anni)	11.8(7.1 – 15.3)
Durata della dialisi (anni)[1]	1.03(0.49 – 1.96)
Trapianti preemptive	1248(23.7)
Tipo di donatore	
Vivente	1780(33.7)
Malattia di base	
Rischio recidiva Alto	728(13.8)
FSGS	392(53.9)
MPGN	87(12.0)
PH	46(6.3)
HUS	203(27.9)
Trapianti falliti nei 5 anni	714(13.5)

I dati sono presentati come mediana (range interquartile) per le variabili continue e numero (%) per quelle categoriche.
FSGS: Glomerulosclerosi Focale e Segmentale. MPGN: Glomerulonefrite Membranoproliferativa. PH: Iperossaluria. HUS: Sindrome Emolitica-Uremica.
[1] n = 4027 pazienti non riceventi un trapianto preemptive.

glia pari al 3% dell'intero campione iniziale [Abu-Hanna *et al.*, 2010; Nannings *et al.*, 2008].

Alberi con parametri pre-trapianto

Il primo albero ha selezionato otto sottogruppi utilizzando 5 delle variabili inserite nel modello: durata della dialisi, età al trapianto, rischio di recidiva, tipo di donatore e sesso del ricevente. In particolare il gruppo con la miglior sopravvivenza del trapianto a 5 anni (90.4%) era quello dei pazienti con un età dialitica minore di 3 mesi, mentre nel peggiore erano inclusi riceventi con età dialitica > 2.26 e un età al trapianto maggiore di 13.5 anni (51.7%).

Il secondo albero nella figura 7 a pagina 31 è stato sviluppando definendo un criterio di split inter-nodo, il log-rank test. Questo secondo modello ha individuato sette sottogruppi di pazienti. Il nodo 14, che rappresenta il gruppo con la peggiore sopravvivenza del trapianto a 5 anni stimata (47.4%), è quasi equivalente al nodo del primo albero, mentre il sottogruppo con il tasso inferiore di fallimento (87.1%) del trapianto è individuato da un età dialitica < 0.2 anni è una malattia di base con rischio di recidiva nullo.

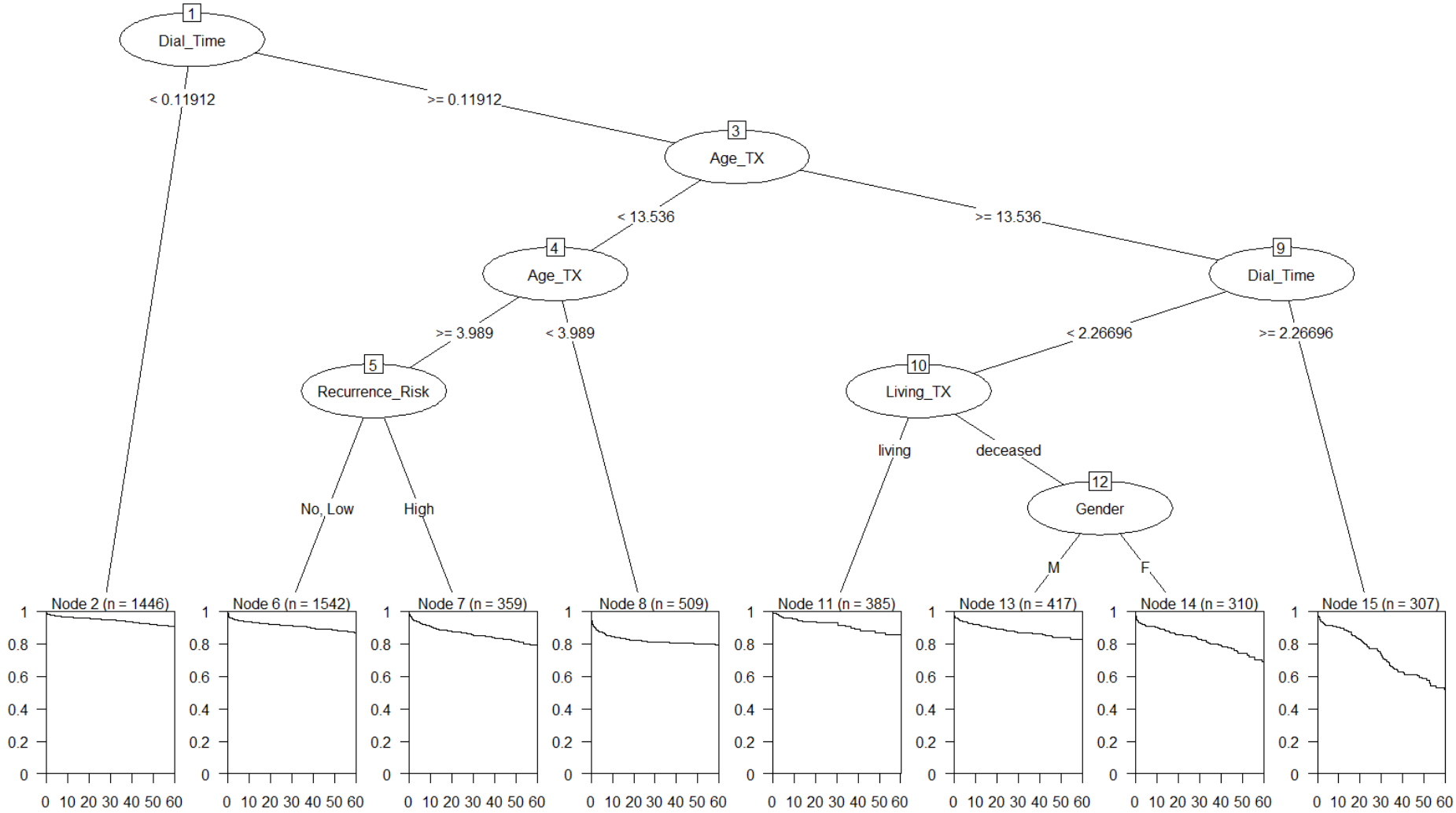


Figura 6: Albero di sopravvivenza a 5 anni del trapianto pediatrico usando parametri pre-trapianto: criterio di split della devianza. Dial_Time: Tempo in dialisi (anni), Age_TX: Età al trapianto (anni), Recurrence_Risk : Rischio di recidiva (No, Low, High), Living_Tx: Donatore vivente (living, deceased), Gender: Sesso del ricevente (M, F)

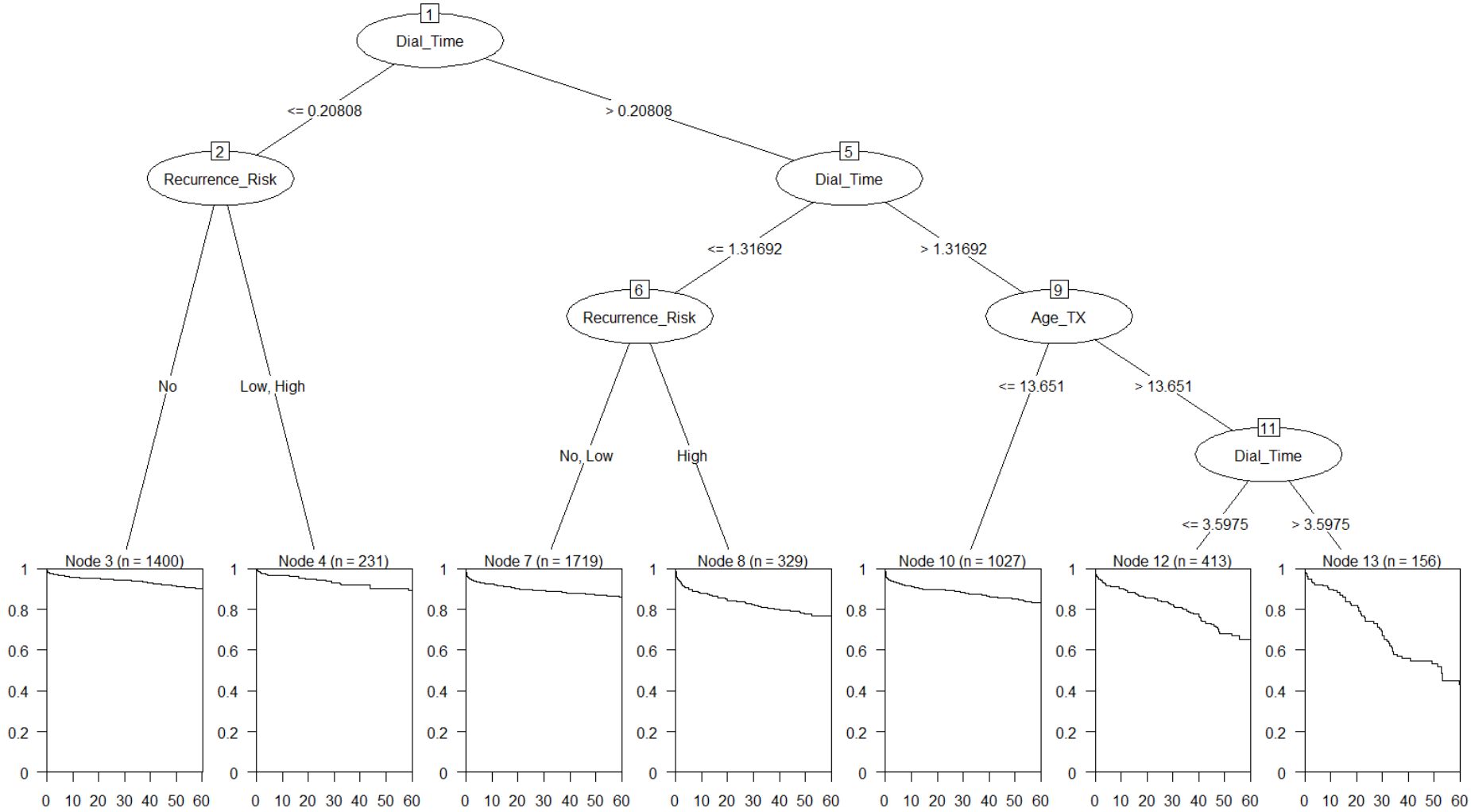


Figura 7: Albero di sopravvivenza a 5 anni del trapianto pediatrico usando parametri pre-trapianto: criterio di split del log-rank test. Dial_Time: Tempo in dialisi (anni), Age_TX: Età al trapianto (anni), Recurrence_Risk : Rischio di recidiva (No, Low, High), Living_Tx: Donatore vivente (living, deceased), Gender: Sesso del ricevente (M, F)

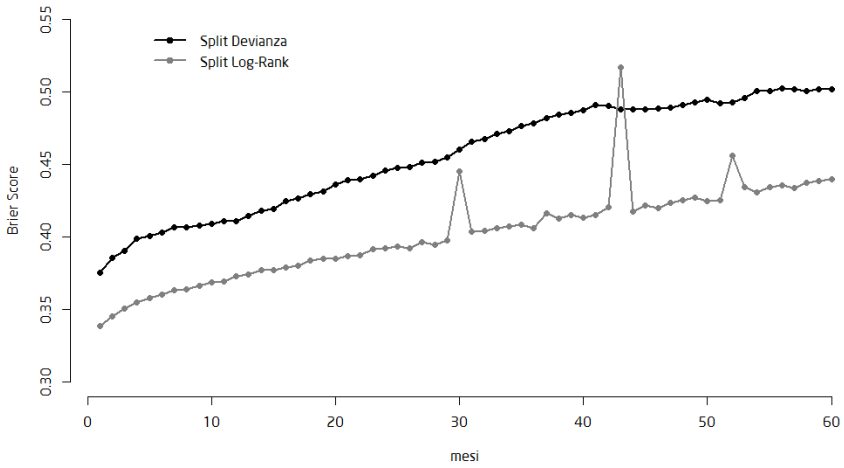


Figura 8: Grafico dello score di Brier per gli alberi di sopravvivenza del trapianto pediatrico

Le prestazioni degli alberi nelle figure 6 e 7 sono state misurate con il Brier score adattato a dati censurati. Nella figura 8 è illustrato il grafico dello score di Brier di entrambi i modelli per i 5 anni di follow-up.

L'albero costruito con il criterio di split della devianza mostra chiaramente migliori prestazioni in termini di previsione della prognosi del trapianto per il tempo considerato.

3.3.2 Integrazione con il modello di Cox

Selezionato l'albero ottimo, è stato costruito un modello che integrasse i modelli ad alberi con la regressione di Cox.

Al set di variabili usate per lo sviluppo degli alberi, è stata aggiunta una variabile dummy che individuasse per ogni paziente il nodo foglio a cui afferiva nell'albero in figura 6 a pagina 30. Per verificare se il framework costruito integrando gli alberi al modello dei rischi proporzionali potesse fornire prestazioni migliori delle analisi convenzionali con regressione di Cox si sono confrontati tre diversi modelli:

- il primo includendo in un modello di Cox solo i fattori pretrapianto (un modello di Cox convenzionale);
- il secondo usando solo la variabile che identificava per ogni paziente in quale dei sottogruppi identificati dall'albero afferisse;
- il terzo utilizzando sia le variabili che i sottogruppi identificati dall'albero.

Tabella 2: Regressione di Cox per la sopravvivenza a 5 anni usando variabili pre-trapianto: Hazard Ratio (95% CI) del modello convenzionale, modello dei sottogruppi e modello integrato

Variabili	*Convenzionale*	*Sottogruppi*	*Integrato*
Femmine	$1.20(1.03-1.39)^1$		$1.12(0.95-1.32)$
Età al trapianto	$1.15(1.01-1.3)^1$		$1.01(0.83-1.24)$
Durata dialisi	$1.13(1.07-1.2)^1$		$1.08(1.01-1.15)$
Preemptive	$0.59(0.47-0.75)^1$		-
Donatore vivente	$0.79(0.66-0.94)^1$		$0.85(0.70-1.02)$
Rischio di recidiva	$1.54(1.28-1.84)^1$		$1.62(1.36-1.95)^c$
Nodo 4		1 (riferimento)	1 (riferimento)
Nodo 6	-	$1.48(1.17-1.88)^1$	-
Nodo 7	-	$2.43(1.79-3.29)^1$	-
Nodo 8	-	$2.67(2.03-3.51)^1$	$1.89(1.42-2.52)^1$
Nodo 11	-	$1.45(0.97-2.15)$	-
Nodo 13	-	$2.08(1.51-2.88)^1$	$1.47(1.06-2.05)^1$
Nodo 14	-	$3.40(2.50-4.61)^1$	$2.11(1.54-2.90)^1$
Nodo 15	-	$5.29(4.01-6.98)^1$	$2.97(2.16-4.07)^1$
C-INDEX	$\mathbf{0.61}(\mathbf{0.59-0.63})$	$\mathbf{0.63}(\mathbf{0.61-0.65})$	$\mathbf{0.65}(\mathbf{0.63-0.67})^2$

[1] $p < 0.001$.
[2] $p < 0.05$ vs. Modelli Convenzionale e Sottogruppi.

Una backward stepwise selection è stata utilizzata per selezionare i modelli di regressioni con le migliori performance predittive. I tre modelli così selezionati sono stati infine comparati usando il *c-index* una statistica equivalente all'area sotto la curva ROC per dati censurati, che misura la concordanza tra la sopravvivenza stimata e quella effettiva [Harrell *et al.*, 1982]. Il c-index finale con l'intervallo di confidenza 95% è stato derivato usando una procedura di bootstrapping su 1000 campioni [Efron e Tibshirani, 1994].

Concordance Index

In tabella 2 sono riportati i risultati delle regressioni di Cox. Dai risultati del C-Index si evince come attraverso l'approccio integrato per l'analisi di sopravvivenza tra alberi e regressione di Cox si ottengano modelli che prevedono in maniera più efficiente la prognosi del trapianto renale.

3.4 DATI CLINICI PRIMO ANNO POST-TRAPIANTO

Lo stesso framework di analisi è stati condotto includendo tra i fattori presi in considerazione, alcuni parametri clinici misurati nel primo anno post-trapianto (mediana 6 mesi, IQR$3.1-8.5$). Natural-

Tabella 3: Dati relativi ai fattori post-trapianto dei pazienti pediatrici riceventi di trapianto renale

Variabili	n=1828
eGFR (ml/min)[1] Imputati 573 (31%)	58.3(47.2 – 69.6)
Emoglobina (d/dl)[1] Imputati 531 (29%)	11.7(10.5 – 12.6)
Pressione Sistolica (SDS)[1] Imputati 567 (31%)	1.06(0.17 – 1.95)
Pressione Sistolica (SDS)[1] Imputati 567 (31%)	1.06(0.17 – 1.95)
Pressione Diastolica (SDS)[1] Imputati 682 (37%)	0.73(0.05 – 1.47)
Altezza (SDS)[1] Imputati 176 (9.6%)	−1.84(−2.72 – −1.00)

I dati sono presentati come mediana (range interquartile).
eGFR: Filtrato Glomerulare stimato. SDS: Standard Deviation Score.
[1] n = 1828 pazienti con dati clinici post-trapianto nel primo anno post-trapianto (dopo una mediana di 6 mesi, IQR 3.1–8.5).

mente tutti i pazienti che avevano perso il trapianto prima della misurazione sono stati esclusi, portando ad una possibile sovrastima della sopravvivenza totale a 5 anni. Inoltre erano stati esclusi tutti i soggetti che non presentavano almeno uno dei parametri clinici analizzati.

Multiple Imputation

In caso di dati mancanti, un algoritmo di imputazione multipla è stato utilizzato come raccomandato dalle linee guida sulla ricerca clinica osservazionale [von Elm *et al.*, 2007]. I dati mancanti sono stati imputati sol nel caso al meno 3 degli altri parametri post-trapianto analizzati fossero presenti. L'algoritmo di imputazione prevedeva la predizione dei dati mancanti attraverso un modello di regressione lineare basato sulle altre variabili presenti e sui valori misurati a tempi successivi. Alla fine del processo di imputazione sono stati selezionati 1828 pazienti dal campione totale.

La tabella 3 mostra i dati relativi ai parametri clinici post-trapianto analizzati sul campione selezionato.

Alberi con parametri post-trapianto

La figura 9 a pagina 36 mostra il primo albero si sopravvivenza costruito analizzando sia informazioni pre- che post-trapianto. Il modello ha selezionato dieci sottogruppi suddivisi utilizzando 5 delle variabili prese in considerazione: durata della dialisi, età al trapianto, età all'inizio della terapia sostitutiva e 2 parametri clinici, il filtrato glomerulare e la pressione diastolica. In questo caso il nodo 4 rappresenta il sottogruppo che mostra la migliore sopravvivenza del

trapianto a 5 anni (97.3%), formato da soggetti riceventi un trapianto dopo un periodo in dialisi < 1.7 anni, un'età al trapianto inferiore agli 8 anni e con un eGFR nel primo anno superiore a 30 ml/min. Il nodo foglia 19 invece è costituito dai pazienti con il rischio più alto di perdita del trapianto a 5 anni (34.7%), formato da soggetti con un lungo periodo pre-trapianto in dialisi, adolescenti e con un filtrato glomerulare inferiore a 63 ml/min.

L'albero costruito usando il log-rank test per implementare il criterio di split è illustrato in figura 10 a pagina 37. Anche in questo caso, il criterio inter-nodo ha selezionato un numero inferiore di sottogruppi, 6 nodi foglia in tutto, utilizzando 3 delle variabili inserite: durata della dialisi, età al trapianto e eGFR. Il gruppo individuato a più basso rischio di perdita del trapianto (sopravvivenza del 95.9%) era quello dei pazienti con età dialitica < 5.5 anni e un'età al trapianto inferiore a 11.6 anni, mentre i soggetti a più alto rischio (sopravvivenza a 5 anni 29.2%) presentavano un periodo in dialisi pre-trapianto più ungo di 5.5 anni e un'età al trapianto > 13.8 anni.

Come chiaramente illustrato in figura 11 a pagina 38, anche inserendo i parametri clinici post-trapianto, il modello che utilizza la devianza come criterio di split mostra delle misure di performance migliori rispetto all'albero costruito con criterio inter-nodo, per tutti i tempi analizzati.

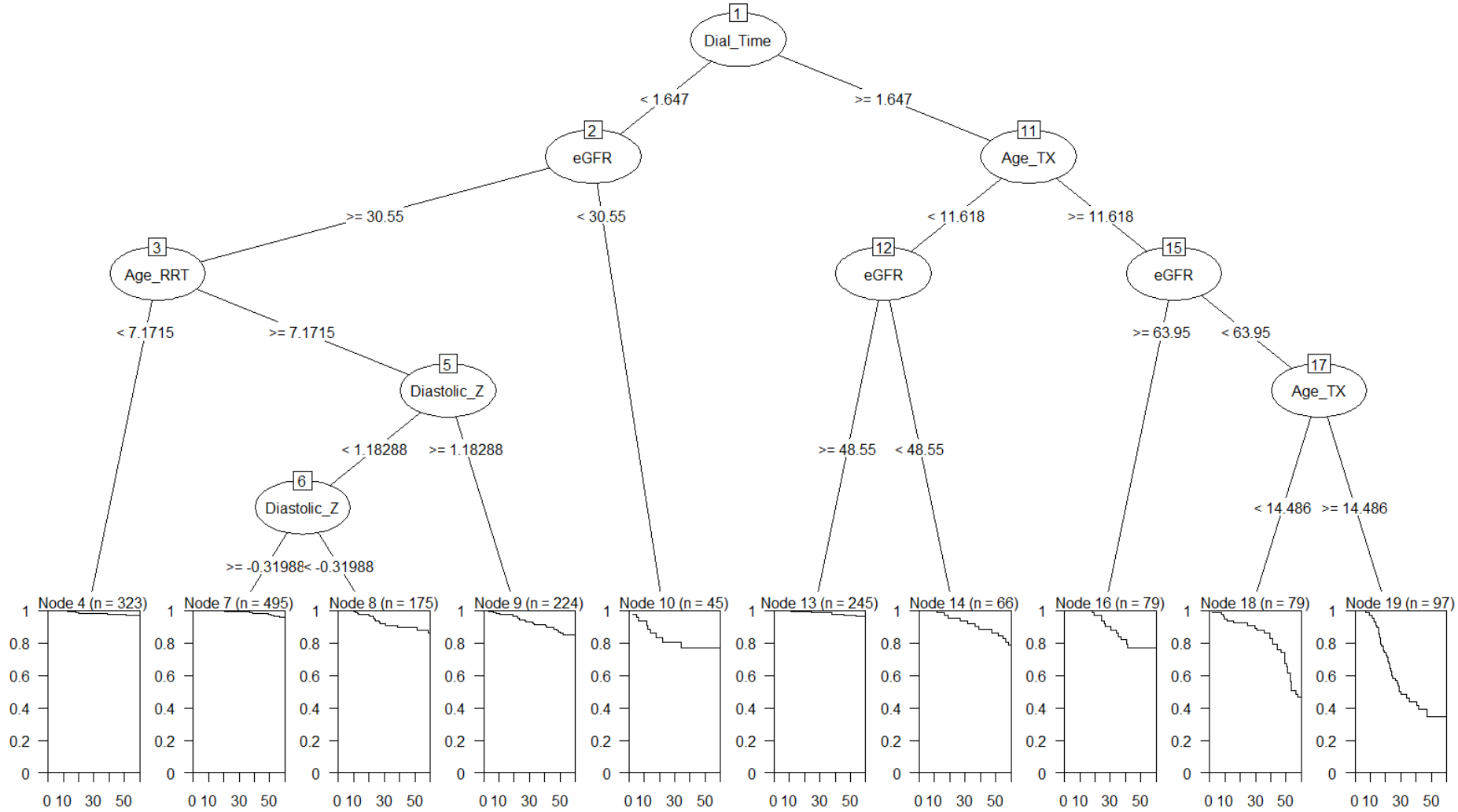


Figura 9: Albero di sopravvivenza a 5 anni del trapianto pediatrico usando anche parametri clinici post-trapianto: criterio di split della devianza. Dial_Time: Tempo in dialisi (anni), Age_TX: Età al trapianto (anni), Age_RRT: Età all'inizio della terapia sostitutiva (anni), eGFR: Filtrato Glomerulare (ml/min), Diastolic_Z: Pressione Diastolica (SDS)

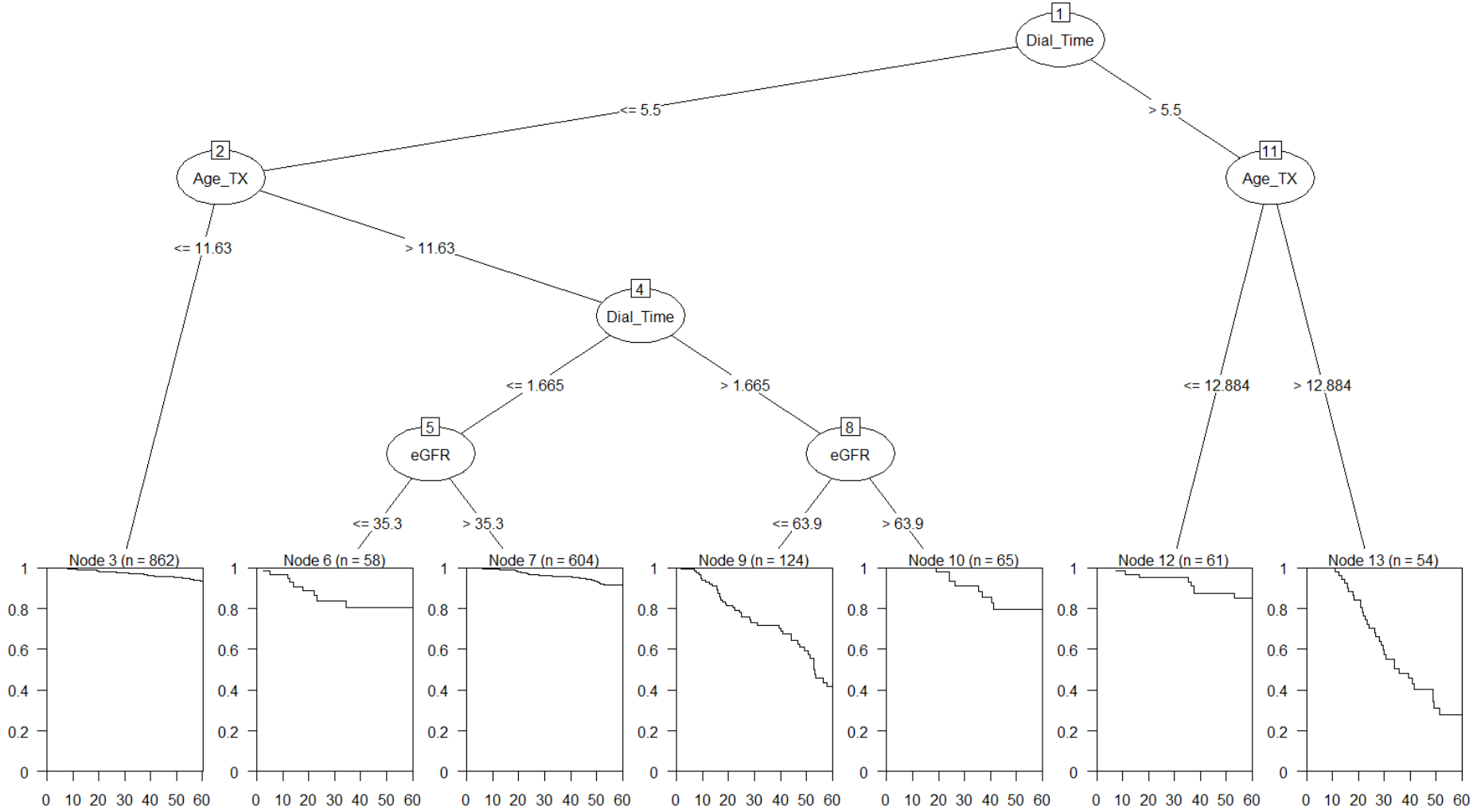


Figura 10: Albero di sopravvivenza a 5 anni del trapianto pediatrico usando anche parametri clinici post-trapianto: criterio di split del log-rank test. Dial_Time: Tempo in dialisi (anni), Age_TX: Età al trapianto (anni), eGFR: Filtrato Glomerulare

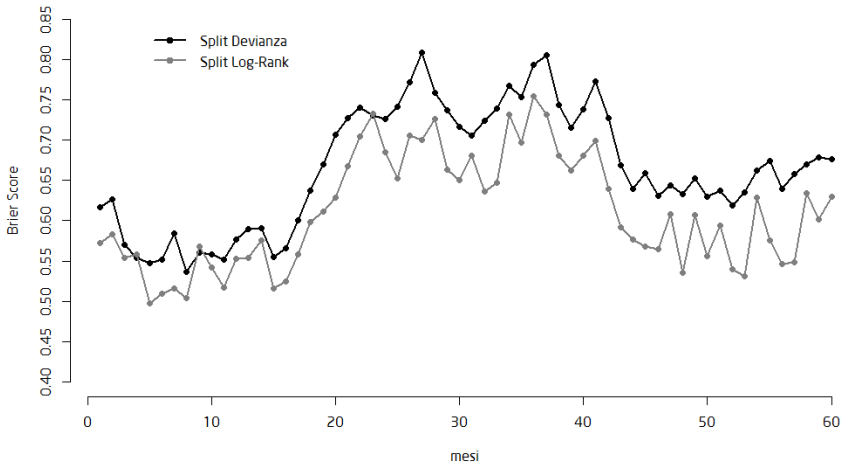


Figura 11: Grafico dello score di Brier per gli alberi di sopravvivenza del trapianto pediatrico

3.4.1 Integrazione con il modello di Cox

Anche per il modello con dati post-trapianto è stato costruito un modello che integrasse i modelli ad alberi con la regressione di Cox.

Come nel caso precedente al set di variabili usate per lo sviluppo degli alberi, è stata aggiunta una variabile dummy che individuasse per ogni paziente il nodo foglio a cui afferiva nell'albero in figura 7 a pagina 31. Si sono infine confrontati tre diversi modelli:

- il modello convenzionale che includeva solo i fattori pre- e post-trapianto;
- il secondo usando solo la variabile che identificava per ogni paziente in quale dei sottogruppi identificati dall'albero afferisse;
- il terzo utilizzando sia le variabili che i sottogruppi identificati dall'albero.

Anche qui si è usata la backward stepwise selection per selezionare i modelli di regressioni con le migliori performance predittive e infine comparati usando il *c-index*.

In tabella 4 nella pagina successiva sono riportati i risultati delle regressioni di Cox. Anche in questo setting sperimentale i risultati del C-Index mostrano come l'approccio integrato alberi-regressione di Cox, anche nel caso di utilizzo di parametri di follow-up si ottengano modelli che prevedono in maniera più efficiente la prognosi del trapianto renale.

Tabella 4: Regressione di Cox per la sopravvivenza a 5 anni usando sia variabili misurate sia pre- che post-trapianto: Hazard Ratio (95% CI) del modello convenzionale, modello dei sottogruppi e modello integrato

Variabili	*Convenzionale*	*Sottogruppi*	*Integrato*
Femmine	$1.20(1.03-1.39)^1$		$1.12(0.95-1.32)$
Età al trapianto	$1.15(1.11-1.20)^1$		$1.02(0.96-1.08)$
Durata dialisi	$1.21(1.16-1.27)^1$		$1.05(0.99-1.12)$
eGFR	$0.98(0.97-0.99)^1$		$0.99(0.98-1.01)$
Pressione Diastolica	–		$0.99(0.83-1.17)$
Nodo 4		1 (riferimento)	1 (riferimento)
Nodo 8	-	$4.66(2.47-8.81)^1$	$4.29(2.15-8.57)^1$
Nodo 9	-	$4.75(2.74-8.22)^1$	$4.42(2.42-8.07)^1$
Nodo 10	-	$8.95(4.31-18.56)^1$	$10.15(4.54-22.72)^1$
Nodo 14	-	$6.37(3.22-12.63)^1$	$4.92(2.34-10.31)^1$
Nodo 16	-	$7.54(3.8-14.95)^1$	$6.73(3.05-14.85)^1$
Nodo 18	-	$17.41(10.37-29.22)^1$	$12.44(6.65-23.25)^1$
Nodo 19	-	$39.4(23.98-64.73)^1$	$25.14(12.48-50.63)^1$
C-INDEX	$\mathbf{0.74(0.70-0.79)}$	$\mathbf{0.79(0.75-0.84)}$	$\mathbf{0.84(0.81-0.87)}^2$

[1] $p < 0.001$.
[2] $p < 0.05$ vs. Modelli Convenzionale e Sottogruppi.

3.5 CONCLUSIONI

In conclusione è possibile affermare che attraverso l'applicazione nel dominio del trapianto renale si è potuto dimostrare come:

- Gli alberi di sopravvivenza siano un tool adatto all'analisi della prognosi del trapianto renale, avendo mostrato flessibilità e robustezza delle soluzioni;
- Che l'approccio integrato Alberi di Sopravvivenza-Regressioni di Cox, non solo offre soluzioni con performance predittive migliori, ma risulta anche di più facile interpretazione per gli esperti di dominio.

BIBLIOGRAFIA

Abu-Hanna, Ameen, Barry Nannings, Dave Dongelmans e Arie Hasman

2010 "PRIM versus CART in subgroup discovery: When patience is harmful", *Journal of Biomedical Informatics*, 43, 5, p. 701-708, ISSN: 1532-0464, DOI: http://dx.doi.org/10.1016/j.jbi.2010.05.009, http://www.sciencedirect.com/science/article/pii/S1532046410000675. (Citato a p. 29.)

Ahn, H. e W. Y. Loh

1994 "Tree-structured proportional hazards regression modeling." eng, *Biometrics*, 50, 2 (giu. 1994), p. 471-485. (Citato a p. 11.)

Bacchetti, P. e M. R. Segal

1995 "Survival trees with time-dependent covariates: application to estimating changes in the incubation period of AIDS." eng, *Lifetime Data Anal*, 1, 1, p. 35-47. (Citato alle p. 13, 23-25.)

Breiman, Leo, Jerome Friedman, Charles J. Stone e R.A. Olshen

1984 *Classification and Regression Trees*, Chapman e Hall/CRC, ISBN: 0412048418. (Citato alle p. 9, 10, 12, 15-17, 19.)

BRESLOW, NORMAN

1970 "A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship", *Biometrika*, 57, 3, p. 579-594, DOI: 10.1093/biomet/57.3.579, http://dx.doi.org/10.1093/biomet/57.3.579. (Citato a p. 13.)

Breslow, Norman

1972 "Contribution to the discussion of paper by D.R. Cox. J R", *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 216-217. (Citato a p. 14.)

BRIER, GLENN W.

1950 "VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY", *Monthly Weather Review*, 78, 1, p. 1-3. (Citato a p. 6.)

Ciampi, A

1995 "Tree-structured prediction for censored survival data and the cox model", *Journal of Clinical Epidemiology*, 48, 5 (mag. 1995), p. 675-689, DOI: 10.1016/0895-4356(94)00164-L, http://dx.doi.org/10.1016/0895-4356(94)00164-L. (Citato a p. 11.)

Ciampi, A., R. S. Bush, M. Gospodarowicz e J. E. Till

1981 "An approach to classifying prognostic factors related to survival experience for non-Hodgkin's lymphoma patients: based on a series of 982 patients: 1967-1975." eng, 47, 3 (feb. 1981), p. 621-627. (Citato a p. 10.)

Ciampi, Antonio, Sheilah A. Hogg, Steve McKinney e Johanne Thiffault

1988 "RECPAM: a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. I. Methods and program features", *Computer Methods and Programs in Biomedicine*, 26, 3 (mag. 1988), p. 239-256, DOI: 10.1016/0169-2607(88)90004-1, http://dx.doi.org/10.1016/0169-2607(88)90004-1. (Citato a p. 11.)

Colvin, Robert B.

2003 "Chronic Allograft Nephropathy", *New England Journal of Medicine*, 349, 24, PMID: 14668453, p. 2288-2290, DOI: 10.1056/NEJMp038178, eprint: http://www.nejm.org/doi/pdf/10.1056/NEJMp038178, http://www.nejm.org/doi/full/10.1056/NEJMp038178. (Citato a p. vii.)

Cox, David R

1972 "Regression models and life-tables", *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 187-220, http://www.jstor.org/stable/10.2307/2985181. (Citato alle p. 4, 13.)

1975 "Partial likelihood", *Biometrika*, 62, 2, p. 269-276, http://biomet.oxfordjournals.org/content/62/2/269.short. (Citato a p. 4.)

Davis, R. B. e J. R. Anderson

1989 "Exponential survival trees." eng, *Stat Med*, 8, 8 (ago. 1989), p. 947-961. (Citato alle p. 11, 20, 22.)

Efron, Bradley

1983 "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation", *Journal of the American Statistical Association*, 78, 382 (giu. 1983), p. 316-331, DOI: 10.1080/01621459.1983.10477973, http://dx.doi.org/10.1080/01621459.1983.10477973. (Citato a p. 20.)

Efron, Bradley e R.J. Tibshirani

1994 *An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*, Chapman e Hall/-CRC, ISBN: 0412042312, http://www.amazon.com/Introduction-Bootstrap-Monographs-Statistics-Probability/dp/

0412042312?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0412042312. (Citato a p. 33.)

Fellström, Bengt, Hallvard Holdaas, Alan G. Jardine, Gudrun Nyberg, Carola Grönhagen-Riska, Sören Madsen, Hans-Hellmut Neumayer, Edward Cole, Bart Maes, Patrice Ambühl e et al.

2005 "Risk Factors for Reaching Renal Endpoints in the Assessment of Lescol in Renal Transplantation (ALERT) Trial", *Transplantation*, 79, 2 (gen. 2005), p. 205-212, DOI: 10.1097/01.TP.0000147338.34323.12, http://dx.doi.org/10.1097/01.TP.0000147338.34323.12. (Citato a p. vii.)

First, M. Roy

2003 "Renal function as a predictor of long term graft survival in renal transplant patients", *Nephrology Dialysis Transplantation*, 18, suppl 1, p. i3-i6, DOI: 10.1093/ndt/gfg1027, eprint: http://ndt.oxfordjournals.org/content/18/suppl_1/i3.full.pdf+html, http://ndt.oxfordjournals.org/content/18/suppl_1/i3.abstract. (Citato a p. vii.)

Fisher, Lloyd D e DY Lin

1999 "Time-dependent covariates in the Cox proportional-hazards regression model", *Annu Rev Public Health*, 20, 1, p. 145-157, http://www.annualreviews.org/doi/abs/10.1146/annurev.publhealth.20.1.145. (Citato a p. 6.)

GEHAN, E. A.

1965 "A GENERALIZED WILCOXON TEST FOR COMPARING ARBITRARILY SINGLY-CENSORED SAMPLES." eng, *Biometrika*, 52 (giu. 1965), p. 203-223. (Citato a p. 13.)

Goldfarb-Rumyantzev, Alexander S, John D Scandling, Lisa Pappas, Randall J Smout e Susan Horn

2003 "Prediction of 3-yr cadaveric graft survival based on pre-transplant variables in a large national dataset", *Clinical Transplantation*, 17, 6, p. 485-497, ISSN: 1399-0012, DOI: 10.1046/j.0902-0063.2003.00051.x, http://dx.doi.org/10.1046/j.0902-0063.2003.00051.x. (Citato a p. vii.)

Gondos, Adam, Bernd Döhler, Hermann Brenner e Gerhard Opelz

2013 "Kidney Graft Survival in Europe and the United States", *Transplantation Journal*, 95, 2 (gen. 2013), p. 267-274, DOI: 10.1097/TP.0b013e3182708ea8, http://dx.doi.org/10.1097/TP.0b013e3182708ea8. (Citato a p. vii.)

Gordon, L. e R. A. Olshen

1985 "Tree-structured survival analysis." eng, *Cancer Treat Rep*, 69, 10 (ott. 1985), p. 1065-1069. (Citato a p. 10.)

Graf, Erika, Claudia Schmoor, Willi Sauerbrei e Martin Schumacher

1999 "Assessment and comparison of prognostic classification schemes for survival data", *Stat Med*, 18, 17-18, p. 2529-2545, http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19990915/30)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5/abstract. (Citato alle p. 6, 7.)

Groothoff, Jaap W., Karlien Cransberg, Martin Offringa, Nicole J. van de Kar, Marc R. Lilien, Jean Claude Davin e Hugo S. A. Heymans

2004 "Long-Term Follow-Up of Renal Transplantation in Children: A Dutch Cohort Study", *Transplantation*, 78, 3 (ago. 2004), p. 453-460, DOI: 10.1097/01.TP.0000128616.02821.8B, http://dx.doi.org/10.1097/01.TP.0000128616.02821.8B. (Citato a p. 27.)

Hariharan, Sundaram, Maureen A Mcbride, Wida S Cherikh, Christine B Tolleris, Barbara A Bresnahan e Christopher P Johnson

2002 "Post-transplant renal function in the first year predicts long-term kidney transplant survival", *Kidney International*, 62, 1 (lug. 2002), p. 311-318, DOI: 10.1046/j.1523-1755.2002.00424.x, http://dx.doi.org/10.1046/j.1523-1755.2002.00424.x. (Citato a p. vii.)

Harrell, FE Jr, RM Califf, DB Pryor, KL Lee e RA Rosati

1982 "EValuating the yield of medical tests", *JAMA*, 247, 18, p. 2543-2546, DOI: 10.1001/jama.1982.03320430047030, eprint: /data/Journals/JAMA/9062/jama_247_18_030.pdf, +%20http://dx.doi.org/10.1001/jama.1982.03320430047030. (Citato a p. 33.)

HARRINGTON, DAVID P. e THOMAS R. FLEMING

1982 "A class of rank test procedures for censored survival data", *Biometrika*, 69, 3, p. 553-566, DOI: 10.1093/biomet/69.3.553, http://dx.doi.org/10.1093/biomet/69.3.553. (Citato a p. 12.)

Hothorn, Torsten, Berthold Lausen, Axel Benner e Martin Radespiel-Tröger

2004 "Bagging survival trees", English, *Statistics in Medicine*, 23, 1, p. 77-91. (Citato a p. 6.)

Huang, Xin, Shande Chen e Seng-jaw Soong

1998 "Piecewise Exponential Survival Trees with Time-Dependent Covariates", English, *Biometrics*, 54, 4, p. 1420-1433, ISSN: 0006341X, http://www.jstor.org/stable/2533668. (Citato alle p. 22, 23.)

Kaplan, Edward L e Paul Meier

1958 "Nonparametric estimation from incomplete observations", *J Am Stat Assoc*, 53, 282, p. 457-481, http://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452. (Citato a p. 3.)

Kasiske, Bertram L., Robert S. Gaston, Sita Gourishankar, Philip F. Halloran, Arthur J. Matas, John Jeffery e David Rush

2005 "Long-Term Deterioration of Kidney Allograft Function", *American Journal of Transplantation*, 5, 6, p. 1405-1414, ISSN: 1600-6143, DOI: 10.1111/j.1600-6143.2005.00853.x, http://dx.doi.org/10.1111/j.1600-6143.2005.00853.x. (Citato a p. vii.)

Krikov, Sergey, Altaf Khan, Bradley C. Baird, Lev L. Barenbaum, Alexander Leviatov, James K. Koford e Alexander S. Goldfarb-Rumyantzev

2007 "Predicting Kidney Transplant Survival Using Tree-Based Modeling", *ASAIO Journal*, 53, 5 (set. 2007), p. 592-600, DOI: 10.1097/MAT.0b013e318145b9f7, http://dx.doi.org/10.1097/MAT.0b013e318145b9f7. (Citato a p. vii.)

LeBlanc, M. e J. Crowley

1992 "Relative risk trees for censored survival data." eng, *Biometrics*, 48, 2 (giu. 1992), p. 411-425. (Citato alle p. 9, 11, 12, 14, 16, 17, 19.)

1993 "Survival Trees by Goodness of Split", *Journal of the American Statistical Association*, 88, 422 (giu. 1993), p. 457-467, DOI: 10.1080/01621459.1993.10476296, http://dx.doi.org/10.1080/01621459.1993.10476296. (Citato alle p. 11, 12, 15-17, 19, 20.)

Mantel, N.

1966 "Evaluation of survival data and two new rank order statistics arising in its consideration." eng, *Cancer Chemother Rep*, 50, 3 (mar. 1966), p. 163-170. (Citato a p. 13.)

Marubini, E., A. Morabito e M. G. Valsecchi

1983 "Prognostic factors and risk groups: some results given by using an algorithm suitable for censored survival data." eng, *Stat Med*, 2, 2, p. 295-303. (Citato a p. 10.)

Meier-Kriesche, Herwig-Ulf, Jesse D. Schold, Titte R. Srinivas e Bruce Kaplan

2004 "Lack of Improvement in Renal Allograft Survival Despite a Marked Decrease in Acute Rejection Rates Over the Most Recent Era", *American Journal of Transplantation*, 4, 3, p. 378-383, ISSN: 1600-6143, DOI: 10.1111/j.1600-6143.2004.00332

.x, http://dx.doi.org/10.1111/j.1600-6143.2004.00332.x. (Citato a p. vii.)

Moeschberger, Melvin L e John P Klein

2003 *Survival analysis: Techniques for censored and truncated data*, Springer. (Citato alle p. 2, 4, 5.)

Morgan, James N e John A Sonquist

1963 "Problems in the analysis of survey data, and a proposal", *J Am Stat Assoc*, 58, 302, p. 415-434, http://amstat.tandfonline.com/doi/full/10.1080/01621459.1963.10500855. (Citato a p. 10.)

Nannings, Barry, Ameen Abu-Hanna e Evert de Jonge

2008 "Applying PRIM (Patient Rule Induction Method) and logistic regression for selecting high-risk subgroups in very elderly ICU patients", *International Journal of Medical Informatics*, 77, 4, p. 272-279, ISSN: 1386-5056, DOI: http://dx.doi.org/10.1016/j.ijmedinf.2007.06.007, http://www.sciencedirect.com/science/article/pii/S1386505607001220. (Citato a p. 29.)

Negassa, Abdissa, Antonio Ciampi, Stanley Abrahamowicz Michal and Shapiro e Jean-François Boivin

2005 "Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria", *Statistics and Computing*, 15, 3 (lug. 2005), p. 231-239, DOI: 10.1007/s11222-005-1311-z, http://dx.doi.org/10.1007/s11222-005-1311-z. (Citato a p. 20.)

Pascual, Manuel, Tom Theruvath, Tatsuo Kawai, Nina Tolkoff-Rubin e A. Benedict Cosimi

2002 "Strategies to Improve Long-Term Outcomes after Renal Transplantation", *New England Journal of Medicine*, 346, 8, PMID: 11856798, p. 580-590, DOI: 10.1056/NEJMra011295, eprint: http://www.nejm.org/doi/pdf/10.1056/NEJMra011295, http://www.nejm.org/doi/full/10.1056/NEJMra011295. (Citato a p. vii.)

Peterson Jr, Arthur V

1977 "Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions", *J Am Stat Assoc*, 72, 360a, p. 854-858, http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1977.10479970. (Citato a p. 3.)

Ponticelli, Claudio e Giorgio Graziani

2012 "Education and counseling of renal transplant recipients", *Journal of Nephrology*, 25, 6 (nov. 2012), p. 879-889, DOI: 10.5301/jn.5000227, http://dx.doi.org/10.5301/jn.5000227. (Citato a p. vii.)

Quiroga, Isabel, Philip McShane, Dicken D. H. Koo, Derek Gray, Peter J. Friend, Susan Fuggle e Christopher Darby

2006 "Major effects of delayed graft function and cold ischaemia time on renal allograft survival", *Nephrology Dialysis Transplantation*, 21, 6, p. 1689-1696, DOI: 10.1093/ndt/gfl042, eprint: http://ndt.oxfordjournals.org/content/21/6/1689.full.pdf+html, http://ndt.oxfordjournals.org/content/21/6/1689.abstract. (Citato a p. vii.)

RABBAT, CHRISTIAN G., KEVIN E. THORPE, J. DAVID RUSSELL e DAVID N. CHURCHILL

2000 "Comparison of Mortality Risk for Dialysis Patients and Cadaveric First Renal Transplant Recipients in Ontario, Canada", *Journal of the American Society of Nephrology*, 11, 5, p. 917-922, eprint: http://jasn.asnjournals.org/content/11/5/917.full.pdf+html, http://jasn.asnjournals.org/content/11/5/917.abstract. (Citato a p. vii.)

Rao, Panduranga S., Douglas E. Schaubel, Mary K. Guidinger, Kenneth A. Andreoni, Robert A. Wolfe, Robert M. Merion, Friedrich K. Port e Randall S. Sung

2009 "A Comprehensive Risk Quantification Score for Deceased Donor Kidneys: The Kidney Donor Risk Index", *Transplantation*, 88, 2 (lug. 2009), p. 231-236, DOI: 10.1097/TP.0b013e3181ac620b, http://dx.doi.org/10.1097/TP.0b013e3181ac620b. (Citato a p. vii.)

Rao, Panduranga S., Douglas E. Schaubel, Xiaoyu Jia, Shiqian Li, Friedrich K. Port e Rajiv Saran

2007 "Survival on Dialysis Post–Kidney Transplant Failure: Results From the Scientific Registry of Transplant Recipients", *American Journal of Kidney Diseases*, 49, 2, p. 294-300, ISSN: 0272-6386, DOI: http://dx.doi.org/10.1053/j.ajkd.2006.11.022, http://www.sciencedirect.com/science/article/pii/S0272638606016970. (Citato a p. vii.)

Rao, Panduranga S., Douglas E. Schaubel e Rajiv Saran

2005 "Impact of graft failure on patient survival on dialysis: a comparison of transplant-naïve and post-graft failure mortality rates", *Nephrology Dialysis Transplantation*, 20, 2, p. 387-391, DOI: 10.1093/ndt/gfh595, eprint: http://ndt.oxfordjournals.org/content/20/2/387.full.pdf+html, http://ndt.oxfordjournals.org/content/20/2/387.abstract. (Citato a p. vii.)

Schnuelle, P, D Lorenz, M Trede e F J Van Der Woude

1998 "Impact of renal cadaveric transplantation on survival in end-stage renal failure: evidence for reduced mortality risk compared with hemodialysis during long-term follow-up." *Journal of the American Society of Nephrology*, 9, 11, p. 2135-41, eprint: http://jasn.asnjournals.org/content/9/11/2135.full.pdf+html, http://jasn.asnjournals.org/content/9/11/2135.abstract. (Citato a p. vii.)

Schoop, R., E. Graf e M. Schumacher

2008 "Quantifying the Predictive Performance of Prognostic Models for Censored Survival Data with Time-Dependent Covariates", *Biometrics*, 64, 2 (giu. 2008), p. 603-610, DOI: 10.1111/j.1541-0420.2007.00889.x, http://dx.doi.org/10.1111/j.1541-0420.2007.00889.x. (Citato a p. 7.)

Segal, Mark Robert

1988 "Regression Trees for Censored Data", *Biometrics*, 44, 1 (mar. 1988), p. 35, DOI: 10.2307/2531894, http://dx.doi.org/10.2307/2531894. (Citato alle p. 11, 12.)

1997 "Features of tree-structured survival analysis", *Epidemiology*, 8, 4 (lug. 1997), p. 344. (Citato a p. 21.)

TARONE, ROBERT E. e JAMES WARE

1977 "On distribution-free tests for equality of survival distributions", *Biometrika*, 64, 1, p. 156-160, DOI: 10.1093/biomet/64.1.156, http://dx.doi.org/10.1093/biomet/64.1.156. (Citato alle p. 11-13.)

Terasaki, Paul I. e Miyuki Ozawa

2004 "Predicting Kidney Graft Failure by HLA Antibodies: a Prospective Trial", *American Journal of Transplantation*, 4, 3, p. 438-443, ISSN: 1600-6143, DOI: 10.1111/j.1600-6143.2004.00360.x, http://dx.doi.org/10.1111/j.1600-6143.2004.00360.x. (Citato a p. vii.)

Therneau, Terry, Beth Atkinson e Brian Ripley

2013 *rpart: Recursive Partitioning*, R package version 4.1-3, http://CRAN.R-project.org/package=rpart. (Citato alle p. 11, 28.)

Toma, Hiroshi, Kazunari Tanabe, Tadahiko Tokumoto, Tomokazu Shimizu e Hiroaki Shimmura

2001 "TIME-DEPENDENT RISK FACTORS INFLUENCING THE LONG-TERM OUTCOME IN LIVING RENAL ALLOGRAFTS", *Transplantation*, 72, 5 (set. 2001), p. 941-947, DOI: 10.1097/00007890-200109150-00033, http://dx.doi.org/10.1097/00007890-200109150-00033. (Citato a p. vii.)

Von Elm, Erik, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gotzsche e Jan P. Vandenbroucke

2007 "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies", *Annals of Internal Medicine*, 147, 8, p. 573-577, DOI: 10.7326/0003-4819-147-8-200710160-00010, http://dx.doi.org/10.7326/0003-4819-147-8-200710160-00010. (Citato a p. 34.)

Wolfe, Robert A., Valarie B. Ashby, Edgar L. Milford, Akinlolu O. Ojo, Robert E. Ettenger, Lawrence Y.C. Agodoa, Philip J. Held e Friedrich K. Port

1999 "Comparison of Mortality in All Patients on Dialysis, Patients on Dialysis Awaiting Transplantation, and Recipients of a First Cadaveric Transplant", *New England Journal of Medicine*, 341, 23, PMID: 10580071, p. 1725-1730, DOI: 10.1056/NEJM199912023412303, http://www.nejm.org/doi/full/10.1056/NEJM199912023412303. (Citato a p. vii.)