

Statistica di base degli studi Real-life I

Scardapane Marco

Diapositiva preparata da MARCO SCARDAPANE e ceduta alla Società Italiana di Diabetologia.
Per ricevere la versione originale si prega di scrivere a siditalia@siditalia.it



Center for Outcomes Research and Clinical Epidemiology

Gli alberi di regressione

- Gli **alberi di regressione**, o tecniche di **regressione ad albero**, sono metodi originariamente nati nel campo del data mining e della statistica esplorativa;
- Offrono una presentazione semplice ed intuitiva della relazione tra una serie di predittori (covariate) rispetto all'outcome;
- Gli alberi “segmentano” il campione attraverso una sequenza di split in sottogruppi di pazienti eterogenei fra di loro e omogenei al loro interno rispetto all'outcome...
- ... a differenza dei classici modelli di regressione lineare, che con una sola equazione matematica, esprimono associazioni di covariate con un outcome overall, cioè valide generalmente;
- Nel corso degli anni sono state sviluppate svariate tecniche tree-structured, ognuna con le sue opzioni, funzionalità e limiti.

L'algoritmo RECPAM

- Ideato da Antonio Ciampi, **RECPAM** (**R**ecursive **P**artitioning and **A**malgamation) è un algoritmo che generalizza il celebre metodo CART (Classification and Regression Trees), inventato da Leo Breiman nel 1984;
- Il pregio maggiore infatti è la possibilità di sfruttare dati di qualsiasi natura: continui, dicotomici, tempo all'evento etc.

Outcomes

Continui
Dicotomici
Sopravvivenza
...

Modelli statistici

Lineare
Logistico
Cox
...

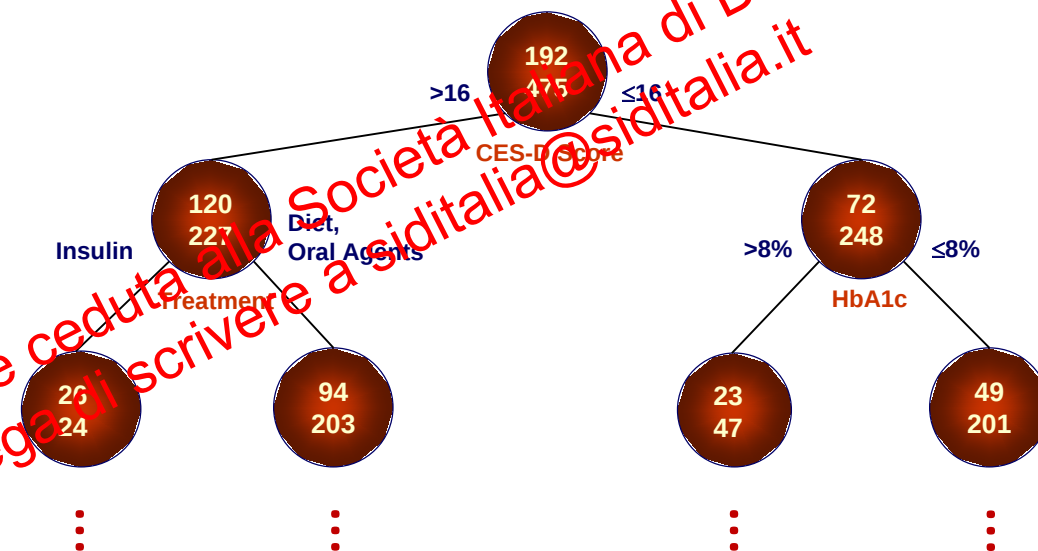
Misure di efficacia

Mean change
Odds Ratio
Hazard Ratio
...

- Ogni albero, incluso RECPAM, è costituito da 3 ingredienti fondamentali: **splitting**, **stopping**, **pruning**.

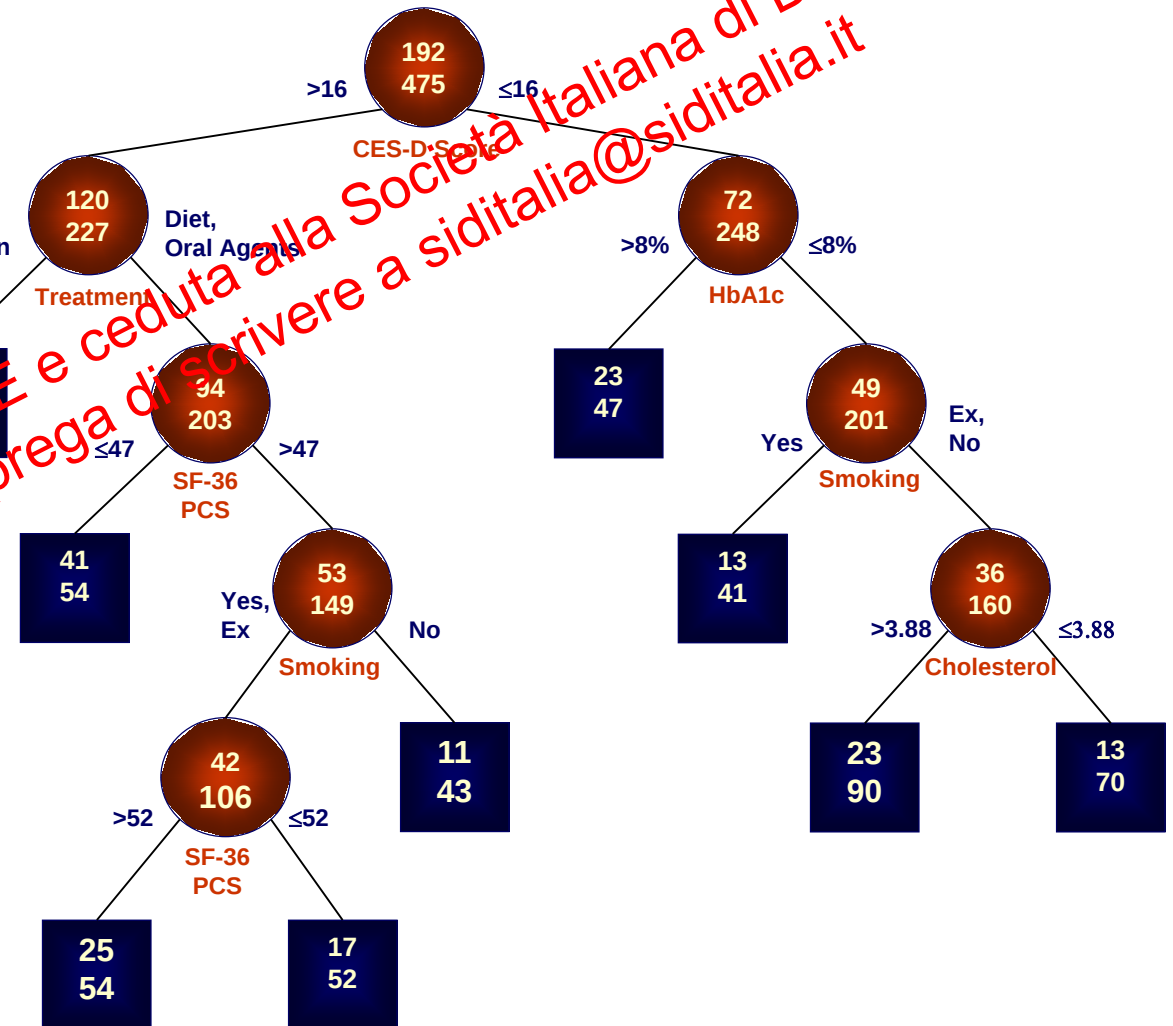
Come funziona / 1: splitting

- Le covariate da dare in pasto all'algoritmo possono essere continue, nominali o ordinali;
- Ad ogni split l'algoritmo sceglie la variabile e la condizione che meglio separa (rispetto all'outcome) i due sottogruppi che andrà a creare;
- Ad ogni split tutto il set di covariate è passato in rassegna → una variabile può essere scelta più volte!
- Se la variabile scelta è continua l'algoritmo sceglie anche il cut-off ottimale esaminando tutti i decili (o percentili);
- Se è nominale esamina tutte le combinazioni possibili;
- Se è ordinale solo le combinazioni contigue.



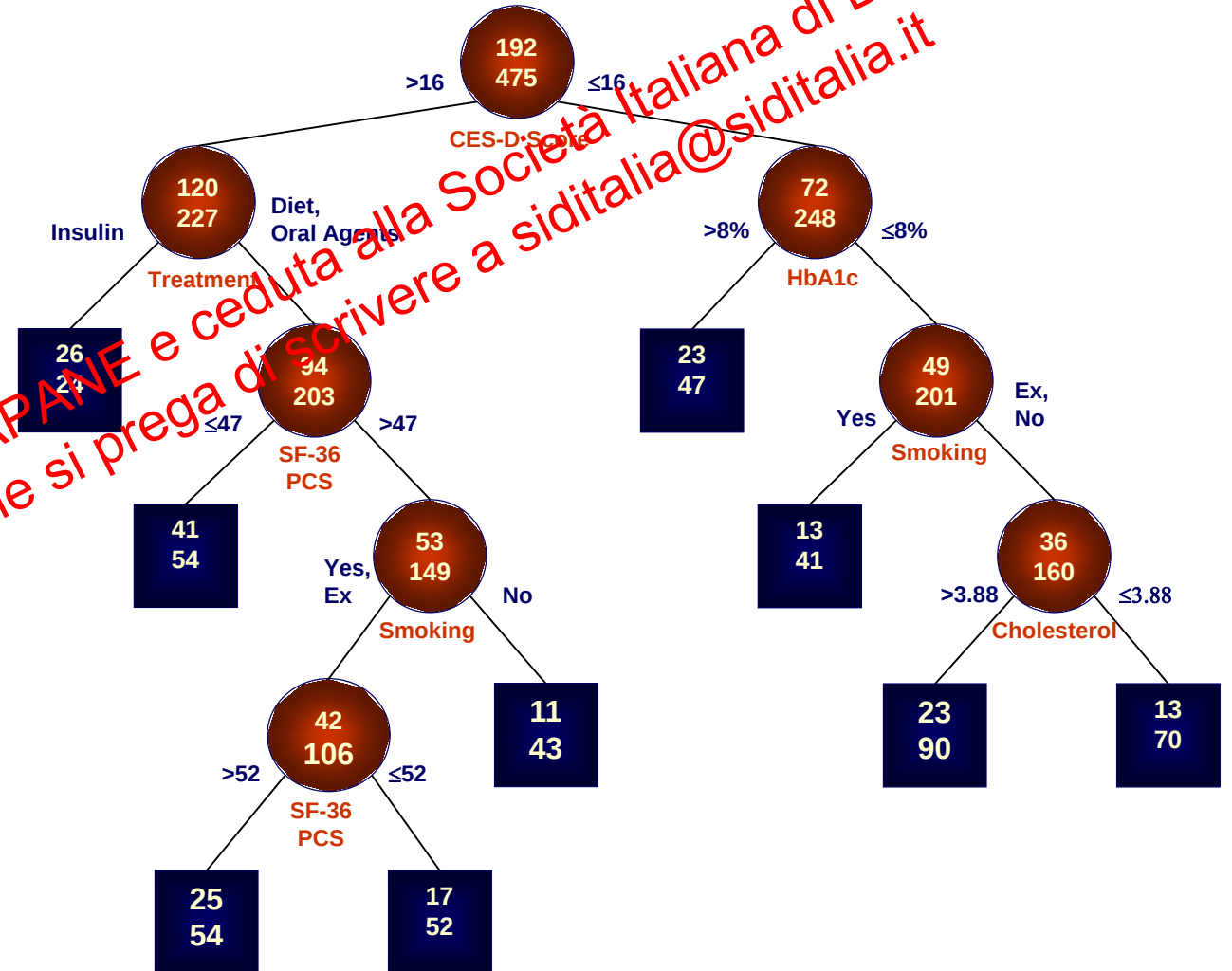
Come funziona / 2: stopping

- Come tutti gli algoritmi, è necessario specificare all'albero quando fermarsi.
- Due motivi:
 1. Impedire la crescita abnorme dell'albero;
 2. Ottenere risultati finali abbastanza robusti.
- La regola di **stopping** è il raggiungimento del numero minimo di soggetti e/o eventi in un nodo.
- Se lo split di un nodo comporta la nascita di due nodi che violano la regola, lo split si interrompe e si prova a splittare altri nodi, o si passa alla fase successiva di **pruning**.



Come funziona / 3: pruning

- Malgrado le stopping rules, l'albero completo (large tree) è ancora eccessivamente grande.
- Serve un'ulteriore fase, il **pruning**.
- Nel pruning possiamo uno dopo l'altro i rami via via creati fino a tornare al nodo origine.
- Della sequenza di alberi, scegliamo quello che rappresenta il compromesso migliore fra perdita di «informazione» e leggibilità clinica.
- Il pruning è essenziale per evitare il problema dell'**overfit**.



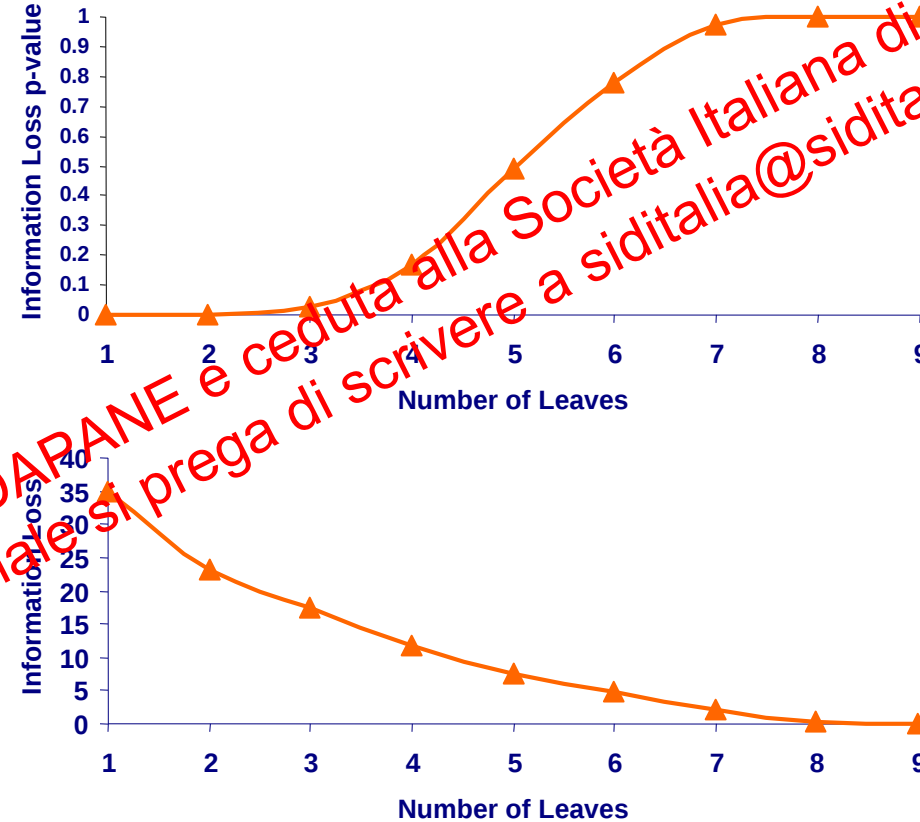
Il problema dell'overfitting

- Generalmente, i fenomeni che cerchiamo di spiegare o predire, sono costituiti da una parte **sistematica** («spiegabile») e da una **casuale** («inspiegabile»), il famoso errore.
- Lo scopo di un modello statistico è fornire una rappresentazione (stima) della prima parte e diminuire la seconda, MAI eliminarla!
- Se si introducono troppe variabili in un modello, esso tende a adattarsi non solo la prima parte ma anche l'errore, e si crea overfit;
- Ciò mina la bontà del modello e fa sì che il nostro modello sia «buono» (troppo) quando applicato ai nostri dati, e «pessimo» per dati esterni;
- Come evitarlo? Per la regressione classica si usa il principio di **parsimonia**, per quelle ad albero il pruning;
- Se lo scopo del modello è la predizione è buona norma usare anche un metodo di validazione.

Diapositiva preparata da MARCO SCARDAPANE e ceduta alla Società Italiana di Diabetologia.
Per ricevere la versione originale si prega di scrivere a siditalia@siditalia.it

Pruning: criteri statistici

Step #	Leaves # Pruned	IC	IL	IL p-value	
0	9	35.0	0	1	
1	8	30 31	34.6	0.3	1
2	7	44 45	32.8	2.1	0.97
3	6	14 15	30.2	4.7	0.78
4	5	22 23	27.6	7.4	0.49
5	4	10 11	23.4	11.5	0.17
6	3	4 5	17.6	17.3	0.02
7	2	6 7	11.9	23.08	0.003
8	1	2 3	0	35.0	<0.001



N.B: non dimenticate di unire le valutazioni statistiche alle considerazioni di tipo clinico (e.g. plausibilità biologica)!

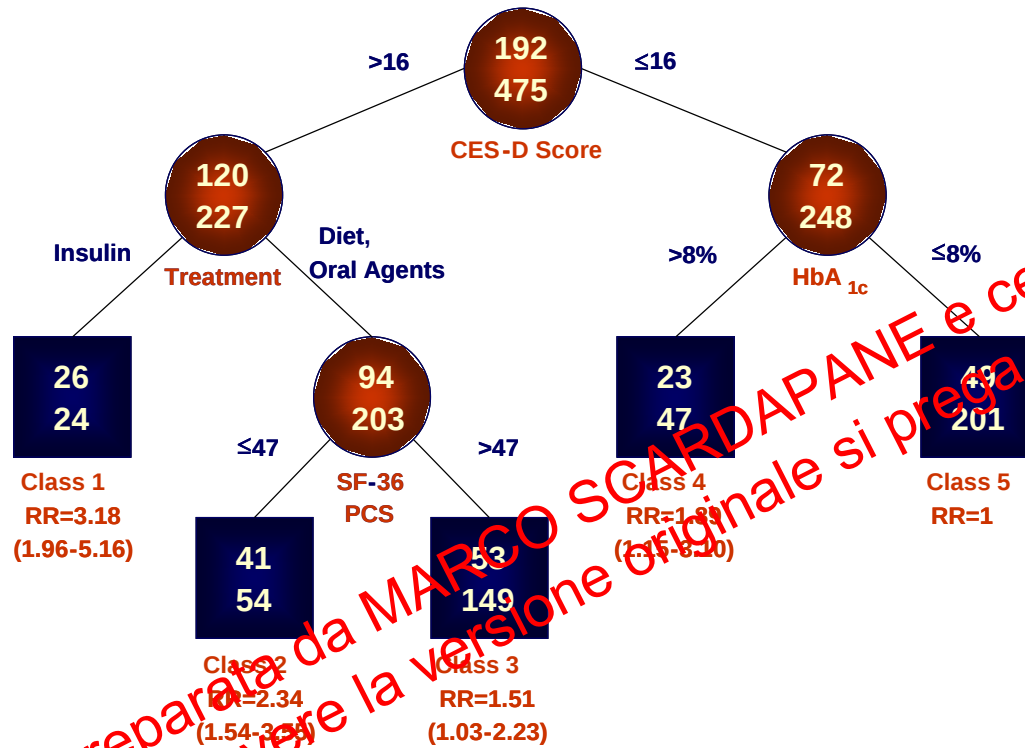
Diapositiva preparata da MARCO SCARDAPANE e ceduta alla Società Italiana di Diabetologia. Si prega di scrivere a siditalia@siditalia.it

L'albero finale

- Una volta scelto l'albero finale RECPAM prende come riferimento la classe più a destra (quella col minor numero di pazienti "suscettibili" di sviluppare l'outcome) e calcola i rischi relativi rispetto a questa;
- Ad ogni stima associa un p-value e un CI, come qualunque modello di regressione classico;
- È evidente come RECPAM riesca a scovare potenziali effetti locali altrimenti invisibili usando un modello lineare classico;
- Migliore è anche la sua capacità nel rilevare interazioni ed effetti non lineari (e.g. se una stessa covariata compare in più split);
- Per tali motivi RECPAM viene spesso chiamata tecnica di **interaction detecting**, **prognostic stratification** (o segmentation), e **prediction structure**.



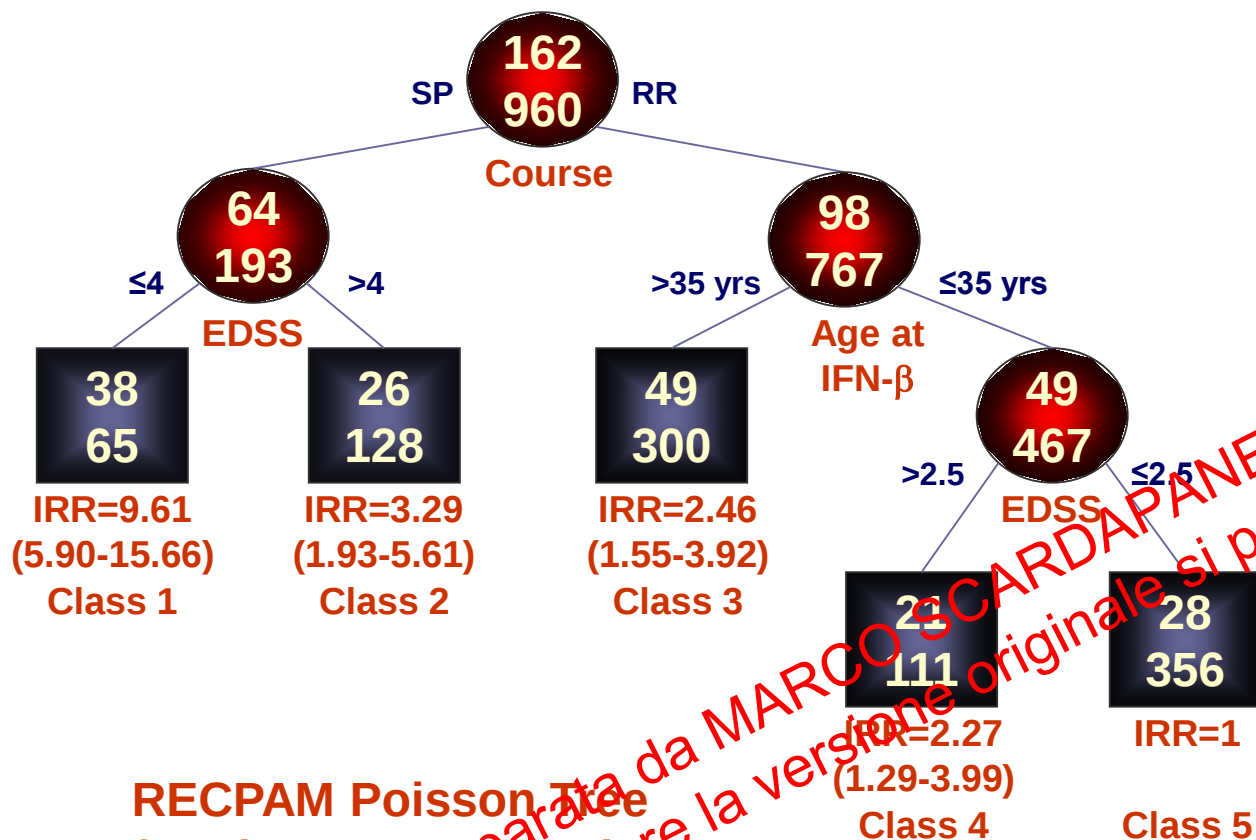
Con un po' di immaginazione...



Variable	RR	95% CI
Age (yrs)		
1	1.05	1.03-1.07
RECPAM class		
1	3.16	1.93-5.16
2	2.08	1.36-3.17
3	1.45	0.98-2.15
4	1.84	1.12-3.02
5	1.00	-
Cholesterol (mmol/L)		
<3.88	1.00	-
3.88-5.17	2.72	0.98-7.55
>5.17	2.65	0.96-7.34
Neuropathy		
No	1.00	-
Yes	1.54	1.00-2.36

Diapositiva preparata da MARCO SCARDAPANE e ceduta alla Società Italiana di Diabetologia.
Per ricevere la versione originale si prega di scrivere a siditalia@siditalia.it

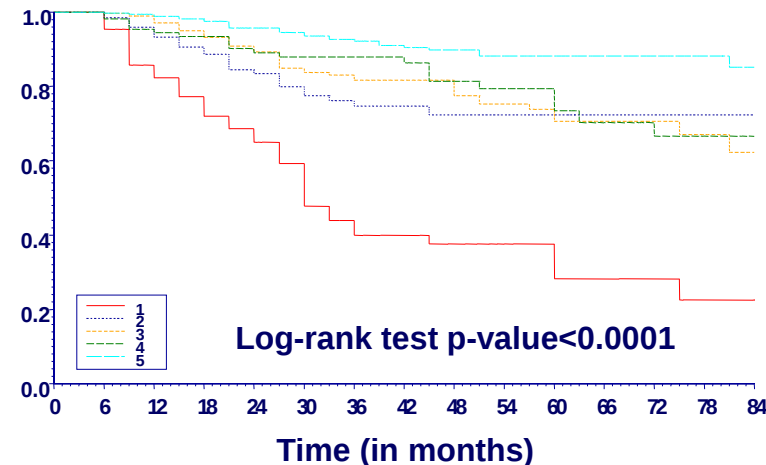
Con un po' di immaginazione...



% of patients not in progression

RECPAM class	Months	12	24	36
1	65	48	28	
2	83	72	72	
3	89	77	71	
4	89	81	67	
5	96	90	88	

Kaplan-Meier curves for Time to progression by RECPAM class



RECPAM Poisson Tree
for Time to Progression

Diapositiva preparata da MARCO SCARDAPANE e ceduta alla Società Italiana di Diabetologia.
Per ricevere la versione originale si prega di scrivere a siditalia@siditalia.it