

XLIX CONGRESSO 2025

Oltre la Molteplicità: Strategie per la valutazione dei rischi occupazionali nella sorveglianza dei tumori professionali in Italia

Giorgia Stoppa, Annibale Biggeri, Claudio Gariazzo, Stefania Massari, Dolores Catelan

Unità di Biostatistica, Epidemiologia e Sanità Pubblica, Dipartimento di Scienze Cardio-Toraco-Vascolari e Sanità Pubblica, Università degli Studi di Padova
Dipartimento di medicina, epidemiologia, igiene del lavoro e ambientale, Istituto Nazionale Assicurazione Infortuni sul Lavoro (INAIL), Roma

BIG DATA E DEEP LEARNING
NELLA SORVEGLIANZA DEI
TUMORI PROFESSIONALI

Il Progetto BEST

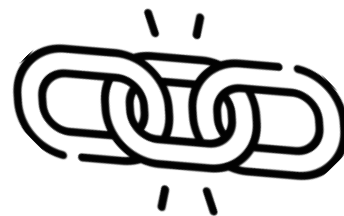


Il progetto **Big Data e Deep Learning nella sorveglianza dei tumori professionali (BEST)** ha come obiettivo lo sviluppo di un sistema di sorveglianza dei rischi occupazionali, sfruttando il record-linkage tra archivi amministrativi sanitari ISTAT e l'archivio INPS delle storie contributive.



Registro di causa di morte ISTAT 2005-2018

Identificati i decessi per causa tumorale, ricavata la sede specifica e mantenuti i record relativi ad undici organi predefiniti



Archivio INPS delle storie contributive 1974-2018

Individuato il comparto lavorativo prevalente per ciascun soggetto, considerando una latenza minima di cinque anni antecedenti al decesso.



Il Progetto BEST



Per ogni combinazione di sede tumorale e comparto, separatamente per sesso e per Blu o Blu & White collar, è stato calcolato il **Cancer Mortality Odds Ratio**.

I soggetti con sede tumorale diversa da quella di volta in volta esaminata sono stati utilizzati come controlli.

Variabili considerate per l'aggiustamento:

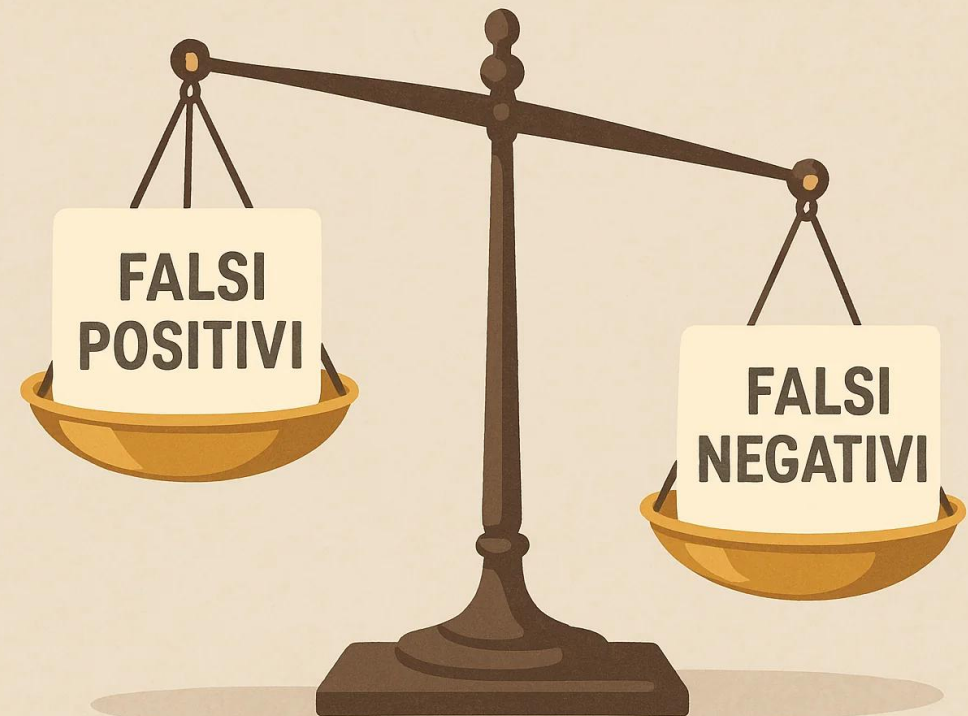
- **Età** in classi quinquennali
- **Titolo studio:** Low, Middle, High
- **Regione di residenza**
- **Anno del decesso**

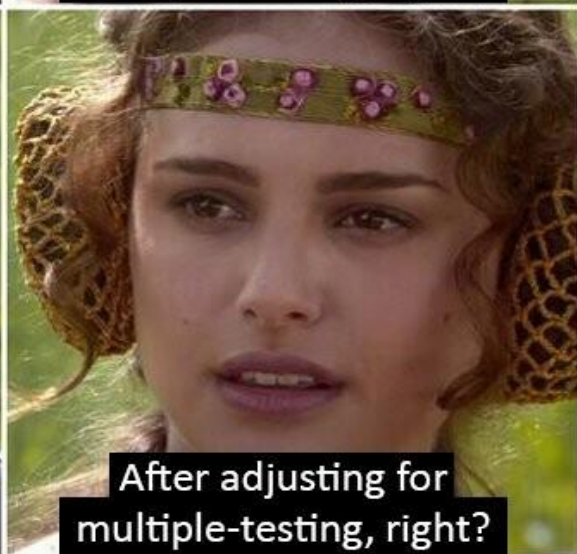
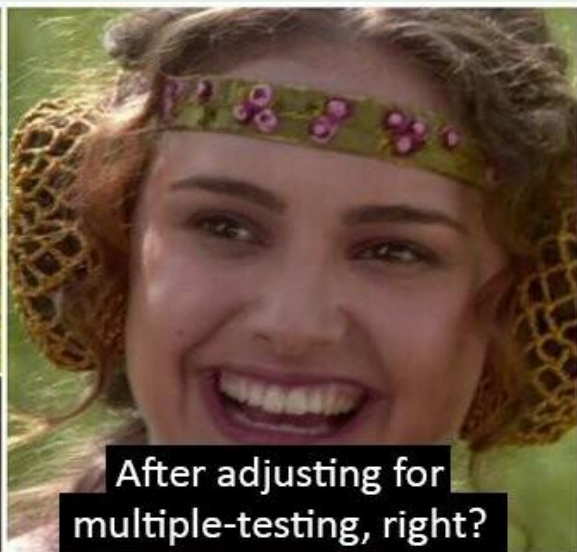


Obiettivo: è migliorare l'interpretazione dei rischi occupazionali attraverso l'applicazione di approcci statistici innovativi, come l'uso di modelli bayesiani gerarchici, per identificare le priorità sanitarie relative ai comparti lavorativi e alle malattie e gestire adeguatamente la molteplicità dei test statistici e il loro impatto sull'inferenza.

Razionale: molte stime, molte sfide

Il problema della molteplicità dei test è stato raramente affrontato nell'epidemiologia occupazionale, dove la priorità è spesso evitare falsi negativi.





Metodi

QQ-Plot con Guide Rails

I QQ-plot mostrano i quantili dei valori di $\log(\text{MOR})$ standardizzati rispetto al relativo errore standard, confrontati con i quantili attesi di una distribuzione normale standard.

Guide rails

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

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

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

evidenziano le aree in cui ci si aspetterebbe di trovare le osservazioni se fossero tratte da una distribuzione normale.

Positive False Discovery Rate (pFDR)

Il pFDR è calcolato a partire dal q-value:

$$\text{q-value} = \text{Prob}(H_0 \mid \text{Test} \geq \text{Test}_{\text{obs}})$$

che rappresenta il minimo FDR accettabile per rifiutare l'ipotesi nulla.

Gli intervalli di confidenza sono corretti per la selezione utilizzando la formula:

$$ss \pm z(1 - \alpha'/2) \times \text{se}(ss) \quad \text{dove } \alpha' = R \times \alpha / m$$

dove:

- R è il numero di test significativi
- m è il numero totale di test eseguiti

Graduatorie Basate sui Ranghi e loro incertezza

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

$$R_k = \text{rank}(\mathcal{G}_k) = \sum_{k'=1}^K I_{(\mathcal{G}_k \geq \mathcal{G}_{k'})}$$

$I(\cdot)$ è la funzione indicatrice

$\mathcal{G}_k$ il parametro di interesse (RR)

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

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

e la matrice $\mathbf{S}_{I \times J}$ dei ranghi le cui colonne rappresentano le **firme** (**signatures**) dei comparti per malattia, cioè i vettori

$\mathbf{S}_{\cdot j}=(S_{1j}, \dots, S_{ij}, \dots, S_{Ij})$ con

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

$$Y_{ij} \sim N(\vartheta_{ij} E_{ij})$$

$i = 1, \dots, I$ comparto; $j = 1, \dots, J$ malattie
 ϑ_{ij} rischio relativo

E_{ij} errore standard di $Y_{ij} - \log \text{MOR}$

$$\log(\boldsymbol{\vartheta}_i) \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$\begin{matrix} 1 \times J & J \times J \end{matrix}$

$$\mu_j \sim N(0, 0.0001)$$

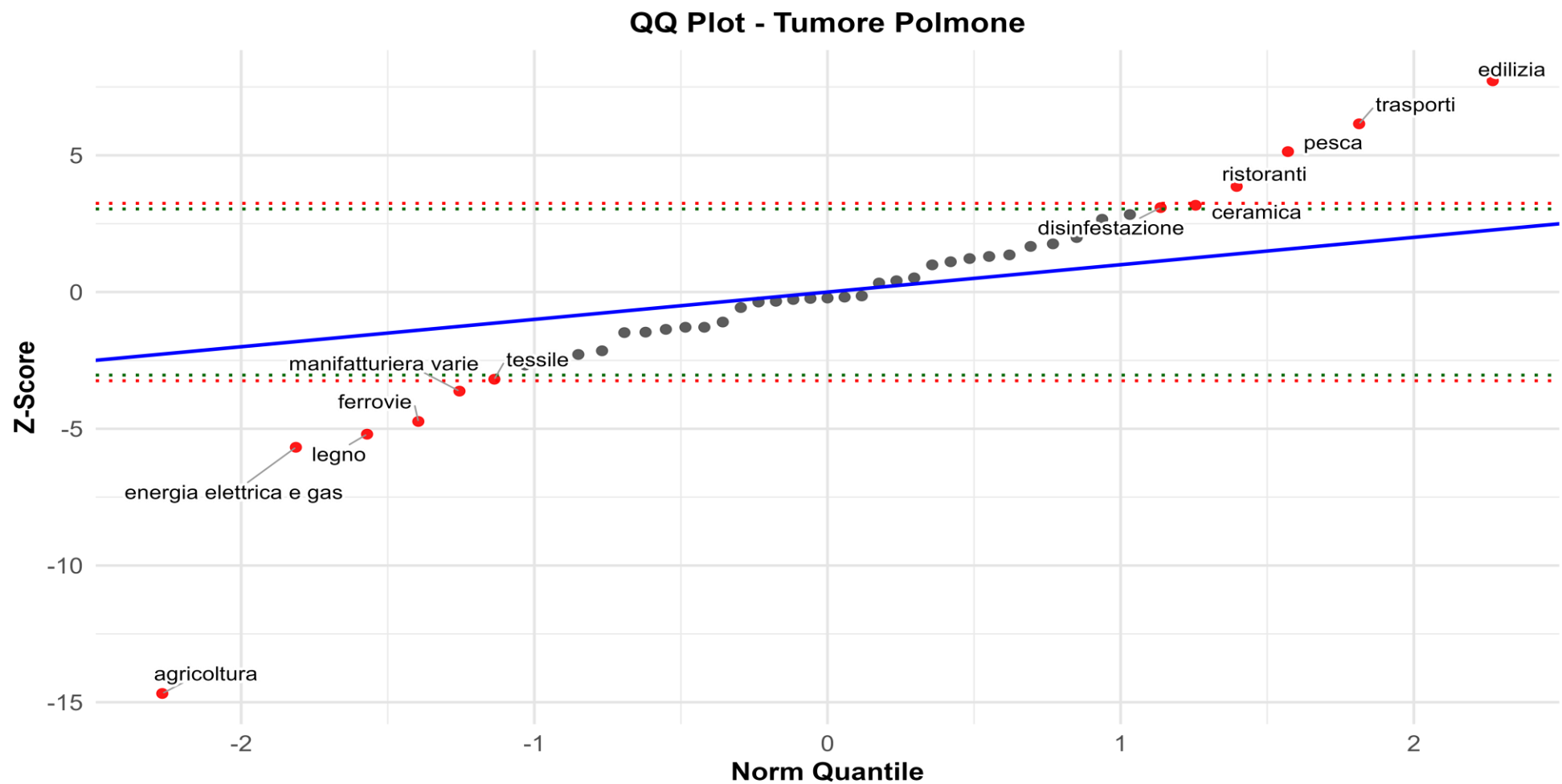
$$\boldsymbol{\Sigma} \sim W\text{ishart}(\boldsymbol{\Omega}, \nu)$$

Risultati

Comparto	Tumore al polmone				
	<u>Totale</u>	<u>Uomini</u>		<u>Donne</u>	
	n (%)	n (%)	CMOR (90% IC)	n (%)	CMOR (90% IC)
Tutti	215'576 (100)	183'968 (85.34)	-	31'608 (14.66)	-
Parrucchieri	597 (0.28)	266 (44.56)	1.14 (1.01-1.29)	331 (55.44)	1.26 (1.13-1.40)
Smaltimento rifiuti e acque di scarico	811 (0.38)	765 (94.33)	1.13 (1.05-1.21)	46 (5.67)	1.22 (0.93-1.6)
Trasporti marittimi	1'283 (0.60)	1'266 (98.67)	1.10 (1.04-1.17)	17 (1.33)	1.09 (0.71-1.69)
Pulizia e disinfestazione	4'402 (2.04)	2'086 (47.39)	1.09 (1.04-1.14)	2'316 (52.61)	1.12 (1.07-1.17)
Trasporti	9'859 (4.57)	9'621 (97.59)	1.09 (1.06-1.11)	238 (2.41)	1.15 (1.02-1.30)
Ristoranti	5'285 (2.45)	3'262 (61.72)	1.09 (1.05-1.13)	2'023 (38.28)	1.16 (1.11-1.22)
Edilizia	39'199 (18.18)	38'764 (98.89)	1.09 (1.07-1.11)	435 (1.11)	1.02 (0.92-1.12)



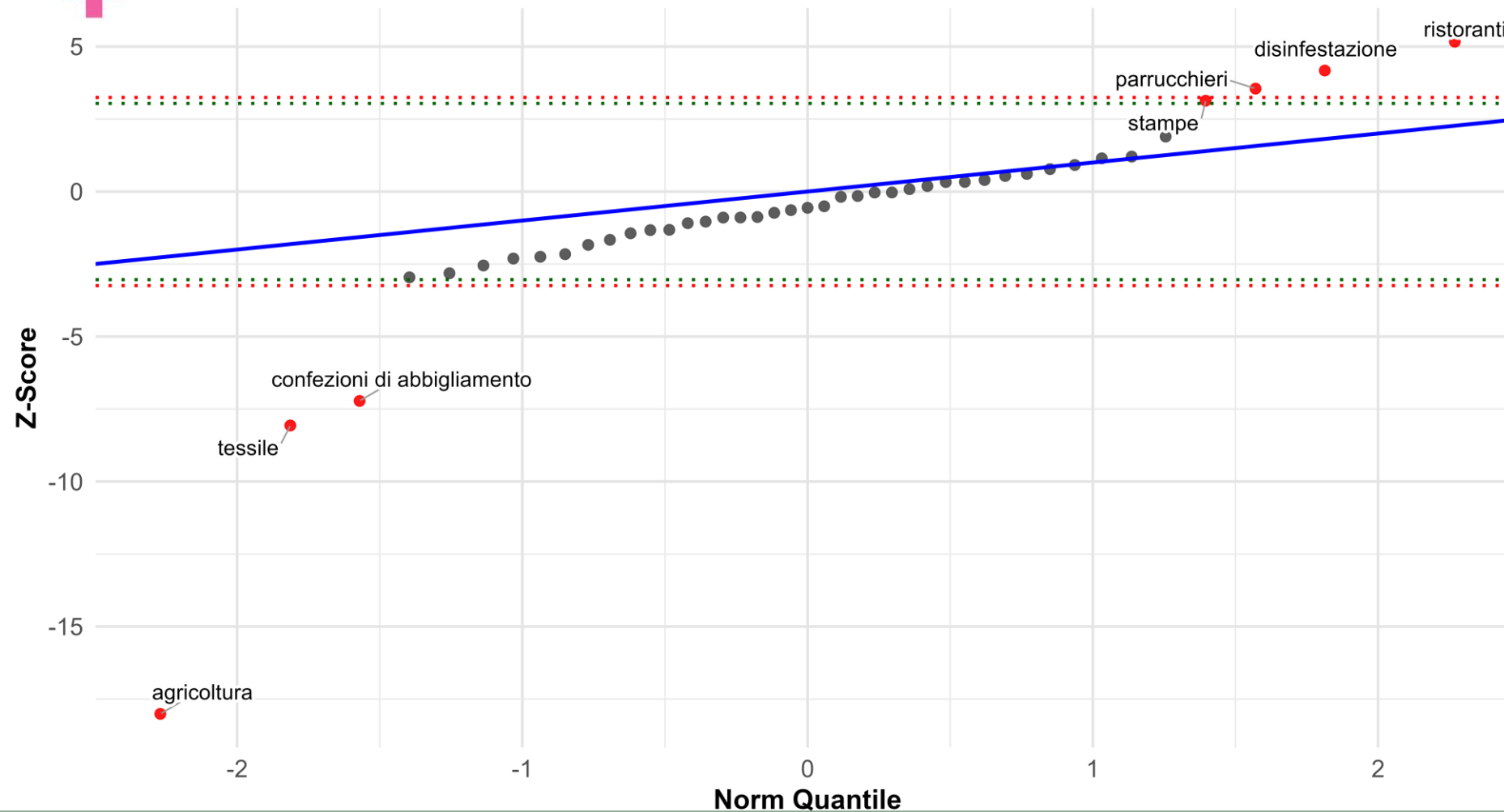
UOMINI



DONNE



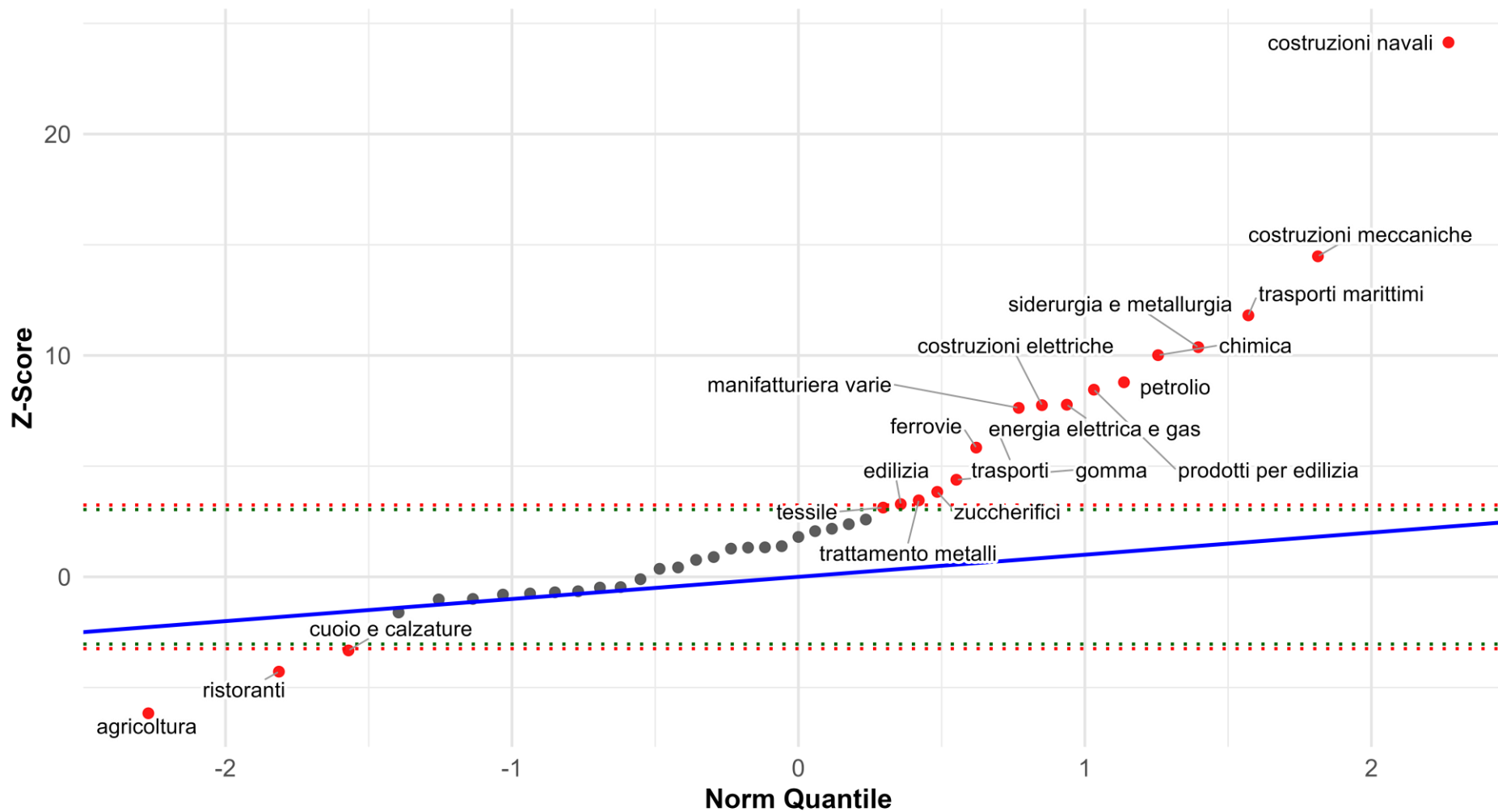
QQ Plot - Tumore Polmone



UOMINI



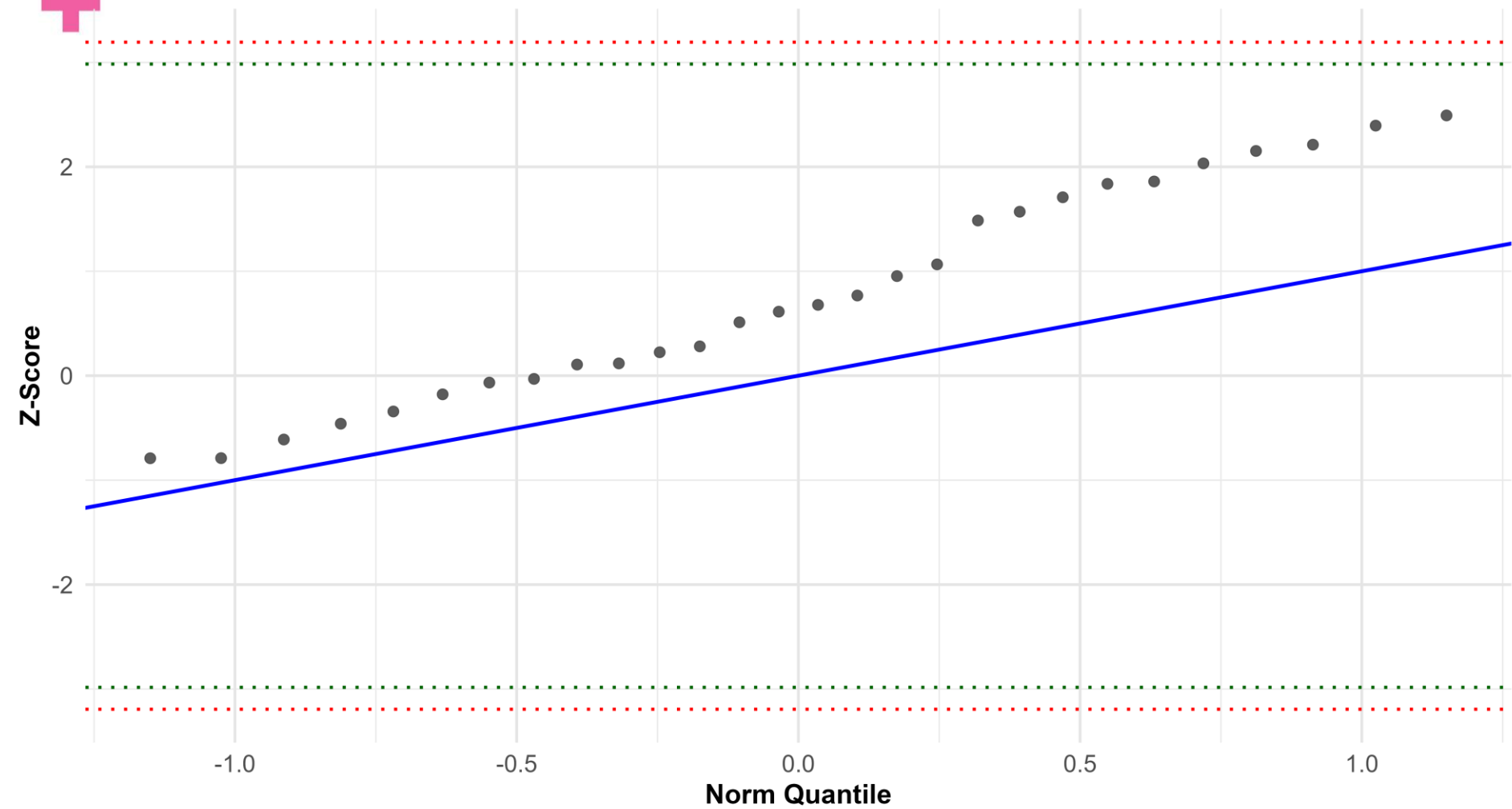
QQ Plot - Tumore Pleura,peritoneo,pericardio



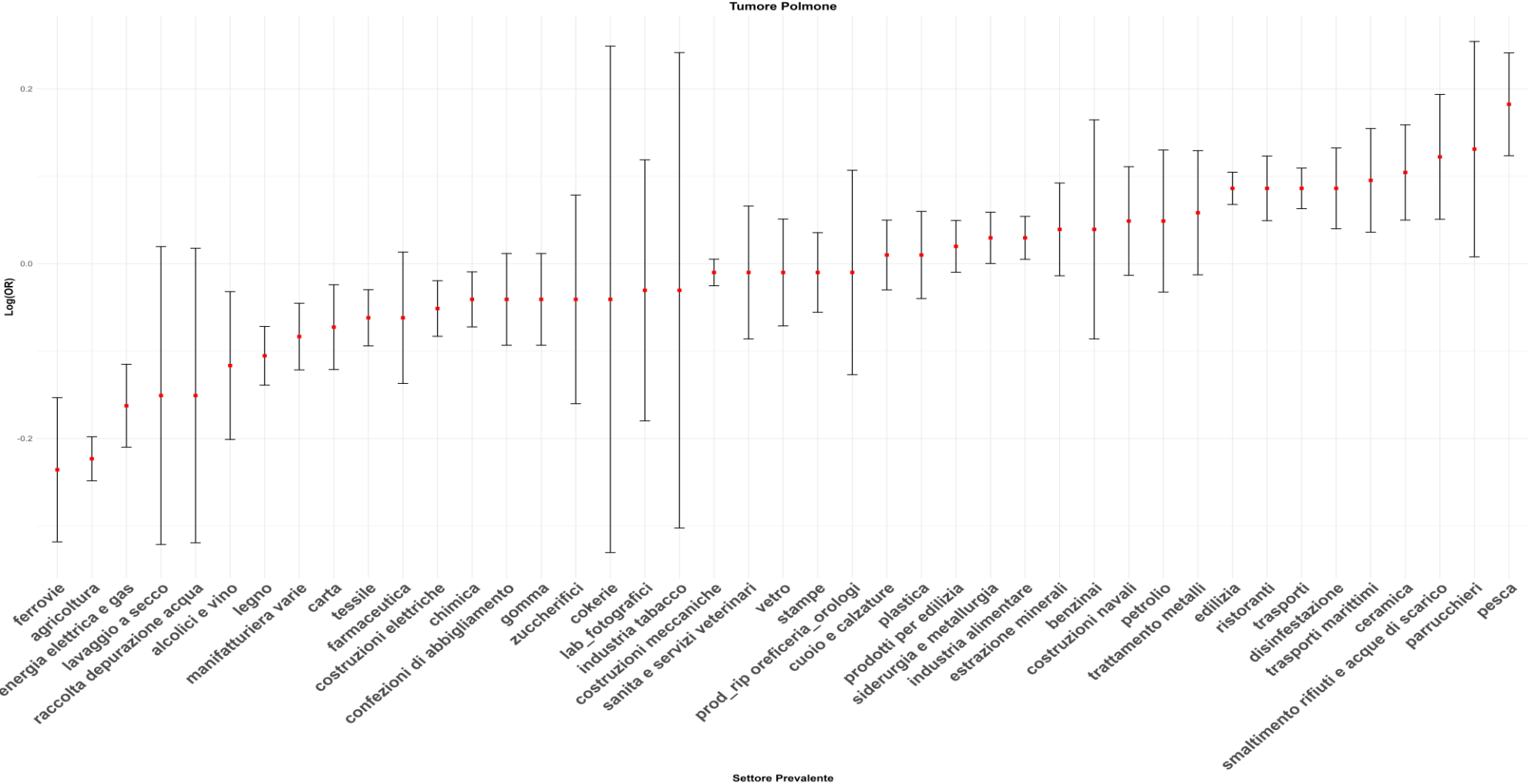
DONNE



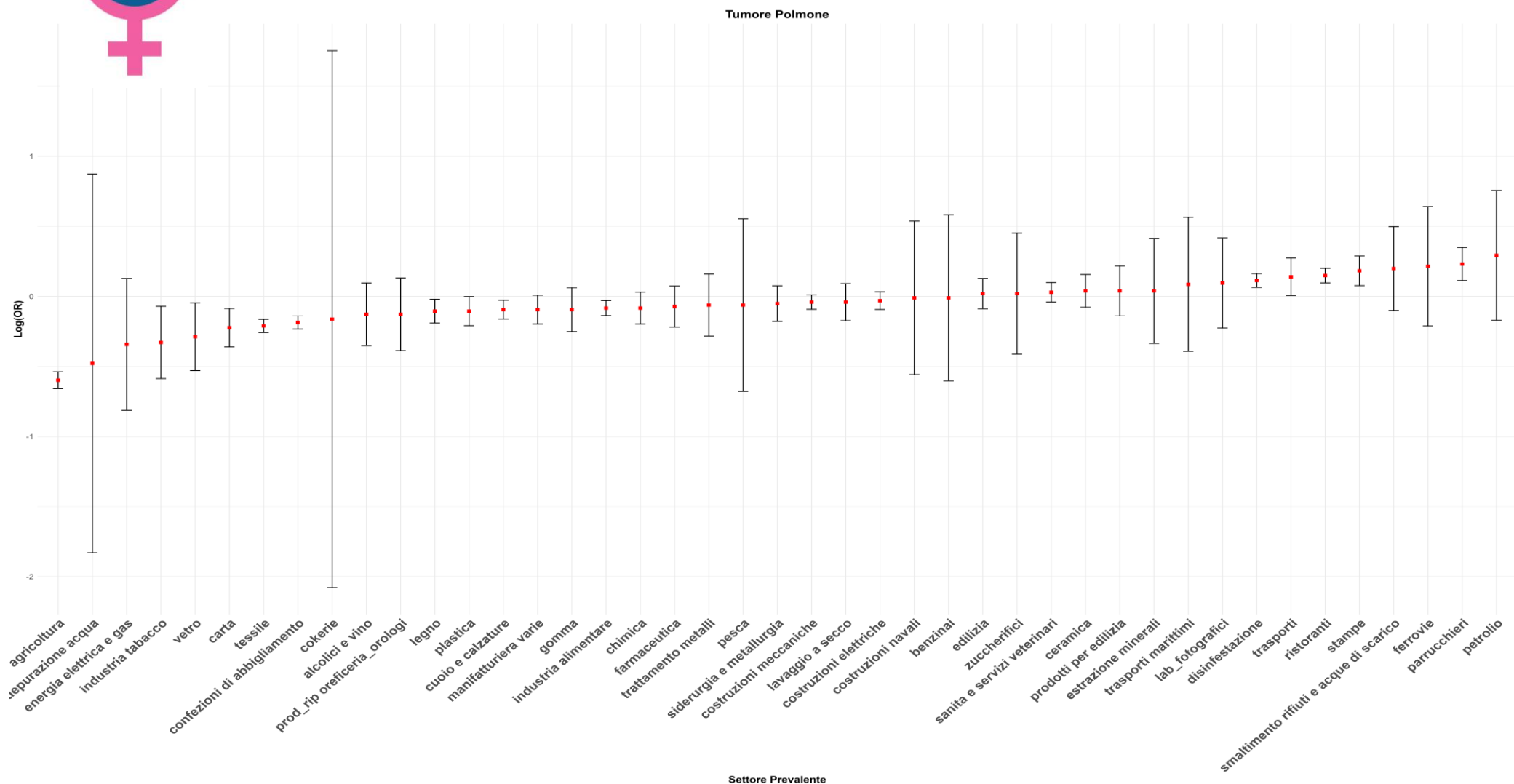
QQ Plot - Tumore Pleura,peritoneo,pericardio



UOMINI



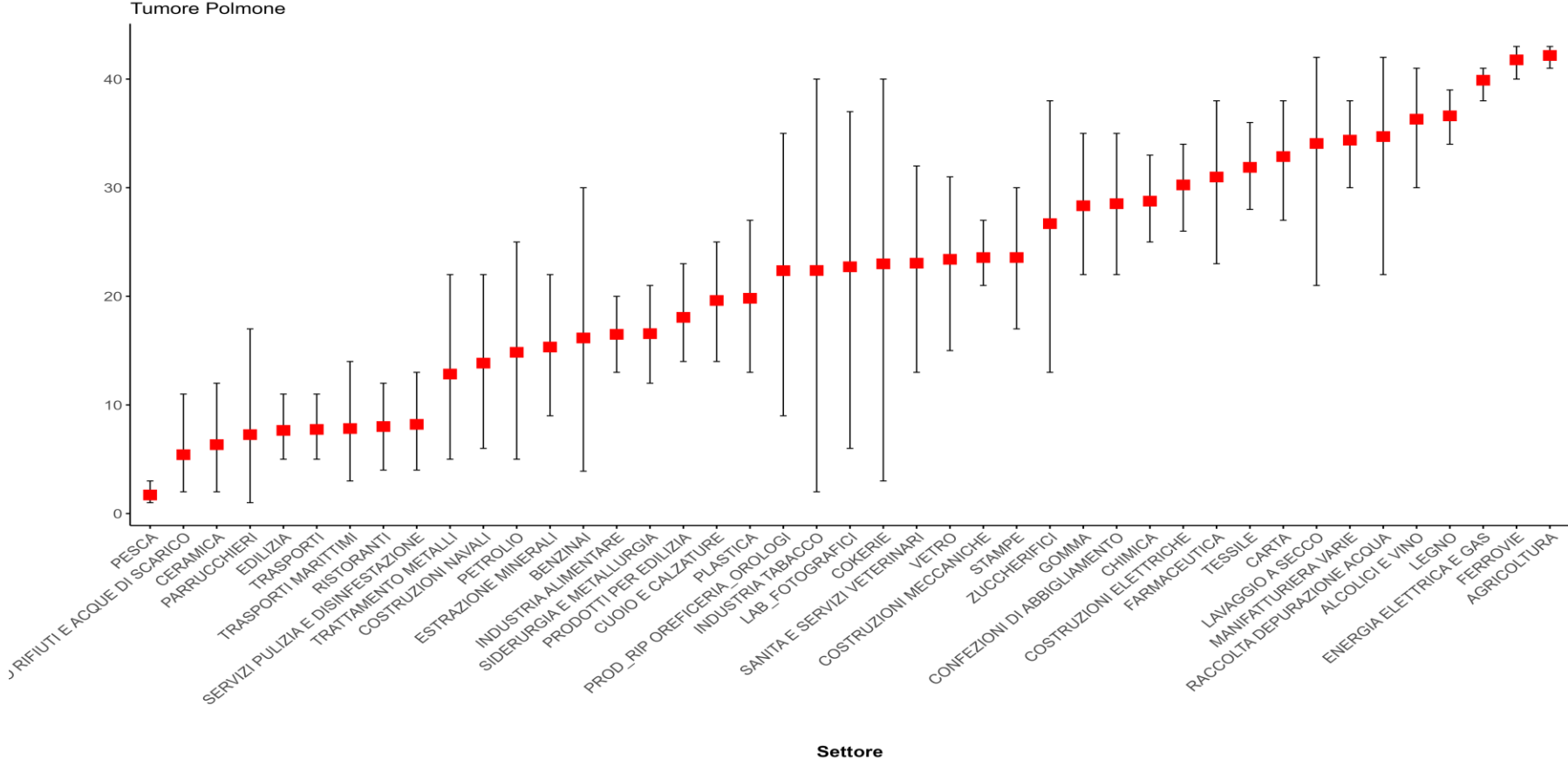
DONNE



UOMINI



Ranghi a posteriori e IC 80%

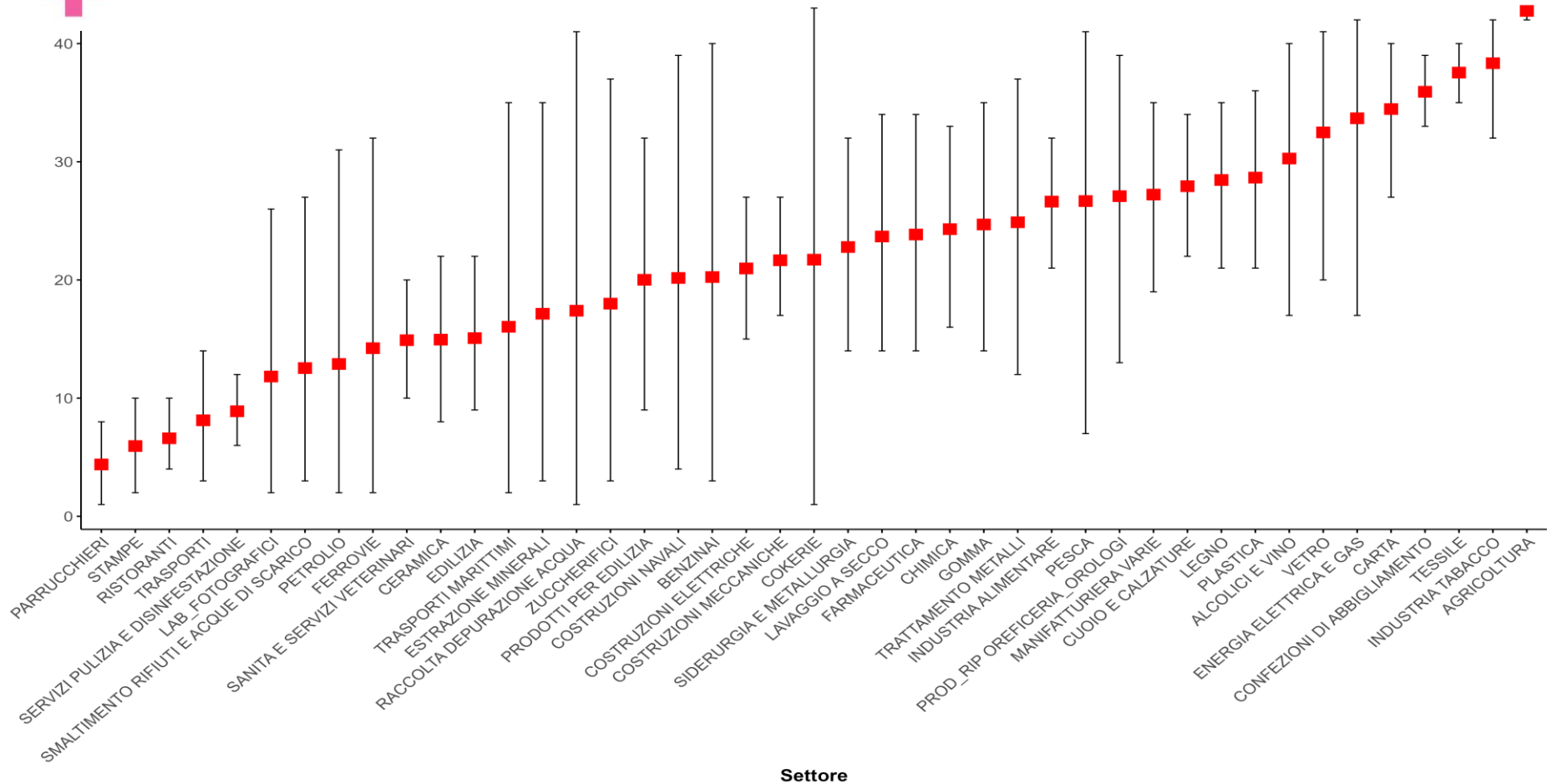


DONNE



Polmone

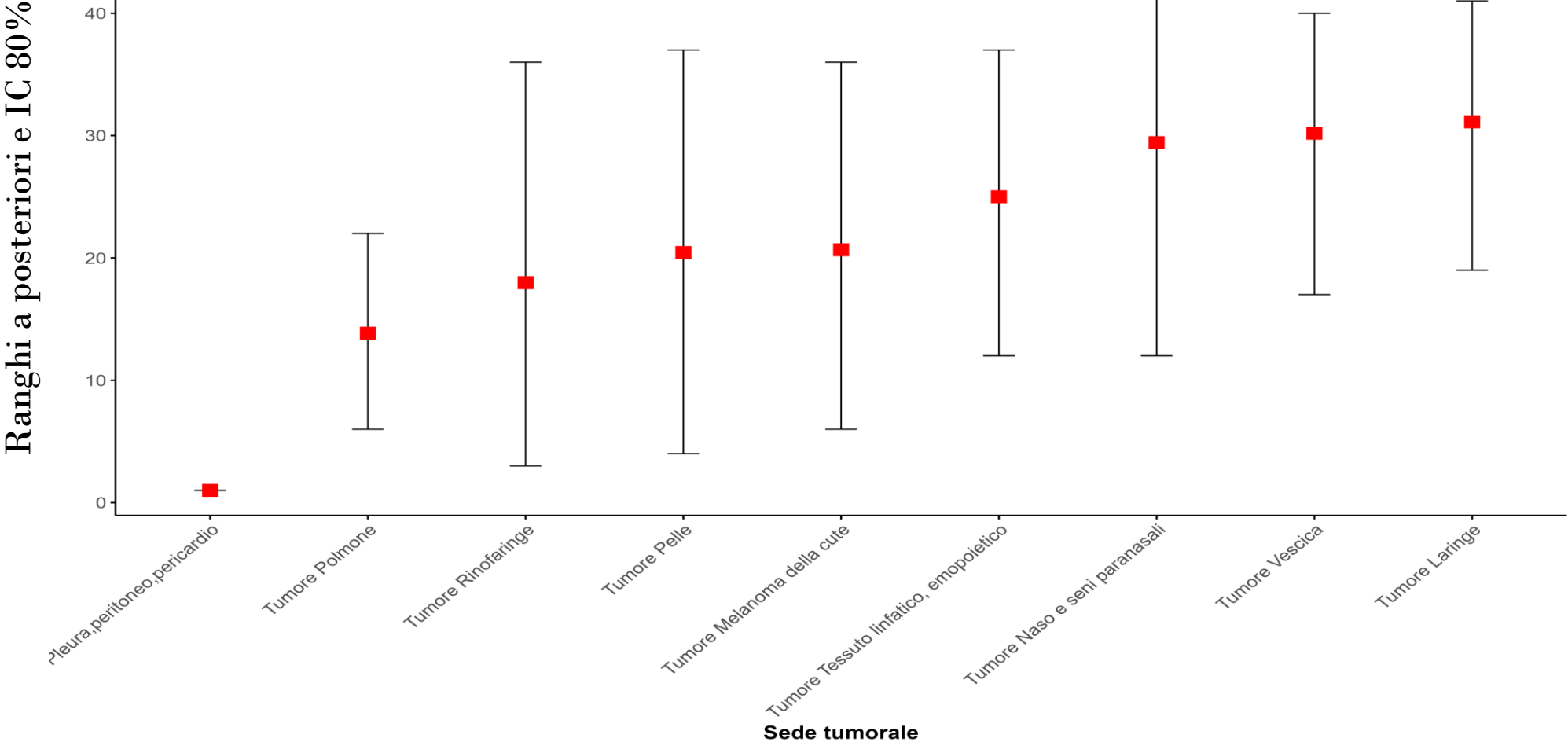
Ranghi a posteriori e IC 80%



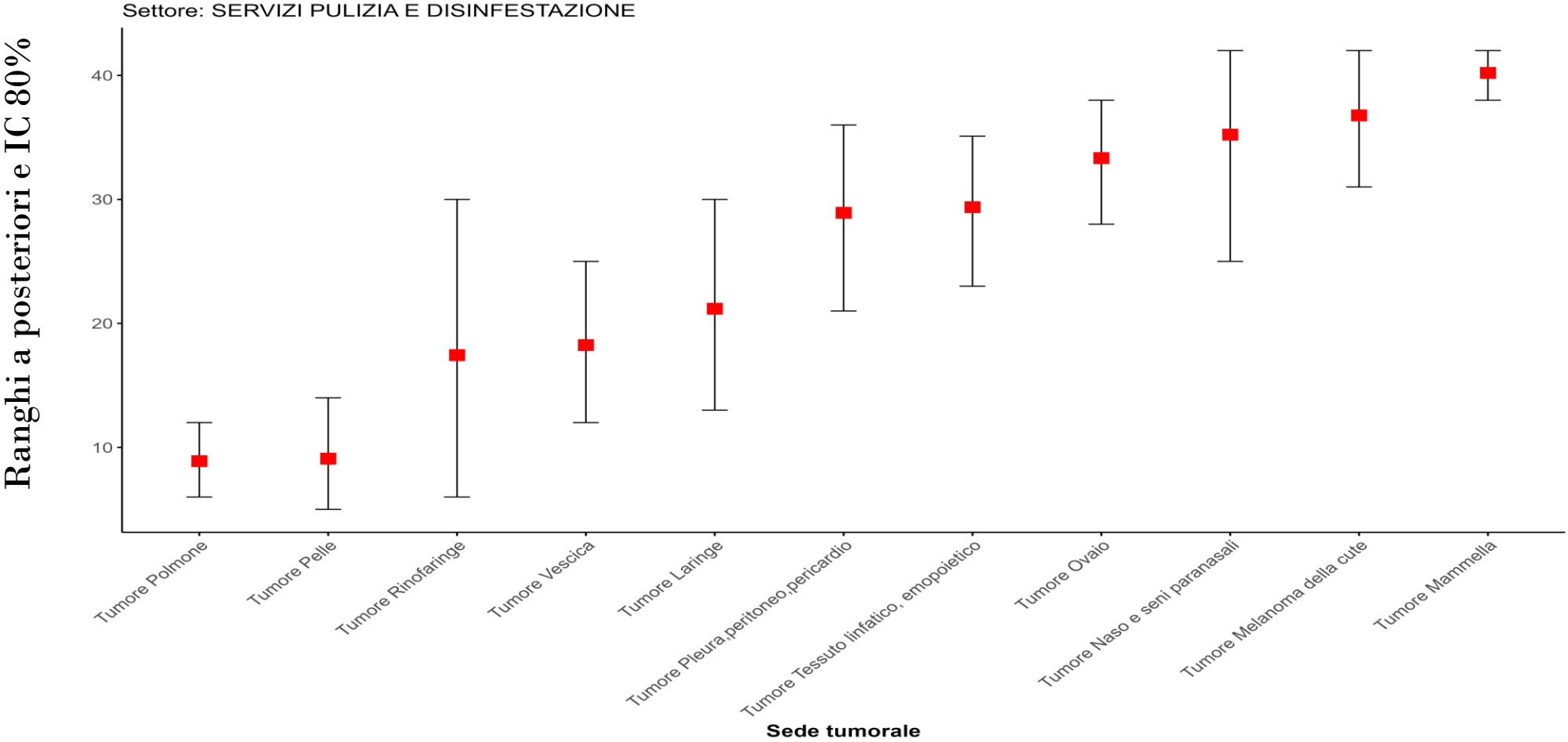
UOMINI



Settore: COSTRUZIONI NAVALI



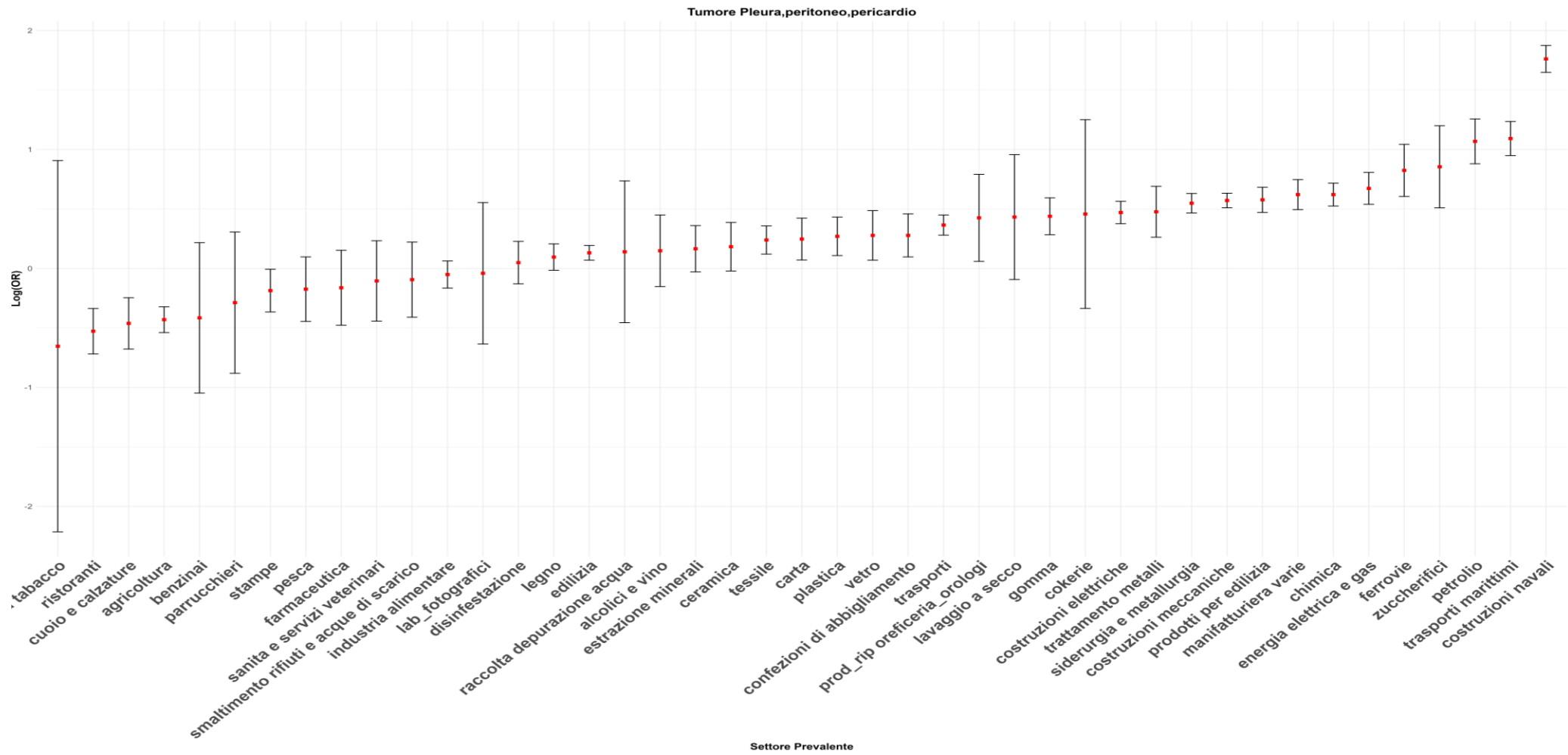
DONNE



UOMINI



Ranghi a posteriori e IC 80%



Conclusioni

- Nella stesura di rapporti che presentano informazioni epidemiologiche relative a molte malattie per molte popolazioni diverse può essere di estremo interesse fornire una sintesi dei risultati che sia di aiuto nell'identificare alcune priorità generali per azioni di sanità pubblica.
- I risultati evidenziano l'importanza di una sintesi dell'informazione che tenga conto della molteplicità dei test e della selezione delle osservazioni rilevanti.
- L'uso di strumenti grafici e metodi bayesiani gerarchici per le graduatorie migliora la comunicazione dei risultati, facilitando la formulazione di politiche di intervento basate su un'analisi approfondita dei rischi occupazionali.

Cosa c'è di sbagliato nel trovare sempre qualcosa di significativo? Nulla, se si adotta una procedura appropriata di controllo della probabilità di falsi positivi tra tutti i confronti dichiarati significativi.

Grazie

Giorgia Stoppa

giorgia.stoppa@unipd.it