Progetto Big Data in Public Health

Julia Bui Xuan 882385
Simone Farallo 889719
Michele Salvaterra 891109
Luca Sammarini 884591
Eugenio Tarolli 889038

Corso di laurea magistrale - Data Science

Prof.ssa Maria Grazia Valsecchi
Prof.ssa Paola Rebora
Dott.ssa Anita Andreano

Dataset

German Health Registry – GermanH.csv

Estratto del German health registry per l'anno 1984 relativo ad una piccola cittadina (dataset reale openData modificato)

idnum:

codice identificativo del soggetto

smoke (yes or no):

se il soggetto fuma

sex (Female or Male)

married (yes or no):

se il soggetto è sposato

age (numerica, in anni)

kids (yes or no):

se il soggetto ha figli

work(yes or no):

se lavora

education:

no/low = da nessuna a diploma scuola media inferiore;
medium/high = diploma scuola media superiore o oltre

Dataset

Registro tumori tedesco - Cancerregister.csv

Registro tumori tedesco (dataset simulato) relativo al mese di Gennaio 1984

idnum:

codice identificativo del soggetto

stadio (I, II, III, IV):

stadio del tumore alla diagnosi

incidenza:

data di diagnosi del tumore

tipo di tumore:

seno, polmone, colon, altro

geneticm:

fattore genetico (1=positivo, 0=negativo)

Dataset

Schede di dimissione ospedaliera – SDO.csv

Schede di dimissione ospedaliera dei soggetti ricoverati in Germania tra gennaio 1984 e ottobre 1984 per trattamenti oncologici (dataset simulato)

idnum:

codice identificativo del soggetto

prestazione:

tipo di trattamento ricevuto durante il ricovero (chirurgico, chemioterapico o radioterapico)

data prestazione:

data del trattamento

ospedale:

codice univoco dell'ospedale

dimissione:

data di dimissione dall'ospedale

Dataset

Registro di mortalità – Deathregister.csv

Estratto del Registro di mortalità della cittadina tedesca che riporta la mortalità dal 1984 al 1988 e lo stato in vita alla fine del 1988 (dataset simulato)

idnum:

codice identificativo del soggetto

dead:

stato in vita alla data enddate

enddate:

data di ultima osservazione
(se *dead*=1, data di morte)

1. Esaminare i dati per ciascun dataset e procedere alla rimozione di eventuali dati errati

E’ stata svolta una fase di cleaning, in particolare sono stati controllati e rimossi:

- **valori nulli**
- **duplicati**
- **date incongruenti**

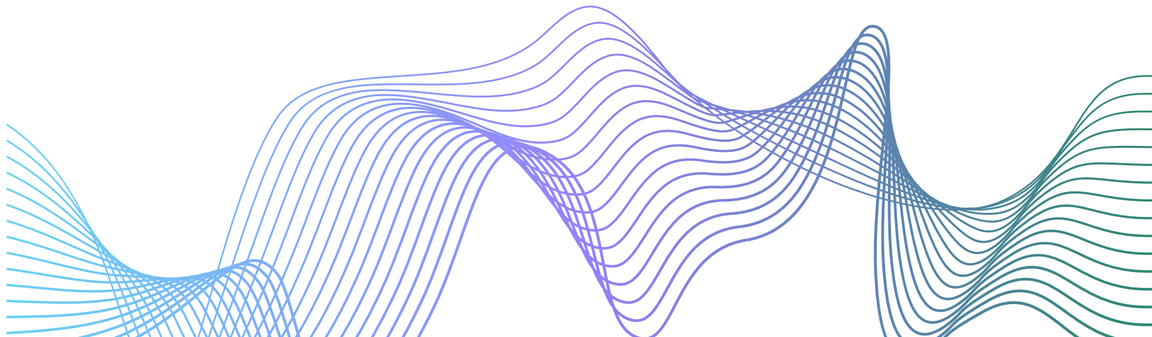
Dataset	Oss. iniziali	N. valori nulli	N. duplicati	Date incongruenti	Oss. post-cleaning
Cancerregister	10005	15	3	0	9993
Deathregister	7748	0	0	0	7748
GermanH	7748	72	0	0	7676
SDO	10002	4	0	42	9955

2. Costruire l’indicatore ‘Intervento chirurgico di asportazione del tumore al seno entro 60 giorni dalla data di diagnosi’ su base mensile per i casi incidenti nel mese di gennaio 1984

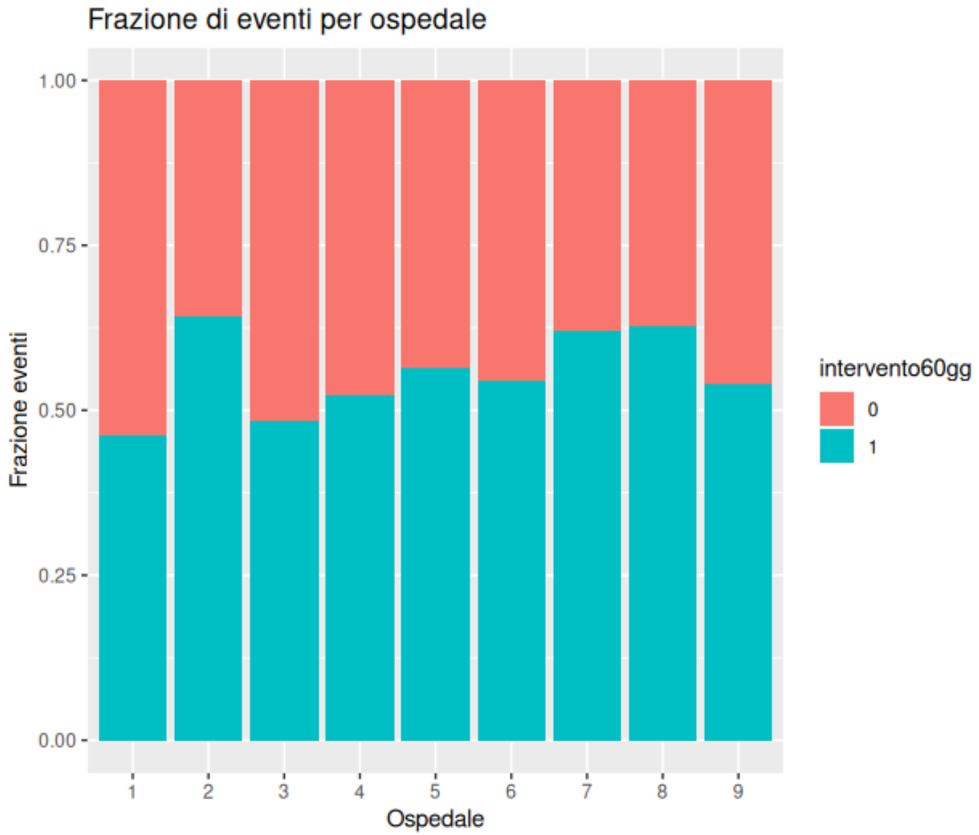
Denominatore: tutte le pazienti di sesso femminile con tumore al seno insorto tra 01/01/1984 e 31/01/1984, in stadio I o II, che hanno subito un intervento chirurgico.
Numeratore: tutte le pazienti al denominatore con intervallo tra la data d’incidenza e la data dell’intervento ≤ 60 giorni.

Criteri di inclusione per il denominatore:
1. sex: female
2. tipotumore: seno
3. stadio: Stadio I, Stadio II
4. incidenza: dal 1984-01-01 al 1984-01-31
5. prestazione: chirurgica

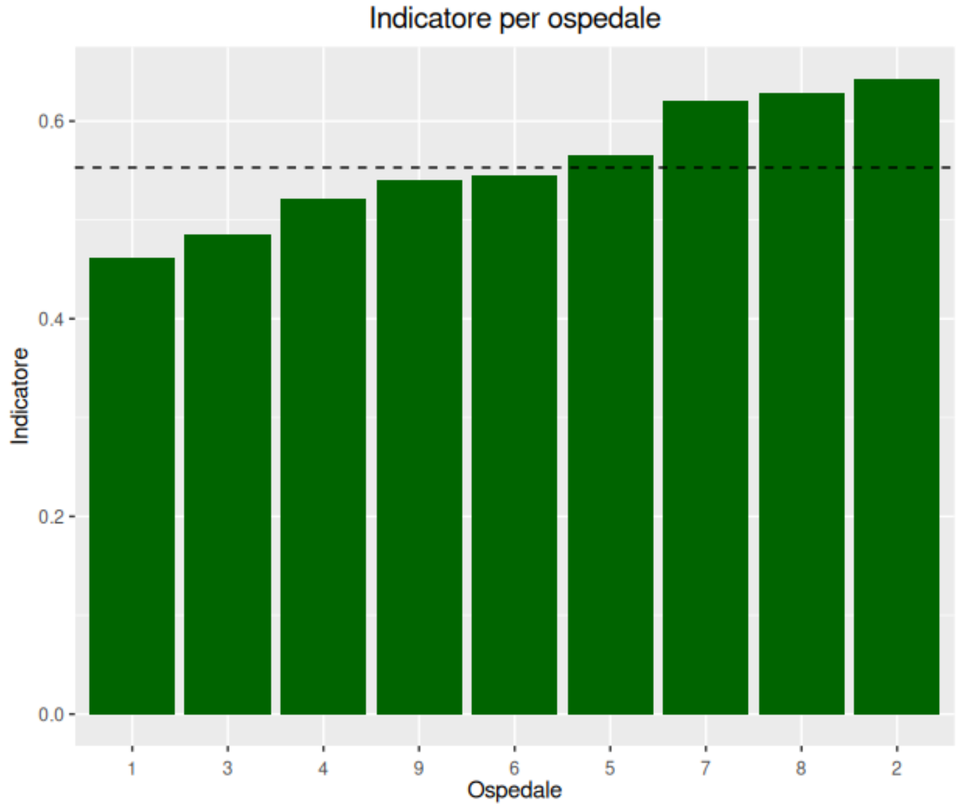
L’indicatore è pari a 0.55. La percentuale di interventi entro 60 giorni dalla data di diagnosi è pari al 55%.



3. Calcolare l’indicatore ‘Intervento chirurgico di asportazione del tumore al seno entro 60 giorni dalla data di diagnosi’ per ospedale



Gli ospedali 2, 7, 8 presentano la percentuale maggiore di eventi.



Il valore di riferimento (linea tratteggiata) rappresenta l’indicatore calcolato sull’intero dataset.

4. Associazione a livello individuale tra il livello di educazione ed il valore dell’indicatore

Si vuole confrontare la probabilità che una donna con un basso livello di educazione abbia un intervento chirurgico entro 60 giorni, con la probabilità che una donna con un medio/alto livello di educazione abbia lo stesso intervento entro lo stesso periodo.

Misura di effetto: ODDS RATIO
0.97 (0.51, 1.87)

L’OR non è significativamente diverso da 1.
Il livello di educazione non influenza il valore dell’indicatore.

5. OR aggiustato per la variabile *working* tramite il metodo di Mantel Haenszel

Si costruisce la tabella di contingenza per i due gruppi in base allo stato lavorativo.

Il test di BreslowDay non rifiuta l'ipotesi nulla di omogeneità; si può proseguire nel calcolo dell'OR.

Il test di MH presenta un p-value del 0.93.

Non è possibile rigettare l'ipotesi nulla; l'OR di Mantel Haenszel non è significativamente diverso da 1. Il rapporto fra l'OR crudo e l'OR di MH è pari circa a 1. **Lo stato lavorativo non è un confondente.**

6. Stima dell’associazione fra tutte le variabili ritenute come possibili confondenti attraverso un modello di regressione logistica

Variabili analizzate come possibili confondenti:

- **geneticm**
- **stadio**
- **age**, suddivisa in intervalli regolari

Il rapporto fra l’OR crudo e l’OR di MH risulta essere circa pari a 1 per tutte le variabili considerate.
Le variabili *geneticm*, *stadio* ed *age* non sono confondenti.

La variabile *stadio* risulta positivamente associata alla presenza di evento: le pazienti con tumore in stadio più avanzato sono generalmente operate prima.

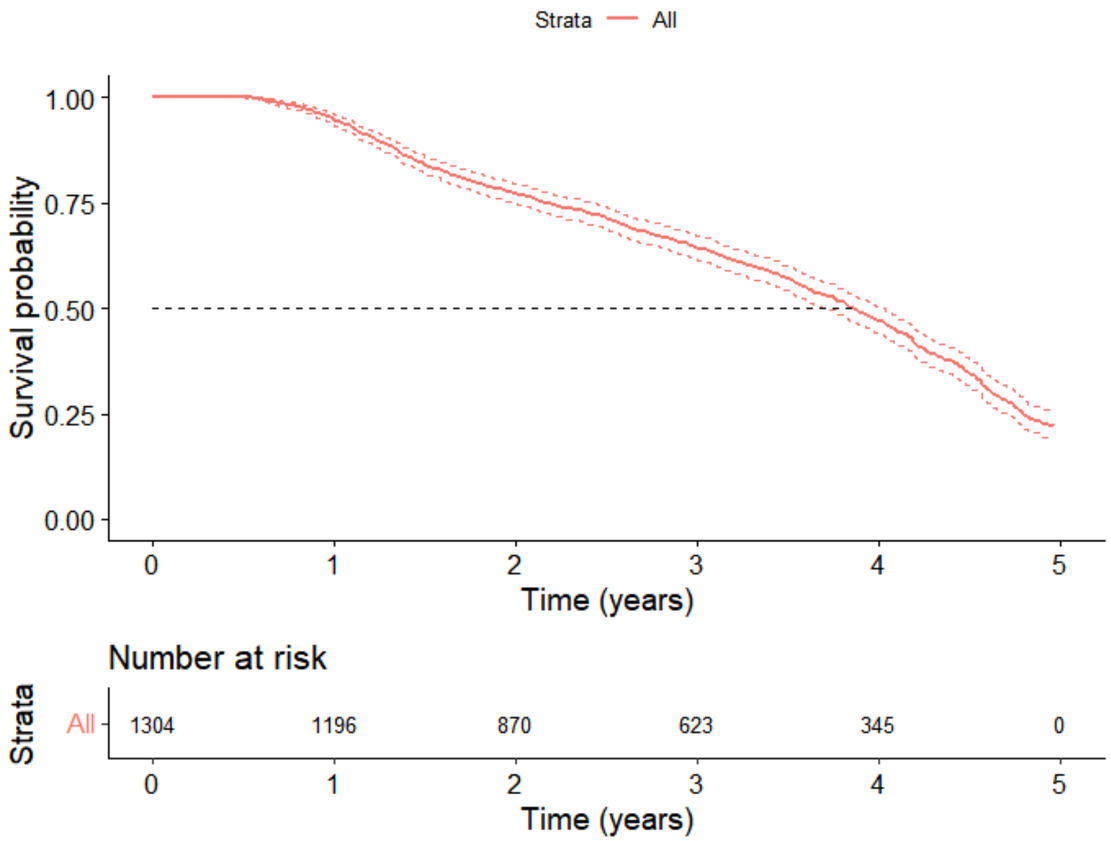
7. Stimare la sopravvivenza a 5 anni tramite lo stimatore di Kaplan-Meier per i casi di tumore al colon. Stimare approssimativamente la sopravvivenza mediana.

Nell'analisi sono inclusi **1304** soggetti.

Nel periodo di interesse sono morti **707** pazienti.

Si stima la sopravvivenza di **Kaplan-Meier**, che viene rappresentata graficamente.

La sopravvivenza mediana, come si osserva nel grafico, è poco inferiore a 4 anni.



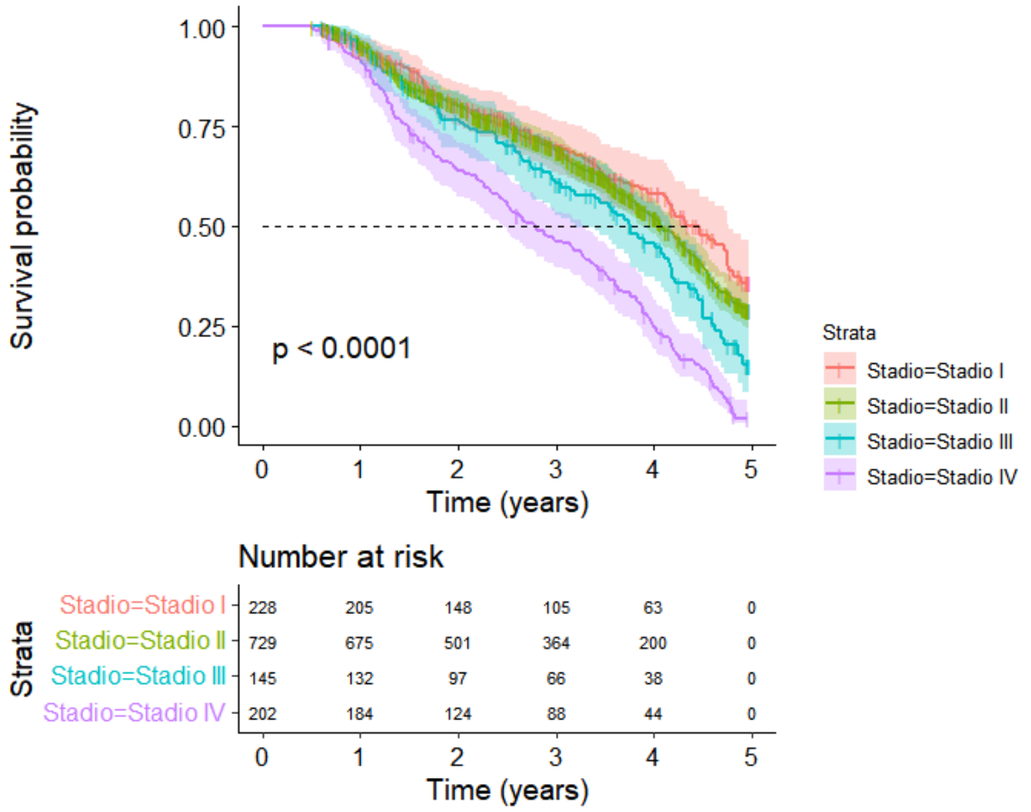
8. Stimare la sopravvivenza nei primi 5 anni dalla diagnosi per *stadio* ed effettuare un test d'ipotesi per verificare se l'azzardo di morte sia diverso per stadio di malattia alla diagnosi.

Si stima la sopravvivenza di **Kaplan-Meier** per stadio, che viene rappresentata graficamente.

Dal grafico risulta piuttosto evidente una differenza sul trend della sopravvivenza per stadio.

Si esegue un **test di ipotesi log-rank** per valutare se la differenza è significativa.

Il test ha un p-value inferiore a 2e-16, dunque si rigetta l'ipotesi nulla: **la differenza tra gli azzardi è statisticamente significativa**.



9. Applicare un modello per valutare l’associazione tra sesso e mortalità e interpretare la misura di effetto stimata.

Modello scelto:
Modello di regressione logistica

La misura di effetto ricavata dal modello è l’ODDS Ratio, pari a **0.65**, prendendo come riferimento il livello degli uomini.
Ciò significa che le donne hanno un valore di ODDS di morte pari al 65% rispetto a quello degli uomini.

10. Quali variabili sono associate alla mortalità? Riportare le relative stime di effetto con gli intervalli di confidenza.

Modello scelto:
Regressione logistica

Variabili analizzate:
sex, stadio, geneticm, smoke, married, kids, work, education, age.

Variabili significative:
sex, stadio, geneticm, smoke, education, age.

Variabile	ODDS Ratio	IC: 2.5	IC: 97.5
sexFemale	0.586	0.448	0.765
StadioStadio II	1.732	1.207	2.486
StadioStadio III	3.085	1.881	5.058
StadioStadio IV	28.426	15.597	51.808
geneticm1	11.932	6.705	21.231
smokeyes	1.653	1.186	2.304
educationmedium/high	2.142	1.339	3.425
age	1.109	1.092	1.127

Le variabili hanno una associazione significativa con la mortalità con un **OR** significativamente **diverso da 1**.

11. Valutare la presenza di confondenti e/o modificatori di effetto tra le variabili disponibili nel German health register e nel registro tumori nella valutazione dell'associazione tra sesso e mortalità. Se si identifica un'interazione tra sesso e un'altra variabile, riportare le stime di effetto per maschi e femmine separatamente e commentare il tipo di interazione trovato.

Tramite **stratificazione e modelli di regressione logistica** si studia la presenza di confondenti e modificatori di effetto per l'associazione tra mortalità e sesso.

Nessuna delle variabili testate risulta confondente.

Le variabili ***geneticm***, ***education*** e ***stadio*** risultano modificatori di effetto. Studiando separatamente gli effetti per sesso, risulta che, per tutti e 3 i modificatori, **le donne subiscono un'interazione positiva rispetto agli uomini.**

12. Scegliere un modello finale per valutare i fattori di rischio della mortalità dopo diagnosi di tumore al colon e commentare i risultati

Si utilizza il **modello di Cox** per studiare il rischio di mortalità tenendo conto del tempo di osservazione.

Modello 1

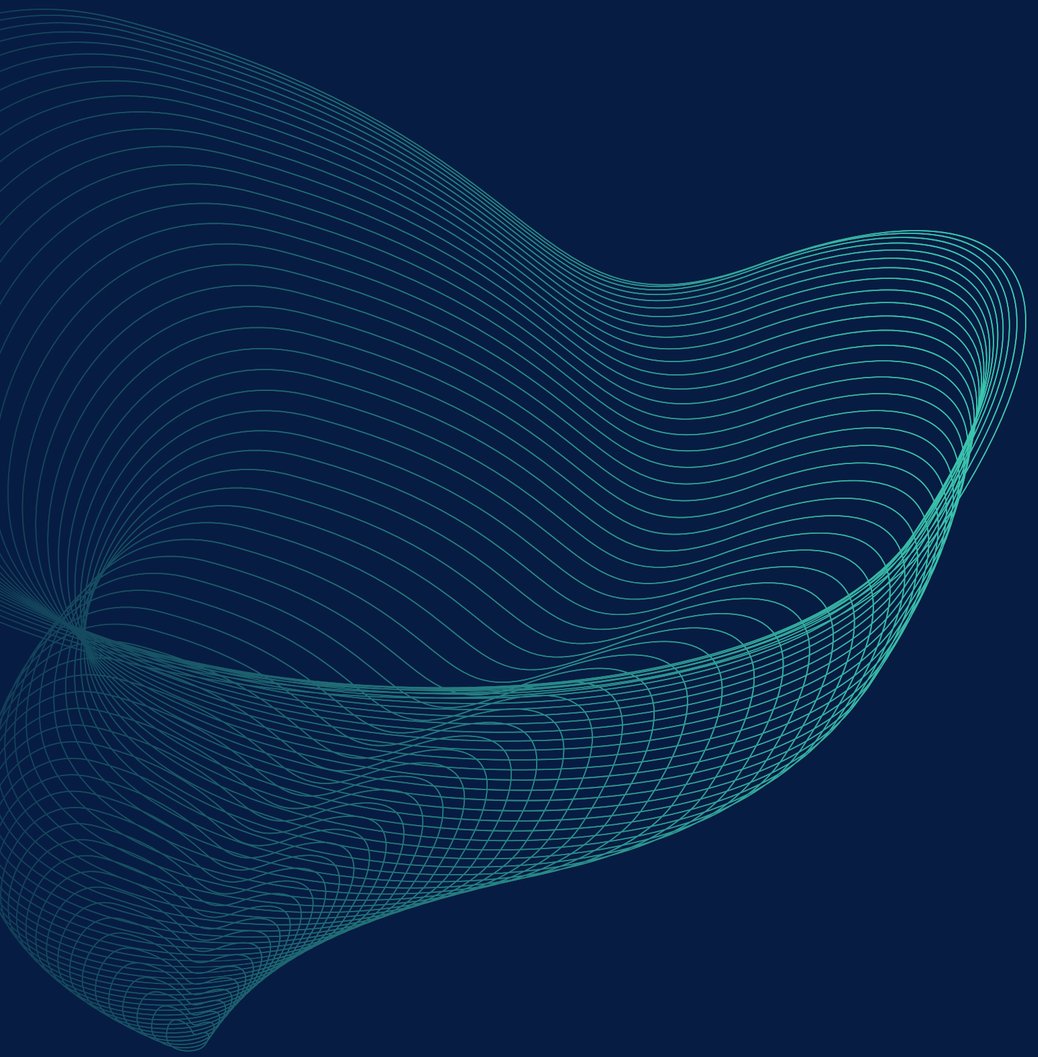
Variabili analizzate: variabili risultate significative al punto 10
Variabili significative: *sex*, *geneticm*, *stadio III*, *stadio IV*, *age*
Variabili non significative: *education*, *smoke*

Modello 2

Variabili analizzate: variabili risultate significative al punto 10 + interazioni trovate al punto 11
Variabili significative: *sex*, *geneticm*, *stadio IV*, *age*, interazione tra *sex* e *stadio II*, interazione tra *sex* e *stadio III*
Variabili non significative: *education*, *smoke*

In entrambi i modelli, l'assunto di costanza dell'Hazard Ratio è verificato e i due modelli risultano significativamente diversi.

Variabile	Hazard Ratio M1	Hazard Ratio M2
sexFemale	0.795 (0.686 - 0.922)	0.429 (0.277 - 0.665)
geneticm1	1.918 (1.561 - 2.358)	1.553 (1.164 - 2.072)
StadioStadio II	1.246 (0.990 - 1.569)	0.946 (0.706 - 1.267)
StadioStadio III	1.751 (1.301 - 2.356)	1.289 (0.872 - 1.906)
StadioStadio IV	2.630 (2.044 - 3.384)	2.267 (1.640 - 3.133)
age	1.031 (1.025 - 1.037)	1.031 (1.024 - 1.037)
sexFemale:StadioStadio II	/	1.899 (1.172 - 3.077)
sexFemale:StadioStadio III	/	2.040 (1.111 - 3.746)



Grazie per l'attenzione

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA