UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA SPECIALISTICA IN SCIENZE STATISTICHE DEMOGRAFICHE E SOCIALI

TESI DI LAUREA

CURVE R.O.C. PER LA VALUTAZIONE DI TEST DIAGNOSTICI PER EVENTI NEL TEMPO

RELATORE: Prof.ssa MONICA CHIOGNA

LAUREANDA: ERICA BINO

ANNO ACCADEMICO 2009-2010

Alla mia famiglia

in particolare al mio papà

Indice

1 Le curve di R.O.C (Receiver Operating Characteristic)

1.1 Introduzione

In tutti i campi della scienza vengono sistematicamente messe a punto e utilizzate procedure più o meno complesse e della più svariata natura, ma sempre ben codificate, allo scopo di verificare un'ipotesi. Tali procedure sono comunemente dette "test". In particolare, in epidemiologia, i test rappresentano lo strumento di base nelle operazioni di *screening*, eseguite cioè su popolazioni presuntivamente sane (e nelle quali la prevalenza della malattia in studio è ignota) allo scopo di identificare precocemente la presenza di malattie. Anche nell'attività diagnostica di *routine* i test rappresentano elementi fondamentali, e spesso determinanti, nel processo decisionale volto a confermare (o escludere) la presenza di una determinata malattia già sospettata in base ai dati clinici.
In base alla tipologia di responso fornito, i test possono essere classificati in due categorie:

- i test "qualitativi", ossia che restituiscono un *output* (risposta) dicotomico (es. positivo/negativo, vero/falso ecc.);
- i test "quantitativi", ossia che producono risultati sotto forma di variabili numeriche "discrete" o "continue"

Per i test quantitativi, l'ottenimento di risultati affidabili è subordinato alla condizione che il parametro misurato possieda una distribuzione approssimativamente unimodale sia nella classe dei soggetti sani che in quella degli ammalati, ovviamente con medie differenti per ciascuna classe. Tale ipotesi di distribuzione viene detta "bi-normale".
È evidente che, per i test quantitativi (siano essi discreti o continui), occorre individuare sulla scala di lettura un valore-soglia ("*cut off*") che discrimini i risultati da dichiarare "positivi" da quelli "negativi". Ciò consente di categorizzare in "positivi" e "negativi" la gamma di tutti possibili risultati e di equiparare l'interpretazione di un test quantitativo a quella di un test qualitativo.

Il problema di base che genera incertezza nell'interpretazione di un test risiede nel fatto che – nella grande maggioranza dei casi – esiste una zona di sovrapposizione fra le distribuzioni dei risultati del test medesimo applicato in popolazioni di soggetti rispettivamente sani e ammalati.
Infatti, se le due popolazioni restituissero valori separati (Figura 1) allora sarebbe facile individuare sull'asse delle ascisse il valore di cut off capace di discriminare con precisione assoluta le due popolazioni.

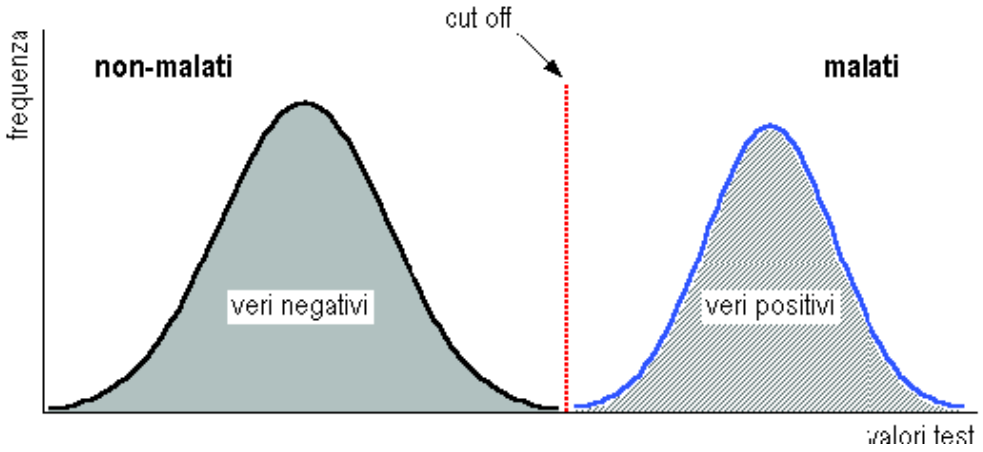


Figura 1: Distribuzione degli esiti di un ipotetico test nelle classi di individui malati e non malati, senza sovrapposizione inter-classe.

Purtroppo, invece, nella pratica si verifica sempre una sovrapposizione più o meno ampia delle due distribuzioni (Figura 2) ed è perciò impossibile individuare sull'asse delle ascisse un valore di cut off che consenta una classificazione perfetta, ossia tale da azzerare sia i falsi positivi che i falsi negativi.

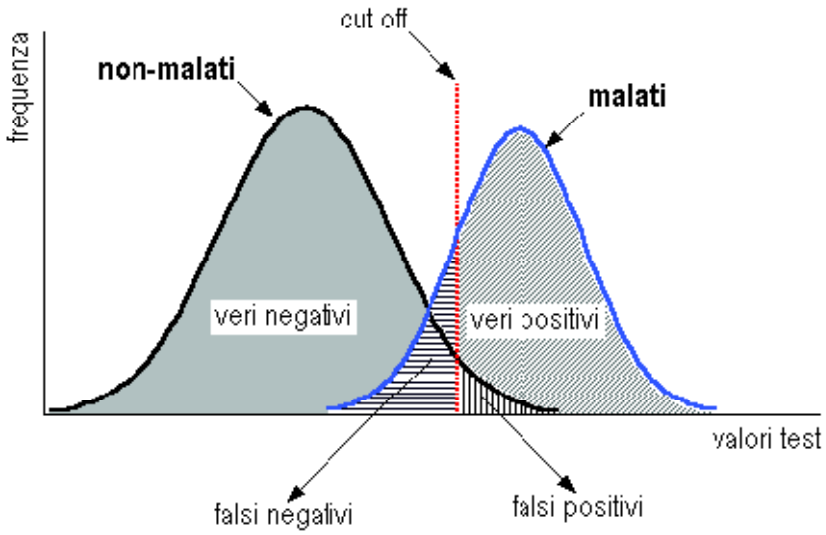


Figura 2: Distribuzione degli esiti di un ipotetico test nelle classi di individui malati e non malati, con sovrapposizione inter-classe.

La capacità diagnostica (o validità, o *performance*) di un test ad un determinato valore di *cut off* rappresenta la capacità di condurre ad una diagnosi positiva nei soggetti affetti da una determinata malattia e ad una diagnosi negativa nei soggetti non ammalati. Essa può essere valutata attraverso una semplice tabella di contingenza (Tabella 1), confrontando l'*output* del test in esame con il vero stato dei soggetti. Quest'ultimo può essere già noto in partenza oppure può essere stabilito per mezzo di un test di riferimento provvisto della più alta attendibilità ("test aureo"), possibilmente basato su un principio biologico diverso rispetto al test da valutare. In genere i "test aurei" presentano alcuni svantaggi (es. difficile somministrazione, costo elevato, ecc.) che li rendono inapplicabili di *routine* o in operazioni di *screening*. Ai fini del raffronto con il test in studio, nell'ipotesi più semplice si assume che il "test aureo" fornisca risultati perfettamente corrispondenti alla verità.

	POSITIVO	NEGATIVO
POSITIVO	a	b
NEGATIVO	c	d

Tabella 1: Tabella di contingenza

Il confronto fra i risultati del test in esame e l'autentico stato di ogni individuo consente di stimare due importanti parametri: la sensibilità (*Se*), ossia la probabilità che un soggetto malato risulti test-positivo, e la specificità (*Sp*), ossia la probabilità che un soggetto sano risulti test-negativo:

$$Se = a/(a+c), \quad (1)$$

$$Sp = d/(d+b). \quad (2)$$

Soprattutto nel campo dell'epidemiologia clinica, e cioè quando i test vengono utilizzati a scopo diagnostico e non in operazioni di *screening*, ancor più interessanti risultano altri due parametri e cioè il valore predittivo di un test positivo (o valore predittivo positivo, *VP+*) ed il valore predittivo di un test negativo (o valore predittivo negativo, *VP-*). Contrariamente a *Se* e *Sp*, che esprimono probabilità pre-test, questi due parametri rappresentano invece probabilità post-test, nel senso che individuano, a fronte di un certo risultato del test, la probabilità che il soggetto in questione sia realmente provvisto (o meno) del carattere ricercato. In particolare, *VP+* indica la probabilità che un test-positivo sia effettivamente ammalato e, viceversa, *VP-* indica la probabilità che un test-negativo sia effettivamente sano. Essi possono quindi essere stimati, rispettivamente, dalla proporzione di veri positivi sul totale dei positivi al test (*VP+*) e dalla proporzione di veri negativi sul totale dei negativi al test *(VP-*):

$$VP+ = a/(a+b), \quad (3)$$

$$VP- = d/(c+d). \quad (4)$$

È facile verificare, osservando ad esempio la Figura 2, che *Se* e *Sp* sono fra loro inversamente correlate in rapporto alla scelta del valore di *cut off*. L'adozione di una soglia che offre un'elevata *Se* comporta

una perdita di *Sp* e viceversa. È possibile dimostrare che, quando la distribuzione dei valori nelle due classi malati-sani è di tipo normale, la "soglia discriminante ottimale", ossia il valore di *cut off* che minimizza gli errori di classificazione, è pari al valore in ascissa corrispondente al punto di intersezione delle due distribuzioni. Tuttavia, la scelta del *cut off* non può essere dettata soltanto da considerazioni di ordine probabilistico volte a minimizzare la proporzione di classificazioni errate, ma è necessario basarsi anche sul prevedibile impatto di tipo sanitario, economico, sociale, ecc. di ciascuno dei due tipi di errata -classificazione (falsi positivi e falsi negativi). Ad esempio, per malattie ad alta contagiosità potrebbe essere opportuno minimizzare la quota di falsi negativi, e quindi privilegiare la sensibilità a scapito della specificità. Viceversa, in altre situazioni (es. malattie non contagiose, trattabili soltanto con una terapia molto costosa) il prezzo di un falso positivo sarà verosimilmente superiore rispetto a quello di un falso negativo, e quindi il *cut off* verrà determinato in modo da privilegiare la specificità. Un metodo empirico comunemente utilizzato per la scelta del *cut off* consiste nel fissare a priori il valore desiderato di specificità (generalmente >0.9) e quindi nel calcolare la corrispondente sensibilità del test nella suddetta condizione. Questo approccio genera tuttavia due effetti collaterali negativi:

- il test in questione può produrre risultati complessivamente migliori attraverso l'adozione di un *cut off* diverso da quello scelto;
- l'impossibilità di effettuare un raffronto affidabile fra la *performance* di due o più test valutati in base ad un singolo valore di *cut off*.

Pertanto, si configura un evidente e non trascurabile inconveniente pratico quando si tratta di scegliere fra uno o più test e, in subordine, vengono ostacolati gli studi di meta-analisi nei quali, come è noto, vengono effettuate comparazioni qualitative fra risultati ottenuti in studi diversi sullo stesso argomento. Alle difficoltà ora accennate, è da sovrapporre un ulteriore elemento che ostacola sia la scelta del *cut off* ottimale per un singolo test che il raffronto fra le *performances* di test diversi. Tale elemento è costituito dal fatto che i valori predittivi dipendono, oltre che dalla *Se* e *Sp* del test, anche dalla prevalenza della malattia nella popolazione studiata. Infatti è intuitivo che, all'aumentare della frazione dei malati nel campione sottoposto al test, la proporzione dei malati positivi aumenti nell'insieme dei positivi al test. Al contrario, per una patologia poco rappresentata tenderà ad

aumentare la frazione dei falsi positivi sul totale dei positivi al test. Tali relazioni possono essere ricavate analiticamente applicando alcune regole di base di calcolo delle probabilità. Le proprietà delle curve ROC consentono di valutare la performance di un test diagnostico basato su valori soglia in modo semplice, efficace e costruito su una ormai consolidata teoria statistica nonchè di eseguire di conseguenza confronti tra test diversi. L'analisi ROC rappresenta una famiglia di metodologie statistiche estremamente versatili, cui si è potuto solo brevemente accennare nella presente trattazione, e che ha dimostrato la sua utilità in diversi campi della scienza, compresa la medicina.

1.2 Le curve ROC: il principio di base

L'analisi delle curve ROC (*Receiver Operating Characteristic* o *Relative Operating Characteristic*) è una metodologia sviluppata per la prima volta durante la II Guerra mondiale per l'analisi delle immagini radar e lo studio del rapporto segnale/disturbo. Essa venne ben presto applicata in altri campi della tecnica e, a partire dagli anni '70, anche in campo medico (Lusted, 1971) inizialmente allo scopo di quantificare l'attendibilità dei responsi di immagini radiografiche interpretate da operatori diversi (Goodenough e coll., 1974; Hanley e McNeil, 1982). In tempi più recenti, l'utilizzo delle curve ROC si è fatto relativamente comune per la valutazione non solo delle immagini, ma anche dei più svariati test sia nel settore medico, con particolare riguardo alla valutazione dei test clinici di laboratorio (Erdrich 1981, Henderson, 1993) che, in minor misura, in quello veterinario (Greiner, Pfeiffer e Smith, 2000).

L'analisi ROC viene effettuata attraverso lo studio della funzione che – in un test quantitativo – lega la probabilità di ottenere un risultato vero-positivo nella classe dei malati-veri (ossia la sensibilità) alla probabilità di ottenere un risultato falso-positivo nella classe dei non-malati (ossia 1-specificità). In altre parole, vengono studiati i rapporti fra "allarmi" veri (*hit rate*) e falsi "allarmi". La relazione tra i suddetti parametri può venire raffigurata attraverso una linea che si ottiene riportando, in un sistema di assi cartesiani e per ogni possibile valore di *cut off*, la proporzione di veri positivi in ordinata e la proporzione di falsi positivi in ascissa. Se il risultato del test è riportato su scala continua, si possono

calcolare i valori di sensibilità e 1-specificità per ogni valore registrato (oppure, in modo del tutto equivalente, utilizzando la media tra ogni valore e quello precedente). Un altro approccio, applicabile anche a dati ordinali, consiste nel suddividere l'intera gamma di valori restituiti dal test nelle due classi (malati e non-malati) in una serie di *k* intervalli, per *k* variabile in rapporto al numero di dati disponibili (ampiezza del set di dati) e della risoluzione della curva che si desidera ottenere. Quest'ultimo approccio consente di ottenere una curva con risoluzione ottimale compatibilmente con lo scarso numero di dati disponibili. L'unione dei punti ottenuti riportando nel piano cartesiano ciascuna coppia (*Se*) e (1-*Sp*) genera una curva spezzata con andamento a scaletta (ROC *plot*). Per interpolazione (*smoothing*), è possibile eliminare la scalettatura ed ottenere una curva (ROC *curve*) che rappresenta una stima basata sui parametri del data set sperimentale (Figura 3).

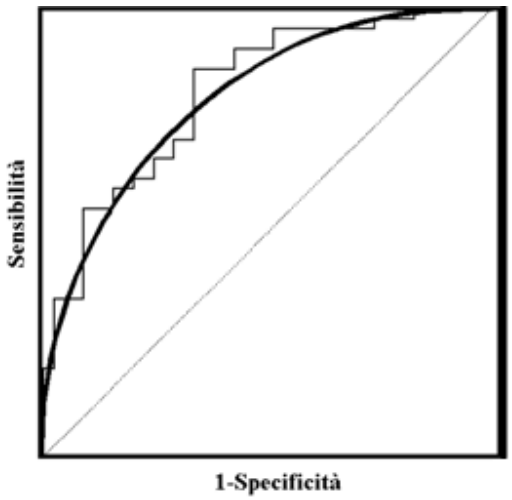


Figura 3: Curva ROC prima (linea spezzata) e dopo (linea continua) interpolazione .

1.3 Valutazione della capacità discriminante di un test e scelta di cut off ottimali

La capacità discriminante di un test, ossia la sua attitudine a separare propriamente la popolazione in studio in "malati" e "sani" è proporzionale all'estensione dell'area sottesa alla curva ROC (Area Under Curve, AUC) ed equivale alla probabilità che il risultato di un test su un individuo estratto a caso dal gruppo dei malati sia superiore a quello di uno estratto a caso dal gruppo dei non-malati (Bamber, 1975; Zweig e Campbell, 1993). Per chi conosce i metodi di statistica non parametrica, risulta evidente la stretta relazione che lega la AUC alla statistica *U* di Wilcoxon.

La statistica *U*, ideata dal chimico Wilcoxon e perfezionata dal matematico Mann-Whitney, rappresenta una delle più note tecniche di statistica non parametrica. Viene utilizzata per il confronto della distribuzione di una variabile continua tra due gruppi e per testare l'ipotesi nulla che i due gruppi presentino la stessa mediana. Tale ipotesi è del tutto equivalente a testare che un soggetto estratto a caso da un gruppo *X* abbia la stessa probabilità di presentare un valore della variabile superiore ad un valore predefinito di quello di un soggetto estratto a caso dall'altro gruppo *Y*. Nella sua formulazione originaria, il test si basa sul numero U delle coppie di valori (*X,Y*), tali che *X* > *Y*.

La stima campionaria di *U* è:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{5}$$

dove n_1 e n_2 sono le numerosità campionarie dei due gruppi, mentre R_1 è la somma dei ranghi nel gruppo a numerosità n_1.

Il valore atteso di *U* è:

$$E(U) = \mu_u = \frac{n_1 n_2}{2} \tag{6}$$

Bamber (1975) ha dimostrato l'equivalenza tra l'area AUC sottesa ad una curva ROC, costruita per dati su scala continua, e la statistica *U*. La relazione che lega i due parametri è la seguente:

$$AUC = \frac{U}{n_1 n_2}, \tag{7}$$

da cui:

$$E(AUC) = \mu_{AUC} = \frac{1}{2}. \tag{8}$$

Nel caso di un test perfetto, ossia che non restituisce alcun falso positivo né falso negativo (capacità discriminante = 100%), la AUC passa attraverso le coordinate {0;1} ed il suo valore corrisponde all'area dell'intero quadrato delimitato dai punti di coordinate (0,0), (0,1), (1,0) (1,1), che assume valore 1 corrispondendo ad una probabilità del 100% di una corretta classificazione. Si noti che, in tale caso limite che corrisponde ad una distribuzione separata della variabile nei due gruppi a confronto (Figura 1), i valori predittivi non dipendono più dalla prevalenza. Al contrario, la ROC per un test assolutamente privo di valore informativo è rappresentata dalla diagonale ("*chance line*") che passa per l'origine, con AUC=0.5.

In una curva ROC esistono in genere due segmenti di scarsa o nulla importanza ai fini della valutazione dell'attitudine discriminante del test in esame. Essi sono rappresentati dalle frazioni di curva sovrapposte rispettivamente all'asse delle ascisse ed all'asse delle ordinate. Infatti, i corrispondenti valori possono essere scartati in quanto esistono altri valori di *cut-off* che forniscono una migliore *S*p senza perdita di *Se* o, viceversa, una migliore *Se* senza perdita di *Sp*. Infine è da ricordare che la valutazione di un test attraverso l'AUC viene compiuta attribuendo ugual importanza alla *Se* e alla *Sp*, mentre in molti casi è necessario, nella pratica, differenziare il peso da assegnare ai suddetti parametri. Nella maggioranza degli studi, l'individuazione del *cut off* ottimale viene effettuata assumendo una distribuzione normale per la variabile in studio e si raggiunge adottando un valore pari a: $[media\ aritmetica + 2\ deviazioni\ standard]$ dei risultati generati dal gruppo di soggetti sani di

referenza. Questo approccio rigido corrisponde ad ottenere un test con specificità pari a 97.5% (Barajas-Rojas e coll., 1993) e presenta lo svantaggio di trascurare completamente il valore della sensibilità. Un'altra possibilità è quella di selezionare un livello di *cut off* sulla base dei percentili della distribuzione dei non-malati (ad esempio, il 90° percentile), e di considerare come potenzialmente malati i soggetti con valori superiori. Tale metodo, cui si è accennato in introduzione, corrisponde a fissare a priori la specificità del test (si noti, infatti, che il 90° percentile nella distribuzione dei non-malati corrisponde a settare la specificità al 90%). Un approccio più adeguato può essere adottato tenendo in considerazione la relazione che lega sensibilità e specificità, ovvero studiando la curva ROC. L'utilizzo della curva ROC rappresenta infatti un criterio più "flessibile", in quanto offre la possibilità di visualizzare, dato un valore a scelta di *Sp*, la corrispondente *Se* e viceversa (Schäfer, 1989).

Come regola generale, si può affermare che il punto sulla curva ROC più vicino all'angolo superiore sinistro rappresenta il miglior compromesso fra sensibilità e specificità. Tuttavia, in condizioni ottimali, la procedura di selezione del *cut off* consiste in un percorso decisionale molto più complesso che deve tener conto, come già ricordato in precedenza, sia della situazione epidemiologica nella popolazione da studiare (con particolare riferimento alla prevalenza della malattia) che dell'esame comparativo delle conseguenze pratiche derivanti dall'ottenimento di risultati falsi positivi e falsi negativi in quella particolare situazione contingente.

1.4 Valutazione della performance di un singolo test mediante una curva ROC

L'area sottesa ad una curva ROC rappresenta un parametro fondamentale per la valutazione della *performance* di un test, in quanto costituisce una misura di accuratezza non dipendente dalla prevalenza ("*pure accuracy*"). Poiché l'AUC rappresenta una stima da popolazione campionaria finita, risulta quasi sempre necessario testare la significatività della capacità discriminante del test, ovvero se l'area sotto la curva eccede significativamente il suo valore atteso di 0.5. Tale procedura corrisponde a verificare se la proporzione dei veri positivi è superiore a quella dei falsi positivi.

Dalle proprietà della statistica *U*, AUC può essere considerata una variabile normale, per cui si può costruire un test *Z* nella seguente maniera: ove σ^2 rappresenta la varianza di AUC.

$$Z = \frac{AUC - 0.5}{\sqrt{\sigma^2\, AUC}} \tag{9}$$

Se, ad esempio, il valore di *Z* eccede il valore critico di 1.96, si può affermare che il test diagnostico presenta una *performance* significativamente superiore a quella di un test non discriminante, con α <0.05. Se il test *Z* risulta invece significativamente inferiore (curva ROC al di sotto della *chance line*), occorre invertire il criterio di classificazione, in quanto il marcatore evidenziato dal test presenta valori mediamente più elevati nella popolazione dei non-malati.

1.5 Stima dell'area sottesa ad una curva ROC

Il calcolo dell'AUC per una curva empirica (cioè ottenuta da un campione finito) può venire effettuato semplicemente connettendo i diversi punti del ROC *plot* all'asse delle ascisse con segmenti verticali e sommando le aree dei risultanti poligoni generati nella zona sottostante. Come sopra accennato, questa tecnica, detta "regola trapezoidale", può fornire risultati sistematicamente distorti per difetto. I metodi di stima dell'area "vera" e di interpolazione delle curve ROC per dati ordinali vanno oltre i limiti della presente trattazione.

Per quanto riguarda l'interpretazione del valore di AUC, si può tenere presente la classificazione della capacità discriminante di un test proposta da Swets (1998).

Essa è basata su criteri largamente soggettivi ed avviene secondo lo schema seguente:

- AUC=0.5 test non informativo;
- 0.5<AUC≤0.7 test poco accurato;
- 0.7<AUC≤0.9 test moderatamente accurato;
- 0.9<AUC<1.0 test altamente accurato;
- AUC=1.0 test perfetto.

1.6 La comparazione di due test mediante analisi ROC

Sotto l'ipotesi bi-normale sopra descritta, due test possono essere quindi confrontati tra di loro comparando le accuratezze stimate mediante l'area sottesa alle corrispondenti curve ROC (Figura 4). Un test *Z* (cioè basato sulla distribuzione normale standardizzata) può essere eseguito rapportando la differenza delle due aree all'errore standard di tale differenza. Nel caso di indipendenza dei due test, tale parametro viene facilmente stimato dalla radice quadrata della somma della varianza di ogni area.

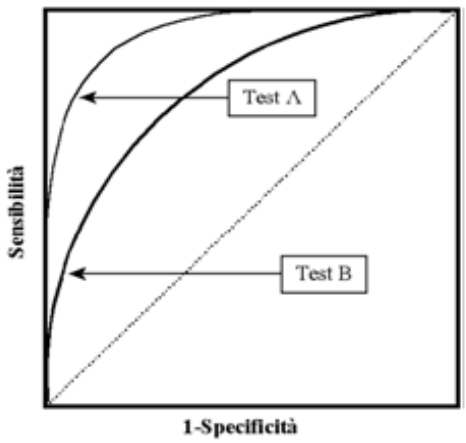


Figura 4: Confronto tra due test diagnostici mediante analisi ROC. Sotto l'ipotesi bi-normale (curve ROC proprie), tale confronto corrisponde a testare la differenza tra le rispettive aree. Risulta evidente la superiorità del test A la cui curva ROC teorica si trova interamente al di sopra di quella corrispondente al test B.

Il Test *Z* per il confronto tra due curve ROC indipendenti assume forma:

$$Z = \frac{AUC_1 - AUC_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{10}$$

dove σ_1^2 e σ_2^2 rappresentano le varianze delle due curve.

Nel caso i due test non siano indipendenti (situazione che può verificarsi quando essi vengono applicati agli stessi soggetti), l'errore standard della differenza delle due aree viene a dipendere dalla correlazione *r* esistente tra esse:

$$Z = \frac{AUC_1 - AUC_2}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2r\sigma_1\sigma_2}} \tag{11}$$

La stima di *r*, sia nel caso di variabili continue che categoriche ordinali, è stata illustrata in dettaglio da Hanley e McNeil (1982). In pratica, il primo passaggio consiste nel calcolare il coefficiente di correlazione dei valori dei due test separatamente nel gruppo dei malati e in quello dei non-malati; nel caso di dati ordinali occorre utilizzare il coefficiente τ di Kendall. La correlazione media tra i due test può essere quindi stimata dalla media dei due coefficienti così ottenuti. Infine, da questo valore si può ricavare la correlazione tra le due AUC (che è anche funzione della media delle due aree).

1.7 Violazioni dell'ipotesi bi-normale: curve ROC non proprie

Lo scostamento dall'ipotesi bi-normale produce curve ROC non proprie, ovvero si assiste ad una perdita della concavità oppure della simmetria rispetto alla diagonale discendente. Un caso piuttosto frequente consiste nell'incrocio della curva con la *chance line*; ciò può indicare l'esistenza di una distribuzione bimodale all'interno di uno dei due gruppi a confronto, come illustrato nella Figura 5.

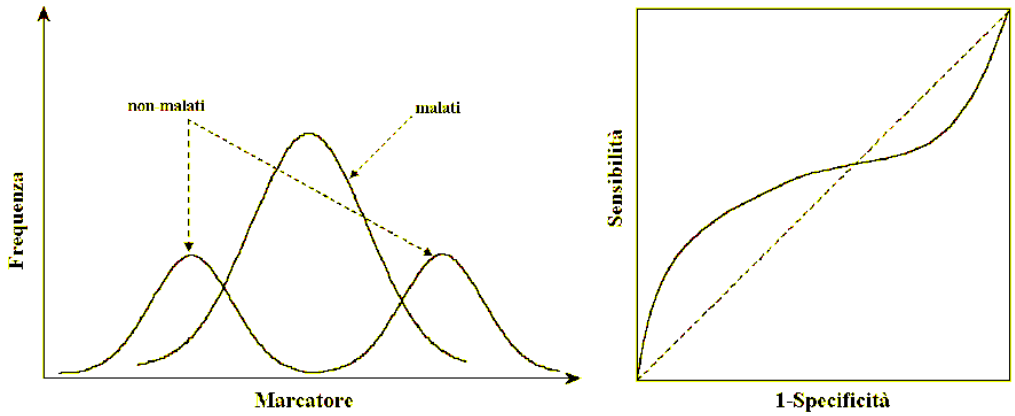


Figura 5: Distribuzione bimodale di uno dei gruppi a confronto e corrispondente Curva ROC non propria

In tal caso risulta che la popolazione dei non malati è costituita da (almeno) due diverse sottopopolazioni di cui una presenta un valore di marcatore mediamente più elevato rispetto al gruppo dei malati, e l'altra mediamente più basso. In linea teorica si potrebbe quindi generare una regola di classificazione basata su due valori diversi di *cut off*. Tuttavia, visto che la prevalenza delle diverse sottopopolazioni nel gruppo dei non-malati non è nota, in genere un simile risultato induce al rigetto del test.

2 Definizione di ROC per risultati "event time"

2.1 I dati di sopravvivenza

L'analisi della sopravvivenza è una parte della statistica inferenziale che mette in rapporto un certo evento con il fattore tempo. Essa infatti riguarda tutti gli studi in cui si vuole analizzare l'incidenza di un determinato evento in un certo "periodo di rischio" ovvero il periodo in cui, almeno logicamente, è possibile sperimentare un determinato evento. Pertanto, il tempo di sopravvivenza è il tempo che intercorre tra l'inizio dello studio e il verificarsi dell'evento di interesse.

Nei soggetti in cui non si verifica l'evento, il tempo di sopravvivenza corrisponde a quello compreso tra l'inizio dello studio e la fine dell'osservazione. La funzione di sopravvivenza si definisce come la probabilità che un soggetto non sperimenti l'evento in un determinato intervallo o in un qualsiasi altro periodo di tempo antecedente al verificarsi dell'evento stesso.

Per esempio, la probabilità che un soggetto sopravviva dopo tre giorni dall'ingresso in uno studio è condizionata dal fatto che il soggetto sia sopravvissuto nei due giorni precedenti. Questa probabilità viene anche definita probabilità cumulativa o sopravvivenza cumulativa.

Se si indica con p_1 la probabilità che ha un soggetto di sopravvivere al primo giorno, con p_2 la probabilità di sopravvivere al secondo giorno e con p_3 la probabilità di sopravvivere al terzo giorno, la sopravvivenza cumulativa è data dal prodotto di queste tre probabilità ($P = p_1 * p_2 * p_3$) *(regola moltiplicatoria delle probabilità).*

Ovviamente la funzione di sopravvivenza varia tra 0, valore iniziale e 1, valore che assume quando nessuno sperimenta l'evento.

Un altro elemento di misurazione è il Rischio Osservato ovvero il rischio misurato analizzando ciò che accade nella realtà dell'esperimento. Si supponga di seguire, per un anno, un gruppo di soggetti con

l'obiettivo di misurare il rischio relativo per un dato evento associato alla presenza di un certo fattore di rischio. Si costruisca una tabella di contingenza (2*2) e sulle colonne si riporti il fattore di rischio (presente / assente) mentre sulle righe l'evento oggetto dello studio (SI / NO):

		FATTORE DI RISCHIO	
		presente	assente
EVENTO	SI	a	b
	NO	c	d
TOT		n	m

Tabella 2: tabella di contingenza

Per calcolare il rischio relativo di eventi a cui sono esposti i soggetti del gruppo con il fattore di rischio presente rispetto al gruppo dove il fattore di rischio è assente la formula da utilizzare è $(c/n)/(d/m) = K$. Ciò vuol dire che i soggetti per i quali il fattore di rischio è presente hanno un rischio osservato di accadimento dell'evento *K*-volte superiore rispetto ai soggetti senza fattore di rischio.

2.2 Lo Stimatore di Kaplan-Meier

Come abbiamo visto nel primo capitolo la sensibilità e la specificità sono due importanti parametri per la costruzione delle curve ROC. Utilizzando il teorema di Bayes si può riscrivere la sensibilità e la specificità per dati di sopravvivenza come:

$$P\{X > c|D(t) = 1\} = \frac{\{1 - S(t|X > c)\}\, P(X > c)}{1 - S(t)}, \qquad (12)$$

$$P\{X \leq c|D(t) = 0\} = \frac{\{S(t|X \leq c)\}\, P(X \leq c)}{S(t)}, \qquad (13)$$

dove $S(t) = P(T > t)$ *e* $S(t|X > c)$ è la funzione di sopravvivenza condizionale per la parte dei dati definita da $X > c$.

Nella descrizione di questo stimatore si farà riferimento ai dati censurati. Infatti, non sempre l'osservazione permette di poter disporre di informazioni complete circa il tempo di permanenza in uno stato, da parte di un individuo, e di conoscere l'esatta durata del periodo di esposizione al rischio di subire l'evento oggetto di studio. I soggetti che vengono censurati sono quelli che "sopravvivono" fino alla fine del periodo di osservazione, oppure che escono dallo studio per cause diverse da quelle di interesse.

Lo stimatore di Kaplan - Meier ha un trattamento ottimale di tutta l'informazione disponibile (senza approssimazioni): gli intervalli, infatti, non sono definiti dal ricercatore ma dai tempi di accadimento degli eventi. Ciò lo rende particolarmente adatto per il trattamento di piccoli data-set con durate misurate.

Kaplan e Meier nel 1958 dimostrarono che la funzione di sopravvivenza, stimata con questo metodo, è la stima di massima verosimiglianza della funzione di sopravvivenza. Per contro, il metodo non permette una facile rappresentazione tabellare dei dati a causa del numero elevato di intervalli. Questo strumento, invece di raggruppare gli eventi per costruire intervalli, costruisce gli intervalli in relazione alle diverse durate che si osservano, in modo che ciascun intervallo contiene solo un tempo osservato di accadimento dell'evento. Pertanto, ciascun intervallo di Kaplan-Meier inizia con il tempo di accadimento dell'evento e termina appena prima che accada l'evento successivo. Per convenzione, esiste anche l'intervallo iniziale t0 (spesso = 0) che si conclude appena prima che accada il primo evento.

L'ultimo intervallo inizia:

- appena dopo l'accadere del penultimo evento e termina con l'accadimento dell'ultimo evento, se non ci sono censure di durata più lunga;
- oppure con l'infinito in caso di censura più lunga.

Inoltre, se un individuo è censurato in corrispondenza di un tempo in cui accade un evento, si assume che l'evento precede il caso censurato (cioè si assume che la censura accada un infinitesimo di tempo immediatamente dopo il momento in cui accade l'evento): ciò fa sì che il caso censurato cada nell'intervallo di tempo dell'evento osservato.

La sopravvivenza nel periodo è la percentuale di pazienti vivi in quel periodo. Perciò per calcolare la sopravvivenza nel periodo si utilizza la formula: [1- il rapporto tra il numero di eventi nell'intervallo di tempo "*t*" e il numero di persone a rischio nell'intervallo di tempo "*t*"]. La probabilità di sopravvivere in un certo intervallo di tempo è "condizionata" dalla probabilità di sopravvivere nell'intervallo di tempo precedente. La sopravvivenza cumulativa è data dalla probabilità di sopravvivenza in un dato intervallo moltiplicata per la probabilità di sopravvivenza nell'intervallo precedente (la probabilità cumulativa del primo intervallo coincide con la probabilità di sopravvivenza; nel secondo intervallo la sopravvivenza cumulativa è la probabilità di sopravvivere nel primo intervallo per la probabilità di sopravvive vere nel secondo intervallo e così via). Quando si è in possesso di questi dati è possibile costruire la curva di sopravvivenza, riportando, su un sistema di assi cartesiani, sull'asse delle ordinate la sopravvivenza cumulativa e sull'asse delle ascisse l'intervallo di tempo.

Un semplice stimatore per la sensibilità e la specificità al tempo t è quindi dato dalla combinazione dello stimatore di Kaplan Meier e della funzione di distribuzione empirica del marcatore*, X,* come:

$$\hat{P}_{KM}\{X > c|D(t) = 1\} = \frac{\{1 - \hat{S}_{KM}(t|X > c)\}\{1 - \widehat{F_x}\,(c)\}}{1 - \hat{S}_{KM}(t)}, \tag{14}$$

$$\hat{P}_{KM}\{X \leq c|D(t) = 0\} = \frac{\{\hat{S}_{KM}(t|X \leq c)\}\{\widehat{F_x}\,(c)\}}{\hat{S}_{KM}(t)}, \tag{15}$$

dove $\widehat{F_x} = \frac{1}{n}\sum I(X_i \leq c)$ con $I(X_i \leq c)$ funzione di identità.

Questo metodo però non garantisce la monotonicità e un secondo potenziale problema è che lo stimatore condizionale Kaplan-Meier $\{\hat{S}_{KM}(t|X \leq c)\}$ assume che il processo di censura dei dati non dipende da *X*. Questa assunzione, però può essere violata nella pratica, quando l'intensità del follow-up è influenzata dalle misurazioni diagnostiche del marcatore al tempo 0.

2.3 Lo stimatore del vicino più vicino (Nearest Neighbor Estimation – NNE)

Un altro metodo con il quale si può ottenere una buona curva ROC al tempo *t* è utilizzare uno stimatore della funzione di distribuzione bivariata $F(c,t) = P\,(X \leq c, T \leq t)$ o equivalentemente, $S(c,t) = P(X > c, T < t)$, fornita da Akritas (1994). Questo stimatore è basato sulla rappresentazione: $S(c,t) = \int_c^{\infty} S(t|X = s)dF_x\,(s)$, dove $F_x(s)$ è la funzione di distribuzione di X. Questo stimatore può essere fornito da:

$$\hat{S}_{\lambda_N}(c,t) = \frac{1}{n}\sum_i \hat{S}_{\lambda_N}\,(t|X = X_i)I(X_i > c) \tag{16}$$

dove $\hat{S}_{\lambda_n}(t|X = X_i)$è uno stimatore "desiderabile" della funzione di sopravvivenza condizionale caratterizzata dal parametro λ_n. A meno che *X* non sia discreta e che ci siano sufficienti osservazioni per ogni valore di X_i, sarà necessario utilizzare un lisciamento per stimare $S(t|X = X_i)$. Possiamo definire lo stimatore Kaplan-Meier ponderato come:

$$\hat{S}_{\lambda_n}(t|X = X_i) = \prod_{s \in T_n, S \leq t}\left\{1 - \frac{\sum_j k_{\lambda_n}(X_j, X_i)\; I\;(Z_j = s)\delta_j}{\sum_j k_{\lambda_n}(X_j, X_i)\; I\;(Z_j \geq s)}\right\} \tag{17}$$

dove $k_{\lambda_n}(X_j, X_i)$ è il nucleo della funzione che dipende dal parametro di lisciamento λ_n, e T_n è l'insieme dei valori che ha assunto Z_i per gli eventi osservati, δ_i*=1*. Akritas (1994) ha usato un nucleo

che può valere 0 o 1 per il "vicino più vicino", $k_{\lambda_n}(X_j, X_i) = I\{-\lambda_n < \hat{F}_x(X_i) - \hat{F}_x(X_j) < \lambda_n\}$, dove $2\lambda \in (0,1)$ rappresenta la percentuale di osservazioni che sono incluse in ciascun intorno (eccetto che per gli estremi della distribuzione di *X*). Questa particolare scelta del nucleo e l'utilizzo dell'approccio del "vicino più vicino" comporta che le risultanti stime ROC siano invarianti rispetto a trasformazioni monotone del marcatore. Le stime di sensibilità e specificità sono date da:

$$\hat{P}_{\lambda_n}\{X > c|D(t) = 1\} = \frac{\{1 - \widehat{F_x}(c)\} - \hat{\mathrm{S}}_{\lambda_\mathrm{n}}(\mathrm{c},\mathrm{t})}{1 - \hat{\mathrm{S}}_{\lambda_\mathrm{n}}(t)} \tag{18}$$

$$\hat{P}_{\lambda_n}\{X \leq c|D(t) = 0\} = 1 - \frac{\hat{\mathrm{S}}_{\lambda_\mathrm{n}}(\mathrm{c},\mathrm{t})}{\hat{\mathrm{S}}_{\lambda_\mathrm{n}}(t)} \tag{19}$$

dove $\hat{S}_{\lambda_n}(t) = \hat{S}_{\lambda_n}(-\infty, t)$. Contrariamente all'utilizzo di uno stimatore Kaplan-Meier, queste stime comportano la monotonicità della sensibilità e della specificità e permettono che il processo di censura possa dipendere dal marcatore diagnostico *X*. Questo deriva dal fatto che in ogni possibile intorno di $X = x$ vengono utilizzati solamente stimatori locali di Kaplan-Meier. Dato che negli studi longitudinali l'intensità dei controlli sarà più alta nei soggetti che presentano valori del marcatore che sembrano portare alla comparsa dell'evento, questa flessibilità nel meccanismo di censura può essere utile in casi reali.

Un recente articolo nel "New England journal of medicine" sui bio-indicatori per eventi cardiovascolari (Wang et al. 2006) stabilisce che "non esistono eventi standard per stimare curve ROC per dati del tipo 'time to event' ". Infatti, sono stati proposti diversi approcci; tuttavia la letteratura non è stata ancora sistematizzata e un approccio standard non è tuttora emerso.

In questo capitolo verranno rivisti i metodi esistenti e sono discussi i casi che spingono a preferire un metodo rispetto ad un altro.

2.4 Definizione di curva ROC per esiti "event time"

2.4.1 Veri positivi "time dependent"

Si consideri innanzitutto lo scenario più semplice, nel quale l'indicatore *Y* è binario e misurato al tempo base (*t=0*). I casi sono i soggetti che sperimentano l'evento di interesse e *T* indica il tempo dell'evento per un caso. Una definizione dei veri positivi "*time dependent*" è:

$$VP_{(t)} = Prob(Y = 1|T = t) \tag{20}$$

Questa definizione consente alla "sensibilità" dell'indicatore di dipendere dal tempo nel quale si verifica l'evento. Ovvero, è probabile che il rischio associato ai veri positivi sia più alto per eventi anticipati piuttosto che per eventi posticipati. Ad esempio, nei casi di screening per il cancro, i bio-indicatori tendono ad avere livelli più alti in soggetti con un tumore subclinico più grande e che probabilmente si manifesteranno clinicamente prima. Perciò, verosimilmente, i veri positivi saranno una funzione decrescente di *t*.

Si supponga ora che l'indicatore sia misurato al tempo *s,* come negli studi longitudinali, o quando non esistano tempi base.

Si indica con

$$VP_{(t,s)} = Prob(Y_{(s)} = 1|T = t + s) \tag{21}$$

la "sensibilità" dell'indicatore agli eventi che avvengono *t*-unità di tempo dopo che *Y* è misurata.

In alcune applicazioni questa sensibilità potrebbe non dipendere solo dall'intervallo di tempo *t*, ma anche da *s*, tempo assoluto di misurazione. Ad esempio, se *s* denota l'età o il tempo dopo un intervento, allora i veri positivi potrebbero dipendere da *t* e *s*, e si scriverà:

$$VP_{(t,s)} = Prob\big(Y_{(s)} = 1 \big| T = t + s\big). \tag{22}$$

Heagerty e Zheng (2005) introdussero una tassonomia per le misure d'accuratezza delle variabili "time dependent". La formula $VP_{(t)}$ sopra definita indica l'incidenza dei veri positivi ed è la versione adottata dalla maggior parte dei metodologisti. (Heagerty and Zheng 2005; Etzioni et al. 1999; Cai et al. 2006; Heagerty and Zheng 2004; Song and Zhou, in stampa). Una versione alternativa sono i veri positivi cumulati:

$$VP_{c\,(t,s)} = Prob\big(Y_{(s)} = 1 \big| s < T \leq t + s\big), \tag{23}$$

la quale valuta la sensibilità per gli eventi che avvengono nell'intervallo di tempo *(s, s+t]*, come opposti agli eventi che avvengono a *t*-unità di tempo dopo che *Y* è misurata. Questa definizione è usata in molti articoli applicativi perché è facilmente stimabile empiricamente. In particolare, per i dati non censurati, una stima è la semplice proporzione dei soggetti con la presenza dell'evento nell'intervallo di tempo e che hanno indicatori con valori positivi *t*-unità di tempo prima che accada l'evento. Cai et al.(2006) si focalizzarono sull'incidenza dei veri positivi ma notarono che i veri positivi cumulati, che potrebbero essere clinicamente interessanti, possono essere calcolati direttamente come:

$$VP_{c\,(t_1,s)} = \int VP_{(u,s)} dF_t\,(u))/[F_T(t+s) - F_T(s)], \tag{24}$$

dove T_f è la distribuzione cumulata del tempo dell'evento. D'altra parte, stimare l'evento sulla base di una stima di veri positivi cumulati è più difficile, perché la differenziazione richiede stime basate

sull'integrazione. Inoltre, accumulando tutti gli eventi in (*s, s+t*], non si distingue tra la sensibilità degli eventi che avvengono prima e quella degli eventi che avvengono più tardi nell'intervallo di tempo. Inoltre, una serie di veri positivi cumulati mostra informazioni ridondanti nel senso che *VPc* (t_2,*s*) include *VPc* (t_1,*s*) se $t_1<t_2$, in quanto una serie di veri positivi rileva informazioni essenzialmente diversi da altre. In alcuni contesti i tempi esatti degli eventi sono sconosciuti, mentre invece sono conosciute le versioni degli intervalli censiti. Per esempio, il test di una condizione subclinica può essere fatto mensilmente. Così *T* può essere considerata una variabile discreta, che indica gli intervalli di tempo. Con dati non censurati, i veri positivi vengono calcolati come una proporzione.

2.4.2 I Falsi Positivi e le curve ROC

Nell'ambientazione classica con risultati binari, i falsi positivi, *FP=P(Y=1/D=0*), sono la frazione di controlli con test positivo. Chi sono i controlli quando l'esito e' un tempo di guasto? Una definizione sensata è considerare controlli quegli individui per i quali un test positivo è un errore. Un gruppo di controllo naturale emerge in alcuni scenari; in pratica, la possibilità di esaminare il corso della vita è raramente disponibile per tutti i soggetti, così i soggetti senza l'evento di interesse, al tempo dell'analisi, sono spesso considerati un gruppo di controllo approssimativo.

La definizione dello status di controllo risulta più problematica quando tutti i soggetti presentano l'evento d'interesse. Una possibilità è scegliere un ampio punto di riferimento temporale τ, e definire come controlli soggetti con $T>\tau$. La scelta ottima per τ dovrebbe dipendere dal contesto. Per esempio, se l'intenzione è quella di monitorare i soggetti in un intervallo di tempo, δ, considerando che l'intervento sarà adeguato se amministrato al tempo δ prima dell'evento, allora la scelta di $\tau=2\delta$ sarebbe sufficiente. I soggetti per i quali $T>2\delta$ non hanno bisogno di avere un test positivo perché possono ancora essere testati e trattati adeguatamente con un monitoraggio futuro. Con questa scelta di τ i controlli sono meglio descritti come un gruppo di riferimento rispetto al quale comparare soggetti con eventi anticipati, piuttosto che come un vero e proprio gruppo di controllo.

Heagerty and Zheng (2005) definiscono i falsi positivi, sopra descritti, ovvero

$$FP_{(s)} = Prob\big(Y_{(s)} = 1 \big| T > s + \tau\big), \tag{25}$$

i falsi positivi statici. Un alternativa è permettere ai falsi positivi di variare in base all'intervallo di tempo: la proporzione di test positivi tra soggetti senza eventi a *t*-unità di tempo dopo il tempo di misurazione dell'indicatore, cioè *FPd(t,s)=Prob(Y(s)=1|T>s+ t),* è chiamato falsi positivi dinamici. A volte però, questa quantità può non rappresentare l'accuratezza del bio-indicatore. Si considerino, ad esempio, i soggetti con un evento poco dopo il tempo *t*, contati come controlli dinamici a *t*; un test positivo per questi soggetti è contato come un falso positivo. Forse potrebbe essere valutato a favore dell'abilità del test a diminuire eventi futuri.

Per un indicatore continuo *Y(s)*, la curva ROC della variabile "time dependent" compara i casi con eventi al tempo *t* con i controlli. In particolare, la proporzione dei casi con indicatori che superano la soglia *c(s),* il quantile *(1-f) di Y(s)* nei controlli, è mostrata contro *f*, la frazione dei falsi positivi. Un altro problema pratico con i falsi positivi dinamici è che, dato che i gruppi di controllo variano con il tempo, così fa anche l'ordinata delle corrispondenti curve ROC. Perciò diventa più difficile interpretare le tendenze riguardo il tempo nelle curve ROC della variabile "time dependent".

Con un gruppo di controllo statico, il trend del tempo nelle curve ROC si riferisce al trend nell'indagine degli eventi. Comunque, tali trend possono essere determinati dalla combinazione del cambiamento dei gruppi di controllo e con il cambiamento delle proprietà di indagine quando le curve ROC usano controlli dinamici. Indubbiamente, considerare questo, quando si usano i falsi positivi dinamici, anche se i veri positivi associati ad una specifica soglia, *I (Y(s)>=c)*, sono costanti nel tempo, determinerà un aumento, delle curve ROC, con un *t* più elevato, nel momento in cui i gruppi di controllo escludono soggetti che presentano eventi con più ampi valori di *Y(s)*.

Per questo motivo, ci si focalizzerà sui falsi positivi statici. Per le applicazioni, dove c'è un gruppo di controllo naturale che non è definito solamente dal tempo, verrà assegnato ai controlli un tempo dell'evento fittizio più ampio di τ, un modo da poter usare una notazione uniforme.

Se *F* denota la funzione di ripartizione per *Y(s)* nel gruppo di controllo, la curva ROC della variabile "time dependent", è scritta matematicamente come:

$$ROC_{t,s}(f) = Prob(Y(s) \geq c(s)|T = s + t), \quad (26)$$

dove $c(s) = F^{-1}(1-f),\ f \in (0,1)$. Ovvero la curva $Roc_{t,s}(f)$ riporta i $VP_{(t,s)}$corrispondenti ai $FP(s) = f$.
Si provi ora a considerare il fatto che l'indicatore al tempo *s*, *Y(s)*, potrebbe essere una funzione della storia dell'indicatore fino al tempo *s*, e non necessariamente il valore di una singola misurazione al tempo *s*. Inoltre, la distribuzione di *Y(s)* può variare con *s* (per esempio quando la scala del tempo *s* è l'età o il tempo dopo un intervento), in questi casi la soglia *c(s)* può dipendere da *s*. In alcune applicazioni la discriminazione raggiunta con l'indicatore può dipendere dalla scala assoluta del tempo *s* come dall'intervallo di tempo *t*, e la nostra annotazione permette questo livello di generalità.

2.4.3 Censurare e contendere eventi-rischio

Censurare è spesso un problema negli studi di previsione con risultati "event time". La censura, infatti, è un disturbo nei dati e chiaramente non dovrebbe influenzare la definizione di veri positivi e falsi positivi.

La semplice pratica comune di includere tutti i soggetti senza eventi nel calcolo dei falsi positivi è errata perché alcuni dei soggetti censurati potrebbero presentare l'evento, quindi contaminare il gruppo di controllo. D'altra parte, gli eventi che comportano rischio sono il vero fenomeno che avviene nella popolazione e dovrebbero perciò impattare nelle definizioni (*VP* e *FP*).

I soggetti con eventi che comportano il rischio dovrebbero essere considerati casi o controlli? In uno studio, è possibile che i bio-indicatori siano predittori di eventi che concorrono al rischio. Li si potrebbe considerare un secondo gruppo di casi e valutare gli indicatori separatamente. Così, è possibile stimare due curve ROC separate: una, di principale interesse, che compara i soggetti che presentano l'evento con il gruppo di controllo e una, d'interesse secondario, che confronta i soggetti a rischio con i controlli. Un'alternativa è inserirli nel gruppo di controllo. Ciò sarebbe giustificato solo se indicare tali soggetti come positivi conducesse a decisioni clinicamente errate. Anche così, dovrebbe essere ugualmente interessante compararli con gli altri controlli in modo da poter interpretare i falsi positivi in termini di componenti da ogni singolo tipo di gruppo di controllo.

2.5 Stima dai dati

Gli approcci alla stima dei parametri di performance dei bio-indicatori possono essere classificati genericamente come prospettici o retrospettici. Ogni classe ha i propri punti di forza.

2.5.1 Metodi retrospettici

Si consideri il semplice contesto in cui *Y(s)* è un indicatore binario e non c'è censura. Si scrivano i dati per i controlli come:

$$\{Y_j(s_{jk}), s_{jk}; k = 1, \dots n_j; j = 1, \dots n_{\overline{D}}\}, \tag{27}$$

e i dati per i casi come:

$$\{Y_i(s_{ik}), s_{jk}; k = 1, \dots n_{ij}; T_i; i = 1, \dots n_D\}. \tag{28}$$

Leisenring et al. (1997), per questo scenario, proposero metodi di regressione binaria semplice.
Si può considerare *FP(s)* come una funzione parametrica di s usando i dati di controllo come corrispondenti parametri del modello. Il controllo *j*-esimo contribuisce con n_j dati della forma $(Y_j(s_{jk}), s_{jk})$.
Similmente *VP(t,s)* può essere considerata come funzione parametrica di (*t,s*) e corrispondente ai dati dei casi. Ogni caso contribuisce con n_j dati nella formula $Y_i(s_{ik}), s_{ik}, t_{ik})$. dove l'intervallo di tempo è $t_{ik} = T_i - s_{ik}$. L'intervallo di tempo varia rispetto ai record di dati secondo il tempo di misurazione del bio-indicatore s_{ik}.
I soggetti non censurati entrano nell'analisi, nell'approccio Leinsering et al., come caso se $0<T-s\leq\tau$ e come controllo se $T-s>\tau$. I soggetti censurati a *X* entrano come un controllo se $X-s>\tau$, altrimenti come medie pesate di casi e controlli. Quando $X-s \leq \tau$, si ha:

$$Prob(Y(s) = 1|T > X) = FP(s)P(T > s + \tau|T > X) + \int_{X-s}^{\tau} VP(s,t)dP(T \leq t + s|T > X). \tag{29}$$

Quindi il contributo alla verosimiglianza per *Y(s)* è una media pesata di *FP(s)* e *VP(s,t)* per l'intervallo di tempo *t* in (*X-s, τ*). I pesi sono facilmente determinati stimando la distribuzione di *T* con i metodi standard per il fallimento dei dati di durata. Se esistono gli eventi che concorrono al rischio, la distribuzione dovrebbe essere stimata con metodi di incidenza cumulativa, piuttosto che trattarli come eventi censurati.

Le stime degli errori standard dei parametri sono basate sulle stime dell'intervallo di varianza per calcolare la correlazione tra record dello stesso soggetto. Ad esempio nelle applicazioni ad un nuovo test per l'infezione da cytomegalovirus nei pazienti che hanno subito il trapianto di midollo osseo, sebbene *Y(s)* vari in base al tempo, la distribuzione di *Y(s)* non varia con *s*, tempo dal trapianto. Essi perciò riportavano tutti i falsi positivi e la funzione monotona decrescente dei veri positivi è così modellizzata:

$$VP_{(t)} = g(\alpha + \beta\eta(t)) \tag{30}$$

dove g^{-1}, funzione di collegamento, è logistica e *η(t)* è una base polinomiale.

Per un indicatore continuo, in assenza di censura, Etzioni et al. (1999), estendevano l'approccio della regressione binaria. Per semplificare, si supponga, come in Etzioni et al., che le curve ROC della variabile "time dependent" non dipendano da *s*. Essi costruirono il modello come segue:

$$ROC_t(f) = g\big(h(f) + \beta\eta(t)\big) \tag{31}$$

dove g^{-1}, è la funzione di collegamento e *g(h(f))* è la curva ROC base a *t=0*.

Essi implementarono il metodo con i dati provenienti da uno studio sul cancro alla prostata, stimando la distribuzione di *Y(s)* non parametricamente nei controlli e usando una forma parametrica per *h*, del tipo $h(f) = \alpha_0 + \alpha_1\Phi^{-1}(f)$.

Cai e Pepe (2002), invece, assegnarono alla funzione *h* una base non parametrica. Cai et al. (2006), offrirono il più esauriente tra gli approcci retrospettici esistenti, includendo i metodi precedenti ed estendendoli ai dati temporali censurati. Con gli indicatori binari, le funzioni da stimare sono *FP(s)* e *VP (t,s)*.

Per i bio-indicatori continui, Cai et al., adottano il modello ROC-GLM (31) con curve ROC non parametriche come baseline. Analogamente a Etzioni et al., l'approccio è non parametrico rispetto alla distribuzione di *Y(s)* anche nei controlli. Il modello può essere implementato sostituendo ogni record di bio-indicatori, *Y(s)*, con una serie di *p* record di variabili binarie nella forma $I\big(Y(s) > c_p\big)$, corrispondenti alle soglie del bio-indicatore $c_1 \dots \dots c_p$. L'algoritmo per indicatori binari è poi applicato con una serie di *FP*, *(*FP_1*(s),* FP_2*(s),….,* FP_p*(s)),* corrispondenti alle soglie stimate attraverso questo approccio. In più, sono stimate una serie di intercette in (1), $\{\alpha_1,\alpha_2,\dots..,\ \alpha_p\}$ che corrispondono alle *p* soglie.

2.5.2 Metodi prospettici

Le tecniche del rischio di regressione sono ben collaudate per modellare dati "event-time" ed esse naturalmente ammettono dati censurati. Dopo aver stimato un modello prospettico, lo si può combinare con distribuzioni predittrici osservate per calcolare i parametri *FP* e *VP*.

Heagerty and Zheng (2005) impiegano un modello Cox per un indicatore base, *Y*:

$$\lambda(t) = \lambda_0(t)\exp(\gamma(t)Y), \tag{32}$$

dove il parametro di regressione γ può dipendere da *t*. Adeguando il modello a un campione casuale semplice $\{(Y_i\ , T_i\ ), i = 1, \dots, n\}$ essi individuarono questo come un indicatore binario e determinarono il set rischio a *t*, con *R(t)*,

$$\widehat{VP}(t) = \frac{\sum_{i\epsilon R(t)} Y_i \exp(\hat{\gamma}(t)Y_i)}{\sum_{i\epsilon R(t)} \exp(\hat{\gamma}(t)Y_i)} \tag{33}$$

dove $\widehat{VP}(t)$ è una stima consistente di *VP(t)*.

Ciò che segue dall'osservazione di Xu and O'Quigley (2000) è che sotto il modello Cox, la distribuzione di $Y\ exp(\gamma(t)Y)$ per soggetti nel set rischio *R(t)* è uguale alla distribuzione condizionale di $Y|T = t$. Per stimare i *FP* essi impiegano la stima empirica nei controlli nel set rischio a τ:

$$\widehat{FP} = \sum_{i\epsilon R(\tau)} {}^{Y_i}/_{n(\tau)}, \tag{34}$$

dove *n(τ)* è la misura dell'insieme di rischio. Con bio-indicatori continui, sia *Fτ* la distribuzione empirica di *Y* per soggetti nel set rischio a *τ*, quindi:

$$\widehat{ROC_{(t)}}(f) \equiv \frac{\sum_{i\epsilon R(t)} I(Y_i \geq \widehat{F_\tau^{-1}}(1-f))exp\{\hat{\gamma}(t)Y_i\}}{\sum_{i\epsilon R(t)} exp\{\hat{\gamma}(t)Y_i\}}. \tag{35}$$

Song and Zhou impiegano la stessa struttura di dati, ma un modello semplificato con parametri non "time dependent", cioè $\gamma(t) = \gamma$. Gli autori utilizzano il teorema di Bayes per scrivere *VP(t)* e *FP* per un indicatore binario:

$$VP(t) = P(Y = 1)P(T = t|Y = 1)/P(T = t)$$

$$= P(Y = 1)\frac{\lambda_0(t)\exp(\gamma)\exp(-\Lambda_0(t)\exp(\gamma))}{\lambda_0(t)\exp(\gamma)\exp(-\Lambda_0(t)\exp(\gamma))P(Y = 1) + \lambda_0(t)\exp\big(-\Lambda_0(t)\big)P(Y = 0)} \tag{36}$$

$$= logit^{-1}\{logit\ P(Y = 1) + \gamma + \Lambda_0(t)(1 - \exp(\gamma))\},$$

dove $logit^{-1}(x) = \exp(\mathrm{x})/(1+\exp(x))$ e Λ_0 è la funzione cumulativa di base;

$$FP = P(Y=1)P(T>\tau|Y=1)/P(T>\tau)$$

$$=logit^{-1}\{logit\, P(Y=1) + \Lambda_0(t)(1-\exp(\gamma))\} \tag{37}$$

Si osservi che se $\gamma = 0$ allora *VP(t) = FP = P(Y=1),* che è un risultato intuitivo.

Assumendo che valori più grandi di *Y* sono associati ad un tasso di rischio più grande, γ>0, si ha che una *baseline* con valori più grandi conduce a *FP* e *VP* più piccoli. D'altra parte se gli eventi sono rari $\Lambda_0(\tau) \approx 0$, abbiamo $FP \approx P(Y=1)$, la proporzione degli indicatori positivi nella popolazione alla base e $FP \approx\ logit^{-1}\{logitP(T=1+\ \gamma\}$ che non dipende da *t*.

Con indicatori continui, gli integrali sulla distribuzione di *Y* entrano nelle espressioni *VP* e *FP* corrispondente all'indicatore soglia, $I[Y \geq y]$:

$$VP(t) = 1 - F_{D,t}(y)$$

$$\equiv \frac{\int_y^\infty \exp(\gamma Y)\, exp\{-\Lambda_0(t)\exp(\gamma Y)\}dF(Y)}{\int_{-\infty}^\infty \exp(\gamma Y)\, exp\{-\Lambda_0(t)\exp(\gamma Y)\}dF(Y)} \tag{38}$$

$$FP = 1 - F_\tau(y)$$

$$= \frac{\int_y^\infty exp\{-\Lambda_0(t)\exp(\gamma Y)\, dF(Y)\}}{\int_{-\infty}^\infty exp\{-\Lambda_0(t)\exp(\gamma Y)\, dF(Y)\}}. \tag{39}$$

Song and Zhou sostituiscono $\hat{\gamma}$, $\widehat{\Lambda_0}$ e la distribuzione empirica di *Y*, $\hat{F}$, nelle espressioni di cui sopra, per stimare i *VP* e *FP*. Lo stimatore della curva ROC è calcolato come

$$\widehat{ROC}_t(f) = 1 - \widehat{F_{D,t}}\left(\widehat{F_\tau^{-1}}(1-f)\right). \tag{40}$$

Il metodo di Song and Zhou presenta due vantaggi rispetto all'approccio Heagerty and Zheng. Il primo è stato dimostrato essere più efficiente in studi di simulazione. Ciò probabilmente è dovuto all'impiego del mple (mple: stimatori della massima verosimiglianza parziale) per γ e Λ_0 e così i corrispondenti stimatori di (*VP, FP*) sono anch'essi mple. Per contro, lo stimatore di *VP (t)* impiegato da Heagerty and Zheng non è mple. Inoltre il loro stimatore empirico di *VP (t)* non utilizza la struttura conferita dal modello Cox. Song and Zhou utilizzano questa struttura nella stima dei *FP*.

Il secondo vantaggio riguarda la censura. I metodi di Heagerty and Zheng non permettono alla censura di dipendere da *Y*. I soggetti a rischio al tempo *t* possono essere rappresentativi della popolazione "a rischio" riguardo alla distribuzione predittrice. L'approccio Song and Zhou utilizza solo la distribuzione base degli indicatori e i parametri del modello di rischio. Questi ultimi sono consistentemente stimati sotto assunzioni di censura standard che permettono al follow-up di dipendere dai predittori modellizzati. L'approccio Song and Zhou è valido anche se la censura dipende dall'indicatore *Y*. Comunque, il metodo Song and Zhou è valido soltanto quando l'assunzione proporzionale dei rischi è soddisfatta, mentre Heagerty and Zheng hanno esteso il loro approccio per permettere la stima sotto rischi non-proporzionali.

2.5.3 Comparazione degli attributi

Tra i metodi retrospettici, quello Cai et al.(2006) è il più esauriente. Altri metodi retrospettici possono essere visti come casi specifici del metodo usato da Cai. Quindi sarà comparato con i due approcci prospettici, il metodo Song and Zhou e quello Heagerty and Zheng. Usiamo le notazioni: Cai, S+Z, H+Z per i tre metodi.

2.5.3.1 Prospettico

I Veri e i Falsi positivi sono definiti come quantità retrospettive, nel senso che concernono la distribuzione di *Y* condizionata al risultato. Secondo Pepe et al. 2007, un'analisi retrospettiva sembra l'approccio più naturale e diretto per stimarli. In più, i parametri relativi a *t* nell'approccio retrospettico, β in (30) e in (31), quantificano direttamente come varia la performance con t. L'inferenza riguardo questi parametri è in linea con l'approccio retrospettico. Al contrario, i parametri nei modelli prospettici non quantificano direttamente l'effetto del tempo nella performance del bio-indicatore.

2.5.3.2 Assunzioni di modello

Le assunzioni di modello richieste da Cai sono molto blande. Il metodo è non parametrico rispetto alla distribuzione di *Y*, nei controlli e semiparametrico, nei confronti della distribuzione di *Y*, nei casi. In particolare non si specifica una distribuzione per *Prob(Y|T)*, ma si modella solo l'effetto di *T* sulla sua distribuzione con una forma parametrica.
I metodi prospettici sono basati su assunzioni blande. Usano un modello semiparametrico per *Prob(T|Y)*, ma non specificano la distribuzione di *Y* nella coorte.

2.5.3.3 Censura

La censura, che è non dipendente da *Y* è utilizzata da tutti i metodi.
La censura che dipende da *Y* è usata solamente dal metodo Song and Zhou.
L'estensione agli altri due metodi del contesto più generale di censura, condizionalmente indipendente, è possibile, anche se non proprio banale.
Il problema della verifica del campionamento distorto (errore sistematico), che è ben studiato nella valutazione del test diagnostico, è completamente analogo alla censura del predittore dipendente. La

verifica dei campioni distorti avviene quando il risultato del test diagnostico è usato per selezionare soggetti per l'accertamento del loro reale stato della malattia. L'errore sistematico risultante nelle semplici stime di *FP* e *VP* è chiamato verifica della distorsione. Le stime corrette sono calcolate usando semplici stime di valori predittivi positivi e negativi, i quali non sono distorti, unendoli poi alle frequenze naturali originarie di test positivi e negativi attraverso il teorema di Bayes. Analogamente, quando il follow-up per i risultati di "event time" dipende dal predittore, si possono usare le stime dei parametri prospettici e la distribuzione base dei predittori per calcolare VP e FP con il teorema di Bayes. Vista in questa maniera, il metodo S+Z estende i metodi di correzione della verifica dell'errore sistematico al dato "event time".

2.5.3.4 Eventi che concorrono al rischio

Anche se non è stato affrontato nessuno dei metodi proposti, tutti possono essere estesi in modo da considerare gli eventi che concorrono al rischio. Le funzioni di rischio sono rimpiazzate da funzioni rischio "a causa specifica" e le funzioni di sopravvivenza sono rimpiazzate dalla probabilità che non si verifichi l'evento (di qualsiasi tipo, né eventi di interesse, né eventi che concorrono al rischio). Si può stimare una funzione *VP(t,s)* separata per eventi che concorrono al rischio. Nel metodo Cai si è convenuti ad un modello separato. Negli approcci prospettici, saranno impiegati separatamente modelli di rischio a causa specifica. Uno studio dettagliato di questi metodi è giustificato ma non aderente allo scopo di questo capitolo.

2.5.3.5 Campionamento

I metodi prospettici sono stati proposti per studi di coorte dove sono disponibili i dati di un campione casuale dalla popolazione, anche se possono essere generalizzati. In breve, se il metodo di campionamento permette il calcolo delle stime della funzione di rischio e la distribuzione dei predittori della popolazione, i due approcci prospettici possono essere applicati.

Gli studi del gruppo-caso e i disegni caso-controllo, dove i controlli a *T* sono un campione casuale dalla popolazione a rischio, possono quindi essere adattati. Un disegno alternativo dello studio caso-controllo dove i controlli sono un campione casuale dalla popolazione dei controlli che presentano *T>t*, non dà origine alle stime della funzione rischio, e perciò essi non sono adattati. Al contrario, i metodi retrospettici accolgono naturalmente quest'ultima forma di studio caso-controllo, assumendo che la censura non dipenda da *Y*. Essi tengono in considerazione anche i disegni sperimentali di coorte, *case-cohort* e *case -'risk set control'*, sotto le stesse assunzioni sui dati censurati.

2.3.3.6 Bio-indicatori longitudinali

Cai et al. (2006) hanno sviluppato il metodo retrospettico nel contesto generale dove i dati indicatori sono collocati longitudinalmente rispetto al tempo. Un assunzione implicita è che i dati indicatore a s siano mancanti condizionatamente al successivo dato-evento. I metodi prospettici possono essere generalizzati per adattarsi ai dati longitudinali usando modelli di regressione marginale. Specificamente, ogni misura dell'indicatore *Y(s)* genera un record di dati con tempo di origine per *T* rimesso a *s*. Ciò fa sì che l'evento o il tempo censurato associato a *Y(s)* sia rispettivamente *T- s* o *X - s*, permettendo al corrispondente rischio di base e ai coefficienti di regressione di dipendere da *s*. I *VP* e *FP* possono essere scritti come funzioni di *s* e di *t*.

2.5.3.7 Covariate

Diversi fattori possono influenzare la distribuzione e/o la performance degli indicatori come predittori di eventi. Chiamiamo questi fattori covariate. Le covariate specifiche della malattia identificano le caratteristiche della malattia.

Le covariate specifiche di una malattia associate solo ai casi possono essere modellizzate come parte della funzione *VP* in un approccio d'analisi retrospettico. Tali covariate non sono generalmente

accolte dai metodi prospettici. Comunque, se la covariata è discreta, gli eventi possono essere classificati e trattati come rischi concorrenti. Altre covariate riguardano sia i controlli sia i casi.

2.5.3.8 Comparazione degli indicatori

I metodi retrospettici possono includere indicatori multipli nel contesto del modello di regressione per *VP* e *FP* in una forma simile a quella descritta nel capitolo 3 di Pepe (2003).
I parametri specifici dei modelli che si riferiscono alle differenze nelle performance tra indicatori e l'inferenza comparativa, possono quindi essere calcolati. Si vedano le sezioni 6.4.3 e 6.4.4 di Pepe (2003) per le illustrazioni compresa l'illustrazione con i dati dello screening per il cancro alla prostata. I modelli prospettici attuali non hanno la capacità di fare questo. Comparare i rischi relativi non risolve la questione.

2.5.3.9 Combinazione degli indicatori

I metodi discussi in questo capitolo non concernono la metodologia per combinare insieme i predittori. In ogni caso, quando una combinazione è definita, i metodi discussi in questo paper possono essere utilizzati per valutare le performance della combinazione adoperando un dataset test indipendente.

3 Una Applicazione

Nel seguente capitolo si vuole descrivere una applicazione che esemplifica le questioni chiave nella valutazione delle performance degli indicatori per risultati di "event time".

3.1 indicatori di lesione acuta del rene

I pazienti che si sottopongono a chirurgia cardiaca presentano un alto rischio di soffrire di lesioni renali acute (AKI) a causa dell'interruzione del flusso sanguigno ai reni durante l'intervento.

Un evento AKI avviene quando il livello di creatinina supera del 25% i livelli pre-operatori per 24 ore. I pazienti con AKI sono classificati come evento grave se i livelli superano il 200% del siero di creatinina pre-operatorio durante il corso della cura clinica.

Nei casi in cui ciò non avviene l'evento è classificato come medio.

Approssimativamente l'80% dei pazienti ricoverati per interventi senza AKI o altri eventi gravi sono dimessi 3-5 giorni dopo l'intervento. Ci si aspetta che il 20% proverà un evento AKI, di cui il 12% medio e l'8% grave. Un ulteriore piccolo gruppo di pazienti, meno dell'1%, moriranno per complicazioni associate alla loro malattia o intervento senza riscontrare criteri per AKI. Questi sono chiamati eventi non-AKI.

Sebbene controllare l'aumento del siero di creatinina sia l'approccio standard per prevedere l'AKI, esiste un importante "retroscena". L'aumento del siero di creatinina è tipicamente ritardato di 2-3 giorni a causa di parecchi fattori, "non-renali", che ostentano la sua produzione ed il suo livello standard. Interessa però individuare altri bio-indicatori che possono individuare l'AKI più velocemente rispetto a quanto si otterrebbe con il controllo del siero di creatinina, così da poter iniziare prontamente un trattamento appropriato.

Due indicatori misurati in urine sono correntemente sotto osservazione in uno studio di 1800 pazienti sottoposti a una grossa operazione cardiaca. I campioni di urina sono prelevati a diversi intervalli dopo l'intervento, congelati e immagazzinati. Alla fine dello studio questi campioni saranno successivamente testati utilizzando i nuovi indicatori. Il siero di creatinina è invece monitorato come parte della cura clinica di routine.

Alcuni dei punti di interesse in questi studio sono:

- determinare il numero dei pazienti per i quali la diagnosi dell'AKI può essere anticipata con i nuovi indicatori;
- individuare di quanto tempo si può anticipare la diagnosi
- stimare i costi in termini di false diagnosi.

L'analisi ha bisogno di contemplare il rischio di eventi dovuti a morti non-AKI, i vari gradi di gravità dell'AKI e la natura longitudinale dei bio-indicatori. I dati non saranno censurati se tutti i soggetti sono strettamente seguiti fino alla dimissione o alla morte.

3.2 Studio di simulazione

3.2.1 Descrizione del data-set

3.2.1.1 La generazione dei dati simulati

Di seguito vengono definiti i parametri e la notazione utilizzati nella simulazione

n = dimensione totale del campione (1800)

i = indice identificatore dei soggetti

k = *k*-esimo campione del "prelievo"

s_{ik} = tempo all'osservazione del k-esimo "prelievo" per l'*i*-esimo soggetto

Il campionamento ha la seguente scadenza:

- analisi di un campione di urina dei pazienti approssimativamente ogni 6 ore per 5 giorni a partire da quello dell'operazione subita.

Si genera:

$$s_{ik} = 0.25 + \varepsilon_{ik},$$

con k = 1, …, 20 e $\varepsilon \sim$U(0,0.25).

Questi tempi potenziali per i test dell'urina sono modificati (come di seguito riportato) a seconda dello status del paziente.

Lo status di controlli è attribuito a un gruppo selezionato casualmente composto da 1.400 pazienti. Si è simulata la loro dimissione al 3° giorno di ricovero (30% dei 1.400 pazienti), al 4° (40% dei 1.400 pazienti) e al 5° (30% dei 1.400 pazienti) censurando le misurazioni dei tempi *sik* per i pazienti ricoverati per 3 e 4 giorni rispettivamente nelle percentuale del 30 e 40% nei due sottoinsiemi.

Lo status di morte non dovuta ad AKI è stato attribuito a 18 pazienti. Le misurazioni dopo il primo giorno sono state censurate per 6 di questi pazienti, simulando che l'evento di interesse sia avvenuto al primo giorno di ricovero. Allo stesso modo sono state censurate le osservazioni al giorno 2, 3 e 4 per 3 pazienti per ogni gruppo simulando l'evento al giorno 2 per 3 pazienti, al giorno 3 per 3 pazienti e al giorno 4 per 3 pazienti e al giorno 5 per 3 pazienti.

I rimanenti pazienti (n. 342) hanno sperimentato un evento AKI. A 206 di questi abbiamo assegnato la gravità bassa, e a 135 quella alta. Per i pazienti con AKI grave è stato generato un tempo di osservazione latente e non osservato come segue:

$E^{sev} \sim U(0, 0.25)$ con probabilità 0.6,

$E^{sev} \sim U(0.25, 1.25)$ con probabilità 0.4.

Cioè E^{sev} , tempo di osservazione reale per gli eventi AKI, è distribuito uniformemente tra 0 e 0.25 nel 60% dei pazienti gravi, e distribuito uniformemente tra 0.25 e 1.25 nel 40%. Allo stesso modo il tempo reale di osservazione per pazienti con AKI lieve è tale che:

$E^{mild} \sim U(0, 0.25)$ con probabilità 0.4,

$E^{mild} \sim U(0.25, 2.25)$ con probabilità 0.6.

Il tempo *T* per la diagnosi clinica dell'AKI con il siero di creatinina si genera come:

T = E + V dove V ~ U(0, 2.75).

I valori del biomarker sono distribuiti normalmente con media 0 e varianza 1, con nessun trend in dipendenza del tempo. È stata simulata una struttura autoregressiva:

$Y_{i,1} \sim N(0, 1)$,

$$Y_{i,k} = \alpha Y_{i,k-1} + \sqrt{1 - \alpha^2}\, \varepsilon_{i,k}, k \geq 2, \quad (41)$$

dove ε_{ik} rappresenta l'errore, indipendente e normale standard, e la correlazione dell'autoregressione è rappresentata da α. Abbiamo assegnato a α valore 0.8.

Nei casi, i valori del biomarker sono generati come per i controlli fino al tempo *E* in cui sperimentano l'evento (momento non osservato). Sia s_{ik}* il tempo della prima misurazione dopo *E*. Abbiamo generato:

$$Y_{i,k^*} = \Delta + \alpha Y_{i,k^*-1} + \sqrt{1 - \alpha^2}\varepsilon_{ik^*}, \quad (42)$$

dove $\Delta = \mu + \delta$, con δ indipendente e distribuito secondo una distribuzione normale con media 0 e deviazione 2, e μ, la media di Y_{ik*}, che dipende dalla gravità dell'AKI.

$\mu = 8$ per i soggetti con AKI grave,

$\mu = 4$ per i soggetti con AKI lieve.

Per le misurazioni successive, si è utilizzato

$$Y_{i,k} = \alpha Y_{i,k-1} + \sqrt{1-\alpha^2}\varepsilon_{ik}, k > k^* \quad (43)$$

In sintesi, la struttura del data-set è mostrata in tabella 3

Nome della variabile	Tipo di dato	Formato di visualizzazione	Etichetta per i valori	Etichetta per la variabile
id	int	%8.00g		Identificatore soggetto
y	float	%9.00g		Valore del marcatore
s	float	%9.00g		Tempo di misurazione del marcatore (in giorni)*
tevent	float	%9.00g		Tempo dalla misurazione all'evento*
statusall	byte	%13.00g	Statusall	Stato dell'evento*
tmfull	float	%8.00g		

Tabella 3: il data set

* vedi note

I valori delle etichette per le variabili con codici numerici sono i seguenti:

statusall:

0: controlli

1: AKI grave

2: AKI leggero

9: morti non dovute all'AKI

Note:

dta: Dati simulati del biomarker per il test di individuazione dell'AKI grave in seguito a una operazione cardiochirurgica.

s: Tempo di misurazione del marker rispetto all'operazione.

tevent: Tempo trascorso dalla misurazione del marker all'evento AKI. Valido solamente per statusall = 1, 2, 9.

statusall: Controlli = nessun AKI in seguito all'operazione.

3.2.2 Le Analisi

3.2.2.1 Analisi di sopravvivenza

Si applicano i metodi del Kaplan Meier e del Nearest Neighbor Estimation per costruire le curve di ROC con i dati AKI relativi all'alto rischio di soffrire di lesioni renali a causa dell'interruzione del flusso sanguigno ai reni durante l'intervento di chirurgia cardiaca. Il data set è lo stesso che è stato usato per le elaborazioni di esemplificazione delle questioni chiave nella valutazione delle performance degli indicatori per risultati di "event time" .

Prima di procedere con le analisi di sopravvivenza, è stata fatta una analisi del data set descritto nel paragrafo precedente. È stato deciso di considerare, per l'elaborazione, solamente i dati contenuti

nella prima riga di ogni soggetto in quanto riportavano tutte le informazioni sufficienti per poter costruire le curve di sopravvivenza.

In essa sono contenuti:

- il valore del marcatore;
- tempo di misurazione del marcatore rispetto all'operazione;
- tempo trascorso dalla misurazione del marcatore all'evento AKI;
- lo stato dell'evento (AKI Grave, AKI Leggero, No AKI);
- il tempo di comparsa dell'evento.

Una prima analisi è stata fatta su tutto il campione considerato: il primo passo per procedere con questa elaborazione è stato quello di rendere dicotomico il parametro di interesse. In questo studio il parametro di interesse è lo stato dell'evento che indica se l'evento AKI si è verificato e la gravità di tale evento. Il campione è stato diviso in due gruppi:

- il gruppo che ha sperimentato l'evento senza distinzione della gravità dell'evento - composto da quei soggetti in cui il valore dello "stato dell'evento" è pari a 1 (AKI Grave) o a 2 (AKI Leggero);
- il gruppo che non ha sperimentato l'evento – composto da quei soggetti in cui il valore dello "stato dell'evento" è pari a 0 (controlli ovvero quelle persone che non hanno sperimentato l'evento) o a 9 (morti non dovute all'AKI).

Il secondo passo è stato stimare il modello secondo il metodo del Kaplan-Meier e del NNE. Per la costruzione del modello è stato considerato solo il primo gruppo ovvero quello composto dai soggetti che hanno sperimentato l'evento. I due modelli sono riportati nella Tabella 4, mentre il grafico riportato in Figura 6 indica la curva ROC ottenuta dalle stime dei falsi positivi e falsi negativi.

	KM	NNE
Predict time	5	5
Survival	0.8016025	0.8016025
AUC	0.6691673	0.6597333

Tabella 4: modello costruito con il gruppo di tutti i soggetti che hanno sperimentato l'evento

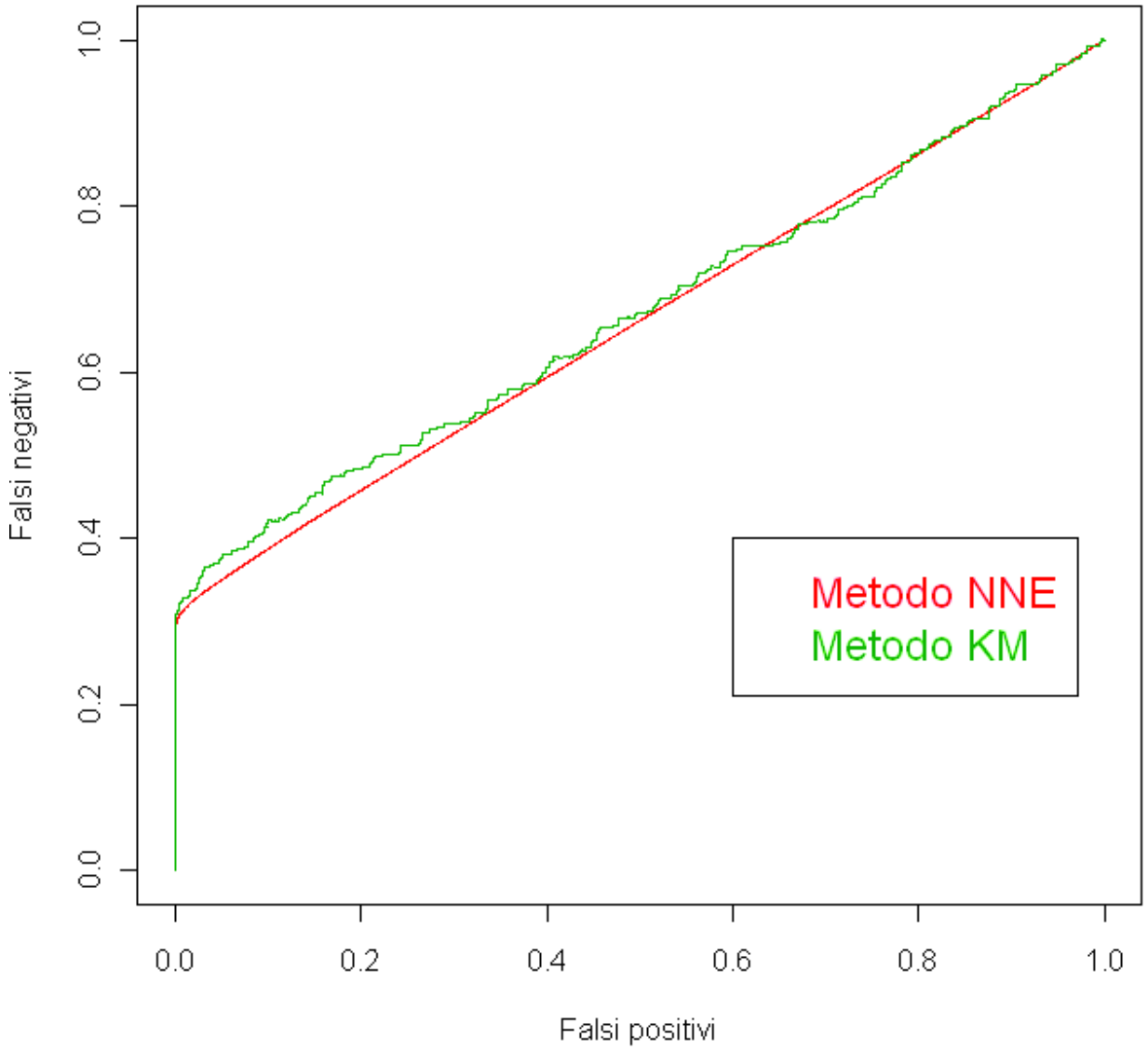


Figura 6: Curve ROC per l'intero campione

Successivamente si è proceduto all'analisi dei soggetti il cui "stato dell'evento" era pari a 1 o a 2.

Il primo gruppo preso in considerazioni è quello composto da soggetti che presentavano un AKI Grave, ovvero quelli con il valore dello "stato dell'evento" pari a 1. Anche per questo gruppo si è dovuto dicotomizzare i parametro e i due gruppi che si sono venuti a formare erano composti:

- il gruppo che ha sperimentato l'evento in cui il valore dello "stato dell'evento" è pari a 1;
- il gruppo di tutti gli altri soggetti (soggetti che non hanno sperimentato l'evento o soggetti il cui valore dello "stato dell'evento" è pari a 2).

Le stime dei modelli costruiti per il gruppo di soggetti che ha sperimentato l'evento sono riportate nella Tabella 5. Per la stima di questi modelli è stata usata la procedura applicata anche per la stima dei modelli che consideravano tutti i soggetti presenti nello studio.

	KM	NNE
Predict time	5	5
Survival	0.9123526	0.9123526
AUC	0.7383302	0.746118

Tabella 5: modello costruito con il gruppo di soggetti che hanno sperimentato l'evento AKI Grave

Il grafico in Figura 7 riporta al curva ROC ottenuta dalle stime dei falsi positivi e falsi negativi di entrambi i modelli stimati.

Curve R.O.C. per la valutazione di test diagnostici per eventi nel tempo

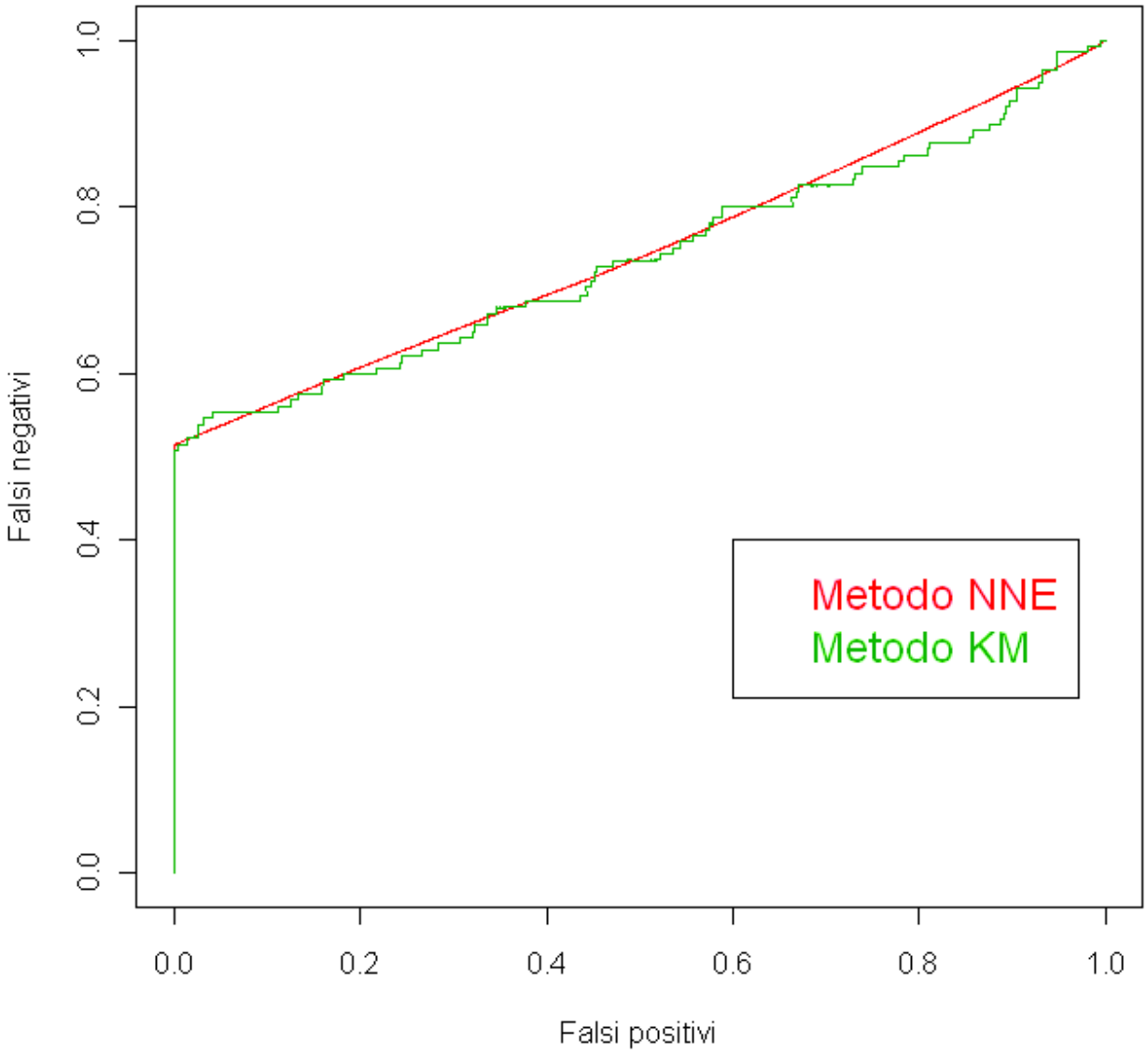


Figura 7: Curve ROC per l'evento AKI Grave

Per le analisi del gruppo con AKI Leggero, sono state applicate le stesse procedure usate per l'analisi del gruppo con AKI Grave. Il gruppo usato per la costruzione dei modelli è composto da tutti i soggetti il cui valore dello "stato dell'evento" è pari a 2.

La Tabella 6 riporta le stime del modello ottenute con i due metodi usati (NNE e KM). Il grafico riportato in Figura 8 mette in relazione stime dei falsi positivi e falsi negativi ottenendo cosi una curva ROC.

	KM	NNE
Predict time	5	5
Survival	0.8677548	0.8677548
AUC	0.6254558	0.603915

Tabella 6: modello costruito con il gruppo di soggetti che hanno sperimentato l'evento AKI Leggero

Curve R.O.C. per la valutazione di test diagnostici per eventi nel tempo

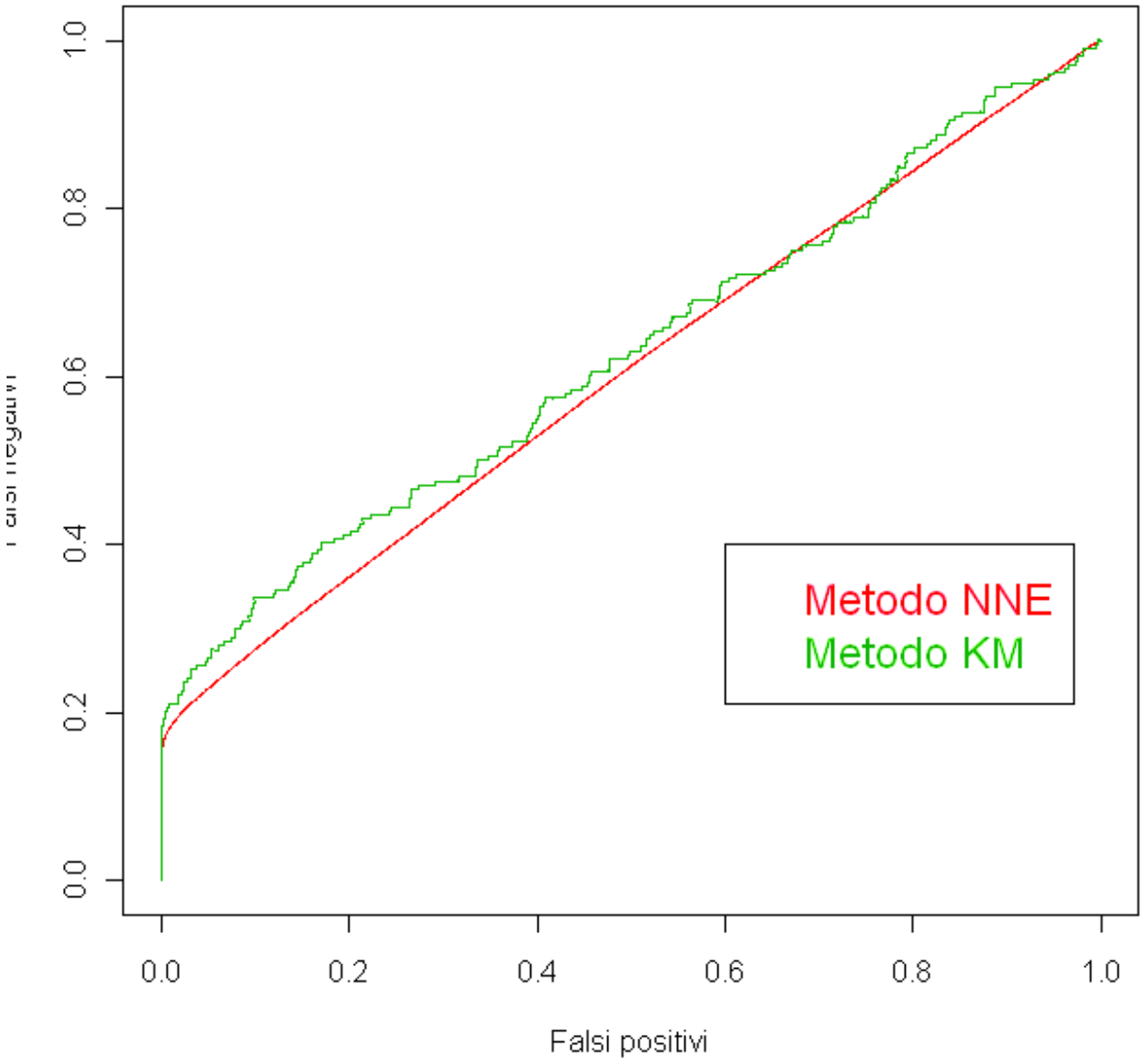


Figura 8: Curve ROC per l'evento AKI Leggero

A conclusione di questa analisi è stato costruito un grafico che mette a confronto i metodi di stima del modello.

Come ci si aspettava, gli andamenti del modello sono molto simili. Una diversità tra questi due metodi è, come già spiegato in precedenza, che il metodo del Kaplan Meier non garantisce la monotonicità delle curve. L'AUC (area sotto la curva), che equivale alla probabilità che il risultato di un test su un individuo estratto a caso dal gruppo dei soggetti che presentano l'evento sia superiore a quello di uno estratto a caso dal gruppo dei soggetti che non presentano l'evento di interesse, è leggermente inferiore nei modelli costruiti con il metodo del Nearest Neighbor Estimation, ciò significa che il modello stimato con questo metodo può risultare leggermente meno accurato del modello costruito con il metodo KM.

Tutto ciò può essere spiegato anche dal fatto che il metodo NNE usa un parametro di lisciamento, il quale può portare ad una sottile perdita di informazione. L'unico gruppo in cui questo non si verifica è quello composto dai soggetti che hanno sperimentato un evento AKI Grave: in questo gruppo l'AUC è maggiore nel modello stimato con il metodo NNE.

In entrambi i grafici in Figura 9 la curva che evidenzia una maggiore AUC, è quella relativa ai soggetti che presentano un AKI grave, nonostante il numero di soggetti che presenta un AKI leggero sia maggiore; ciò significa che il gruppo di soggetti con "stato dell'evento" pari a 2 separa meglio i falsi negativi dai falsi positivi rispetto al gruppo di soggetti il cui stato dell'evento è pari a 1.

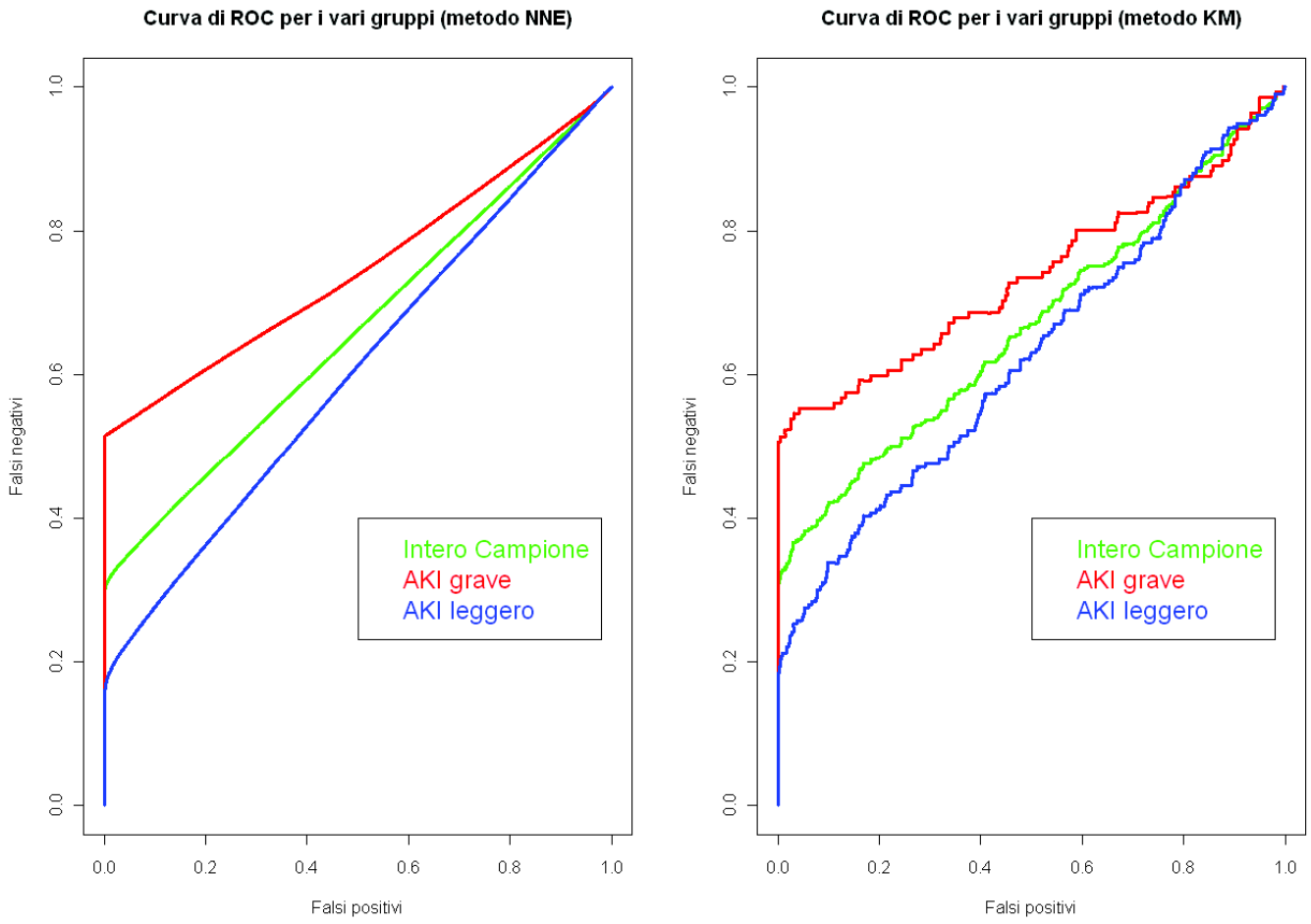


Figura 9: confronto tra i metodi KM e NNE

3.2.2.2 Analisi R.O.C.

Dei 1800 soggetti nello studio, 342 presentano l'AKI, 136 grave e 206 leggero. Inoltre, 18 pazienti morirono per cause apparentemente non collegate al danno renale. In Figura 10 sono riportate le distribuzioni cumulate degli eventi AKI.

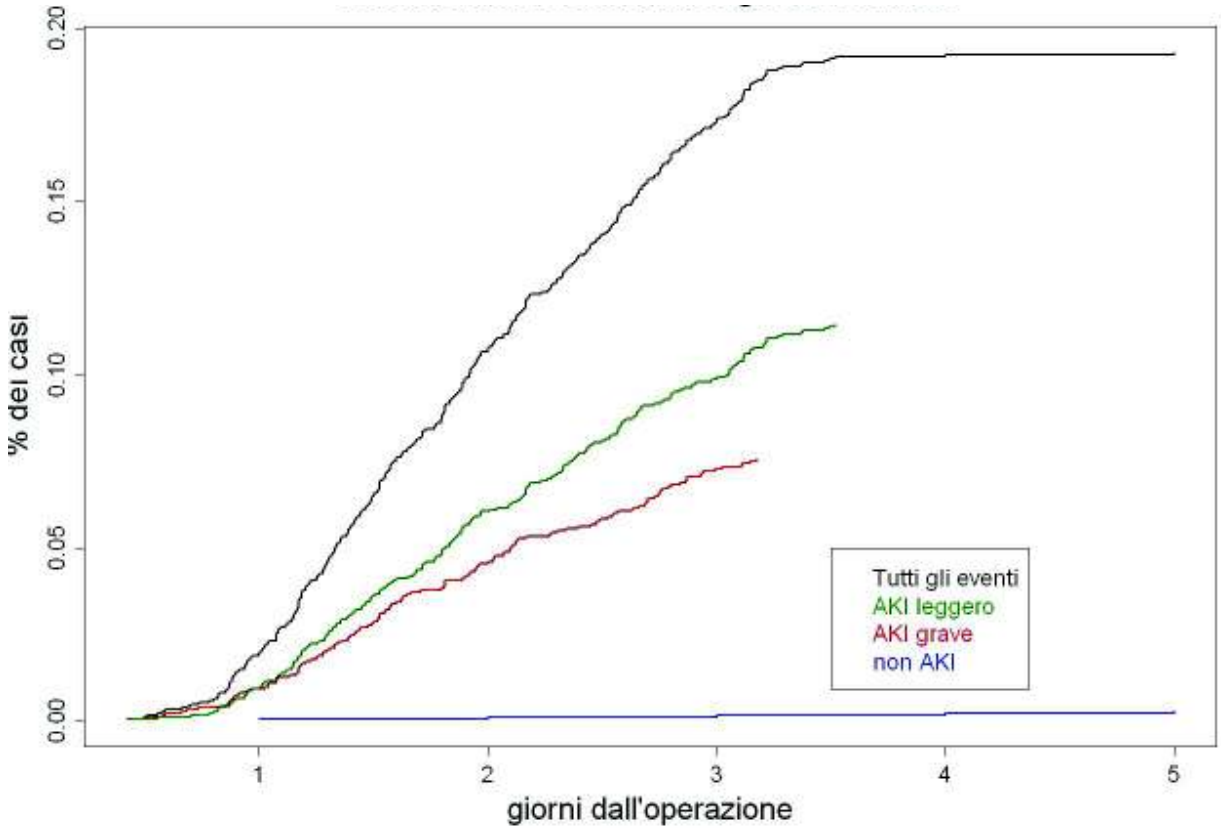


Figura 10:Distribuzione cumulata degli "event times"

Si consideri il bio-indicatore misurato dal primo campione d'urina post-operatorio che chiamiamo bio-indicatore base. La sua distribuzione è mostrata nella Figura 11 per i 4 gruppi di pazienti (AKI grave, AKI leggero, Non AKI e i Controlli). Si vede che, comparati con i controlli, i gruppi AKI hanno generalmente valori più alti dei bio-indicatori base, con il gruppo di AKI grave sono rimossi dai controlli più dei valori dell'AKI leggero.

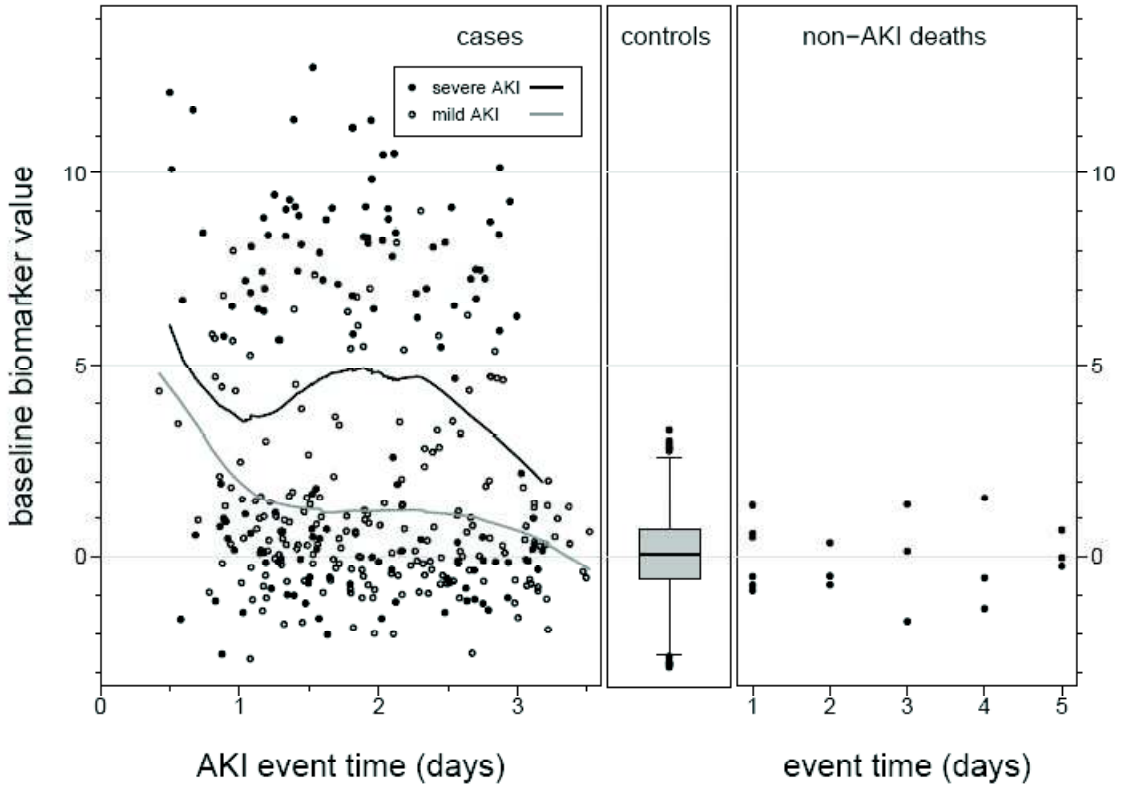


Figura 11: distribuzione dei bio-indicatori base. Le Curve Lowess mostrano l'andamento degli eventi con AKI grave e leggero

I bio-indicatori base nei pazienti che muoiono per eventi non-AKI non sembrano differire dai controlli. Una comparazione formale tra gruppi può essere basata sul *"Wilcoxon rank sum statistic"* che restituisce i seguenti livelli di significatività:

$p<0.001$ per AKI leggero contro controlli;

$p<0.001$ per AKI grave contro controlli;

$p=0.038$ per AKI grave contro AKI leggero;

$p=0.23$ per morti non-AKI contro controlli.

Si ricorda che il *"Wilcoxon rank sum statistic"* è una funzione semplice della stima non parametrica dell'area sotto la curva ROC che compara due gruppi.

Si consideri ora la capacità del bio-indicatore base di diagnosticare gli eventi AKI prima della loro diagnosi clinica con il siero di creatinina. In futuro il trattamento può iniziare sulla base dei bio-indicatori. I clinici avranno bisogno di bilanciare i potenziali benefici del trattamento per soggetti che potrebbero avere un AKI in assenza di trattamento contro i falsi positivi, quelli ovvero coloro che non avrebbero un AKI ma che sono classificati come positivi dal bio-indicatore.

Le curve ROC sommarie, grezze per il bio-indicatore base nella Figura 11, sono state calcolate categorizzando l'asse del tempo dell'evento come presto =(.25,1.5] e medio =(1.5,3]. Le curve ROC che comparano i valori base del bio-indicatore nei controlli con quelli dei soggetti con eventi AKI gravi in ogni intervallo di tempo sono mostrate nella parte sinistra della Figura 12, mentre le curve corrispondenti per soggetti con eventi AKI leggeri sono nella parte destra.

Curve R.O.C. per la valutazione di test diagnostici per eventi nel tempo

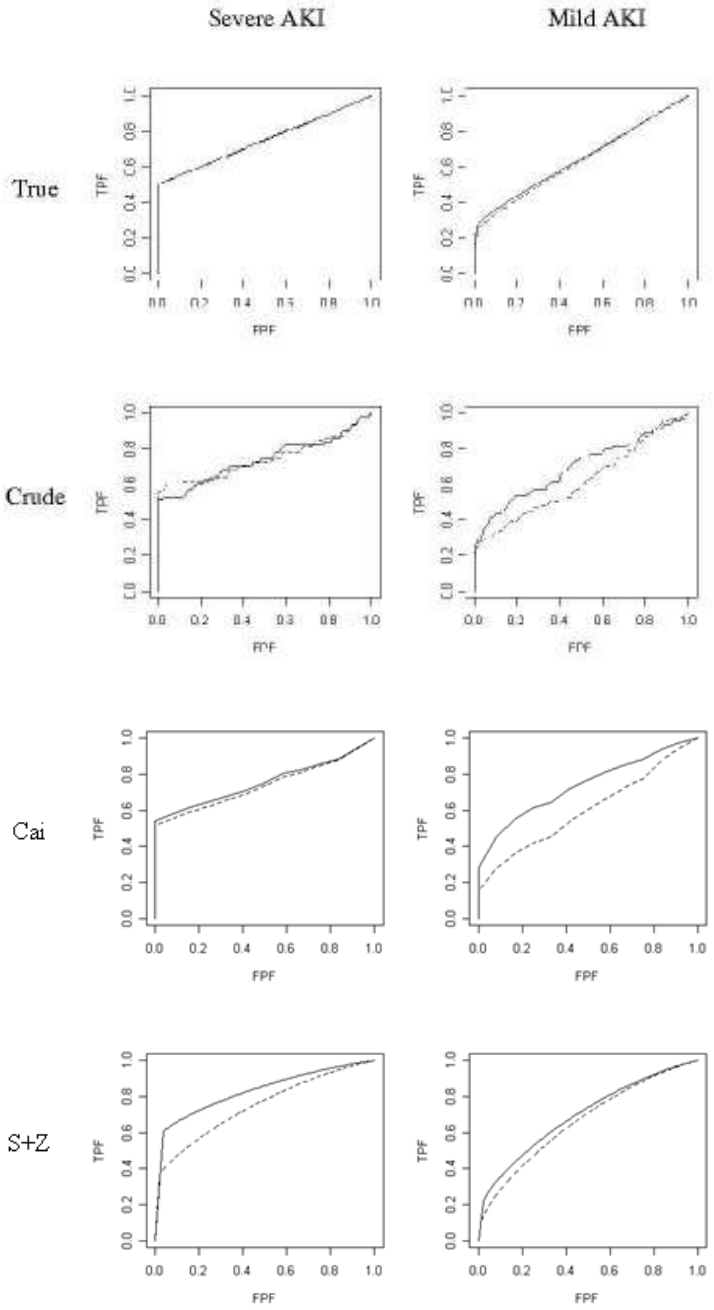


Figura 12: curve ROC e le loro stime per i bio-indicatori base al tempo T=1 e 2 giorni dopo l'intervento

Il metodo Cai è stato implementato per gli indicatori base con il seguente modello di curva ROC "time dependent"

$$logit\{ROC_t(f)\} = h_0(f) + \beta_1 t + \beta_2(t - 1.5)I[t > 1.5] \quad (44)$$

È stata così usata una funzione logistica di collegamento, curve base ROC non-parametriche e abbiamo modellizzato gli effetti del *event time* come uno "*spline*" lineare con un vincolo a *t*=1.5. Modelli separati sono stati adattati per eventi AKI gravi e leggeri, anche se notiamo che un modello che includa entrambi avrebbe potuto essere adeguato includendo interazioni con *"event type*" nella formula ROC-GLM di cui sopra.

È stato anche applicato il metodo Song and Zhou. Per questo sono stati inclusi solo soggetti con eventi gravi e i controlli stimando le curve ROC corrispondenti all'AKI grave contro i controlli e solo gli AKI leggeri contro i controlli nel secondo set di analisi. Il *follow up* è stato terminato, tecnicamente, dopo 5 giorni, termine del tempo di osservazione. Modelli e analisi separati sono stati usati per casi leggeri e gravi. La Figura 12 mostra le curve ROC stimate al tempo *T*=1 e 2 giorni. Con i dati simulati, si è riusciti a calcolare le vere curve ROC *time dependent* simulando un dataset molto ampio, e selezionando casi di ogni gravità con eventi nell'intervallo [*T*-.01, *T*+.01] e controlli, e calcolando le curve ROC empiriche. La Tabella 7 mostra i risultati.

		Severe AKI				Mild AKI			
		True	**Crude**	**Cai**	**S+Z**	**True**	**Crude**	**Cai**	**S+Z**
f=0,05	*T=1*	0,525	0,529	0,565	0,619	0,315	0,354	0,385	0,278
	T=2	0,524	0,600	0,541	0,414	0,289	0,283	0,226	0,199
f=0,20	*T=1*	0,599	0,608	0,631	0,722	0,436	0,538	0,577	0,475
	T=2	0,599	0,613	0,608	0,568	0,413	0,398	0,384	0,414
f=0,50	*T=1*	0,751	0,745	0,755	0,861	0,644	0,754	0,770	0,741
	T=2	0,752	0,725	0,736	0,783	0,634	0,584	0,605	0,710
f=0,80	*T=1*	0,897	0,843	0,875	0,958	0,858	0,892	0,917	0,923
	T=2	0,901	0,863	0,864	0,934	0,857	0,867	0,837	0,914

Tabella 7: confronto delle stime delle Curve ROC per i bio-indicatori base
T indica il tempo (in giorni) dall'intervento alla diagnosi

Si vede, per esempio, che, permettendo un tasso di falsi positivi del 20%, l'indicatore base evidenzia un 59.9% di soggetti che sviluppano un'AKI grave e un 41.3% che sviluppano un'AKI leggero due giorni dopo l'operazione, mentre un giorno dopo l'intervento l'indicatore di base rileva una frazione lievemente più alta, 43.6% di quelli che sviluppano un AKI leggero.

Le vere curve ROC aumentano "a guglia" sulla sinistra e "girano" rapidamente, mentre sono approssimativamente lineari a *VP*=0.5 per AKI grave e a *VP*=0.25 per AKI leggero. La natura non-parametrica della curva ROC base permette, alle curve calcolate con il metodo Cai, di seguire questa forma. Inoltre, le curve stimate con il metodo Cai sono simili alle curve non-parametriche grezze, seguono i dati grezzi piuttosto bene. D'altra parte, le stime Song and Zhou non sono vicine alle curve ROC originali, grezze, presumibilmente, ciò accade perché l'assunzione dei rischi proporzionali non è

idonea. I risultati suggeriscono che l'approccio Song and Zhou dovrebbe essere generalizzato per permettere modelli di rischio non-proporzionali.

Si passa ora ai dati dei bio-indicatori longitudinali. Fino ad ora, è stato esplorato se nei controlli la distribuzione dei bio-indicatori varia con *s*, tempo dall'intervento chirurgico. Questa appare essere stabile nel tempo, in accordo con il modello di simulazione. La Figura 13 è simile alle distribuzioni dei bio-indicatori mostrati in Figura 10, eccetto che tutte le misure dei bio-indicatori sono mostrate, e marginalizzate, sul tempo per i controlli.

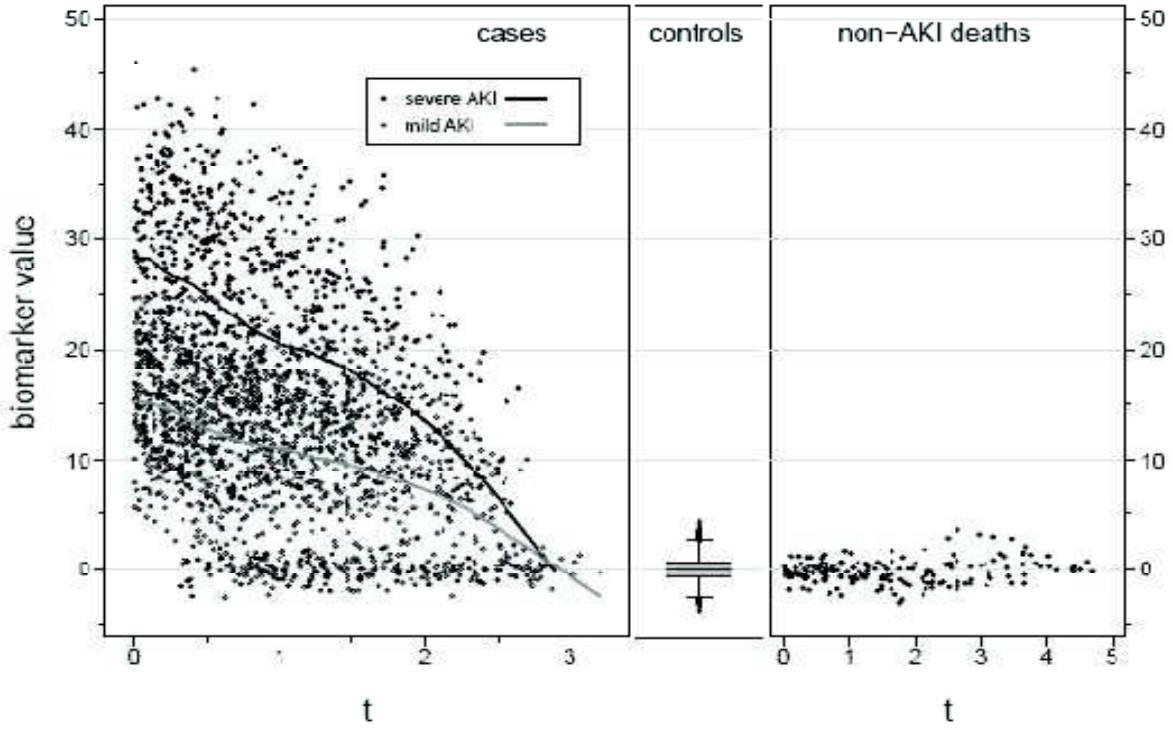


Figura 13: distribuzione dei bio-indicatori nei casi in funzione del tempo tra la misurazione del bio-indicatore e la comparsa dell'evento

L'asse del tempo *t* per i casi è il tempo dalla misurazione dell'indicatore all'evento AKI. Ogni caso ha osservazioni multiple, (*Y,t*), corrispondenti alle varie misurazioni prima del suo evento. Si noti che l'asse del tempo qui differisce da quello usato per gli indicatori base nell'analisi presedente dove *t=T*. Qui si riconosce che l'indicatore base è misurato a qualche tempo nell'intervallo (0,0.25), non a 0. Quindi *t*, tempo dalla misurazione dell'indicatore base all'evento, non è lo stesso del tempo dell'evento, *T*, anche per la misurazione base.

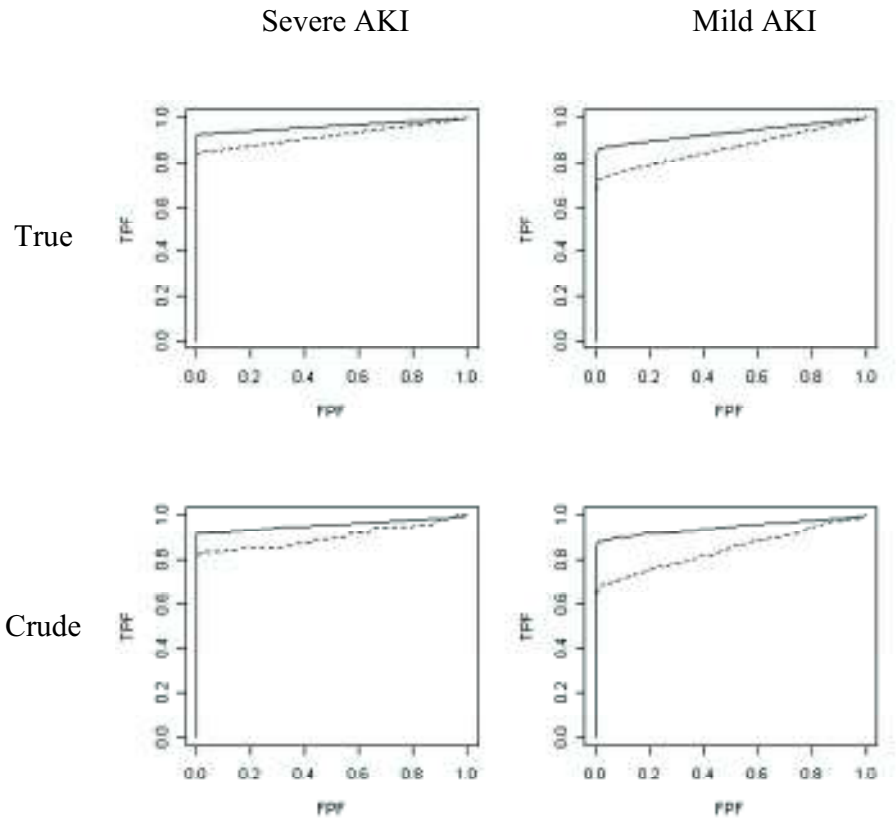
Sono state adattate, usando questi dati longitudinali, curve ROC con gli stessi metodi come descritti precedentemente. I risultati sono mostrati nella Figura 14 e nella Tabella 8.

		Severe AKI				Mild AKI			
		True	**Crude**	**Cai**	**S+Z^a**	**True**	**Crude**	**Cai**	**S+Z^a**
f=0,05	*T=1*	0,932	0,923	0,933	0,645	0,874	0,887	0,880	0,564
	T=2	0,854	0,839	0,804	0,325	0,743	0,696	0,665	0,323
f=0,20	*T=1*	0,942	0,933	0,942	0,717	0,897	0,919	0,917	0,663
	T=2	0,876	0,850	0,828	0,461	0,789	0,753	0,748	0,475
f=0,50	*T=1*	0,965	0,957	0,962	0,839	0,936	0,950	0,953	0,817
	T=2	0,923	0,894	0,882	0,693	0,868	0,853	0,846	0,715
f=0,80	*T=1*	0,986	0,980	0,983	0,943	0,975	0,979	0,983	0,939
	T=2	0,968	0,950	0,944	0,891	0,947	0,948	0,940	0,905

Tabella 8: confronto delle stime delle Curve ROC per i bio-indicatori
T indica l'intervallo (in giorni) trascorso tra l'intervento e la diagnosi

Con il nuovo bio-indicatore quando si accetta un 20% di *FP*, il 94.2% dei soggetti con eventi AKI gravi può essere scoperto un giorno prima della loro diagnosi clinica, e l'87.6% può essere scoperto 2 giorni prima. I numeri corrispondenti per soggetti con AKI leggero sono 89.7% e 78.9%. Si confrontino questi con le proporzioni molto più piccole che possono essere scoperte usando solamente il bio-indicatore base.

Riguardo le stime delle curve ROC "time dependent", il metodo Song and Zhou sembra sottostimare. La sottostima è particolarmente problematica con *FP* più piccoli. Presumibilmente l'assunzione dei rischi proporzionali è ancora inadeguata. Il metodo Cai sembra essere migliore in quanto è vicino alle curve non-parametriche grezze, ma non richiede di scegliere intervalli di tempo per *t* per stimare le curve ROC al tempo *t*.



Curve R.O.C. per la valutazione di test diagnostici per eventi nel tempo

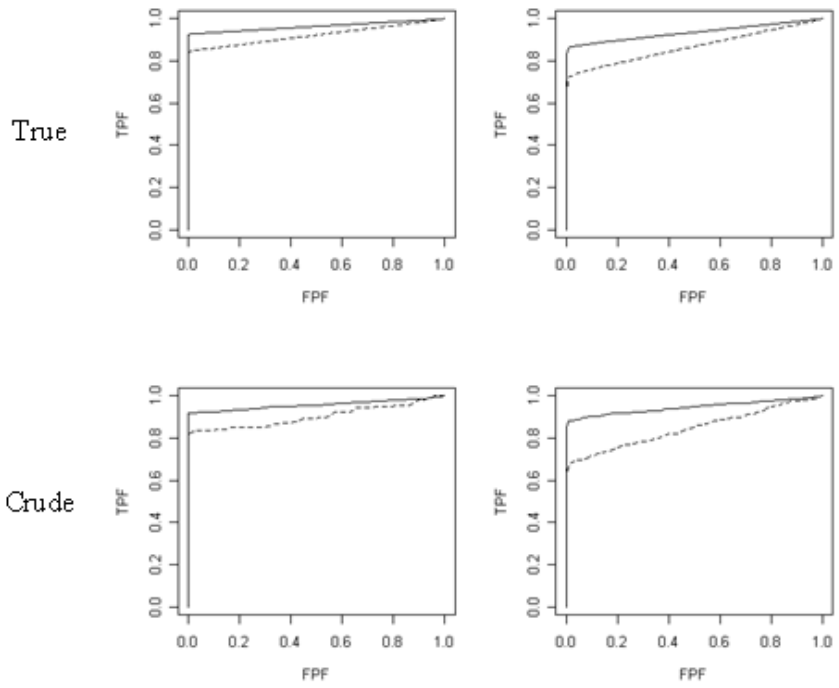


Figura 14: curve ROC e le loro stime per i bio-indicatori longitudinali al tempo T=1 e 2 giorni dopo l'intervento

CONCLUSIONI

In questa tesi sono stati visti i diversi approcci per la valutazione di variabili "event time" usando le curve ROC.

Dopo una breve descrizione di cosa rappresentano e di come si costruiscono le curve di ROC, sono stati messi a confronto due metodi per l'analisi della sopravvivenza: il metodo di Kaplan e Meier ed il Nearest Neighbor Estimation (NNE).

Da questo confronto è emerso che gli andamenti dei modelli sono molto simili, anche se il modello stimato con il metodo NNE può risultare leggermente meno accurato, in quanto l'uso di un parametro di lisciamento porta ad una leggera perdita di informazione. Per contro, il metodo Kaplan e Meier non garantisce la monotonicità delle curve.

Sono state confrontate le curve ROC ottenute con il metodo proposto da Cai e da Song and Zhou. Le curve calcolate con il metodo Cai seguono la forma delle "vere" Curve ROC e seguono i dati grezzi piuttosto bene, mentre le stime Song and Zhou non sono vicine alle curve ROC originali, grezze. Ciò accade, presumibilmente, perché l'assunzione dei rischi proporzionali non è idonea.

Un ultimo confronto è stato effettuato su dati longitudinali, stimando le curve di ROC con i metodi precedentemente utilizzati.

Il metodo Song and Zhou sembra sottostimare le curve ROC "time dependent", mentre il metodo Cai, avvicinandosi alle curve non-parametriche grezze, sembra stimare meglio l'andamento descritto dalle curve ROC.

APPENDICE – comandi usati per le analisi del cap. 3

```
#Caricamento libreria
library(survivalROC)

#Caricamento dati
aki<-read.table("akipulito.csv",sep=";",h=T)


#ANALISI GRUPPO GENERALE

attach(aki)
#Modifica valori dello status (da impostare su 1)
for(i in 1:length(statusall)){
      if(statusall[i]==9) statusall[i]<-0
      if(statusall[i]==2) statusall[i]<-1
}

#Stime del modello

ROC.NNE.G=survivalROC(Stime=tmfull,status=statusall,marker=y,predict.time=5,lambda=0.05)
ROC.KM.G=survivalROC(Stime=tmfull,status=statusall,marker=y,predict.time=5,method="KM")


#plot

plot(c(0,1),c(1,0),type="n",main="Curva ROC per l'intero campione",xlab="Falsi
positivi",ylab="Falsi negativi")
lines(ROC.NNE.G$FP,ROC.NNE.G$TP,col="red")
lines(ROC.KM.G$FP,ROC.KM.G$TP,col="green4")
legend(0.8,0.4,c("Metodo NNE","Metodo KM"),text.col=c("red","green4"),cex=1.5)


detach(aki)
rm(statusall)


#ANALISI GRUPPO 1 (AKI GRAVE)

#Selezione del gruppo 1 (AKI grave) e del gruppo 0 (non AKI)
aki_1<-aki[aki$statusall==1|aki$statusall==0,]
attach(aki_1)



#Stime del modello

ROC.NNE.1=survivalROC(Stime=tmfull,status=statusall,marker=y,predict.time=5,lambda=0.05)
ROC.KM.1=survivalROC(Stime=tmfull,status=statusall,marker=y,predict.time=5,method="KM")


#plot

plot(c(0,1),c(1,0),type="n",main="Curva ROC per l'AKI grave",xlab="Falsi
positivi",ylab="Falsi negativi")
```

```
lines(ROC.NNE.1$FP,ROC.NNE.1$TP,col="red")
lines(ROC.KM.1$FP,ROC.KM.1$TP,col="green4")
legend(0.8,0.4,c("Metodo NNE","Metodo KM"),text.col=c("red","green4"),cex=1.5)


detach(aki_1)
rm(statusall)



#ANALISI GRUPPO 2 (AKI LIEVE)

#Selezione del gruppo 1 (AKI grave) e del gruppo 0 (non AKI)
aki_2<-aki[aki$statusall==2|aki$statusall==0,]
attach(aki_2)


statusall_bis<-statusall
for(i in 1:length(statusall)){
      if(statusall[i]==2) statusall_bis[i]<-1
}

#Stime del modello
ROC.NNE.2=survivalROC(Stime=tmfull,status=statusall_bis,marker=y,predict.time=5,lambda=0.05
)
ROC.KM.2=survivalROC(Stime=tmfull,status=statusall_bis,marker=y,predict.time=5,method="KM")


#plot

plot(c(0,1),c(1,0),type="n",main="Curva ROC per l'AKI leggero",xlab="Falsi
positivi",ylab="Falsi negativi")
lines(ROC.NNE.2$FP,ROC.NNE.2$TP,col="red")
lines(ROC.KM.2$FP,ROC.KM.2$TP,col="green4")
legend(0.8,0.4,c("Metodo NNE","Metodo KM"),text.col=c("red","green4"),cex=1.5)


detach(aki_2)
rm(statusall_bis)


Grafici affiancati:
par(mfrow=c(1,2))

#PLOT CHE CONFRONTA I VARI GRUPPI (NNE)

plot(c(0,1),c(1,0),type="n",main="Curva ROC per i vari gruppi (metodo NNE)",xlab="Falsi
positivi",ylab="Falsi negativi")
lines(ROC.NNE.G$FP,ROC.NNE.G$TP,col="green3",lwd=3)
lines(ROC.NNE.1$FP,ROC.NNE.1$TP,col="red", lwd=3)
lines(ROC.NNE.2$FP,ROC.NNE.2$TP,col="blue",lwd=3)
legend(0.5,0.4,c("Intero Campione","AKI grave","AKI
leggero"),text.col=c("green3","red","blue"),cex=1.5)

#PLOT CHE CONFRONTA I VARI GRUPPI (KM)
```

```
plot(c(0,1),c(1,0),type="n",main="Curva ROC per i vari gruppi (metodo KM)",xlab="Falsi
positivi",ylab="Falsi negativi")
lines(ROC.KM.G$FP,ROC.KM.G$TP,col="green3",lwd=3)
lines(ROC.KM.1$FP,ROC.KM.1$TP,col="red", lwd=3)
lines(ROC.KM.2$FP,ROC.KM.2$TP,col="blue",lwd=3)
legend(0.5,0.4,c("Intero Campione","AKI grave","AKI
leggero"),text.col=c("green3","red","blue"),cex=1.5)
```

Bibliografia

- Lusted L.B. (1971). Signal detectability and medical decision-making. Science,171, 1217-19.
- Goodenough D.J., Rossmann K., Lusted L.B. (1974). Radiographic applications of receiver operating characteristic (ROC) analysis. Radiology, 110, 89-95.
- Hanley J., McNeil B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143, 29-36.
- Erdreich L.S. (1981). Use of Relative Operating Characteristic analysis in Epidemiology. Am. J. Epidemiol., 114, 649-62.
- Henderson A.R. (1993). Assessing test accuracy and its clinical consequences: a primer for receiver operating characteristic curve analysis. Ann. Clin. Biochem., 30, 521-39.
- Greiner M., Pfeiffer D., Smith R.D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Prev. Vet.
- Bamber D. (1975). The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph. J. Math. Psychol., 12, 387-415.
- Zweig H.H., Campbell G. (1993). Receiver Operating Characteristic (ROC) plots: a fundamental evolution tool in medicine. Clin. Chem., 39, 561-77.
- Barajas-Rojas J. A., Riemann H. P., Franti C. E. (1993). Notes about determining the cut-off value in enzyme-linked immunosorbent assay (ELISA). Prev. Vet. Med., 15, 231-3.
- Schäfer H. (1989). Constructing a cut-off point for a quantitative diagnostic test. Statistics in Medicine, 8, 1381-91.
- Swets J.A. (1998). Measuring the accuracy of diagnostic systems. Science, 240, 1285-93.

- Cai T, Pepe MS, Zheng Y, Lumley T, Jenny NS (2006). The sensitivity an specificity of markers for event time, Biostatistics 7:182-197.
- Cai T, Pepe MS (2002). Semiparametric ROC analysis to evaluate biomarkers for disease. J Am Stat Assoc 97:1099-1107
- Etzioni R, PepeM, Longton G, Hu C, Goodman G (1999). Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. Med Decis Making 19:242-251
- Heagerty PJ, Zheng Y (2005). Survival model predictive accuracy and ROC curves. Biometrics 61:92-105
- Heagerty PJ, Lumley T, Pepe MS (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 56:337-344
- Leisering W, Pepe MS, Longton G (1997). A marginal regression modeling framework for evaluating medical diagnostic tests. Stat Med 16:1263-1281
- Pepe MS (2003). The statistical evaluation of medical tests for classiocation and prediction. Oxford University Press, New York
- Song X, Zhou XH (in press) A semiparametric approach for the coviariate specific ROC curve with survival outcome. Stat Sinca.
- Ang TJ, GonaP, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robinson SJ, Benjamin EJ, D'Agostino RB, Vasan RS (2006) Multiple biomarkers for the prediction of first major cardiovascular events and death. N Engl J Med 355:2631-2639
- Xu R, O' Quigley J (2000). Propotional hazard estimate of the conditional serviva function, J Roy Stat Soc Ser B 62: 667-680
- Zheng Y, Heagerty PJ (2004). Semiparametric estimation of time dependent ROC curves for longitudinal marker data. Biostatistic5:615-632
- Zheng Y, Heagerty PJ (2007). Prospective accuracy for longitudinal markers. Biometrics 63:332-341
- Zheng Y, Cai T, Feng Z (2006). Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. Biometrics 62:279-287
- Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distibution under random censoring. Annals of statistic, 22: 1299-1327
- http://www.sin-italy.org/Governo_Clinico/pdf/Analisi_di_sopravvivenza_metodo_Kaplan-Meier.pdf

Indice delle Tabelle

Indice delle Figure