UNIVERSITA' DEGLI STUDI DI MILANO

Scuola di dottorato in **STATISTICA BIOMEDICA (XXIV ciclo)**

Dipartimento di **Medicina del Lavoro "Clinica del Lavoro Luigi Devoto"**

Il fumo di sigaretta e il rischio di tumore del pancreas: diversi approcci di analisi in una *pooled-analysis*

Dottorando

Dott.ssa Ersilia Lucenteforte

Correlatore

Dott.ssa Cristina Bosetti

Tutor

Prof. Adriano Decarli

Coordinatore del Dottorato

Prof. Silvano Milano

INDICE

INTRODUZIONE

Il fumo di sigaretta è il principale fattore di rischio per il tumore del pancreas, ed è responsabile di circa il 15-25% di tutti i casi [1,2]. Una meta-analisi di 82 studi di coorte e caso-controllo pubblicati tra il 1950 e il 2007 ha mostrato un rischio relativo pari a 1,7 (intervalli di confidenza -IC- al 95%: 1,6-1,9) per i fumatori correnti e di 1,2 (IC 95%: 1,1-1,3) per gli ex-fumatori [3]. Nella stessa meta-analisi è stato mostrato che il rischio persiste anche quando si è ex-fumatori da meno di 10 anni.

Si è ancora incerti sulla quantificazione della relazione dose-rischio e sul ruolo di vari fattori temporali quali la durata dell'esposizione e, per gli ex-fumatori, il tempo dalla cessazione. Inoltre, ancora poche sono le ipotesi sul tipo di relazione.

Per chiarire questa importante questione, è stata condotta una *pooled-analysis* di 12 studi caso-controllo sul tumore del pancreas, all'interno del *International Pancreatic Cancer Case-Control Consortium (PanC4)* [4-7].

Una struttura generale per la conduzione di *pooled-analysis* consiste nel formulare i criteri di inclusione per gli studi, individuare tutti quelli che rispondono a tali criteri, ottenere i dati originali, creare un dataset standardizzato e ottenere delle stime "esposizione-malattia" riassuntive.

L'approccio standard per le analisi statistiche è quello di condurre *analisi aggregate*, cioè considerare i dati come un unico dataset e calcolare gli odds ratios (ORs), e i rispettivi IC al 95% attraverso modelli di regressione logistica multivariabile aggiustati per la variabile che identifica lo studio, oltre che per i potenziali fattori confondenti [8]. Un altro approccio è quello di condurre *analisi a due stadi* [9], cioè calcolare gli OR studio-specifico usando modelli di regressione logistica aggiustati per i potenziali fattori confondenti [8] e poi riassumere i rischi studio-specifici usando tecniche meta-analitiche [10]. In alternativa, per tener conto del naturale raggruppamento dei dati per studio, si possono condurre *analisi a livelli* [11-14].

Utilizzando questi metodi si ottengono stime riassuntive una volta categorizzate le variabili di interesse. La categorizzazione ha il vantaggio di ottenere una misura facilmente interpretabile del rischio, senza nessuna ipotesi sull'andamento del rischio. Ciò, però, potrebbe oscurare importanti differenze di rischio all'interno o attraverso le categorie considerate, proprio per i cut-off scelti. In alternativa, si possono considerare le variabili di interesse in continuo e studiarle attraverso modelli di regressione lineare standard. Si impone, così, un effetto dose-rischio lineare che, ancora una volta, potrebbe oscurare importanti differenze di rischio associate a differenti livelli della misura in esame.

Lo scopo principale del presente studio è quello di esplorare l'effetto di alcune variabili relative al fumo di sigaretta (quali dose, durata e tempo dalla cessazione) sul rischio di

tumore del pancreas applicando l'*analisi a due stadi*, l'*analisi a livelli*, ed esplorare la relazione sul rischio non imponendo nessuna predeterminata costrizione sui dati, attraverso l'utilizzo di *modelli additivi generalizzati*. I risultati ottenuti saranno confrontati con quelli che si ottengono con l'*analisi standard aggregata*.

ANALISI A DUE STADI

L'analisi a due stadi consta di due passi [9].

Calcolare il rischio relativo studio-specifico usando modelli di regressione logistica [15] definiti dalla trasformata *logit* data da

$$\pi_i(x) = \frac{e^{g_i(x)}}{1 - e^{gi(x)}} \forall i = 1,....,k$$

dove k è il numero di studi e $g_i(x) = \beta_{i0} + \beta_{i1}x_{i1} + ... + \beta_{ip}x_{ip}$, con x_{ij} confondenti studio-specifici.

Poi riassumere i rischi studio-specifici usando modelli ad effetti casuali [10].

1.1. Inferenza

Si consideri una serie di k studi che abbiano $Y_1,...,Y_k$ come stima degli *effect size* e siano $\theta_1,...,\theta_k$ i veri valori degli *effect size*, ciascuno stimatore può essere espresso come lo scostamento dal valore vero $Y_i = \theta_1 + e_i$ dove $e_i = N(0, \sigma^2_i)$ per $i=1,...,k$.

Gli e_i costituiscono le deviazioni dal vero valore e si assume che siano indipendenti, con media 0 e varianza σ^2_i. Questo implica che le stime degli *effect size* siano a loro volta normalmente distribuite, con stessa varianza e media pari a σ_i, quindi $Y_i = N(\theta_1, \sigma^2_i)$.

Le stime possono essere qualunque misura dell'effetto, purché soddisfino l'assunto di normalità (anche approssimativamente): esempi sono i *log-Odds Ratio* o la differenza fra medie. In generale, il parametro di interesse è μ, ovvero l'effetto complessivo (*overall effect*) determinato in modi diversi secondo la tecnica utilizzata.

1.2. Test di omogeneità

Testare l'omogeneità fra gli studi equivale a verificare la seguente ipotesi:

$$\begin{cases} H_o : \theta_1 = \theta_2 = ... = \theta_k \\ H_1 : \textit{almeno una delle uguaglianze non è verificata} \end{cases}$$

Sotto l'ipotesi nulla e per numerosità campionarie elevate la statistica di omogeneità si distribuisce come un χ^2_{k-1}, ovvero

$$Q_W = \sum W_i (Y_i - \hat{\theta}_{MLE})^2 = \chi^2_{k-1}$$

dove $\hat{\theta}_{MLE} = \frac{\sum_{i}^{k} W_i Y_i}{\sum_{i}^{k} W_i}$ e $W_i = \frac{1}{s_i^2}$.

Il rifiuto dell'ipotesi nulla porta alla conclusione che l'effetto dei singoli studi proviene da due o più popolazioni distinte. A questo punto si possono identificare le covariate che identificano gruppi omogenei di studi o si può applicare un modello ad effetti casuali.

Se non viene rifiutata l'ipotesi nulla, i *k* studi possono ritenersi omogenei e condividono un unico parametro comune θ, la cui stima sarà $\hat{\theta}_{MLE}$.

1.3. Modello ad effetti fissi

Il modello ad effetti fissi presuppone che ogni studio abbia lo stesso *underlying effect*, ovvero

$$\theta_i = \mu$$

e quindi $Y_i = N(\mu, \sigma^2_i)$.

In altre parole, tutti gli studi provengano dalla stessa popolazione, anche se non sono identicamente distribuiti (infatti possono avere diversa varianza).

Lo stimatore della media globale è generalmente una semplice media pesata dei singoli effetti stimati

$$\mu = \sum \omega_i Y_i / \sum \omega_i,$$

dove i pesi ottimali sono proporzionali all'inverso della varianza

$$\omega_i = 1/var(Y_i).$$

Non essendo nota la varianza si utilizza la sua stima $\hat{\sigma}_i^2$ sia per ottenere $\hat{\mu}$ che per ottenere $\hat{\text{var}}(\hat{\mu})$ e quindi risulta che

$$\hat{\mu} = \sum (Y_i / \hat{\sigma}_i^2) / \sum (1 / \hat{\sigma}_i^2)$$

$$\hat{\text{var}}(\hat{\mu}) = 1 / \sum (1 / \hat{\sigma}_i^2)$$

1.4. Modello ad effetti casuali

Il modello ad effetti casuali presuppone che il vero parametro di ciascun studio non sia lo stesso, ma vari attorno al vero parametro μ distribuendosi come una normale di varianza pari a τ^2:

$$Y_i = \theta_i + e_i \quad \text{dove} \quad e_i = N(0, \sigma_i^2)$$

$$\theta_i = \mu + \varepsilon_i \quad \text{dove} \quad \varepsilon_i = N(0, \tau^2)$$

si assume che i termini di errore e_i e ε_i siano indipendenti: questi esprimono rispettivamente la variazione dell'effetto stimato Y_i attorno al valore vero θi e la variazione di quest'ultimo attorno all'effetto globale μ.

La varianza τ^2 del modello ad effetti casuali è una misura dell'eterogeneità fra gli studi (un modello ad effetti fissi è il caso particolare per cui $\tau^2=0$). Normalmente, in caso di eterogeneità, dovrebbe essere utilizzato il modello ad effetti casuali. L'eterogeneità viene comunemente testata con la statistica di Cochran $Q_\omega = \sum \omega_i (Y_i - \hat{\mu})^2$ che va a verificare il sistema di ipotesi

$$\begin{cases} H_o : \tau^2 = 0 \\ H_1 : \tau^2 > 0 \end{cases}$$

Se σ^2_i sono note, allora sotto H_0, la statistica è $Q_w = \chi^2_{k-1}$. Nella pratica si disporrà della stima della varianza e quindi di $\hat{\omega}_i = 1/\hat{\sigma}_i^2$ che danno una statistica $Q_{\hat{w}}$ approssimativamente distribuita come un χ^2_{k-1} sotto H_0. Valori elevati di $Q_{\hat{w}}$ indicano una elevata variabilità tra gli studi, quindi l'ipotesi di omogeneità viene rifiutata e generalmente viene applicato il metodo degli effetti casuali.

Poiché il test di Cochran può essere caratterizzato da una potenza bassa, l'ipotesi di omogeneità non è sempre correttamente accettata e l'uso del modello ad effetti casuali dovrebbe avere un'applicazione più frequente di quanto suggerito dal test.

Le stime del modello ad effetti fissi sono immediatamente ottenibili. Per il modello ad effetti casuali sono stati proposti diversi metodi che vengono di seguito descritti: il metodo dei momenti [10], e i metodi basati sulla funzione di verosimiglianza.

Considerato il modello ad effetti casuali è possibile esprimere le stime dei singoli effetti direttamente come funzione della media globale:

$$Y_i = \mu + e_i + \varepsilon_i \quad \text{dove} \quad e_i = N(0, \hat{\sigma}_i^2) \quad \text{e} \quad \varepsilon_i = N(0, \tau^2)$$

e quindi, per l'indipendenza dei due termini d'errore,

$$Y_i = N(\mu, \hat{\sigma}_i^2 + \tau^2).$$

Anche in questo caso la stima del parametro globale viene ricavata tramite una media pesata delle singole stime

$$\hat{\mu}_\tau = \frac{\sum \hat{\omega}_i(\tau) Y_i}{\sum \hat{\omega}_i(\tau)} \quad \text{con varianza pari a } \hat{\text{var}}(\hat{\mu}_{\hat{\tau}}) = 1/\sum \hat{\omega}_i(\hat{\tau})$$

dove i pesi degli effetti casuali sono $\hat{\omega}_i(\tau) = 1/\left(\tau^2 + \hat{\omega}_i^{-1}\right) = 1/(\tau^2 + \hat{\sigma}_i^2)$ ed $\hat{\omega}_i = 1/\hat{\sigma}_i^2$ come definiti sopra.

Supponendo che τ^2 sia noto, la stima globale si distribuisce come una normale tale che:

$$\hat{\mu}_\tau = N\left(\mu, \frac{1}{\sum \hat{\omega}_i(\tau)}\right)$$

Poiché $\tau^2 \geq 0$ si deduce che $\hat{\omega}_i(\tau) \leq \hat{\omega}_i$ e quindi $\text{var}(\hat{\mu}_\tau) \geq \text{var}(\hat{\mu})$: questo significa che gli intervalli di confidenza di μ con il modello ad effetti casuali sono generalmente più ampi di quelli costruiti con il modello ad effetti fissi.

Nella pratica però, τ^2 è ignoto ed è quindi necessario definire un metodo per trovarne una stima. Lo stimatore più comunemente usato si basa sul metodo dei momenti che ricava la stima confrontando il valore atteso ed osservato della statistica $Q_{\hat{w}}$.

$$E(Q_{\hat{w}}) = E\left[\sum \hat{\omega}_i (Y_i - \hat{\mu})^2\right] = \sum \hat{\omega}_i E(Y_i^2) - \sum \hat{\omega}_i E(\hat{\mu}^2) = k - 1 + \tau^2 \left[\sum \hat{\omega}_i - \frac{\sum \hat{\omega}_i^2}{\sum \hat{\omega}_i}\right] = q_{\hat{w}}$$

La stima di τ^2 può essere ricavata risolvendo l'equazione precedente:

$$\hat{\tau}^2 = \begin{cases} t & if\ t > 0 \\ 0 & if\ t \leq 0 \end{cases}$$

dove $t = \dfrac{q_{\hat{\omega}} - (k-1)}{\sum \hat{\omega}_i - \sum \hat{\omega}_i^2 / \sum \hat{\omega}_i}$.

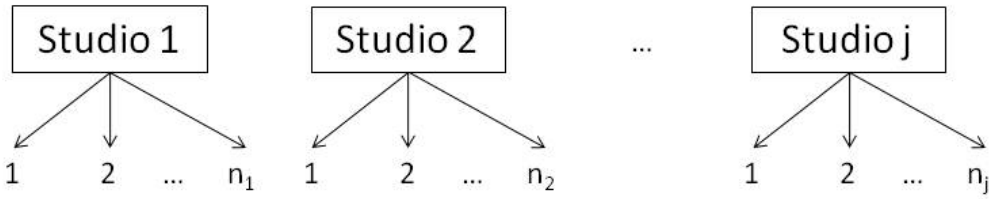
La stima $\hat{\tau}^2$, poiché tronca, risulta essere una stima distorta di τ^2.

DerSimonian and Laird [10] propongono di sostituire la stima trovata $\hat{\tau}^2$ direttamente negli effetti casuali $\hat{\omega}_i(\tau)$ e quindi nelle stime di $\hat{\mu}_{\hat{\tau}}$ e $\hat{\text{var}}(\hat{\mu}_{\hat{\tau}})$ come se τ^2 fosse un valore noto. Nel calcolo degli intervalli di confidenza per μ, vengono inoltre mantenute le assunzioni di normalità per lo stimatore $\hat{\mu}_{\hat{\tau}}$ nonostante l'uso di $\hat{\sigma}_i^2$ e $\hat{\tau}^2$ al posto di σ_i^2 e τ^2.

E' importante notare che l'assunto di normalità comporta dei problemi innanzitutto di validità e inoltre di verificabilità: risulta infatti difficile verificare la normalità per meta-analisi di pochi studi. In particolare la normalità di ε_i non è facilmente verificabile o giustificata. Inoltre, la variabilità tra il vero effetto dei diversi studi viene raccolta e descritta attraverso l'inclusione del parametro τ^2 nei pesi: questa però risulta essere solo una stima della varianza e non viene presa in considerazione l'incertezza associata alla sua stima $\hat{\tau}^2$ ed in particolare non viene in alcun modo modificata la distribuzione di $\hat{\mu}_{\hat{\tau}}$.

MODELLO DI REGRESSIONE LOGISTICA MULTILIVELLO

Lo scopo principale del presente studio è stato quello di analizzare il rischio di tumore del pancreas in accordo con il consumo di sigarette utilizzando un insieme di 12 studi caso-controllo. Un approccio standard sarebbe potuto essere quello delle *analisi aggregate*, cioè considerare i dati come un unico dataset e calcolare gli odds ratios (ORs), e i rispettivi intervalli di confidenza (CI) al 95% attraverso modelli di regressione logistica multivariabile aggiustati per la variabile studio, oltre che per i potenziali fattori confondenti [16]. Questo tipo di approccio non considera però il naturale raggruppamento dei dati per studio, ovvero la loro struttura a livelli.
I dati del PanC4 possono essere rappresentati graficamente come segue:



Ovvero, con una struttura a due livelli: il livello uno è il soggetto, mentre il livello due è lo studio. Se non si tenesse conto di tale struttura si assumerebbe di avere N osservazioni indipendenti, dove N è il numero di soggetti totali ottenibile come somma degli N_j (per i che varia da 1 a 12) e rappresenta la numerosità di ciascuno studio incluso.
I modelli *multilivello* sono modelli statistici che possono essere usati per tener conto della variabilità associata ad ogni livello della struttura. Essi sono anche detti gerarchici, misti, o a coefficienti casuali [11-14]. Quando si ha una variabile risposta binaria, come nel nostro caso, si parla di regressione logistica multilivello.

1.5. Modello ad un solo livello

Sia y una variabile risposta binaria (il paziente è un caso di tumore di pancreas o è un controllo) con distribuzione Bernulliana, ovvero $y \approx Bin(1,\pi)$, un modello ad un solo livello può essere considerato del tutto analogo a quello di regressione logistica "ordinario". Ha quindi la seguente formula [17]:

$$y_{ij} = \pi_{ij} + e_{ij}$$
$$Logit(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha + \beta x_{ij} \qquad (2.1)$$

dove $\mathbf{i} = 1,...,\mathbf{I_j}$ indica il livello paziente, $j = 1,...,J$ indica il livello studio, e π_{ij} è la probabilità, del paziente i-esimo nello studio j-esimo, di ammalarsi di tumore del pancreas condizionata dal fattore di rischio x.

Assumendo che gli e_{ij} hanno una distribuzione binaria con $E(e_{ij})=0$ e $Var(e_{ij})=\sigma_e^2=\pi_{ij}(1-\pi_{ij})$, si ha che la funzione di probabilità ha la seguente formula:

$$\pi_{ij}=\frac{\exp(\alpha+\beta x_{ij})}{1+xp(\alpha+\beta x_{ij})}$$

Questo modello è ad un solo livello e non tiene conto dello studio, quindi non considera dunque la variabilità tra studi e quella dei pazienti all'interno di ogni studio.

1.6. Modello multilivello

Ci sono vari modi per estendere un modello ad un solo livello ad uno a più livelli: uno di questi è quello di aggiungere una variabile all'equazione di regressione logistica (2.1) in modo tale che ogni studio abbia una sua intercetta e le intercette degli studi vengono usate per misurare le differenze tra gli studi stessi. La formula (2.1) diventa quindi:

$$Logit(\pi_{ij})=\alpha_j+\beta x_{ij}$$

Le intercette degli studi, le α_j, si possono considerare sia ad effetto casuale che ad effetto fisso [18]. Considerandole ad effetto fisso, ci si può ricondurre ad un modello ad un solo livello. Considerandole ad affetto casuale, con una loro specifica distribuzione di probabilità, si ha che:

$$\alpha_j=\alpha+u_j$$

ovvero le α_j sono combinazioni lineari di due termini, un primo termine (α) che rappresenta la media totale delle α_j ed un secondo termine (u_j) che rappresenta l'errore associato ad ogni singolo α_j, dove $u_j \approx N(0,\sigma_u^2)$ per ogni j e gli u_j sono indipendenti dagli errori casuali e_{ij}.

I modelli di regressione logistica a due livelli hanno, quindi, la seguente formula:

$$Logit(\pi_{ij})=\alpha+u_j+\beta x_{ij}$$

Tali modelli vengono detti misti perché hanno sia l'effetto fisso (α,β), sia l'effetto casuale (u_j).

Lo studio come fattore di rischio a sé stante

Il modello misto non tiene conte del fatto che lo studio, oltre a rappresentare il livello più alto di gerarchia, potrebbe essere un fattore di rischio a sé: ad esempio, una determinata caratteristica degli studio (disegno dello studio, tipo di raccolta delle informazione) può avere un impatto sulla stima del rischio di tumore.

Supponiamo che la variabile binaria z indichi se uno studio è caso-controllo o coorte. A questo scopo, si aggiunge un effetto fisso nel secondo livello, ovvero si considera la formula:

$$Logit(\pi_{ij}) = \alpha_j + \beta x_{ij}$$
$$\alpha_j = \alpha + \gamma z_j + u_j$$

Quindi le intercette α_j diventano combinazioni lineare di tre termini: la media α generale, l'effetto fisso dello studio γ e l'effetto casuale dello studio u_j.

Si ha quindi la seguente formula:

$$Logit(\pi_{ij}) = \alpha + \gamma z_j + u_j + \beta x_{ij}$$

Le interazioni tra livelli

Se si è interessati a valutare l'interazione tra una o più variabili esplicative, per esempio l'interazione tra il tipo di studio (studio con controlli ospedalieri e provenienti dalla popolazione generale, ad esempio) e una determinata caratteristica del paziente, allora si ha il seguente modello:

$$Logit(\pi_{ij}) = \alpha_j + \beta_j x_{ij}$$
$$\alpha_j = \alpha + \gamma z_j + u_j$$
$$\beta_j = \beta + \theta z_j$$

Quindi β_j è la combinazione lineare della media β e dell'effetto del tipo di studio z_j, e θ rappresenta il parametro del termine di interazione $z_j x_{ij}$. Il modello allora ha la seguente formula:

$$Logit(\pi_{ij}) = \alpha + \gamma z_j + u_j + \beta x_{ij} + \theta z_j x_{ij}$$

MODELLO ADDITIVO GENERALIZZATO

Prima di introdurre il concetto di modello additivo generalizzato è necessario introdurre il concetto di *smoother* [19].

1.7. Lo Smoothing

Uno *smoother* è un oggetto matematico che rappresenta l'andamento che le misure di una variabile risposta Y assumono come funzione di una o più variabili indipendenti $X_1,\ldots,X_p$. Esso produce una stima del trend che è meno variabile della variabile risposta stessa, da qui il nome *smoother*. Un'importante caratteristica dello *smoother* è che ha una natura non parametrica. Questo è il motivo per cui spesso ci si riferisce allo *smoother* come ad un oggetto per regressioni non parametriche. Le stime prodotte da uno *smoother* sono dette *smooth.* Il caso di un singolo predittore è quello più comune e viene chiamato *scatterplot smoothing*.

Si supponga di avere delle misure $\mathbf{y}=(y_1,\ldots,y_n)^T$ di risposta in corrispondenza dei punti $\mathbf{x}=(x_1,\ldots,x_n)^T$. Si assuma che ogni **y** e **x** rappresentino le misure delle variabili Y e X. Uno *scatterplot smoothing* si definisce come una funzione di **x** e **y**, il cui risultato è una funzione s con lo stesso dominio dei valori di **x**: $s=S(y|x)$.

Gli *smoother* hanno due principali utilizzi: il primo è descrittivo ed aiuta a capire l'andamento dei dati, il secondo è modellistico ed aiuta a stimare la dipendenza del valore atteso della variabile risposta in funzione delle variabili indipendenti.

Il più semplice esempio di *smoother* si ha nel caso di dati categorici e consiste nella semplice operazione di media all'interno di ciascuna categoria. Sostanzialmente lo *smoothing* è una sorta di media sulle categorie, una media locale, ovvero una media dei valori della variabile risposta in corrispondenza dei valori delle variabili indipendenti che si avvicinino ad un valore di riferimento. Quindi, bisogna determinare come effettuare l'operazione di media in ciascun "intorno" e quanto considerare grande tale "intorno".

L'operazione di media è strettamente legata al tipo di *smoother* considerato.

La dimensione dell'intorno è espressa generalmente con un parametro *smoothing*. Intuitivamente, ad intorni grandi corrispondono stime dell'andamento della variabile risposta con varianza minore ma potenzialmente distorsione maggiore, viceversa per intorni piccoli. Questa relazione tra distorsione e varianza è ciò su cui agisce il parametro *smoothing.* Il concetto è analogo al concetto di quante variabili indipendenti inserire in una equazione di regressione.

Di seguito alcuni esempi di *smoother*.

Bin smoothers

Un *bin smoother* mima uno *smoother* categorico che partiziona i valori predittori in un certo numero di regioni disgiunte e effettua la media dei valori della variabile risposta all'interno di ogni regione.

Si supponga di avere $c_0<...<c_K$ valori di cut-off della variabile indipendente dove $c_0 = -\infty$ e $c_K = +\infty$, si definisce

$$R_k = \{i; c_k \le x_i \le c_{k+1}\}, \text{ dove } k = 0, \ldots, K\text{-}1$$

l'insieme degli indici dei dati in ciascuna regione. Allora $s=S(y|x)$ è data da $s(x_o) = media_{i \in R_k}(y_i)$ se $x_0 \in R_k$. In genere si sceglie un numero contenuto di regioni e si determinano gli estremi in modo tale che ciascuna regione contenga lo stesso numero di punti. La stima che si ottiene in realtà presenta tante discontinuità quanti sono i punti di cut-off.

Running-mean smoother

Si supponga che il valore di riferimento x_0 sia pari ad uno degli x_j, ad esempio x_i. Se si hanno repliche di x_i , tutti ciò che bisogna fare è calcolare la media dei valori di Y in corrispondenza degli x_i per ottenere la stima $s(x_i)$. Se, invece, non si hanno repliche bisogna considerare i valori situati nell'intorno di x_i e bisogna stabilire un criterio con cui stabilire quali punti possano essere considerati nell'intorno. Un modo semplice è quello di scegliere lo stesso x_i e i k punti che si trovano a destra e a sinistra di x_i. In tal modo, si determina un intorno simmetrico $N^S(x_i)$. I *running-mean smoother* si definiscono come

$$s(x_i) = media_{j \in N^S(x_i)}(y_i).$$

Una definizione formale di un intorno simmetrico è espressa come

$$N^S(x_i) = \{\max(i-k,1), \ldots i-1, \ldots, \min(i+k,n)\}$$

Seppure questo metodo sia facilmente implementabile, esso fornisce stime distorte in prossimità degli estremi.

Un metodo per ovviare alla distorsione è quello di generalizzare il concetto di *running-mean smoother* calcolando la retta di regressione invece che la media, ovvero:

$$s(x_i) = \hat{\alpha} + \hat{\beta}(x_i)x_i$$

dove $\hat{\alpha}$ e $\hat{\beta}$ sono le stime ai minimi quadrati dei coefficienti dei dati contenuti in $N^S(x_i)$. Si parla in questo caso di *running-line smoother*. Grandi valori di k tendono a

produrre curve poco frastagliate, mentre piccoli valori tendono a produrre curve più frastagliate.

Se si definisce la quantità $\omega = (2k+1)/n$, detta *span*, ovvero la proporzione di punti contenuti in ciascun intorno, allora i casi estremi sono quando $\omega = 2$, quindi l'intorno contiene tutti i punti e lo *smoother* non è altro che la retta di regressione, e $\omega = 1/n$, quindi l'intorno del punto è il punto stesso e lo *smoother* coincide con la retta interpolante i dati.

Kernel smoother

Un *kernel smoohter* usa un insieme di pesi locali, definiti del *kernel*, e produce una stima in corrispondenza di ogni punto di riferimento. Il peso attribuito al *j*-esimo punto è

$$S_{0j} = \frac{c_0}{\lambda} d\left(\left|\frac{x_0 - x_j}{\lambda}\right|\right)$$

dove *d(t)* è una generica funzione pari e decrescente in |*t*|. Il parametro λ determina la larghezza di banda, mentre c_0 è scelta in modo tale da normalizzare i pesi a 1.

Quando *d(t)* è la funzione densità di una gaussiana, abbiamo il *kernel smoother guassiano*. Altri esempi sono il *kernel smoother di Epanechnikov* [20], ovvero

$$d(t, |t| \le 1) = \frac{3}{4}(1 - t^2) \text{ e } d(t, altrimenti) = 0$$

che minimizza l'errore quadratico medio, oppure il *kernel smoother minimo*, ovvero

$$d(t, |t| \le 1) = \frac{3}{8}(3 - 5t^2) \text{ e } d(t, altrimenti) = 0$$

che minimizza la varianza.

Dal punto di vista computazionale, il *kernel smoohter* può essere rappresentato come

$$s(x_0) = \frac{\sum_{i=1}^{n} d\left(\frac{x_0 - x_i}{\lambda}\right) y_i}{\sum_{i=1}^{n} d\left(\frac{x_o - x_i}{\lambda}\right)}.$$

Si noti che sia il numeratore che il denominatore sono delle convoluzioni, quindi si potrebbe utilizzare una trasformata di Fourier così da avere calcoli meno complessi.

Regression spline

Le regressioni polinomiali hanno il limite dovuto alla loro natura globale in contrasto con la natura locale degli *smoother*. Le *regression spline* invece offrono la possibilità di fittare i dati con regressioni polinomiali a tratti.

In generale, le regioni dei vari tratti sono separate da una sequenza di nodi e in corrispondenza di tali nodi si forza la continuità delle polinomiali. Nonostante sia possibile scegliere vari tipi di polinomiali (il *bin smoother* ne è un esempio), di solito si considerano le polinomiali che abbiano continuità nei nodi fino alla terza derivata. Le *regression spline* sono molto usate perché è possibile applicare modelli lineari standard, ma la difficoltà principale è quella di selezionare il numero e la posizione dei nodi.

Cubic Smoothing Spline

Si supponga che tra tutte le funzioni f(x) con derivata del secondo ordine continua quella che minimizza la somma dei quadrati dei residui penalizzata sia la seguente

$$\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda\int_a^b \{f''(t)\}^2\,dt$$

dove λ è una costante fissa e $a \leq x_1 \leq \ldots \leq x_n \leq b$. Il primo termine misura la vicinanza ai dati, mentre il secondo penalizza le curvature. È possibile dimostrare che esiste un'unica funzione che minimizza la formula su detta e tale funzione è la *cubic smoothing spline* naturale con nodi negli x_i e parametro λ è il parametro di *smoothing*, e ha la stessa funzione dello *span* nelle *running-line smoother*: grandi valori di λ producono curve *smoother*, mentre piccoli valori producono curve "sinuose".

Smoother per predittori multipli

Nel caso in cui si ha più di un predittore, X_1, …, X_p, bisogna fittare una superficie *p*-dimensionale sui dati. La difficoltà sta nel definire il concetto di intorno, più che nella procedura di "media" dei punti. La forma dell'intorno dipende dalla metrica che si adotta.

1.8. Modello additivo

Il modello additivo è una generalizzazione del modello lineare.

Si supponga di avere *n* osservazioni di una variabile dipendente *Y*, indicate con $\boldsymbol{y}=(y_1,\ldots,y_n)^T$, e che tali *n* osservazioni siano le misure di *n* vettori $\boldsymbol{x}^i=(xi_1,\ldots,xi_p)$. Lo

scopo è quello di modellare la dipendenza di Y sulle $X_1, \ldots, X_p$ dal punto di vista descrittivo, inferenziale e predittivo.

Si consideri, dunque, il modello di regressione lineare multipla

$$Y = \alpha + X_1\beta_1 + \ldots + X_p\beta_p + \varepsilon$$

dove $E(\varepsilon)=0$ e $var(\varepsilon)=\sigma^2$. Questo modello fa una forte assunzione sulla dipendenza di $E(Y)$ sulle $X_1, \ldots, X_p$, ovvero considera una dipendenza lineare su ciascuna delle X_i. Se così fosse, anche solo in parte, l'utilizzo di tali modelli ha il vantaggio di soddisfare tutti gli obiettivi proposti (descrizione, inferenza e predizione).

Pur mantenendo i vantaggi della modellizzazione tramite regressione lineare (additività) senza sottostare a vincoli di restrittivi sulla forma di dipendenza, una buona generalizzazione è l'uso dei modelli additivi

$$Y = \alpha + \sum_{j=1}^{n} f_j(X_j) + \varepsilon$$

dove gli errori, con valore atteso nullo e varianza pari a σ^2, risultano indipendenti dalle variabili indipendenti X_j. Le f_j sono generiche funzioni univariate, una per ogni variabile indipendente X_j. Si può pensare che ogni f_j sia uno *smoother* e che ognuna sia stimata da uno *scatterplot smoothing*.

Siccome la superficie di risposta fornita da un modello additivo è il risultato della somma delle funzioni relative ad ogni variabile indipendente, allora ogni f_j può essere rappresentata singolarmente in quanto non è prevista nessuna correlazione tra variabili nella modellizzazione della variabile dipendente. Ciò ha un costo: i modelli additivi sono un'approssimazione della vera superficie di regressione.

Stimare le funzioni f_j da un modello additivo è come stimare i coefficienti in una regressione lineare, quindi anche l'analisi inferenziale sarà analoga.

Un esempio dei modelli additivi: la regressione logistica

Un esempio semplice ed efficace per capire come i modelli additivi siano la naturale generalizzazione dei modelli lineari è rappresentato dalla regressione logistica.

Sia Y una variabile risposta dicotomica, quindi il suo valore atteso $E(Y|X=x)$ rappresenta la proporzione di successi attesi. L'approccio standard per modellare linearmente dati binari è la regressione logistica che modella il *LOGIT* della probabilità di risposta con la seguente forma lineare

$$LOGIT\{P(X)\} = \log\left\{\frac{P(X)}{1-P(X)}\right\} = X\beta$$

dove $P(X)=pr(Y=1|X)$. Ci sono molte ragioni per cui tale formulazione è molto utilizzata, ma la più importante è che tale modello garantisce che $P(X)$ cade nell'intervallo $[0,1]$ senza nessun vincolo su $X\beta$.

La generalizzazione consiste nel sostituire a $X\beta$ un termine additivo, ovvero

$$LOGIT\{P(X)\} = \log\left\{\frac{P(X)}{1-P(X)}\right\} = \alpha + \sum_{j=1}^{p} f_j(X_j)$$

1.9. Modello additivo generalizzato

Come i modelli additivi sono una generalizzazione dei modelli lineari, quelli additivi generalizzati sono una generalizzazione dei modelli lineari generalizzati. Occorre dunque richiamare i concetti di base dei modelli lineari generalizzati.

Un modello lineare generalizzato è composto da tre componenti: una *random*, una *sistematica* e una funzione *link*. Inoltre si assume che la variabile risposta *Y* abbia una forma esponenziale la cui densità può essere scritta come

$$\rho_Y(y;\theta;\phi) = \exp\left\{\frac{y_\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

dove θ è il parametro naturale e ϕ è il parametro di dispersione. Questa è la componente *random* del modello. Si assume che il valore atteso $\mu=E(Y)$ sia legato alle covariate *X1, ..., Xp* da $g(\mu)=\eta$ dove $\eta=\alpha+X_1\beta_1+\ldots+X_p\beta_p$. La componente *sistematica* del modello è η, anche detto predittore lineare, mentre la funzione *link* è *g*. Si noti che $\mu=b'(\theta)$. Inoltre, una banale funzione *link*, chiamata *link* canonica, è quella funzione tale per cui $\eta=\theta$.

Dati la componente *sistematica*, una funzione *link*, un vettore di *n* osservazioni **y** e *p* corrispondenti predittori $\mathbf{x}_1, \ldots, \mathbf{x}_p$, le stime di massima verosimiglianza di $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^T$ sono ricavabili dalle equazioni, dette funzioni *score*, che hanno la seguente forma

$$\sum_{i}^{n} x_{ij}\left(\frac{\delta\mu_i}{\delta\eta_i}\right) V_i^{-1}(y_i - \mu_i) = 0, \quad j = 0,\ 1,\ldots, p$$

dove $V_i=var(Y_i)$. Per la risoluzioni di queste equazioni è necessario un metodo iterativo, che prende il nome di *Fischer scoring*.

La generalizzazione consiste nel sostituire un predittore additivo al posto del predittore lineare. Ovvero, si supponga di avere una variabile risposta Y con distribuzione

$$\rho_Y(y;\theta;\phi) = \exp\left\{\frac{y_\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

e con media $\mu=E(X_1,\ldots,X_p)$ legata ai predittori con la funzione

$$g(\mu) = \alpha + \sum_{j=1}^{n} f_j(X_J)\,.$$

Si può pensare che ogni f_j sia uno *smoother* e che ognuna sia stimata da uno *scatterplot smoothing*.

Per le soluzioni di α e delle f_j è necessario un algoritmo iterativo, chiamato *local scoring*, che sfrutta gli *smoothing* locali per generalizzare la procedura di *Fischer scoring.*

Si noti che, indicando con Y la variabile risposta e con X quella esplicativa, il modello in cui sia la componente lineare che quella *smoothing* sono significative, ha la forma (semi-parametrica)

$$E(Y) = \beta_0 + \beta_1(X) + f_1(X),$$

dove β_1 è la componente parametrica e f_1 è la componente *smoothing*; il modello in cui solo la componente lineare è significativa ha la forma (parametrica)

$$E(Y) = \beta_0 + \beta_1(X);$$

il modello che ha solo a componente *smoothing* significativa ha la forma (non parametrica)

$$E(Y) = \beta_0 + f_1(X).$$

Si noti, inoltre che i gradi di libertà di uno *smoother* possono non essere interi e la definizione più comunemente usata è che sono pari alla traccia della una matrice che ha lo stesso ruolo della matrice di proiezione nella regressione lineare standard.

APPLICAZIONE AI DATI REALI

1.10. Studi inclusi nel PanC4

Lo studio PanC4 include 12 studi caso-controllo sul tumore del pancreas per un totale di 6.507 casi e 12.890 controlli con informazioni sulle abitudini al fumo [21-24,25 ,26-31].

Le caratteristiche dei 12 studi inclusi nella presente *pooled-analysis* sono mostrate nella **Tabella 1**. Otto studi [21-24,25 ,26,27] sono stati condotti in Nord America, due in Europa [28,29], uno in Cina [30] ed uno era uno studio multicentrico condotto in Canada, Europa e Australia [31]. Il numero di casi e controlli si riferisce ai soggetti inclusi nella presente analisi, e in qualche caso differisce dal numero pubblicato, generalmente per la presenza di dati mancanti nelle variabili più importanti. In tutti gli studi, casi e controlli sono stati intervistati direttamente ad eccezione di 63 soggetti dello studio *Toronto* [27], 474 casi e 332 controlli dello studio *SEARCH* (Surveillance of Environmental Aspects Related to Cancer in Humans) [31] e 155 casi e 150 controlli dello studio *Shanghai* [30] per i quali sono stati intervistati i parenti. Una breve descrizione di ogni studio è data di seguito.

Lo studio *LSU* [dati non pubblicati] include 69 casi incidenti di tumore del pancreas e 158 controlli, con età maggiore dei 20 anni, intervistati tra il 2001 ed il 2006. I controlli sono stati identificati nel database delle patenti di guida del Dipartimento della Motorizzazione della Virginia (USA) e nel database di Medicina Generale e sono stati appaiati ai casi per età, etnia e sesso con un tasso uguale a 2 controlli per ogni caso.

Lo studio *Mayo clinic* [21] include 1.137 casi con tumore del pancreas reclutati consecutivamente tra il 2000 e il 2007 nella Mayo clinic (Rochester, Minnesota). I controlli sono 1.291 soggetti appaiati ai casi per sesso, età, etnia e area di residenza, e reclutati nel reparto di Medicina Generale della Mayo clinic. Il tasso di partecipazione è del 62% per i casi e del 56% per i controlli.

Lo studio *MD Anderson* [22] include 874 casi con tumore del pancreas reclutati dal Centro Gastrointestinale dell'MD Anderson tra il 2000 e il 2006. I controlli esono 790 soggetti appaiati ai casi per sesso, età ed etnia e selezionati tra gli accompagnatori (amici o familiari) dei pazienti che si trovavano in altri centri dello stesso Istituto. Il tasso di partecipazione è del 80% per i casi e dell'84% per i controlli.

Lo studio *NCI* [24] include 493 casi incidenti di carcinoma esocrino del pancreas diagnosticato tra il 1986 e il 1989 tra i residenti delle aree geografiche coperte dal registro tumori di Atlanta (Georgia), di Detroit (Michigan) e del New Jersey. I controlli sono 2.146 pazienti sorteggiati tra la popolazione generale, appaiati ai casi per sesso, età, etnia ed area dello studio. I controlli tra i 30-64 anni di età sono stati sorteggiati seguendo numeri casuali, quelli tra i 65-79 anni di età sono un campione casuale stratificato della popolazione generale di età superiore ai 65 anni che era nelle liste amministrative della sanità. Il tasso di partecipazione varia tra il 75 e il 77% per casi e controlli.

Lo studio *UCSF* (University of California, San Francisco) [25] include 527 casi incidenti di adenocarcinoma esocrino del pancreas identificati tra il 1995 e il 1999 dai registri tumori di sei contee dell'area urbana di San Francisco usando l'identificazione veloce di casi gestita dall'Istituto Tumori della California del Nord. I controlli sono 1.679 soggetti appaiati ai casi per sesso ed età e selezionati dalla popolazione generale usando numeri casuali. I controlli più vecchi dei 65 anni di età provengono dalle liste amministrative della sanità. Il tasso di partecipazione è del 67% per i casi.

Lo studio *Toronto* [27] include 540 casi con adenocarcinoma primario o metastatico del pancreas, patologicamente confermato, identificati dal Registro Tumori dell'Ontario, e 313 controlli selezionati, usando numeri casuali, nelle liste del Ministero della Finanza. Sono stati usati questionari postali amministrati personalmente dai soggetti inclusi nello studio.

Lo studio *Italy* [28] include 322 casi con tumore incidente del pancreas (istologicamente o tomograficamente confermato nel 55% dei casi) identificati tra il 1991 e il 2008 nei maggiori ospedali universitari e non della provincia di Pordenone e di Milano, e 652 controlli ospedalieri appaiati ai casi per sesso, età ed area di residenza. I controlli sono pazienti ricoverati negli stessi ospedali dei casi per condizioni acute. Quelli ricoverati per neoplasie maligne, condizioni collegate al fumo di tabacco e all'uso di alcol o a modificazioni della dieta sono stati esclusi dallo studio.

Lo studio *Milan* [32] include 362 casi con tumore del pancreas istologicamente confermato ricoverati nell'Istituto Nazionale dei Tumori, in vari ospedali universitari e nell'Ospedale Maggiore di Milano tra il 1983 e il 1999, e 1.149 controlli ricoverati negli stessi ospedali dove sono stati identificati i casi per condizioni acute non neoplastiche e non collegate al fumo di tabacco e all'uso di alcol o a modificazioni della dieta. Il tasso di partecipazione è pari a più del 95%. Il questionario usato per la collezione dei dati è stato testato nella sua affidabilità [33].

Lo studio *MSKCC* (Memorial Sloan-Kettering Cancer Center) [23] include 874 casi e 348 controlli senza un precedente tumore. Casi e controlli sono stati intervistati tra il 2003 e il 2008. Sono stati esclusi soggetti che non parlavano la lingua inglese. Il tasso di partecipazione è del 79% per i casi e del 59% per i controlli.

Lo studio *Yale* [26] include 413 casi di tumore del pancreas e 715 controlli selezionati tra i residenti del Connecticut arruolati tra il 2005 e il 2009. I casi sono stati identificati in 30 ospedali; i controlli sono stati selezionati dalla popolazione generale usando numeri casuali. Tutti i soggetti sono stati intervistati di persona e il tasso di partecipazione era del 46% per i casi e del 63% per i controlli.

Lo studio *Shanghai* [30] include 451 casi con tumore del pancreas istologicamente confermato identificati dal Registro Tumori di Shanghai tra il 1990 e il 1993, e 1552 controlli selezionati casualmente tra i residenti della zona urbana di Shanghai e appaiati ai casi per sesso ed età. Il tasso di partecipazione è del 78% per i casi.

Lo studio *SEARCH* [31] è uno studio collaborativo della IARC che include 810 casi e 1.679 controlli collezionati negli anni '80 a Toronto e Montreal in Canada, ad Utrecht in Olanda, Opole in Polonia e Adelaide in Australia, selezionati casualmente dalla popolazione a rischio, e appaiati ai casi per sesso ed età. Il tasso di partecipazione nei vari centri varia tra il 50 e l'80%.

1.11. Metodi

I dataset originali sono stati ristrutturati usando un formato unico deciso sia dai responsabili dei singoli studi sia dai coordinatori del PanC4. Sono stati collezionati dati sulle caratteristiche socio-demografiche, misure antropometriche, consumo di tabacco e di bevande alcoliche, storia di diabete e di pancreatite, storia familiare di tumori nei parenti di primo grado e, per i casi, informazioni sull'istologia e sulla topografia del tumore.

I dati di ogni studio sono stati controllati e si è discusso sulla loro consistenza con i relativi responsabili. E' stata definita una lista delle nuove variabili create e, prima di mettere insieme i dati, sono state preparate e spedite ad ogni responsabile le relative tabelle con le distribuzioni sia delle variabili da analizzare sia delle variabili di confondimento.

Tutti gli studi inclusi nella presente *pooled-analysis* hanno raccolto informazioni sull'abitudine al fumo di sigarette (mai fumatore, ex-fumatore e fumatore corrente), il numero di sigarette fumate al giorno, la durata, l'età all'inizio dell'abitudine al fumo e il tempo da quando un ex-fumatore aveva smesso di fumare o l'età in cui aveva smesso. Le domande del questionario relative alle abitudini al fumo sono simili per ogni studio, anche se differiscono per le specifiche parole usate o per la lingua. Quindi è stata prestata molta attenzione alla loro comparabilità.

Per la presente analisi, un fumatore di sigarette è stato definito come un soggetto che ha fumato almeno 100 sigarette durante la propria vita [21-27,30], oppure più di una sigaretta al giorno per almeno un anno [28,29,31]. Un ex-fumatore è stato definito, in tutti gli studi, come un soggetto che ha smesso di fumare da almeno un anno.

Analisi aggregate standard

Per stimare l'associazione tra differenti misure di esposizione al fumo di sigaretta e il rischio di tumore del pancreas, sono state categorizzate le variabili di interesse nel *dataset* aggregato e sono stati calcolati gli OR, e i rispettivi IC al 95%, attraverso modelli di regressione logistica multivariabile aggiustati per sesso, età, etnia, educazione, indice di massa corporea, consumo di alcol, storia di diabete e storia di pancreatite [8].

Analisi a due stadi

Nel primo stadio, l'associazione tra il fumo di sigaretta e il tumore del pancreas è stata valutata stimando gli (OR) e i relativi intervalli di confidenza (IC) al 95% usando modelli di regressione logistica [15]. Tutti i modelli erano aggiustati per sesso, età (< 50, 50-54, 55-59, 60-64, 65-69, 70-74, ≥ 75 anni), etnia (bianchi non ispanici, neri non ispanici, ispanici, altri), educazione (≤ 8, 9 – 11, 13 – 16, ≥ 17 anni), indice di massa corporea (BMI, <20, 20-<25, 25-<30, ≥ 30 kg/m2), consumo di alcol (0-<1, 1-<6, ≥6 bicchieri al giorno), storia di diabete e storia di pancreatite. Nel secondo stadio le stime riassuntive sono state calcolate usando modelli ad effetti casuali [10].

Analisi a livelli

Sono stati calcolati gli OR e i rispettivi IC al 95% attraverso modelli di regressione logistica a due livelli, considerando il soggetto come primo livello e lo studio come secondo livello [11-14]. Al primo livello (livello-paziente) i modelli sono stati aggiustati per sesso, età, etnia, educazione, indice di massa corporea, consumo di alcol, storia di diabete e storia di pancreatite.

Analisi GAM

Sono state studiate le relazioni tra il numero di sigarette, la durata dell'abitudine al fumo e gli anni dalla cessazione e il rischio di tumore del pancreas in ogni singolo studio incluso nel PanC4 e nel dataset aggregato attraverso modelli additivi generalizzati aggiustati per sesso, età, etnia, educazione, indice di massa corporea, consumo di alcol, storia di diabete, e studio [19].

1.12. Risultati

La **Tabella 2** mostra la distribuzione di 6.507 casi di tumore del pancreas e 12.890 controlli per sesso, età, e altre caratteristiche. La distribuzione del sesso è simile per casi e controlli. Rispetto ai controlli, i casi sono moderatamente più vecchi, più frequentemente di etnia bianca non ispanica, riportano un più alto livello di educazione e di indice di massa corporea, un più alto consumo di alcol e più frequentemente una storia di diabete e di pancreatite.

Analisi aggregata standard

La **Tabella 3** riporta gli OR, calcolati con l'analisi aggregata standard, del tumore del pancreas in accordo con le abitudini al fumo. Rispetto ai non fumatori, l'OR è pari a

1,46 (IC 95%: 1,36-1,57) per i fumatori di sigarette, 1,19 (IC 95%: 1,10-1,29) per gli ex-fumatori e 2,00 (IC 95%: 1,83-2,19) per i fumatori correnti. Gli ORs tendono ad aumentare con l'aumento dei livelli di esposizione. Infatti, l'OR è pari a 1,36 (IC 95%: 1,13-1,63) per meno di 10 sigarette al giorno, 1,69 (IC 95%: 1,47-1,94) per 10-<20 sigarette, 2,28 (IC 95%: 2,00-2,59) per 20-<30 sigarette, 3,14 (95% CI: 2,48-3,94) per 30-<40 sigarette e 3,17 (95% CI: 2,46-4,08) per più di 40 sigarette al giorno. Il rischio aumenta in relazione alla durata fino ai 40 anni (OR=2,31; IC 95%: 2,00-2,67), ma non dopo (per ≥40 anni, OR=2,02; IC 95%: 1,80-2,27). Infine, rispetto ai fumatori attuali, vi è una diminuzione del rischio con l'aumentare degli anni da quando un ex-fumatore ha smesso di fumare fino ad una riduzione pari al 0,46 (IC 95%: 0,40-0,52) per 30 anni o più. Nonostante siano presenti trend significativi in relazione a tutte le misure considerate, si notano relazioni non lineari.

Analisi a due stadi

La **Tabella 4** riporta gli OR, calcolati con l'analisi a due stadi, del tumore del pancreas in accordo con le abitudini al fumo. I risultati puntuali sono simili a quelli ottenuti con l'analisi aggregata standard, mentre le stime intervallari risultano essere più larghe rispetto all'analisi aggregata standard.

La **Figure 1** riporta i forest plot dei differenti livelli di intensità del fumo di sigarette, rispetto ai non fumatori. Si nota che gli studi sono eterogenei tra di loro per tutte le categorie tranne che per i fumatori correnti che fumano meno di 10 sigarette al giorno (*p-value*=0,0952, Figura 1a) o più di 40 sigarette al giorno (*p-value*=0,2815, Figura 1e).

Inoltre, si nota eterogeneità anche per i fumatori che fumano da 20-<30 anni (*p-value*=0,2309, **dati non riportati**), e per gli ex-fumatori che hanno smesso da 10-<15 anni (*p-value*=0,0756) o da 15-<20 anni (*p-value*=0,0739).

Analisi a livelli

La **Tabella 5** riporta gli OR, calcolati con l'analisi a due stadi, del tumore del pancreas in accordo con le abitudini al fumo. I risultati sono uguali a quelli ottenuti con l'analisi aggregata standard.

La **Tabelle 6** riportano gli OR per i differenti livelli di intensità del fumo di sigarette in differenti strati si età, sesso e altre selezionate caratteristiche. Si nota che i rischi sembrano essere più alti per le donne, per livelli moderati, moderati/alti e alti dell'intensità dell'abitudine al fumo (*p-value* per l'interazione=0,004, <0,0001 e 0.017 per 20-<30, 30-<40 e ≥ 40 sigarette al giorno, rispettivamente), per i pazienti con meno di 65 anni per livelli moderati/bassi e moderati (*p-value* per l'interazione=0,005 e 0,003 per 10-<20 e 20-<30 sigarette al giorno, rispettivamente), e per quelli che consumano 0-

<1 bicchieri di alcol al giorno per livelli moderati/bassi (*p-value* per l'interazione=0,016 per 10-<20 e 20-<30 sigarette al giorno). Non si nota nessuna eterogeneità per quanto riguarda l'etnia.

Inoltre, i rischi osservati per la durata e la cessazione risultano essere consistenti nei differenti strati considerati (**dati non riportati**).

Analisi GAM

La **Figura 2** mostra le componenti *smoothing* relative al numero di sigarette fumate al giorno tra i fumatori correnti rispetto ai non fumatori, ottenute attraverso modelli additivi generalizzati, per ogni singolo studio incluso nel PanC4. La componente *smoothing* è significativa nello studio Mayo e ha 0,9 gradi di libertà (**Tabella 7**). Per quasi tutti gli altri studi, invece, la non significatività della componente *smoothing* (o il numero molto piccolo - vicino allo zero - dei gradi di libertà), e la significatività del parametro lineare indica che l'andamento è strettamente lineare. Negli studi LSU e MSKCC nessuna delle due componenti (né *smoothing*, né lineare) è significativa.

Considerando i dati di tutti gli studi in un unico dataset, si nota che la componente *smoothing* del numero di sigarette fumate è significativa, e ha 1.1 gradi di libertà, così come lo è la componente lineare (**Tabella 7, Figura 3a**). La **Figura 3b** mostra l'effetto combinato delle due componenti. Si nota, inoltre, che l'OR aumenta molto velocemente se il numero di sigarette fumate al giorno aumenta da 0 a 25, fino a raggiungere un valore di circa 2, e meno velocemente per più alti consumi, ovvero la pendenza della curva tende a diminuire con l'aumentare del numero di sigarette (**Figura 3c)**.

La componente *smoothing* della durata in anni dell'abitudine al fumo non è significativa nella maggior parte degli studi tranne nello studio Mayo (3.1 gradi di libertà), Milan (3.4 gradi di libertà) e SEARCH (1.4 gradi di libertà) (**dati non riportati**). La componente lineare è significativa per quasi tutti gli studi tranne che nello studio Milan. Negli studi LSU e Toronto nessuna della due componenti è significativa.

Considerando tutti gli studi in un unico in un unico dataset, la componente *smoothing* della durata è significativa (1,7 gradi di libertà) così come lo è la componente lineare (**dati non riportati**). Si nota che l'OR aumenta molto velocemente con l'aumento degli anni da 0 a 25, dai 25 ai 35 anni l'aumento dell'OR è meno veloce (la pendenza della curva tende a diminuire), infine l'OR raggiunge un picco intorno ai 35 anni (**Figura 4)**.

Per quanto riguarda la cessazione dell'abitudine al fumo tra gli ex-fumatori rispetto ai fumatori correnti, la componente *smoothing* è significativa per quasi la metà degli studi (LSU, MD Anderson, MSKCC, UCSF, e SEARCH) e, analogamente, la componente lineare è significativa per quasi tutti gli studi tranne che nello studio LSU. Negli studi Italy e Milan nessuna della due componenti è significativa (**dati non riportati**).

Quando si considerano tutti gli studi in un unico dataset la componente *smoothing* degli anni dalla cessazione è ancora una volta significativa (7,8 gradi di libertà) così come lo è la componente lineare (**dati non riportati**). Si nota un andamento periodico dell'OR che diminuisce comunque con l'aumentare degli anni dalla cessazione fino a raggiungere un valore pari allo 0,2 intorno ai 30 anni (**Figura 5)**.

DISCUSSIONE

Lo studio conferma che il fumo di sigarette è associato ad un aumento di rischio di tumore del pancreas e che il rischio aumenta con l'aumento del numero di sigarette e con la durata, anche se l'aumento non sembra essere di tipo lineare. Inoltre, si nota un aumento di rischio per gli ex-fumatori, e una diminuzione del rischio all'aumentare degli anni dalla cessazione.

La diretta associazione tra il fumo di sigarette e il rischio di tumore del pancreas osservata nella presente *pooled-analysis* è consistente con quella riportata da una meta-analisi di 82 studi di coorte e caso-controllo pubblicati tra il 1950 e il 2007 [3], con quella riportata da una *pooled-analysis* di studi di coorte pubblicata nel 2009 *(International Pancreatic Cancer Cohort Consortium*) [34] su 1.481 casi e 1.539 controlli, che mostra un rischio relativo pari a 1,1 (IC 95%: 0,9-1,3) per gli ex-fumatori e 1,8 (IC 95%: 1,4-2,3) per i fumatori correnti, e con quella mostrata dallo studio di coorte *European Prospective Investigation into Cancer and Nutrition* (EPIC) [35] che include 524 casi di tumore del pancreas e riporta un rischio relativo pari a 1,2 (non significativo) per gli ex-fumatori e un aumento di rischio pari a 1,7 (IC 95%: 1,4-2,2) per i fumatori correnti. Il valore leggermente più elevato del rischio riportato nella presente *pooled-analysis*, rispetto a quelli riportato dagli altri studi, potrebbe essere dovuto al fatto che è stata adottata una migliore classificazione dei fumatori correnti ed ex, cosa che non era possibile fare per la meta-analisi di Iodice e colleghi [3], così come per gli studi prospettici [34,35] dove le abitudini al fumo sono generalmente valutate al tempo dell'arruolamento o al tempo dell'ultima intervista e potevano cambiare negli anni successivi [36].

Il presente studio ha osservato una relazione dose-rischio con il numero di sigarette fumato. Un'analoga relazione è stata osservata nell'*International Pancreatic Cancer Cohort Consortium* [34] anche se gli autori non mostrano stime per i soli fumatori correnti.

Il rischio di tumore del pancreas aumenta in relazione alla durata fino ai 40 anni. Ciò conferma l'importanza di tale fattore temporale [1,34]. Inoltre, i risultati del presente studio confermano che il rischio di tumore del pancreas diminuisce con l'aumento degli anni dalla cessazione [3,34].

I risultati ottenuti con l'analisi aggregata erano uguali a quelli ottenuti con l'analisi a due livelli e simili a quelli ottenuti con l'analisi a due stati.

Le relazioni dose-rischio e quelle relative alle variabili temporali osservate nel presente studio sembrano avere un andamento quadratico per il numero di sigarette e la durata, ed uno periodico per gli anni dalla cessazione. L'andamento non lineare delle misure comunque riflette quello osservato nell'analisi per categorie.

Per quanto riguarda il numero di sigarette fumate al giorno, un andamento di tipo "semi-parametrico" (nel senso che vi è una forte associazione sia lineare che non) si nota nello studio Mayo [21]. Tale studio, condotto negli USA, rappresentava lo studio più grande incluso nel PanC4 e differisce dagli altri studi americani per il fatto di includere controlli ospedalieri. Un andamento "parametrico" (ovvero una forte associazione lineare) si nota nella maggior parte degli altri studi, ad eccezione del piccolo studio non pubblicato LSU (dati non pubblicati) e dello studio americano MSKCC [23], per i quali il numero di sigarette sembra non essere associato al rischio di tumore del pancreas.

Rispetto alla durata, per la maggior parte degli studi condotti negli Stati Uniti [22-26], per lo studio italiano Italy [28], e per quello cinese Shanghai [30] si nota un andamento "parametrico". Per lo studio americano Mayo [21] e lo studio multicentrico SEARCH [31] si nota un andamento "semi-parametrico", e per l'altro studio italiano Milan [29] un andamento "non-parametrico". Per lo studio americano LSU (dati non pubblicati) e lo studio canadese Toronto [27] la durata sembra non essere associata al rischio di tumore del pancreas.

Infine, per quanto riguarda gli anni dalla cessazione all'abitudine al fumo, andamenti: "non-parametrici" sono stati notati nel piccolo studio LSU (dati non pubblicati), "parametrici" negli studi americani Mayo [21], NCI [24] e Yale [26], nello studio canadese Toronto [27] e nello studio cinese Shanghai [30]; "semi-parametrici" negli studi americani MD Anderson [22], MSKCC [23] e UCSF [25]. Gli anni dalla cessazione sembrano non essere associati al rischio di tumore del pancreas nei due studi italiani Italy [28] e Milan [29].

Con l'analisi dell'effetto della dose e delle variabili temporali sul rischio di tumore del pancreas attraverso i modelli additivi generalizzati è stato possibile esplorare l'effetto non-lineare di queste misure senza nessuna costrizione sull'andamento di tali fattori. Inoltre, inserendo i parametri *smoothing* nei modelli, si è ottenuto una stima del rischio più precisa.

Il PanC4 include informazioni dettagliate sul fumo di sigarette per più di 6.500 casi di tumore del pancreas e più di 12.800 controlli e quindi fornisce una opportunità unica per studiare la relazione dose-rischio così come la relazione dei vari fattori temporali, quali la durata e il tempo dalla cessazione. Il presente studio, infatti, include un numero relativamente elevato di forti fumatori e di ex-fumatori. Cosa più importante, le variabili di esposizione, i confondenti e l'*outcome* sono stati standardizzati e lo studio è in grado di valutare quanto l'associazione fosse modificata dall'uso di bevande alcoliche a da altri fattori confondenti quali l'educazione, l'indice di massa corporea e una storia di diabete e di pancreatite.

Sia i controlli ospedalizzati che quelli selezionati dalla popolazione generale possono essere fonte di *selection bias*, per esempio è possibile l'esclusione (o l'inclusione) di malattie correlate al fumo, quindi la conseguente sovra-stima della vera associazione, se si selezionano controlli ospedalizzati, o il basso tasso di partecipazione da parte di soggetti fumatori, quindi la conseguente sotto-stima, se si selezionano i controlli dalla

popolazione generale. Proprio per il fatto che il PanC4 include solo studi caso-controllo, sono possibili anche *recall bias* ed errori di classificazione perché il fumo di sigaretta potrebbe essere sotto-riportato [37], anche se è stato dimostrato che le informazioni sul fumo sono riproducibili negli studi caso-controllo [33] e i nostri risultati sono molto simili a quelli riportati dalla *pooled-analysis* di studi di coorte condotta da Lynch e colleghi [34].

Inoltre, siccome l'inclusione di uno studio nella presente *pooled-analysis* è indipendente dalle pubblicazioni dello studio il *publication bias* non dovrebbe influire sui risultati ottenuti, a differenza delle meta-analisi di dati pubblicati per le quali approssimativamente metà dei risultati potrebbe essere inficiata da tale *bias* [38].

RINGRAZIAMENTI

Ringrazio i Professori **Adriano Decarli**, **Carlo La Vecchia** e **Silvano Milani**, **Paolo Boffetta** (*Mount Sinai School of Medicine, New York, NY, USA; International Prevention Research Institute, Lyon, France*), **Cristina Bosetti** (*Istituto di Ricerche Farmacologiche ''Mario Negri'', Milan, Italy*), **Eric J Duell** (*Catalan Institute of Oncology, Barcelona, Spain*) e tutti gli altri componenti del **PanC4**, in particolare P. A. Baghurst (*Women's and Children's Hospital, Adelaide, Australia*), W. R. Bamlet (*Mayo Clinic, Rochester, MN, USA*), P. Bertuccio (*Istituto di Ricerche Farmacologiche ''Mario Negri''; University of Milan, Milan, Italy*), P.M. Bracci (*University of California, San Francisco, CA, USA*), H. B. Bueno-de-Mesquita (*National Institute for Public Health and the Environment, Bilthoven; University Medical Center Utrecht, Utrecht, The Netherlands*), M. Cotterchio (*Dalla Lana School of Public Health, University of Toronto; Cancer Care Ontario, Toronto, Canada*), E. Fontham (*Louisiana State University School of Public Health, New Orleans, LA, USA*), S. Gallinger (*Toronto General Hospital, Toronto, Canada), Y. T. Gao (Shanghai Cancer Institute, Shanghai, China*), P. Ghadirian (*Montreal University Hospital Research Centre, Montreal, Canada*), M. Hassan (*The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA*), E. A. Holly (*University of California, San Francisco, CA, USA*), B. T. Ji (*National Cancer Institute, Bethesda, MD, USA*), R. C. Kurtz (*Memorial Sloan–Kettering Cancer Center, New York, NY, USA*), D. Li (*The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA*), P. Maisonneuve (*European Institute of Oncology, Milan, Italy*), A. B. Miller (*Dalla Lana School of Public Health, University of Toronto, Toronto, Canada*), E. Negri (*Istituto di Ricerche Farmacologiche ''Mario Negri'', Milan, Italy*), S. H. Olson (*Memorial Sloan–Kettering Cancer Center, New York, NY, USA*), G. Petersen (*Mayo Clinic, Rochester, MN, USA*), J. Polesel (*Centro di Riferimento Oncologico-National Cancer Institute, Aviano, Italy*), H. A. Risch (*Yale University School of Medicine, New Haven, CT, USA*), D. T. Silverman (*National Cancer Institute, Bethesda, MD, USA) J. Su (National Cancer Institute, Bethesda, MD, USA*), R. Talamini (*Centro di Riferimento Oncologico-National Cancer Institute, Aviano, Italy*), H. Yu (*Memorial Sloan–Kettering Cancer Center, New York, NY, USA*), W. Zatonski (*Cancer Center and Institute of Oncology, Warsaw, Poland*).

BIBLIOGRAFIA

1. IARC. *IARC Monographs on the evaluation of carcinogenic risks to humans. Vol. 83. Tobacco smoke and involuntary smoking*. Lyon: International Agency for Research on Cancer; 2004.

2. Secretan B, Straif K, Baan R, et al. A review of human carcinogens--Part E: tobacco, areca nut, alcohol, coal smoke, and salted fish. *Lancet Oncol.* Nov 2009;10(11):1033-1034.

3. Iodice S, Gandini S, Maisonneuve P, Lowenfels AB. Tobacco and the risk of pancreatic cancer: a review and meta-analysis. *Langenbecks Arch Surg.* Jul 2008;393(4):535-545.

4. The Pancreatic Cancer Case Control Consortium (PANC4). *Available at: http://panc4.org/.* 2007.

5. Bertuccio P, La Vecchia C, Silverman DT, et al. Cigar and pipe smoking, smokeless tobacco use and pancreatic cancer: an analysis from the International Pancreatic Cancer Case-Control Consortium (PanC4). *Ann Oncol.* Jun 2011;22(6):1420-1426.

6. Lucenteforte E, La Vecchia C, Silverman D, et al. Alcohol consumption and pancreatic cancer: a pooled analysis in the International Pancreatic Cancer Case-Control Consortium (PanC4). *Ann Oncol.* May 2 2011.

7. Bosetti C, Lucenteforte E, Silverman DT, et al. Cigarette smoking and pancreatic cancer: an analysis from the International Pancreatic Cancer case-control Consortium (PanC4). *Ann Oncol.* 2011;in press.

8. Breslow NE, Day NE. *Statistical methods in cancer research. Vol. I. The analysis of case-control studies. IARC Sci Publ No. 32.* Vol IARC Sci Publ No. 32. Lyon, France: IARC; 1980.

9. Smith-Warner SA, Spiegelman D, Ritz J, et al. Methods for pooling results of epidemiologic studies: the Pooling Project of Prospective Studies of Diet and Cancer. *Am J Epidemiol.* Jun 1 2006;163(11):1053-1064.

10. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials.* Sep 1986;7(3):177-188.

11. Breslow NE, Clayton DG. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association.* Mar 1993;88(421):9-25.

12. Longford NT. Random Coefficient Models. *Oxford: Clarendon Press.* 1993.

13. Snijders TAB, Bosker RJ. Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling. *London: Sage.* 1999.

14. Goldstein H. Multilevel Statistical Models (3rd Edition). *London: Edward Arnold: New York, Halstead Press.* 2003.

15. Breslow NE, Day NE. *Statistical methods in cancer research. Volume I - The analysis of case-control studies.* Vol IARC Sci Publ no. 32. Lyon: IARC; 1980.

16. Breslow NE, Day NE. *Statistical methods in cancer research. Volume I - The analysis of case-control studies. IARC Sci Publ no. 32.* Vol IARC Sci Publ no. 32. Lyon: IARC; 1980.

17. Hosmer DW, Lemeshow S. Applied Logistic Regression (2nd Edition). *New York: John Wiley & Sons, Inc.* 2000.

18. Demidenko E. Mixed Models: Theory and Applications. Wiley Series in Probability and Statistics. . *Hoboken, New Jersey: John Wiley & Sons, Inc.* 2004.

19. Hastie T, Tibshirani R. *Generalized additive models*. 1st ed. London ; New York: Chapman and Hall; 1990.

20. Epanechnikov VA. Nonparametric estimation of a multivariate probability density. *Theor Prob Appl.* 1969;14:153-158.

21. McWilliams RR, Bamlet WR, de Andrade M, Rider DN, Cunningham JM, Petersen GM. Nucleotide excision repair pathway polymorphisms and pancreatic cancer risk: evidence for role of MMS19L. *Cancer Epidemiol Biomarkers Prev.* Apr 2009;18(4):1295-1302.

22. Hassan MM, Bondy ML, Wolff RA, et al. Risk factors for pancreatic cancer: case-control study. *Am J Gastroenterol.* Dec 2007;102(12):2696-2707.

23. Olson SH, Orlow I, Simon J, et al. Allergies, variants in IL-4 and IL-4R alpha genes, and risk of pancreatic cancer. *Cancer Detect Prev.* 2007;31(5):345-351.

24. Silverman DT, Dunn JA, Hoover RN, et al. Cigarette smoking and pancreas cancer: a case-control study based on direct interviews. *J Natl Cancer Inst.* Oct 19 1994;86(20):1510-1516.

25. Chan JM, Wang F, Holly EA. Sweets, sweetened beverages, and risk of pancreatic cancer in a large population-based case-control study. *Cancer Causes Control.* Aug 2009;20(6):835-846.

26. Risch HA, Yu H, Lu L, Kidd MS. ABO blood group, Helicobacter pylori seropositivity, and risk of pancreatic cancer: a case-control study. *J Natl Cancer Inst.* Apr 7 2010;102(7):502-505.

27. Anderson LN, Cotterchio M, Gallinger S. Lifestyle, dietary, and medical history factors associated with pancreatic cancer risk in Ontario, Canada. *Cancer Causes Control.* Aug 2009;20(6):825-834.

28. Talamini R, Polesel J, Gallus S, et al. Tobacco smoking, alcohol consumption and pancreatic cancer risk: A case-control study in Italy. *Eur J Cancer.* Sep 24 2010;46(2):370-376.

29. Fernandez E, La Vecchia C, Decarli A. Attributable risks for pancreatic cancer in northern Italy. *Cancer Epidemiol Biomarkers Prev.* Jan 1996;5(1):23-27.

30. Ji BT, Chow WH, Dai Q, et al. Cigarette smoking and alcohol consumption and the risk of pancreatic cancer: a case-control study in Shanghai, China. *Cancer Causes Control.* Jul 1995;6(4):369-376.

31. Boyle P, Maisonneuve P, Bueno de Mesquita B, et al. Cigarette smoking and pancreas cancer: a case control study of the search programme of the IARC. *Int J Cancer.* Jul 3 1996;67(1):63-71.

32. Tavani A, Pregnolato A, Negri E, La Vecchia C. Alcohol consumption and risk of pancreatic cancer. *Nutr Cancer.* 1997;27(2):157-161.

33. D' Avanzo B, La Vecchia C, Katsouyanni K, Negri E, Trichopoulos D. Reliability of information on cigarette smoking and beverage consumption provided by hospital controls. *Epidemiology.* May 1996;7(3):312-315.

34. Lynch SM, Vrieling A, Lubin JH, et al. Cigarette smoking and pancreatic cancer: a pooled analysis from the pancreatic cancer cohort consortium. *Am J Epidemiol.* Aug 15 2009;170(4):403-413.

35. Vrieling A, Bueno-de-Mesquita HB, Boshuizen HC, et al. Cigarette smoking, environmental tobacco smoke exposure and pancreatic cancer risk in the European Prospective Investigation into Cancer and Nutrition. *Int J Cancer.* May 15 2010;126(10):2394-2403.

36. Bosetti C, Negri E, Tavani A, Santoro L, La Vecchia C. Smoking and acute myocardial infarction among women and men: A case-control study in Italy. *Prev Med.* Nov 1999;29(5):343-348.

37. Rebagliato M. Validation of self reported smoking. *J Epidemiol Community Health.* Mar 2002;56(3):163-164.

38. Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. Empirical assessment of effect of publication bias on meta-analyses. *BMJ.* Jun 10 2000;320(7249):1574-1577.

Tabella 1. Descrizione riassuntiva degli studi inclusi nel PanC4 in accordo con il paese di provenienza dello studio.

Paese Studio	Periodo dello studio	Casi			Controlli		
		Uomini: Donne	Età - *range* (mediana)	Provenienza	Uomini: Donne	Età - *range* (mediana)	Provenienza
Nord America							
Louisiana							
LSU (dati non pubblicati)	2001-2006	33: 36	32-86 (68)	Registro tumori	78 : 80	33-90 (67)	Popolazione
Minnesota							
Mayo [21]	2000-2007	624 : 513	29-92 (68)	Ospedale	626 : 665	29-97 (70)	Ospedale
Texas							
MD Anderson [22]	2000-2006	539 : 335	28-87 (63)	Ospedale	495 : 295	31-84 (61)	Ospedale (visitatori)
New York							
MSKCC [23]	2003-2008	264:245	32-89 (64)	Ospedale	142:206	27-84 (58)	Ospedale (visitatori)
Georgia, Michigan, New Jersey							
NCI [24]	1986-1989	250 : 243	32-79 (63)	Registro tumori	1.364 : 782	30-81 (62)	Popolazione
California							
UCSF [25]	1995-1999	287 : 240	32-85 (65)	Registro tumori	879 : 818	32-85 (66)	Popolazione
Connecticut							
Yale [26]	2005-2009	238:175	36-84 (68)	Misto	404:311	35-84 (68)	Popolazione
Canada							
Toronto [27]	2003-2007	302 : 238	20-89 (65)	Registro tumori	177 : 136	40-79 (67)	Popolazione
Europa							
Italy [28]	1991-2008	174 : 148	34-80 (63)	Ospedale	348 : 304	34-80 (63)	Ospedale
Milan [29]	1983-1999	229 : 133	17-86 (60)	Ospedale	1.140 : 409	21-84 (56)	Ospedale
Asia							
Shanghai [30]	1990-1993	264 : 187	31-74 (64)	Registro tumori	851 : 701	30-74 (62)	Popolazione
Internazionale							
Canada, Europa, Australia							
SEARCH [31]	1983-1989	447 : 363	32-86 (65)	Misto	858 : 821	28-87 (65)	Popolazione

LSU, Louisiana School of Public Health; MSKCC, Memorial Sloan-Kettering Cancer Center; NCI, National Cancer Institute; SEARCH, Surveillance of Environmental Aspects Related to Cancer in Humans; UCSF, University of California-San Francisco.

Tabella 2. Distribuzione di 6.507 casi di tumore del pancreas e 12.890 controlli in accordo con sesso, età ed altre caratteristiche (PanC4).

Caratteristiche	Casi		Controlli	
	No.	(%)	No.	(%)
Sesso				
Uomini	3.651	(56,1)	7.362	(57,1)
Donne	2.856	(43,9)	5.528	(42,9)
Età (anni)				
< 50	596	(9,2)	1.770	(13,7)
50 – 54	602	(9,2)	1.385	(10,7)
55 – 59	905	(13,9)	1.816	(14,1)
60 – 64	1.091	(16,8)	1.983	(15,4)
65 – 69	1.148	(17,6)	2.146	(16,6)
70 – 75	1.084	(16,7)	2.041	(15,8)
≥ 75	1.081	(16,6)	1.749	(13,6)
Etnia				
Bianchi non ispanici	5.409	(83,1)	9.478	(73,5)
Neri non ispanici	356	(5,5)	1.119	(8,7)
Ispanici	115	(1,8)	220	(1,7)
altri	622	(9,6)	1.761	(13,7)
Educazione (anni)				
≤ 8	1.291	(19,8)	3.570	(27,7)
9 – 11	823	(12,6)	1.624	(12,6)
12	1.349	(20,7)	2.186	(17,0)
13 – 16	1.991	(30,6)	3.588	(27,8)
≥ 17	1.006	(15,5)	1.835	(14,2)
Indice di massa corporea (kg/m^2)				
< 20	462	(7,1)	1.111	(8,6)
20 - <25	2.396	(36,8)	5.658	(43,9)
25 - <30	2.363	(36,3)	4.473	(34,7)
≥ 30	1.201	(18,5)	1.488	(11,5)
Consumo di alcol (bicchieri/giorno)[a]				
0 - <1	3.855	(59,2)	7.478	(58,0)
1 - <6	1.432	(22,0)	3.563	(27,6)
≥ 6	697	(10,7)	1.492	(11,6)
Storia di diabete				
No	5.052	(77,6)	11.710	(90,8)
Si	1.378	(21,2)	1.109	(8,6)
Storia di pancreatite[b]				
No	4.674	(71,8)	10.703	(83,0)
Si	313	(4,8)	112	(0,9)

LSU, Louisiana School of Public Health; MSKCC, Memorial Sloan-Kettering Cancer Center; NCI, National Cancer Institute; SEARCH, Surveillance of Environmental Aspects Related to Cancer in Humans; UCSF, University of California, San Francisco.

[a] Non era disponibile nessuna informazione nello studio MSKCC. [b] Non era disponibile nessuna informazione nello studio Italy and Mayo.

Tabella 3: analisi aggregata. Distribuzione di 6.507 casi di tumore del pancreas e 12.890 controlli, *odds ratios* (ORs) e intervalli di confidenza (IC) al 95% in accordo con le abitudini al fumo di sigaretta (PanC4).

	Casi		Controlli		OR[a] (IC 95%)
	No.	(%)	No.	(%)	
Non fumatori	2.373	(36,5)	5.557	(43,1)	1[b]
Fumatori di sigarette	3.962	(60,9)	6.980	(54,2)	1,46 (1,36-1,57)
Ex-fumatori	2.327	(35,8)	4.214	(32,7)	1,19 (1,10-1,29)
Fumatori correnti	1.635	(25,1)	2.766	(21,5)	2,00 (1,83-2,19)
Fumatori di altro tipo di tabacco	164	(2,5)	336	(2,6)	1,23 (0,99-1,53)
valori mancanti	*8*	*(0,1)*	*17*	*(0,1)*	
Intensità (sigarette/giorno)[c]					
<10	203	(3,1)	503	(3,9)	1,36 (1,13-1,63)
10-<20	438	(6,7)	868	(6,7)	1,69 (1,47-1,94)
20-<30	631	(9,7)	957	(7,4)	2,28 (2,00-2,59)
30-<40	175	(2,7)	205	(1,6)	3,14 (2,48-3,94)
≥40	140	(2,1)	170	(1,3)	3,17 (2,46-4,08)
valori mancanti	*48*	*(0,7)*	*63*	*(0,5)*	
p-value per il trend					*<0,0001*
Durata (anni)[c]					
<20	92	(1,4)	301	(2,3)	1,35 (1,04-1,74)
20-<30	219	(3,4)	501	(3,9)	1,89 (1,56-2,28)
30-<40	465	(7,1)	684	(5,3)	2,31 (2,00-2,67)
≥40	837	(12,9)	1227	(9,5)	2,02 (1,80-2,27)
valori mancanti	*22*	*(0,3)*	*35*	(0,4)	
p-value per il trend					*<0,0001*
Anni da quando un ex-fumatore ha smesso di fumare					
Fumatori correnti di sigarette	1.635	(25,2)	2.766	(21,5)	1[b]
1-<10	640	(9,8)	1032	(8,0)	0,86 (0,76-0,98)
10-<15	301	(4,6)	525	(4,1)	0,68 (0,57-0,80)
15-<20	267	(4,1)	503	(3,9)	0,54 (0,45-0,64)
20-<30	469	(7,2)	963	(7,5)	0,48 (0,41-0,55)
≥30	616	(9,5)	1136	(8,8)	0,46 (0,40-0,52)
valori mancanti	*33*	*(0,5)*	*55*	*(0,4)*	
p-value per il trend					*<0,0001*

[a] ORs calcolati usando modelli di regressione logistica aggiustati per età, sesso, etnia, educazione, indice di massa corporea, storia di diabete, storia di pancreatite, consumo di alcol e studio. [b] Categoria di riferimento. [c] Solo fumatori correnti.

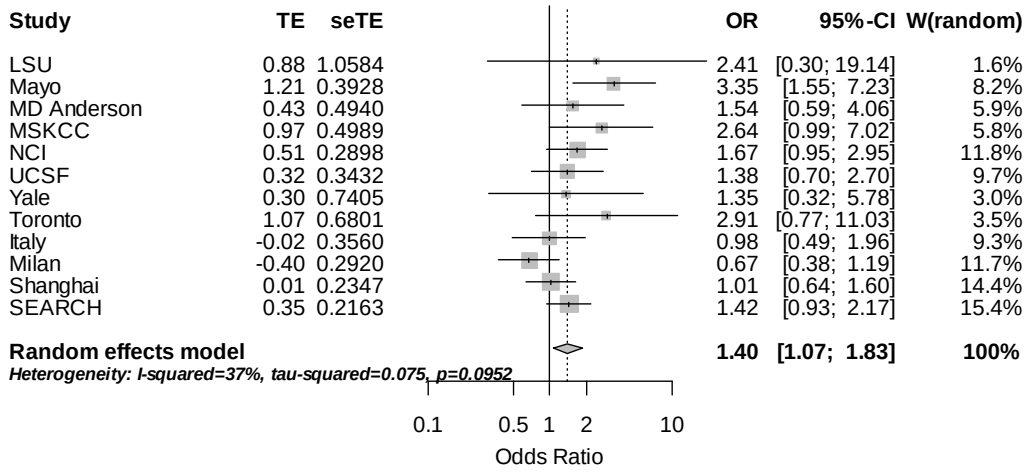
Tabella 4: analisi a due stadi. Distribuzione di 6.507 casi di tumore del pancreas e 12.890 controlli, *odds ratios* (ORs) e intervalli di confidenza (IC) al 95% in accordo con le abitudini al fumo di sigarette (PanC4).

	Casi		Controlli		OR[a] (IC 95%)
	No.	(%)	No.	(%)	
Non fumatori	2.373	(36,5)	5.557	(43,1)	1[b]
Fumatori di sigarette	3.962	(60,9)	6.980	(54,2)	1,40 (1,23-1,58)
Ex-fumatori	2.327	(35,8)	4.214	(32,7)	1,17 (1,02-1,34)
Fumatori correnti	1.635	(25,1)	2.766	(21,5)	2,19 (1,70-2,82)
Fumatori di altro tipo di tabacco	164	(2,5)	336	(2,6)	1,16 (0,80-1,70)
valori mancanti	*8*	*(0,1)*	*17*	*(0,1)*	
Intensità (sigarette/giorno)[c]					
<10	203	(3,1)	503	(3,9)	1,40 (1,07-1,83)
10-<20	438	(6,7)	868	(6,7)	1,77 (1,32-2,38)
20-<30	631	(9,7)	957	(7,4)	2,41 (1,85-3,13)
30-<40	175	(2,7)	205	(1,6)	3,06 (1,91-4,91)
≥40	140	(2,1)	170	(1,3)	3,51 (2,54-4,85)
valori mancanti	*48*	*(0,7)*	*63*	*(0,5)*	
p-value per il trend					*<0,0001*
Durata (anni)[c]					
<20	92	(1,4)	301	(2,3)	1,47 (0,89-2,43)
20-<30	219	(3,4)	501	(3,9)	1,86 (1,45-2,38)
30-<40	465	(7,1)	684	(5,3)	2,41 (1,91-3,05)
≥40	837	(12,9)	1227	(9,5)	2,08 (1,57-2,76)
valori mancanti	*22*	*(0,3)*	*35*	(0,4)	
p-value per il trend					*0,2309*
Anni da quando un ex-fumatore ha smesso di fumare					
Fumatori correnti di sigarette	1.635	(25,2)	2.766	(21,5)	1[b]
1-<10	640	(9,8)	1032	(8,0)	0,74 (0,57-0,96)
10-<15	301	(4,6)	525	(4,1)	0,63 (0,49-0,80)
15-<20	267	(4,1)	503	(3,9)	0,46 (0,35-0,60)
20-<30	469	(7,2)	963	(7,5)	0,41 (0,30-0,56)
≥30	616	(9,5)	1136	(8,8)	0,42 (0,29-0,60)
valori mancanti	*33*	*(0,5)*	*55*	*(0,4)*	
p-value per il trend					*0,0076*

[a] ORs riassuntivi calcolati usando modelli ed effetti casuali. Gli ORs studio-specifico calcolati usando modelli di regressione logistica aggiustati per età, sesso, etnia, educazione, indice di massa corporea, storia di diabete, storia di pancreatite, consumo di alcol e studio. [b] Categoria di riferimento. [c] Solo fumatori correnti.

Figura 1. *Forest plot* del rischio di tumore del pancreas per differenti livelli di intensità del fumo di sigaretta, rispetto ai non fumatori, (PanC4).

a. <10 sigarette al giorno



b. 10-20 sigarette al giorno

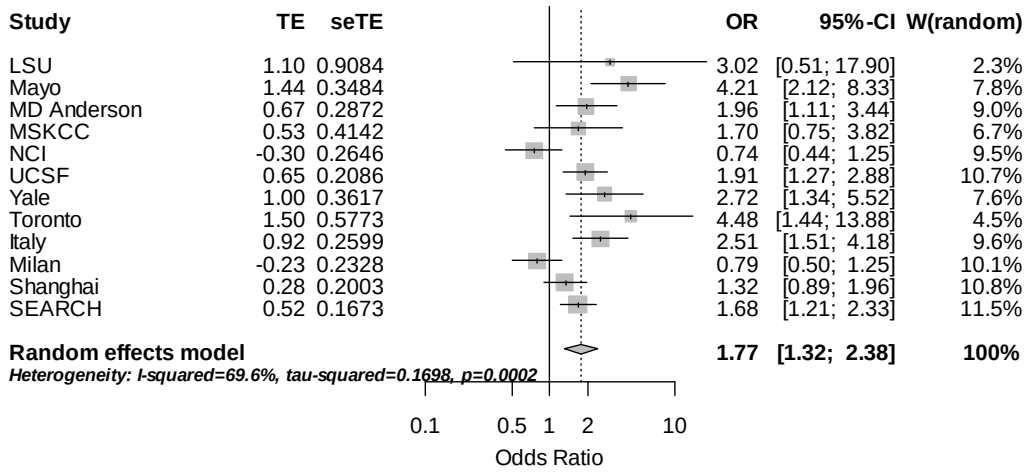
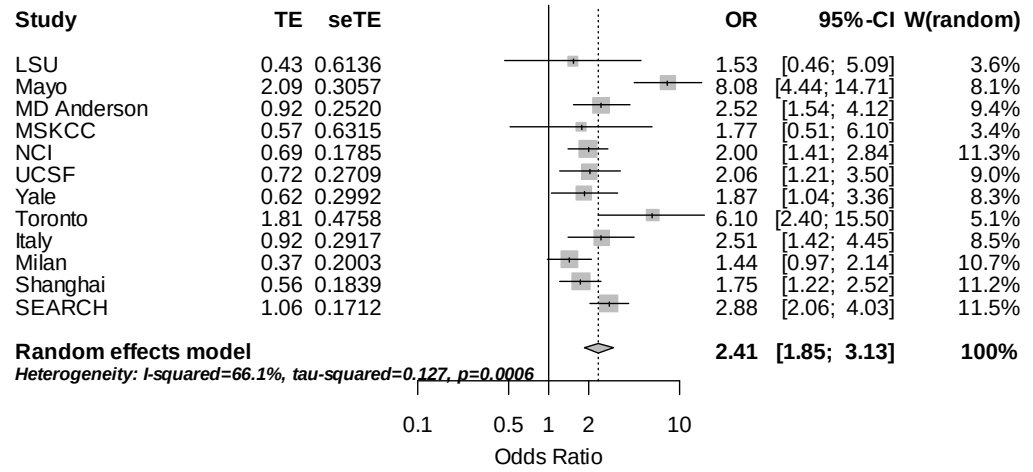


Figura 1. continua

Figura 1. continuava

c. 20-30 sigarette al giorno



d. 30-40 sigarette al giorno

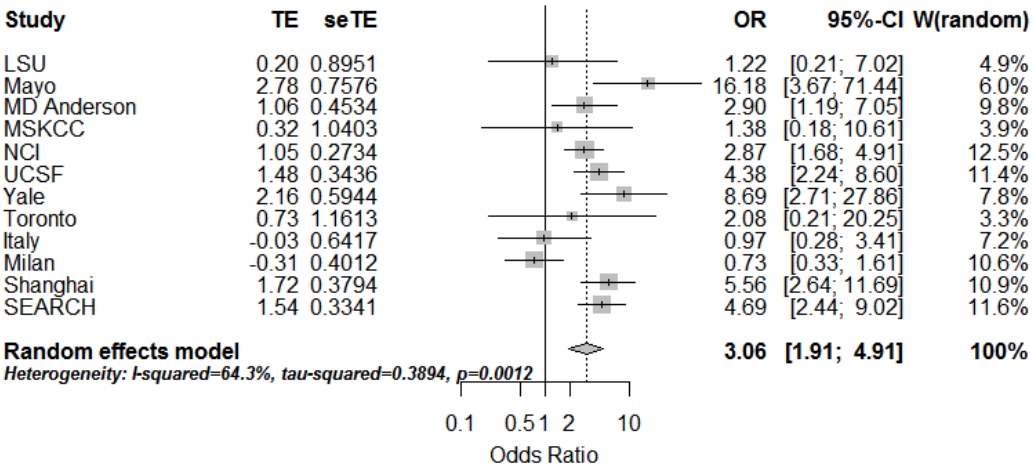
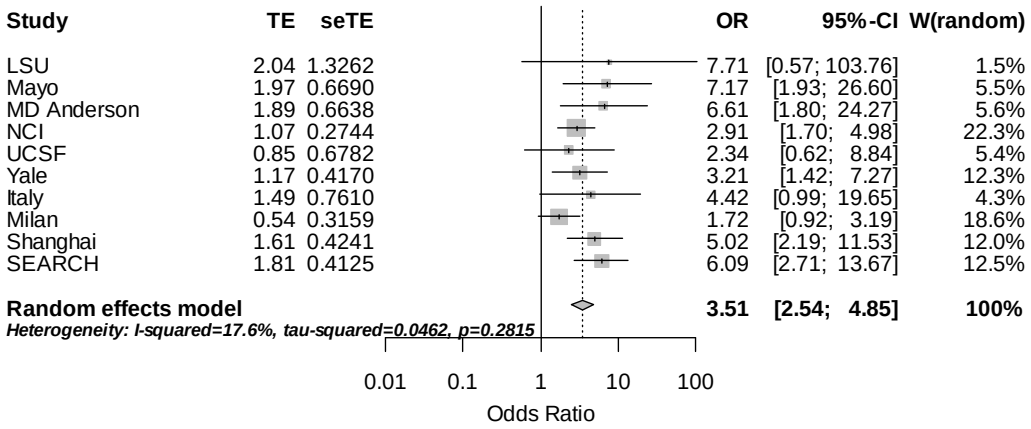


Figura 1. continua

Figura 1. continuava

e. ≥40 sigarette al giorno



LSU, Louisiana School of Public Health; MSKCC, Memorial Sloan-Kettering Cancer Center; NCI, National Cancer Institute; SEARCH, Surveillance of Environmental Aspects Related to Cancer in Humans; UCSF, University of California, San Francisco.

Tabella 5: analisi a due livelli. Distribuzione di 6.507 casi di tumore del pancreas e 12.890 controlli, *odds ratios* (ORs) e intervalli di confidenza (IC) al 95% in accordo con le abitudini al fumo di sigarette (PanC4).

	Casi		Controlli		OR[a] (IC 95%)
	No.	(%)	No.	(%)	
Non fumatori	2.373	(36,5)	5.557	(43,1)	1[b]
Fumatori di sigarette	3.962	(60,9)	6.980	(54,2)	1,46 (1,36-1,57)
Ex-fumatori	2.327	(35,8)	4.214	(32,7)	1,19 (1,10-1,29)
Fumatori correnti	1.635	(25,1)	2.766	(21,5)	2,00 (1,82-2,19)
Fumatori di altro tipo di tabacco	164	(2,5)	336	(2,6)	1,23 (0,99-1,53)
valori mancanti	*8*	*(0,1)*	*17*	*(0,1)*	
Intensità (sigarette/giorno)[c]					
<10	203	(3,1)	503	(3,9)	1,36 (1,13-1,63
10-<20	438	(6,7)	868	(6,7)	1,69 (1,47-1,94)
20-<30	631	(9,7)	957	(7,4)	2,27 (2,00-2,58)
30-<40	175	(2,7)	205	(1,6)	3,13 (2,49-3,93)
≥40	140	(2,1)	170	(1,3)	3,16 (2,46-4,07)
valori mancanti	*48*	*(0,7)*	*63*	*(0,5)*	
p-value per il trend					*0,002*
Durata (anni)[c]					
<20	92	(1,4)	301	(2,3)	1,34 (1,04-1,74)
20-<30	219	(3,4)	501	(3,9)	1,88 (1,56-2,27)
30-<40	465	(7,1)	684	(5,3)	2,30 (1,99-2,66)
≥40	837	(12,9)	1227	(9,5)	2,02 (1,80-2,27)
valori mancanti	*22*	*(0,3)*	*35*	(0,4)	
p-value per il trend					*<0,0001*
Anni da quando un ex-fumatore ha smesso di fumare					
Fumatori correnti di sigarette	1.635	(25,2)	2.766	(21,5)	1[b]
1-<10	640	(9,8)	1032	(8,0)	0,86 (0,76-0,98)
10-<15	301	(4,6)	525	(4,1)	0,68 (0,57-0,80)
15-<20	267	(4,1)	503	(3,9)	0,54 (0,45-0,64)
20-<30	469	(7,2)	963	(7,5)	0,48 (0,41-0,55)
≥30	616	(9,5)	1136	(8,8)	0,46 (0,40-0,52)
valori mancanti	*33*	*(0,5)*	*55*	*(0,4)*	
p-value per il trend					*<0,0001*

[a] ORs calcolati usando modelli di regressione logistica a due livelli, considerando il soggetto come primo livello e lo studio come secondo livello. Il livello paziente è aggiustato per età, sesso, etnia, educazione, indice di massa corporea, storia di diabete, storia di pancreatite e consumo di alcol. [b] Categoria di riferimento. [c] Solo fumatori correnti.

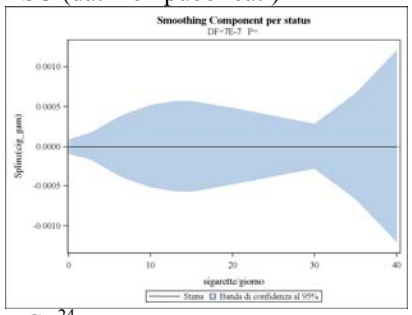
Tabella 6: Analisi a due livelli stratificata. *Odds ratios* (ORs) e intervalli di confidenza (IC) al 95% del tumore del pancreas in accordo ai differenti livelli di intensità del fumo di sigaretta in strati di alcune caratteristiche (PanC4).

	Non-fumatori	**Fumatori correnti di sigarette**									
		<10 sig/giorno		**10-<20 sig/giorno**		**20-<30 sig/giorno**		**30-<40 sig/giorno**		**≥ 40 sig/giorno**	
	ca:co	**ca:co**	**OR (95% CI)**	**ca:co**	**OR (95% CI)**	**ca:co**	**OR (95% CI)**	**ca:co**	**OR (95% CI)**	**ca:co**	**OR (95% CI)**
Totale	2373:5557	203:503	1,36 (1,13-1,63)	438:868	1,69 (1,47-1,94)	631:957	2,27 (2,00-2,58)	175:205	3,13 (2,49-3,93)	140:170	3,16 (2,46-4,07)
Sesso											
Uomini	916:2105	86:291	1,18 (0,90-1,56)	247:574	1,58 (1,31-1,90)	411:771	2,02 (1,71 -2,37)	126:168	1,16 (0,88-1,52)	106:154	2,78 (2,08-3,71)
Donne	1457:3452	117:212	1,55 (1,21-1,99)	191:294	1,81 (1,47-2,23)	220:186	3,03 (2,42-3,80)	49:37	4,23 (2,67-6,70)	34:26	6,48 (3,45-12,18)
(p-value per l'interazione)			*(0,147)*		*(0,3364)*		*(0,004)*		*(<0,0001)*		*(0,017)*
Età (anni)											
< 65	1040:3044	114:302	1,48 (1,16-1,90)	296:561	2,07 (1,74-2,57)	434:656	2,76 (3,35-3,25)	121:158	3,28 (2,49-4,32)	97:129	3,31 (2,43-4,51)
≥ 65	1333:2513	89:201	1,24 (0,94-1,63)	142:307	1,25 (0,99-1,56)	197:301	1,68 (1,36-2,08)	54:47	3,23 (2,11-4,93)	43:41	3,38 (2,14-5,35)
(p-value per l'interazione)			*(0,3361)*		*(0,005)*		*(0,003)*		*(0,949)*		*(0,938)*
Etnia											
Bianchi non ispanici	1918:4041	145:299	1,46 (1,17-1,82)	347:574	1,84 (1,57-2,16)	487:603	2,40 (2,07-2,78)	146:174	2,83 (2,20-3,63)	116:140	2,95 (2,23-3,90)
Altri	455:1516	58:204	1,14 (0,81-1,59)	91:294	1,32 (0,99-1,76)	144:354	2,03 (1,56-2,64)	29:31	4,54 (2,61-7,91)	24:30	3,82 (2,10-6,94)
(p-value per l'interazione)			*(0,222)*		*(0,047)*		*(0,277)*		*(0,127)*		*(0,440)*
Consumo di alcol [c] (bicchieri/giorno)											
0 - <1	1679:4080	117:277	1,37 (1,08-1,75)	242:402	1,97 (1,63-2,37)	316:398	2,47 (2,07-2,95)	83:74	3,81 (2,69-5,41)	60:45	4,50 (2,94.6,87)
1 - <6	375:1024	51:159	1,12 (0,78-1,61)	111:300	1,27 (0,97-1,66)	185:321	1,99 (1,56:2,54)	48:66	2,70 (1,77-4,12)	31:49	2,54 (1,5-4,17)
≥ 6	89:286	17:59	1,30 (0,69-2,44)	59:155	1,32 (0,87-1,98)	117:23	2,07 (1,45-2,96)	41:63	2,43 (1,48-4,00)	48:76	2,64 (1,64-4,25)
(p-value per l'interazione)			*(0,652)*		*(0,016)*		*(0,320)*		*(0,262)*		*(0,141)*

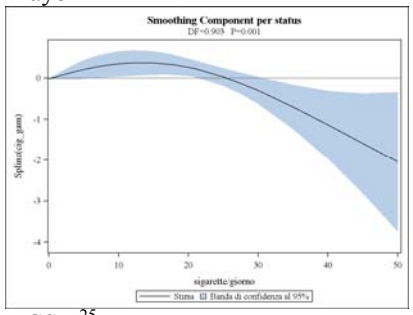
[a] ORs calcolati usando modelli di regressione logistica a due livelli, considerando il soggetto come primo livello e lo studio come secondo livello. Il livello paziente è aggiustato per età, sesso, etnia, educazione, indice di massa corporea, storia di diabete, storia di pancreatite e consumo di alcol. [b] Nessuna informazione nello studio MSKCC.

Figura 2. Componente *smoothing* del numero di sigarette al giorno tra i fumatori correnti, rispetto ai non fumatori, per ogni studio incluso nel PanC4.

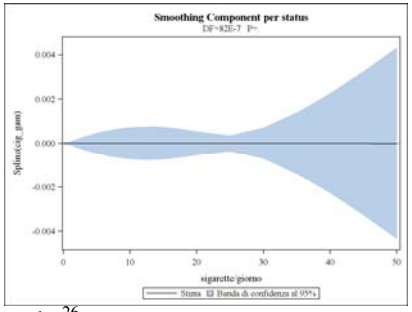
LSU (dati non pubblicati)



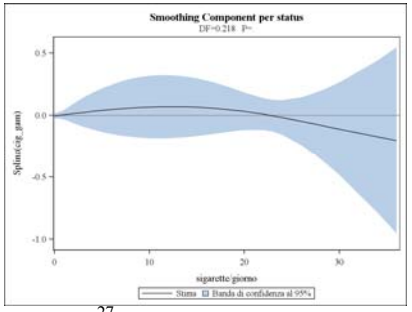
Mayo [21]



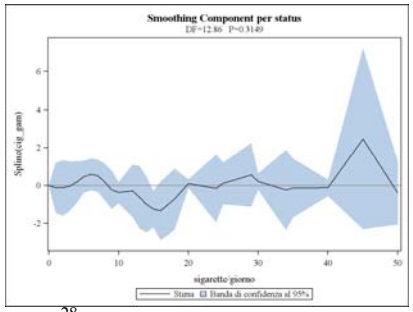
MD Anderson [22]



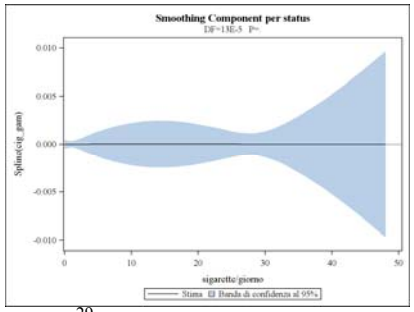
MSKCC [23]



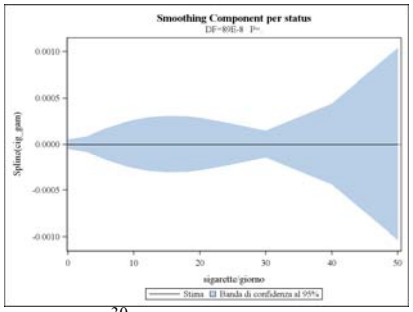
NCI [24]



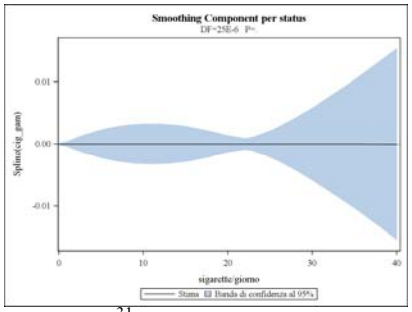
UCSF [25]



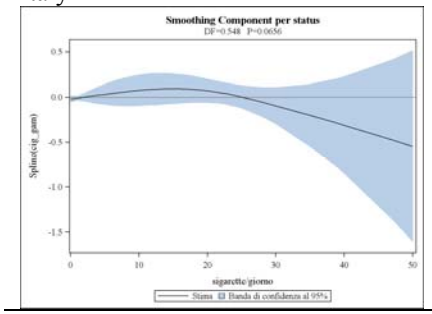
Yale [26]



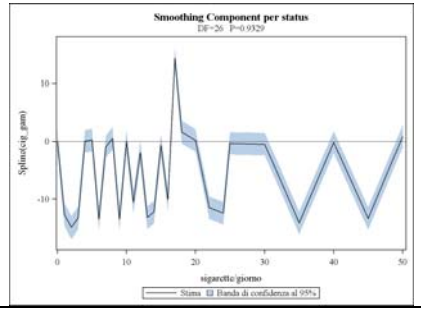
Toronto [27]



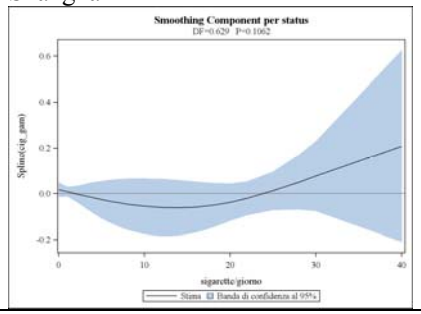
Italy [28]



Milan [29]



Shanghai [30]



SEARCH [31]

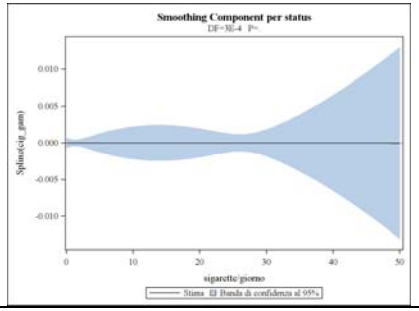


Tabella 7. Analisi del modello di regressione e del modello di *smoothing* del numero di sigarette al giorno tra i fumatori correnti rispetto ai non fumatori, per il *dataset* aggregato PanC4 e per ogni studio incluso.

	Analisi del modello di regressione		**Analisi del modello di *smoothing***	
	Stima parametro lineare	***p-value***	**Gradi di libertà**	***p-value***
PanC4	0,03723	<0,0001	1,06125	0,0038
Nord America				
Louisiana				
LSU (dati non pubblicati)	0,00750	0,7485	<0,5	mancante[a]
Minnesota				
Mayo [21]	0,08608	<0,0001	0,90268	0,0010
Texas				
MD Anderson [22]	0,04256	<0,0001	<0,5	mancante[a]
New York				
MSKCC [23]	0,03085	0,0868	<0,5	mancante[a]
Georgia, Michigan, New Jersey				
NCI [24]	0,03281	<0,0001	12,85666	0,3149
California				
UCSF [25]	0,03750	<0,0001	<0,5	mancante[a]
Connecticut				
Yale [26]	0,03838	<0,0001	<0,5	mancante[a]
Canada				
Toronto [27]	0,08079	<0,0001	<0,5	mancante[a]
Europa				
Italy [28]	0,04056	<0,0001	0,54801	0,0656
Milan [29]	0,01623	0,0146	26,00000	0,9329
Cina				
Shanghai [30]	0,03705	<0,0001	0,62920	0,1062
Internazionale				
Canada, Europa, Australia				
SEARCH [31]	0,04157	<0,0001	<0,5	mancante[a]

Stime ottenute utilizzando modelli additivi generalizzati aggiustati per età, sesso, etnia, educazione, indice di massa corporea, storia di diabete, storia di pancreatite, consumo di alcol (e studio).
[a]Il *p-value* è mancante quando i gradi di libertà sono inferiori allo 0,5.

Figura 3a. Componente *smoothing* del numero di sigarette al giorno tra i fumatori correnti rispetto ai non fumatori (PanC4).

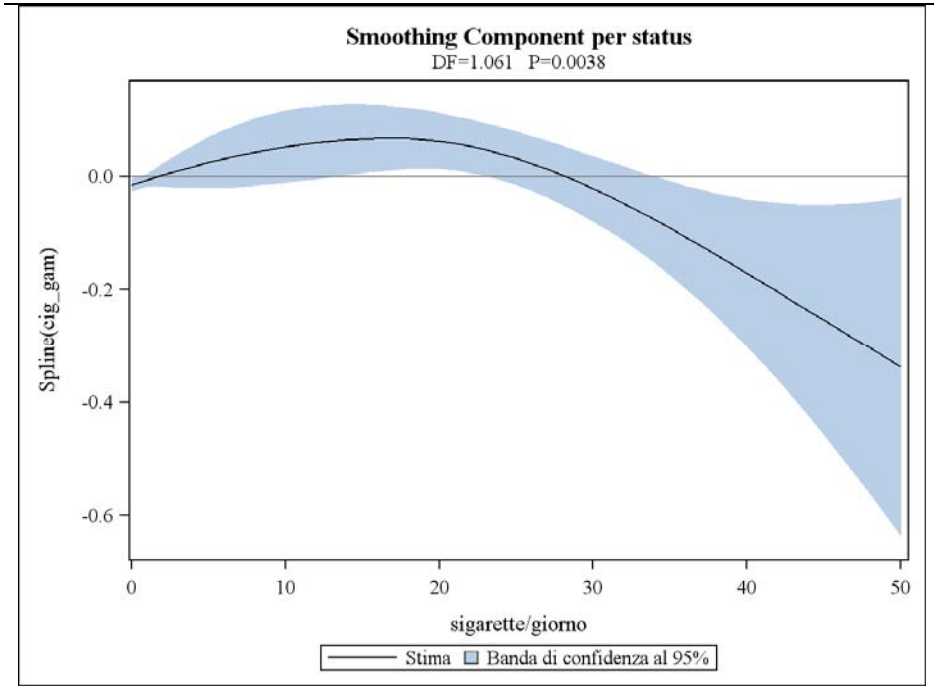


Figura 3b. Combinazione della componente *smoothing* e lineare del numero di sigarette al giorno tra i fumatori correnti rispetto ai non fumatori (PanC4).

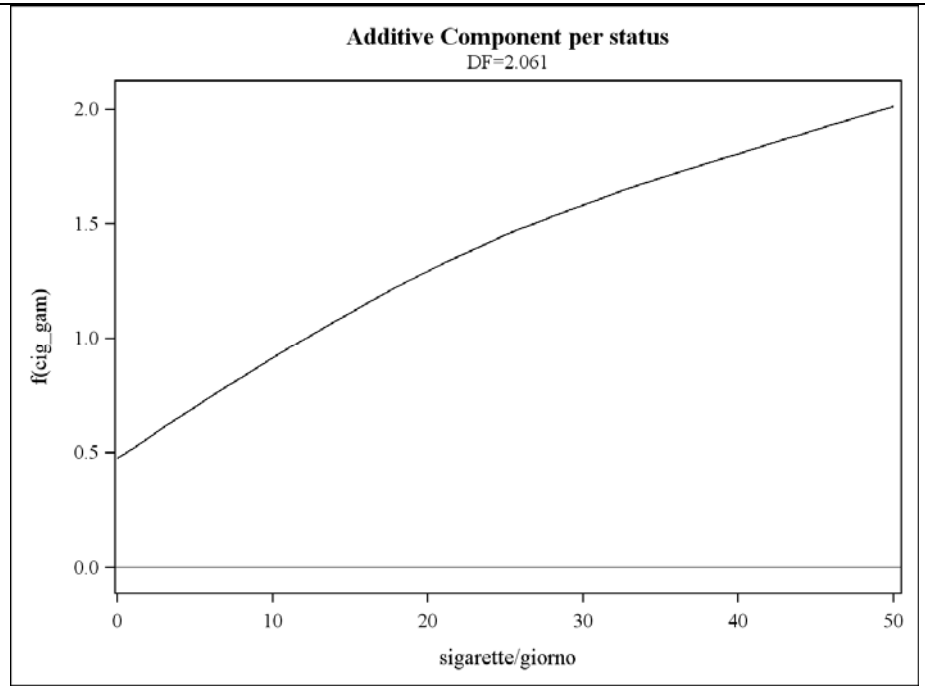


Figura 3c. Relazione tra il rischio di tumore del pancreas e il numero di sigarette fumate al giorno tra i fumatori correnti, rispetto ai non fumatori (PanC4).

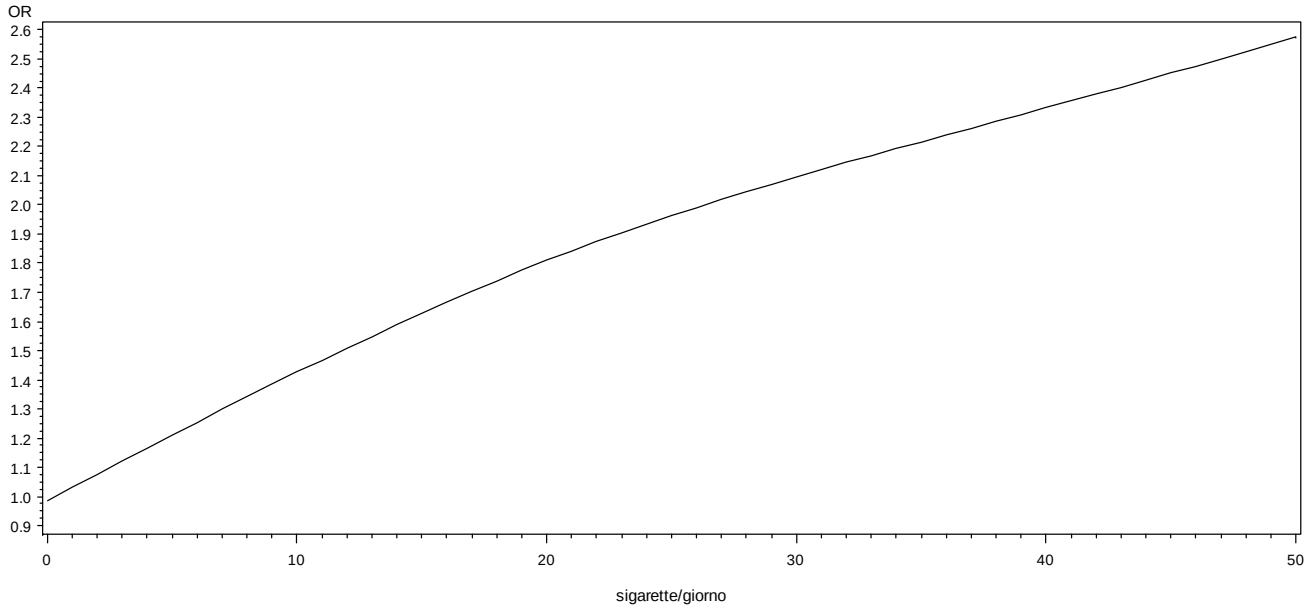


Figura 4. Relazione tra il rischio di tumore del pancreas e la durata in anni dell'abitudine al fumo tra i fumatori correnti, rispetto ai non fumatori (PanC4).

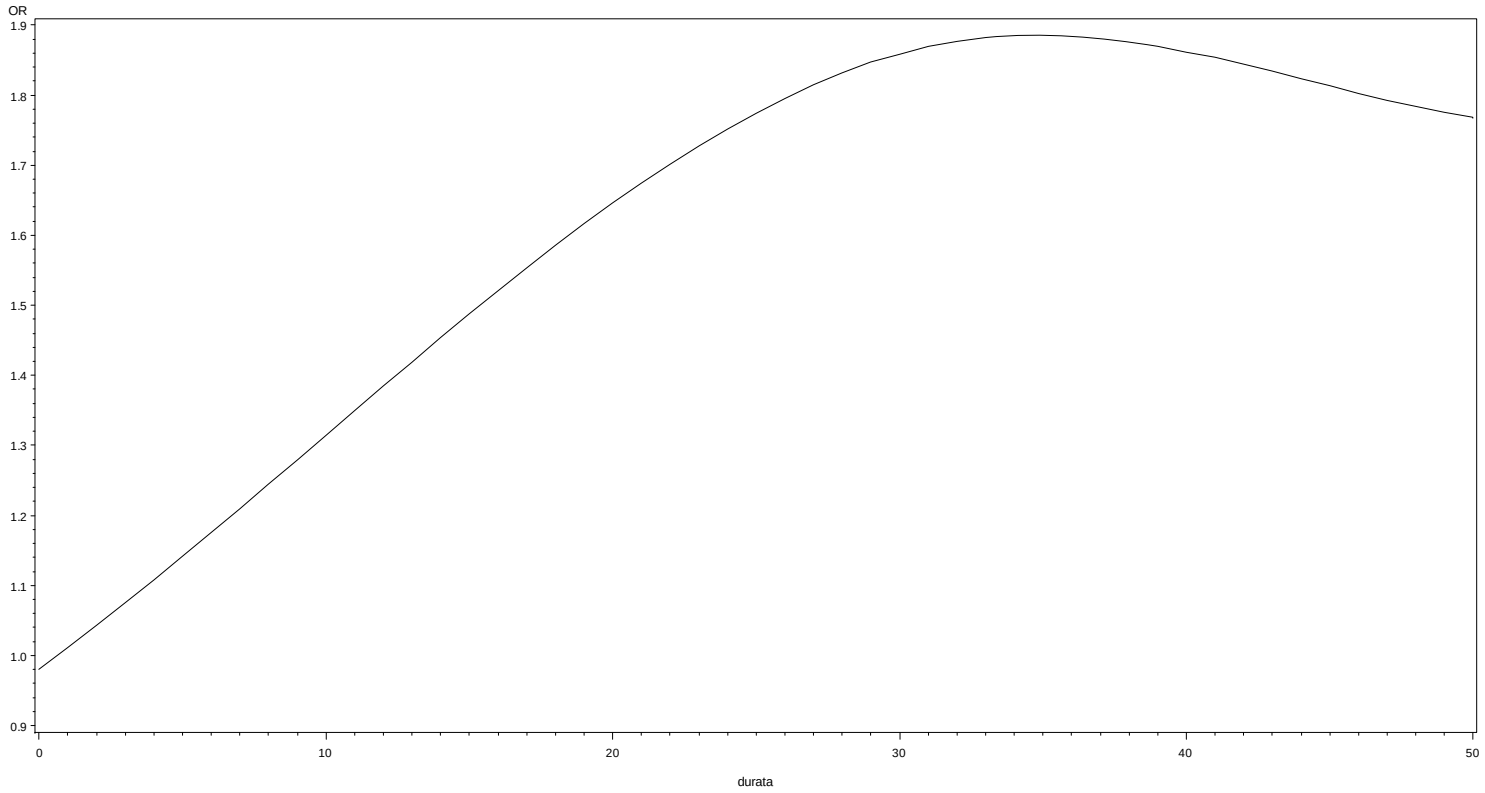


Figura 5. Relazione tra il rischio di tumore del pancreas e gli anni dalla cessazione**,** rispetto ai fumatori correnti (PanC4).