Utilizzo dei dati previdenziali per l'analisi dei rischi. Modelli epidemiologici, risultati ed indicatori di sintesi.

Annibale Biggeri, Università di Padova

Seminario satellite 49° Convegno AIE, 8 aprile 2025

schema

Archivio dati INPS-INAIL (storie contributive) (pres. D’Elia)

Disegno BRIC BEST Analisi esplorativa Machine Learning

Modelli epidemiologici (caso-controllo CMOR)

Metodi statistici (Large Table)

Q-Q Plot e outliers

Eterogeneità (smoothing e ranghi)

Selezione (qvalue)

Casi Attribuibili

E-value

Gli archivi contributivi dell'Inps

I dati sulla storia lavorativa dei lavoratori delle imprese del settore privato, con almeno un dipendente, sono reperibili presso gli archivi informatizzati dell'Inps che riportano, a partire dal 1974, i contributi versati dai lavoratori dipendenti. L'acquisizione di tali dati è possibile grazie a una apposita convenzione siglata da Inail e Inps finalizzata alla realizzazione del Registro nazionale dei tumori professionali.

L'archivio Inps non contiene i lavoratori del settore pubblico, ovvero le amministrazioni centrali (magistratura, autorità indipendenti, agenzie fiscali, ministeri, presidenza del consiglio, carriera diplomatica e prefettizia), le amministrazioni locali (regioni, province, comuni, aziende autonome e altre autonomie locali), i corpi di polizia, le forze armate e i vigili del fuoco, la scuola, sanità, università ed enti di ricerca pubblici, gli enti pubblici non economici, gli enti che producono servizi di pubblica utilità, i lavoratori autonomi, gli artigiani, i lavoratori domestici, i parasubordinati e i lavoratori occasionali.

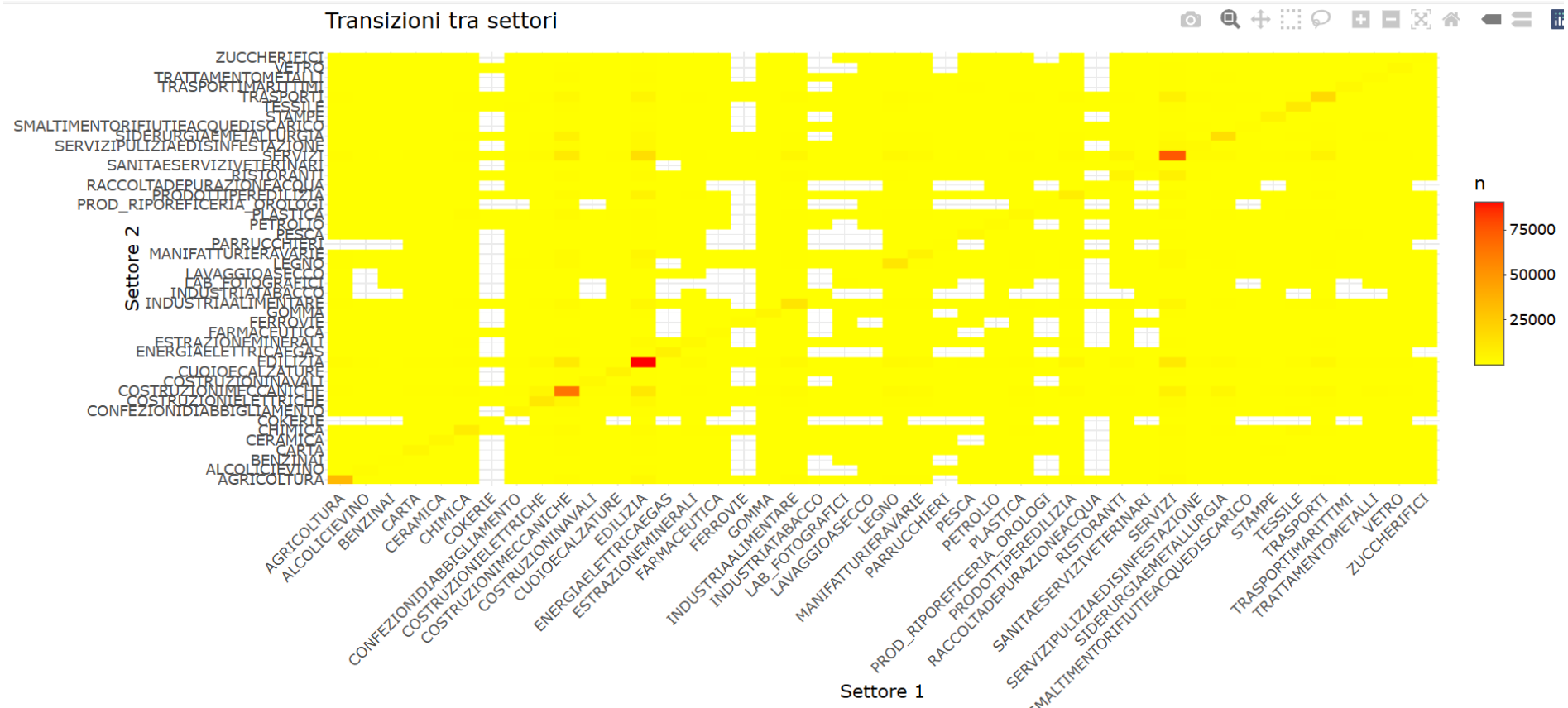
Per ciascun individuo, presente in Inps, è possibile ricostruirne l'intera storia lavorativa con le seguenti informazioni: il nome dell'azienda ove si è svolta l'attività, la sua matricola, la ragione sociale, l'indirizzo, la data d'inizio e di fine lavoro.

La qualifica del lavoratore è presente solo nella forma operaio ('Blue Collar') o impiegato ('White + Blue Collar').

Non è quindi possibile separare, per lo stesso settore, gli addetti con mansioni diverse e in particolare distinguere le figure amministrative da quelle addette alla produzione. Le aziende sono classificate secondo il ramo di attività economica mediante un codice interno Inps chiamato codice statistico contributivo (CSC) che può essere convertito nella classificazione Istat Ateco 81 mediante apposite tabelle.

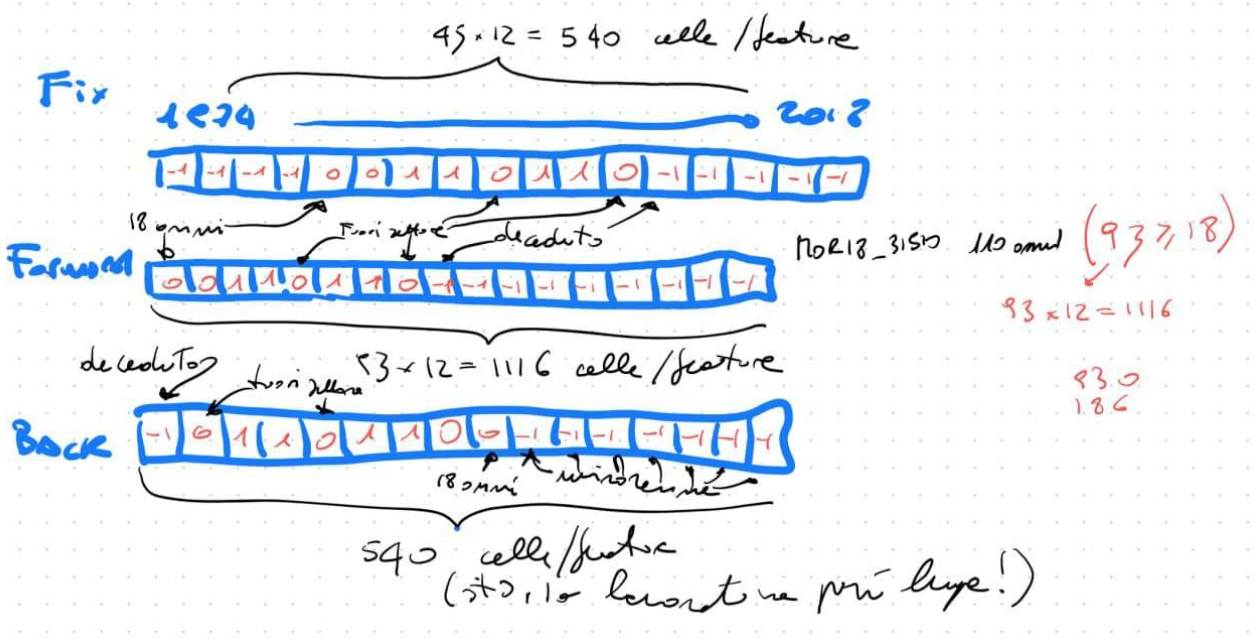
I nostri progetti si basano sul record linkage tra archivio delle storie contributive INPS-INAIL 1974-2018 e l'archivio delle cause di morte ISTAT 2005-2018.

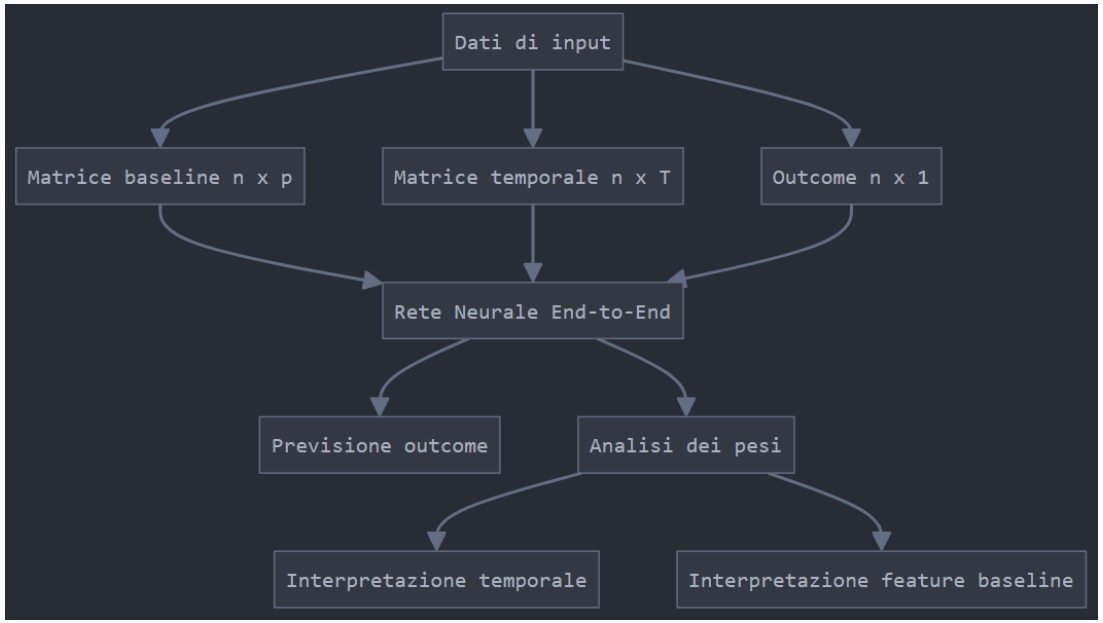
Restringendo l'attenzione alle sole morti per tumore (gli stimatori usati sono il PCMR e il CMOR), il dataset consiste in 689365 decessi e 3917928 periodi di contribuzione. (vedi presentazione D'Elia)



Caratterizzazione dell’esposizione (comparto)

Un individuo viene considerato ‘potenzialmente esposto’ a un comparto produttivo considerando l’attività svolta per il periodo di tempo più lungo nell’ambito dello stesso comparto produttivo, con un minimo di permanenza totale nel comparto di 12 mesi. I periodi si sommano anche se la persona ha lavorato in aziende diverse purché facenti parte dello stesso comparto produttivo. Le esposizioni in altri comparti oltre a quello di maggior durata non vengono considerate nelle analisi, anche se per ogni individuo sono sempre disponibili le storie lavorative complete.





Modelli epidemiologici

Disegno dello studio:

Caso-controllo con soggetti deceduti (vedi progetto OCCAM)

Stimatore:

Odds Ratio di mortalità per tumore

Modello statistico:

Regressione logistica multipla

Using Dead Controls to Adjust for Confounders in Case-Control Studies

Get access >

Geoffrey R. Howe

American Journal of Epidemiology, Volume 134, Issue 7, 1 October 1991, Pages 689–690, https://doi.org/10.1093/oxfordjournals.aje.a116144

Published: 01 October 1991 **Article history** ▾

Cite Permissions Share ▾

Abstract

The use of dead controls in a case-control study, the primary purpose of which is to control for confounding, leads to consistent relative risk estimates within stratum of the confounder, even if the causes of death of controls are associated with that confounder. Dead controls may be more comparable to dead cases in terms of data quality, and therefore, under these circumstances, the use of dead controls may be advantageous

Cause di morte e comparti

icd_chr	icd
Altri tumori maligni	C00-C97
Tumore Faringe	C09-C14 escluso C11
Tumore Rinofaringe	C11
Tumore Stomaco	C16
Tumore Colon-retto	C18-C21
Tumore Fegato	C22
Tumore Naso e seni paranasali	C300, C31
Tumore Laringe	C32
Tumore Polmone	C34
Tumore Pleura,peritoneo,pericardio*	C380,C384,C481,C482,C637
Tumore Melanoma della cute	C43
Tumore Pelle	C44
Mesotelioma*	C45
Tumore Connettivo e tessuti molli	C49
Tumore Mammella	C50
Tumore Ovaio	C56, C570-C574
Tumore Rene	C64-C66,C68
Tumore Vescica	C67,D090,D303,D414
Tumore Melanoma oculare	C699
Tumore Encefalo e Sist. nervoso centrale	C71-C72
Tumore Tessuto linfatico, emopoietico	C81-C86,C90-C95
Altri tumori	C99

Comparti
[1] "EDILIZIA"
[2] "CERAMICA"
[3] "AGRICOLTURA"
[4] "COMMERCIO"
[5] "ALTRE ATTIVITA' DI SERVIZI"
[6] "TESSILE"
[7] "LEGNO"
[8] "PRODOTTI PER EDILIZIA"
[9] "CUOIO E CALZATURE"
[10] "COSTRUZIONI MECCANICHE"
[11] "ESTRAZIONE MINERALI"
[12] "ALCOLICI E VINO"
[13] "RISTORANTI"
[14] "TRASPORTI"
[15] "SIDERURGIA E METALLURGIA"
[16] "COSTRUZIONI NAVALI"
[17] "COSTRUZIONI ELETTRICHE"
[18] "CARTA"
[19] "SANITA E SERVIZI VETERINARI"
[20] "VETRO"
[21] "MANIFATTURIERA VARIE"
[22] "INDUSTRIA ALIMENTARE"
[23] "FARMACEUTICA"
[24] "CHIMICA"
[25] "PLASTICA"
[26] "SERVIZI PULIZIA E DISINFESTAZIONE"
[27] "ENERGIA ELETTRICA E GAS"
[28] "PETROLIO"
[29] "LAB_FOTOGRAFICI"
[30] "CONFEZIONI DI ABBIGLIAMENTO"
[31] "STAMPE"
[32] "PARRUCCHIERI"
[33] "ALBERGHI_CONS_BEVANDE"
[34] "BENZINAI"
[35] "GOMMA"
[36] "TRATTAMENTO METALLI"
[37] "SMALTIMENTO RIFIUTI E ACQUE DI SCARICO"
[38] "ATTIVITA CONNESSA TRASPORTI"
[39] "PROD_RIP OREFICERIA_OROLOGI"
[40] "LAVAGGIO A SECCO"
[41] "ZUCCHERIFICI"
[42] "FERROVIE"
[43] "COKERIE"
[44] "INDUSTRIA TABACCO"
[45] "RACCOLTA DEPURAZIONE ACQUA"
[46] "TRASPORTI MARITTIMI"
[47] "RIPARAZIONI NON ALTROVE CLASSIFICATE"

Si utilizza il criterio della latenza minima:
definiamo un “numero di anni” 5 da scartare a ciascun soggetto (latenza minima), uguale per tutte le sedi tumorali da analizzare (dalla data di morte)

L’analisi tradizionale prevede il criterio OCCAM di definizione del comparto rilevante

Confondenti

Regione, anno decesso, titolo di studio, età

Titolo studio così categorizzato: (a) Low (4+5) middle (3) high (1+2)

dove 1 Laurea; 2 Diploma universitario o laurea breve; 3 Diploma di scuola media superiore; 4 Licenza di scuola media inferiore; 5 Licenza elementare o nessun titolo ; blank Non indicato

Età in classi così suddivise:

20-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79 80-84 85+

Comparto di riferimento: servizi cosi indentificati: "600" "660" "675" "770" "800" "900”

in alternativa

Comparando il comparto j-esimo a tutti gli altri comparti (ad eccezione del j-esimo)

Restrizione Maschi e Blue collar come analisi aggiuntive

Metodi statistici

Questo studio è un esempio di sorveglianza epidemiologica sui rischi occupazionali. Produce una grande tavola di rischi relativi di tumore (22) per comparto (47).

Abbiamo perciò un problema di molteplicità e **selezione**, ad esempio su due test ed un livello di significatività α

$$Prob(1 \cup 2) = 1 - Prob(\bar{1} \cap \bar{2}) = 1 - (1 - \alpha)^2$$

se α=0.05 allora la probabilità che almeno uno dei test risulti significativo sarà 1-(0.95)^2=0.0975 (da cui la correzione di Bonferroni).

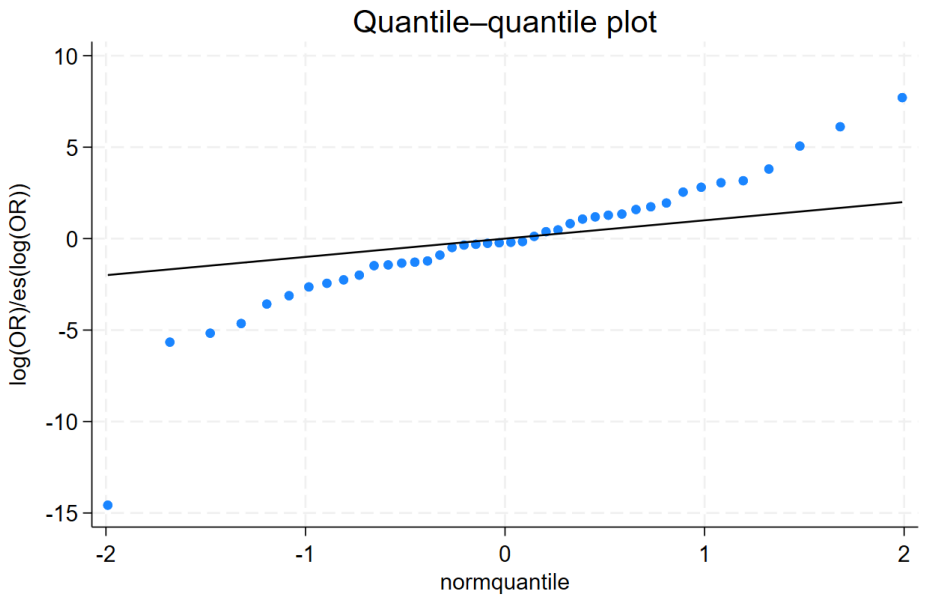
Gli intervalli di confidenza non avranno la copertura richiesta ($\alpha' = R \times \alpha/m$)

(vedi presentazione Stoppa)

Q-Q plot

Una soluzione semplice alla molteplicità è il diagramma quantile-quantile.

L'utilità di questa tecnica esplorativa è data dalla sua capacità di evidenziare valori aberranti, estremi rispetto al corpo delle osservazioni.



$$\Pr(Z_n \le z) = \Phi(z)^n$$

$$1-\alpha = \Phi(z_\alpha)^n$$

$$z_\alpha = \Phi^{-1}\left[(1-\alpha)^{1/n}\right]$$

Si consideri il valore più alto su *n* estrazioni da N(0,1)

Possiamo porre una soglia a $z_\alpha \tilde{\tau}$ ma questo apre il problema di valutare il grado di *eterogeneità*

In realtà i Q-Q plot identificano degli outlier, cioè delle osservazioni incoerenti con il modello che è stato specificato.

Volendo estendere questo approccio dovremo basarci sulle distribuzioni predittive a priori o sulle predittive a posteriori con cross-validazione (tipo *leave-one out*)

- The cross-validation posterior predictive distribution $P(Y_i^{rep}|\boldsymbol{Y}_{-i})$ for a replicate (Y_i^{rep}) of the i-th observation conditional to the remaining data $\boldsymbol{Y}_{-i}$ is frequently used to evaluate **goodness of fit**
- The extremeness is usually measured by some summaries over $P(Y_i^{rep}|\boldsymbol{Y}_{-i})$:
  - **PIT** (probability integral transform, Dawid 1984)

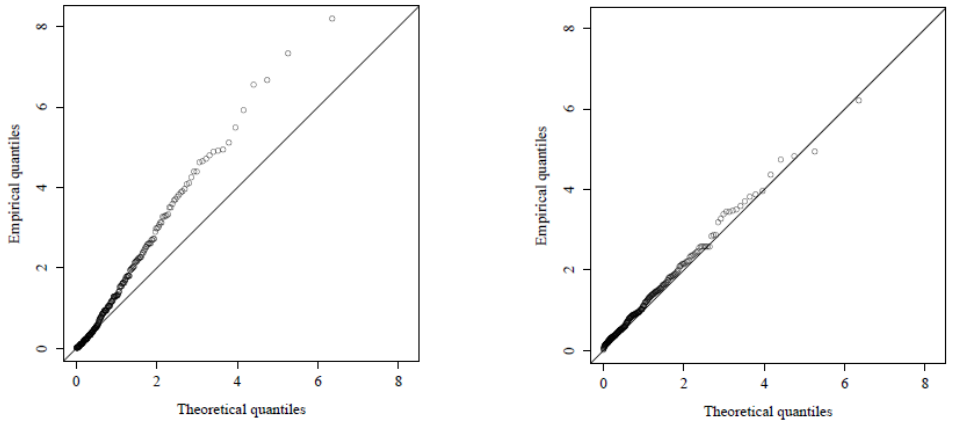
$$P(Y_i^{rep} \leq y_i|\boldsymbol{Y}_{-i})$$

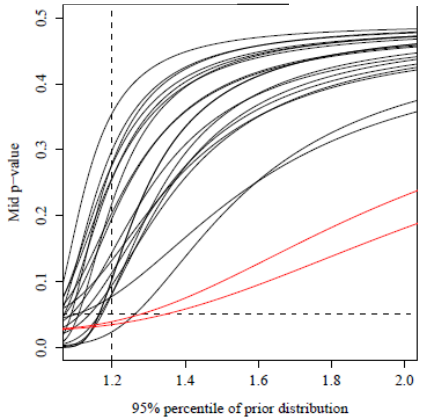
  - **CPO** (conditional predictive ordinate, Geisser 1980)

$$P(Y_i^{rep} = y_i|\boldsymbol{Y}_{-i})$$

$$\text{mid } p\text{-value}_i = P(Y_i^{rep} > y_i|\boldsymbol{Y}_{-i}) + \frac{1}{2}P(Y_i^{rep} = y_i|\boldsymbol{Y}_{-i})$$

$$1 - \textbf{PIT}_i + \textbf{CPO}_i/2$$





Eterogeneità

Analogamente alla metanalisi e al Disease Mapping la presenza di *eterogeneità* implica il riconoscimento di una struttura gerarchica e la necessità di filtrare la variabilità campionaria ottenendo stime *lisciate.*

Siano $\widehat{\beta_i}\widehat{\sigma_i}$ i log(MOR) e relativi errori standard, τ^2 l'eterogeneità tra comparti

Modello a due livelli

$$\hat{\beta}_j \left| \beta_j, \hat{\sigma}_j^2 \sim N\left(\beta_j, \hat{\sigma}_j^2\right) \right.$$

$$\beta_j \left| \beta, \tau^2 \sim N\left(\beta, \tau^2\right) \right.$$

Dist. a posteriori

$p\left(\beta \left| \hat{\beta}_1,\ldots,\hat{\beta}_J, \hat{\sigma}_1^2,\ldots,\hat{\sigma}_J^2 \right.\right)$ dell'effetto globale (non di interesse)

$p\left(\tau^2 \left| \hat{\beta}_1,\ldots,\hat{\beta}_J, \hat{\sigma}_1^2,\ldots,\hat{\sigma}_J^2 \right.\right)$ dell'eterogeneità

$p\left(\beta_h \left| \hat{\beta}_1,\ldots,\hat{\beta}_J, \hat{\sigma}_1^2,\ldots,\hat{\sigma}_J^2 \right.\right)$ MOR lisciati

Eterogeneità

Il problema però è multivariato e nella graduatoria occorre filtrare opportunamente la variabilità campionaria.

$Y_{ij} \sim N(\vartheta_{ij}\ E_{ij})$ i = 1,...,I comparto; j = 1,...,J malattie
ϑ_{ij} rischio relativo
E_{ij} errore standard di Y_{ij} − log MOR

$$\log(\boldsymbol{\vartheta}_i) \sim M\,VN(\underset{1\times J}{\boldsymbol{\mu}}, \underset{J\times J}{\sum})$$
$\mu_j \sim N\,(0,0.0001)$
$\sum \sim W\ ishart(\boldsymbol{\Omega}, \nu)$

Ranghi a posteriori dei rischi relativi.

Il vettore dei ranghi $\mathbf{R}=(R_1,\ldots,R_k,\ldots,R_K)$ è definite come l'insieme di ranghi

$$R_k = \text{rank}(\vartheta_k) = \sum_{k'=1}^{K} I_{(\vartheta_k \geq \vartheta_{k'})}$$

I(·) è la funzione indicatrice

ϑ_k il parametron di interesse (RR)

L'insieme di MOR per comparto e malattia genereno una matrice $\mathbf{P_{I\times J}}$ di ranghi le cui righe rappresentano i ***profili*** di malattia per comparto cioè I vettori $P_{i.}=(P_{i1},\ldots,P_{ij},\ldots,P_{iJ})$ con

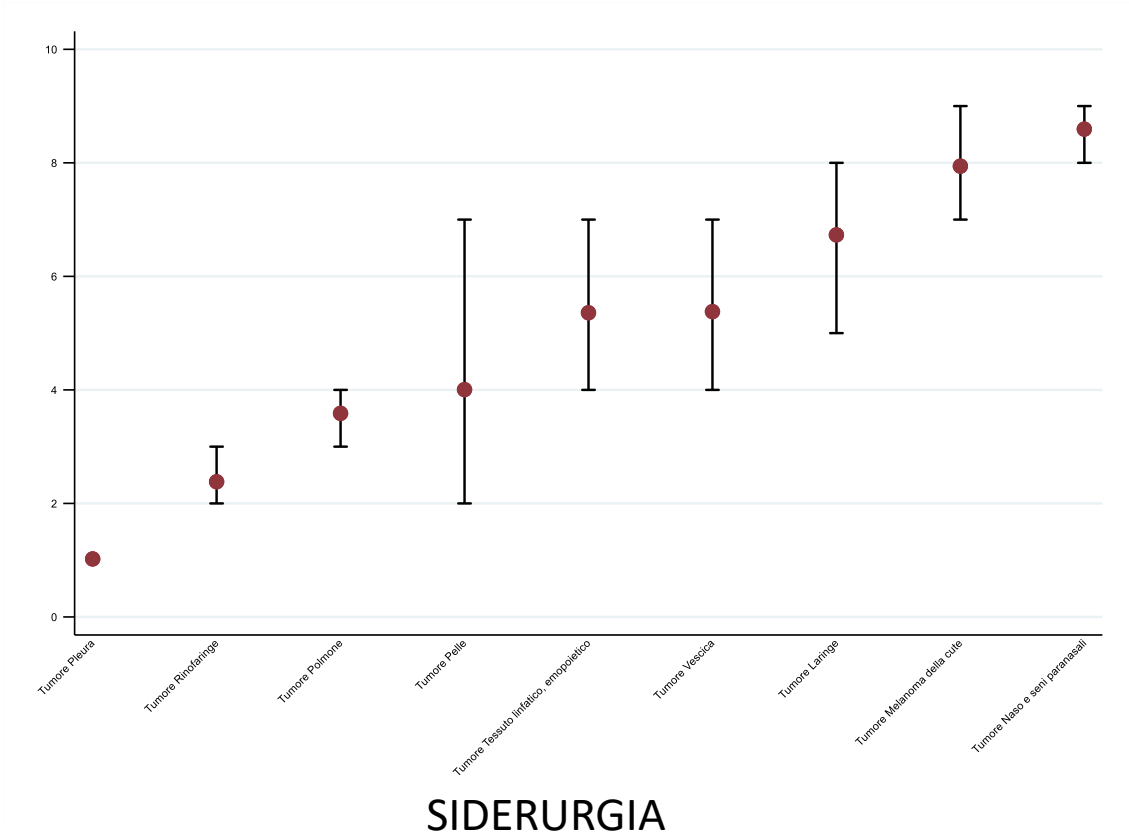
$$P_{ij} = \text{rank}(\vartheta_{ij}) = \sum_{j'=1}^{J} I_{(\vartheta_{ij} \geq \vartheta_{ij'})}$$

e la matrice $\mathbf{S_{I\times J}}$ dei ranghi le cui colonne rappresentano le ***firme (signatures)*** dei comparti per malattia, cioè i vettori $S_{.j}=(S_{1j},\ldots,S_{ij},\ldots,S_{Ij})$ con

$$S_{ij} = \text{rank}(\vartheta_{ij}) = \sum_{i'=1}^{I} I_{(\vartheta_{ij} \geq \vartheta_{i'j})}$$

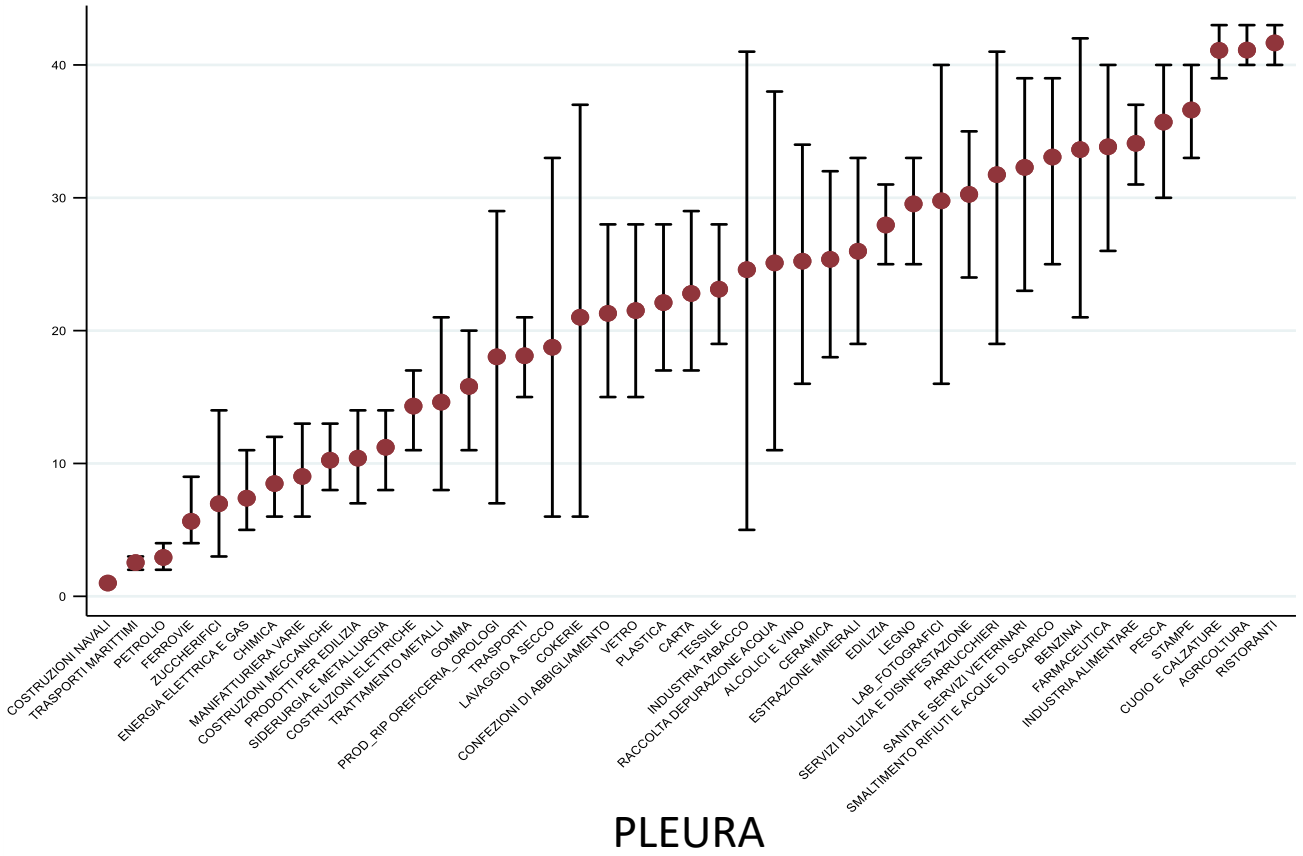
Si veda presentazione Stoppa

RANGHI delle malattie per comparto



SIDERURGIA

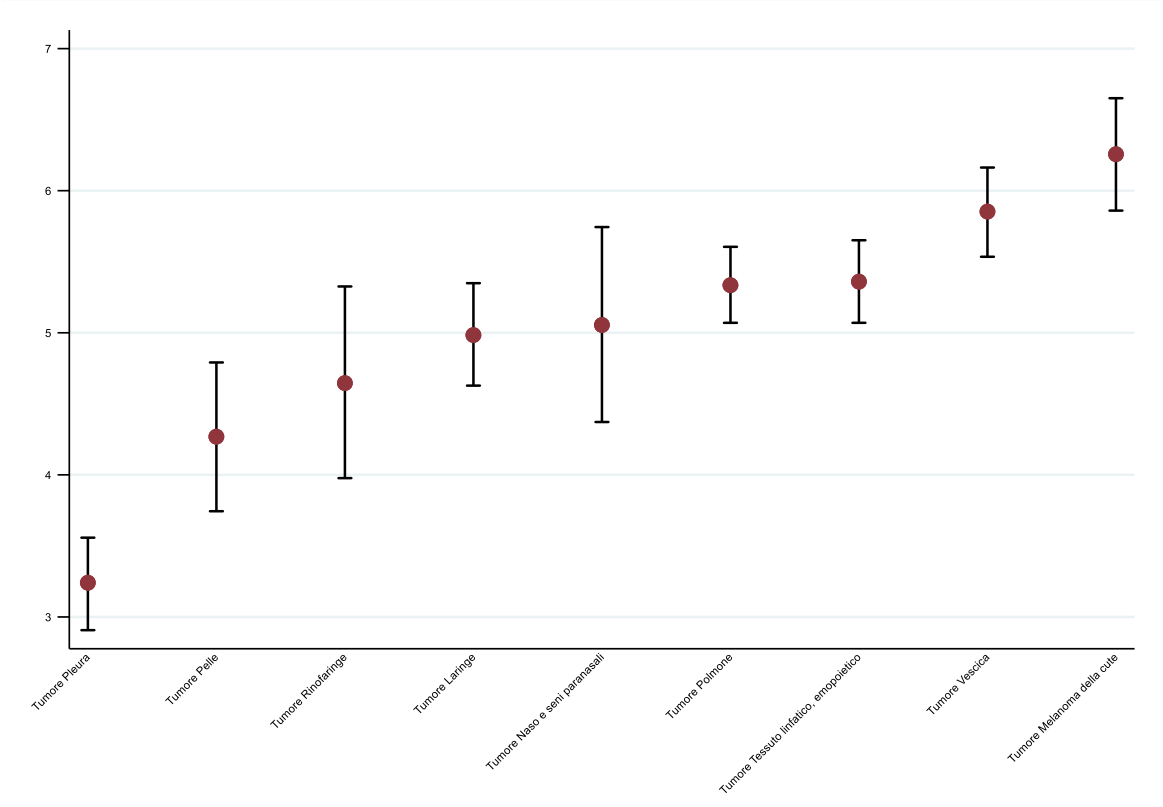
RANGHI dei comparti per malattia



PLEURA

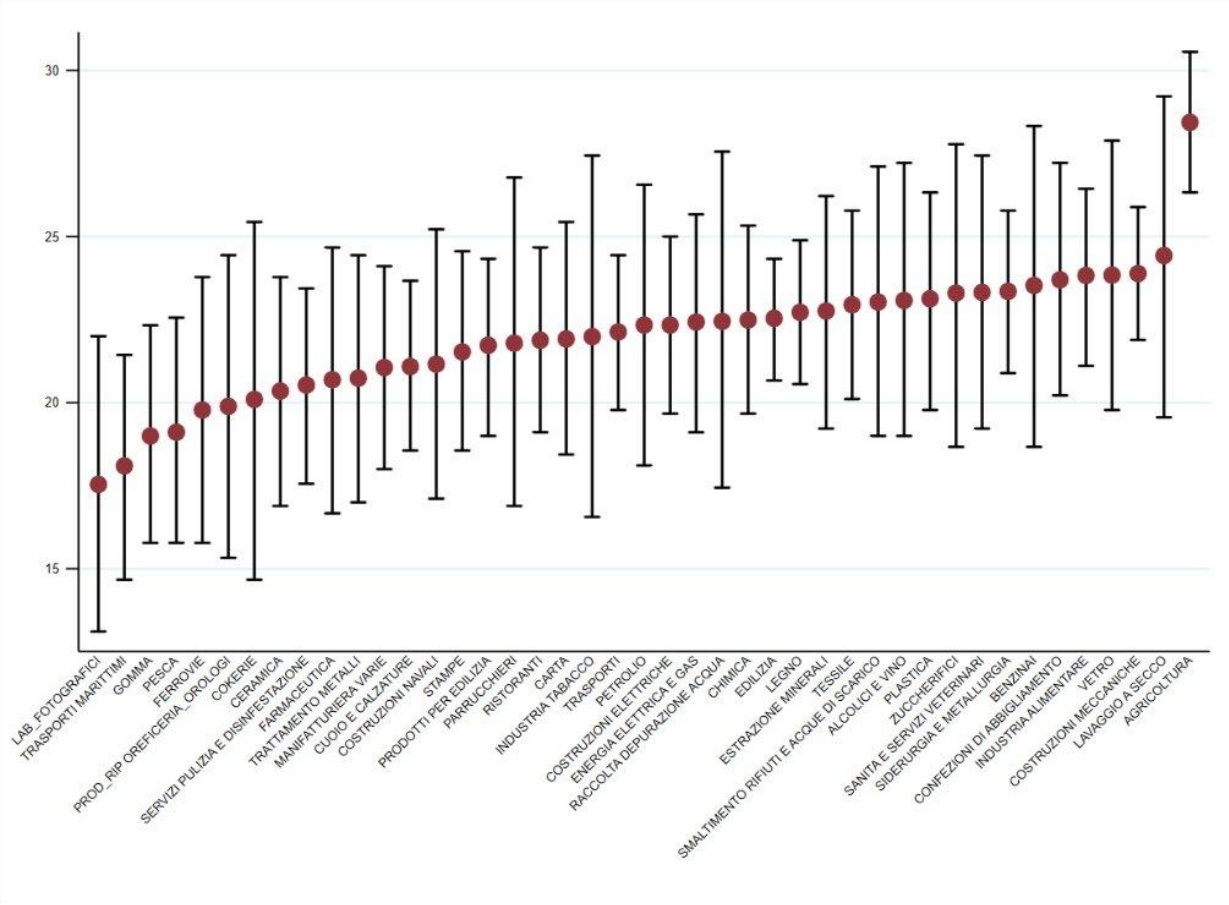
RANGHI MARGINALI

Esiste una malattia che «sta peggio» delle altre



RANGHI MARGINALI

Esiste un comparto che «sta peggio» degli altri



Selezione (molteplicità)

Nel momento in cui si opera una scelta dei comparti in condizioni di incertezza si rende necessario garantire una adeguata copertura agli intervalli di confidenza. La statistica usata è il *qvalue*:

$$Pr(H_0|T) = \frac{Pr(T|H_0)Pr(H_0)}{Pr(T)}$$

$$pvalue = \Pr(T|\, H_0)$$
$$qvalue = \Pr(H_0|\, T)$$

$$\frac{pvalue_{(i)}}{i/m} \rightarrow \Pr(\, H_0|T)$$

disease	pvalue	qvalue
Pneumoconiosis	**0**	**0**
Diabetes	**0.00166**	**0.0151**
Prostate cancer	**0.00202**	**0.0151**
Lung cancer	**0.01180**	**0.0661**
Respiratory acute	**0.01505**	**0.0675**
CHD	**0.01848**	**0.0690**
Liver Cirrhosis	0.03451	0.1098
Ill defined conditions	0.03918	0.1098
Respiratory chronic	0.04689	0.1116
Melanoma	0.04980	0.1116
Larynx cancer	0.06508	0.1326
Myeloma	0.17530	0.3275

Risk Communication

È sempre preferibile evitare percentuali o comunque misure adimensionali.

In Epidemiologia clinica è popolare l'uso del NNT / NNH che nel nostro contesto si traduce in Casi Attribuibili

$$CA = 1/NNH = \Big(\big(PEER \times (OR - 1)\big) \times (1 - PEER)\Big) / \Big(\big(PEER \times (OR - 1)\big) + 1\Big)$$

dove PEER indica il tasso (per cento) nel gruppo di controllo e OR è maggiore di uno.

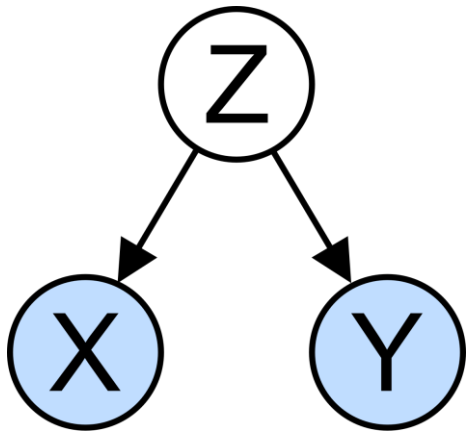
Gli intervalli di confidenza per i casi attribuibili hanno un limite inferiore che non può essere negativo. Inoltre ha senso chiedersi qual è la probabilità $H_0: CA = 0$ che può essere approssimata dal *qvalue* (dall'equivalenza con un modello bayesiano a tre livelli) a una coda.

		Ricoverati (2013 – 2017)			
	Patologia	OSS	**SHR** (IC90%) *		Casi attribuibili (prob casi=0)
UOMINI	Tutte le cause				
	Malattie cardiovascolari	2782	**108,5**	(105,2 – 112,0)	**218** (p<0.01)
	Malattie coronariche	895	**122,2**	(115,6 – 129,2)	**126** (p<0.01)
	Tumori	1477	**98,4**	(94,2 – 102,7)	**0** (p>0,99)
	Tumore del polmone	158	**104,6**	(91,3 – 119,3)	**6** (p=0,59)
	Malattie respiratorie	1373	**92,3**	(88,2 – 96,5)	**0** (p>0,99)
	Malattie respiratorie croniche	52	**65,8**	(51,6 – 82,8)	**0** (p>0,99)
	Malattie sistema digerente	2027	**99,1**	(95,5 – 102,8)	**0** (p>0,99)
	Cirrosi epatica	80	**139,3**	(114,7 – 167,6)	**22** (p<0,01)
	Malattie urinarie	677	**121,7**	(114,1 – 129,6)	**120** (p<0,01)
	Diabete mellito	50	**78,1**	(60,9 – 98,9)	**0** (p>0,99)

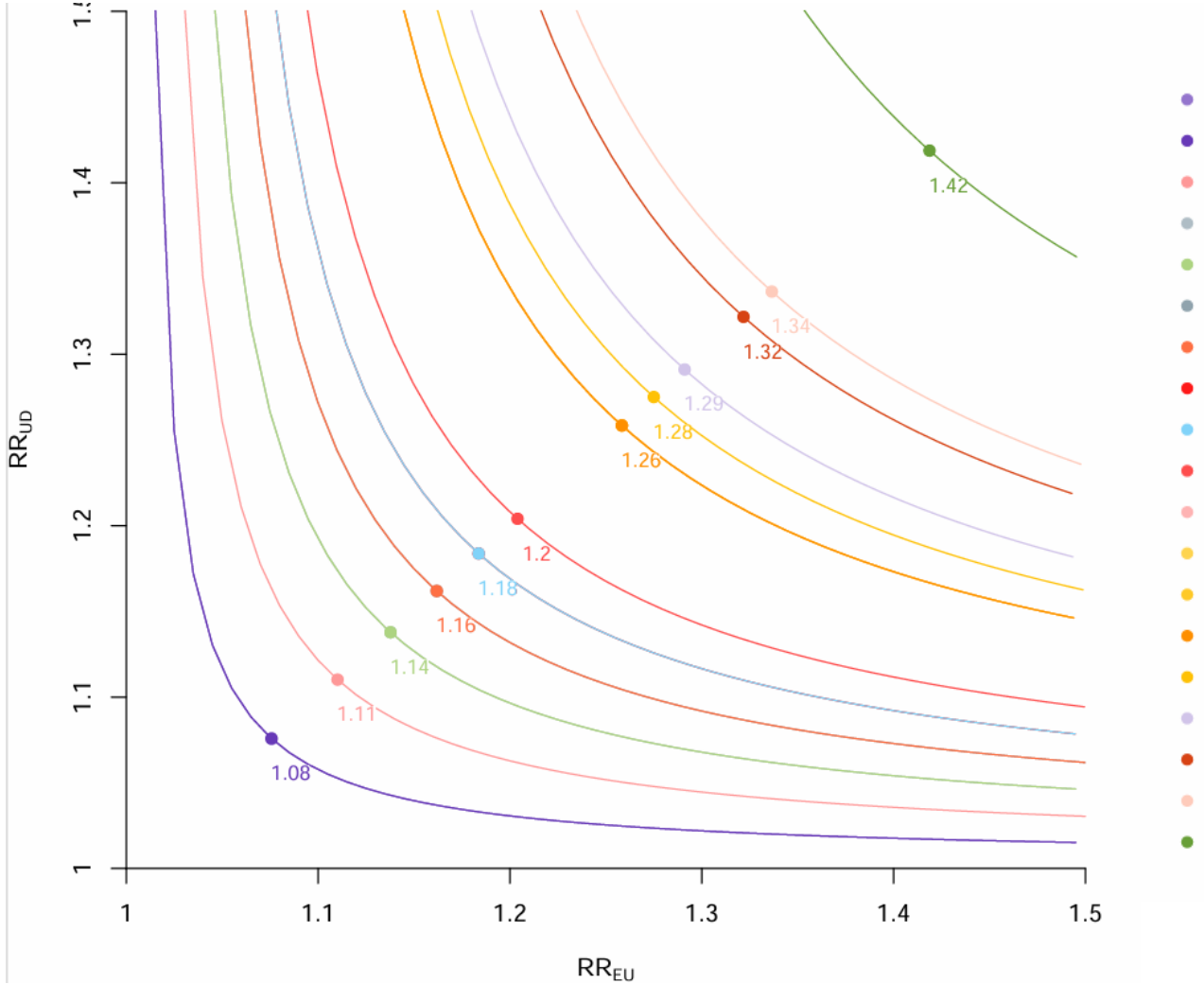
E value

(si veda presentazione Berti)

$$E - \text{value} = \text{RR} + \sqrt{\text{RR} \times (\text{RR} - 1)}$$



Il criterio causale di confondimento è detto *back-door criterion* e stabilisce nel diagramma direzionato aciclico un passaggio indiretto di informazione tra X (esposizione) e Y (malattia) tramite un progenitore di entrambi. E-value è il valore di RR (posto uguale XZ=YZ) che annulla il RR XY marginale (al lordo di Z).



VALORI per COMPARTO

- (OR: 1.01) (E-value: 1.08) CUOIO E CALZATURE
- (OR: 1.01) (E-value: 1.08) PLASTICA
- (OR: 1.02) (E-value: 1.11) PRODOTTI PER EDILIZIA
- (OR: 1.03) (E-value: 1.14) SIDERURGIA E METALLURGIA
- (OR: 1.03) (E-value: 1.14) INDUSTRIA ALIMENTARE
- (OR: 1.04) (E-value: 1.16) ESTRAZIONE MINERALI
- (OR: 1.04) (E-value: 1.16) BENZINAI
- (OR: 1.05) (E-value: 1.18) COSTRUZIONI NAVALI
- (OR: 1.05) (E-value: 1.18) PETROLIO
- (OR: 1.06) (E-value: 1.2) TRATTAMENTO METALLI
- (OR: 1.09) (E-value: 1.26) EDILIZIA
- (OR: 1.09) (E-value: 1.26) RISTORANTI
- (OR: 1.09) (E-value: 1.26) TRASPORTI
- (OR: 1.09) (E-value: 1.26) SERVIZI PULIZIA E DISINFESTAZIONE
- (OR: 1.1) (E-value: 1.28) TRASPORTI MARITTIMI
- (OR: 1.11) (E-value: 1.29) CERAMICA
- (OR: 1.13) (E-value: 1.32) SMALTIMENTO RIFIUTI E ACQUE DI SCARICO
- (OR: 1.14) (E-value: 1.34) PARRUCCHIERI
- (OR: 1.2) (E-value: 1.42) PESCA

Utilizzo dei dati previdenziali per l'analisi dei rischi. Modelli epidemiologici, risultati ed indicatori di sintesi.

Annibale Biggeri, Università di Padova

Seminario satellite 49° Convegno AIE, 8 aprile 2025