

# PROIEZIONE DEL RISCHIO INDIVIDUALE DI TUMORE AL SENO: VALIDAZIONE ED ESTENSIONE DEL MODELLO DI GAIL

Calza S<sup>1</sup>, Ferraroni M<sup>2</sup>, Decarli A<sup>2</sup>

<sup>1</sup>*Dipartimento di Scienze Biomediche e Biotecnologie, Università degli Studi di Brescia,*

<sup>2</sup>*Sezione di Statistica Medica e Biometria, Università degli Studi di Milano*

## Scopo dello studio

Il tumore al seno rappresenta la forma di neoplasia femminile più comune nei paesi industrializzati, con più di 800000 nuovi casi diagnosticati ogni anno a livello mondiale (Landis et al., 1999) mentre in Italia questa malattia è la prima causa di morte nelle donne nella fascia d'età tra i 35 e i 44 anni e in molte zone rappresenta un quarto circa di tutti i tumori di cui soffrono le donne. La diagnosi precoce seguita da trattamenti farmacologici ha permesso di ridurre la mortalità in alcune tipologie di donne affette da tumore al seno, tuttavia, soprattutto a causa dell'elevato numero di soggetti affetti, l'interesse nello sviluppo di strategie di prevenzione primaria è tuttora molto alto. Tra gli interventi preventivi, lo screening mammografico riveste un ruolo di rilievo, con una riduzione della mortalità di circa un terzo in donne di età superiore ai 50 anni. Tale riduzione non è associata a particolari rischi ed ha un costo economico sostenibile. L'utilizzo di tale pratica di screening è invece più controverso nel caso di donne di età inferiore. Diversi fattori sembrano entrare in gioco nell'influenzare il rischio di insorgenza del tumore; fattori irreversibili (e.g. fattori genetici, età al menarca, età al primo figlio e numero di gestazioni a termine), così come fattori legati allo stile di vita, quali dieta, indice di massa corporea, consumo di alcool, ecc. Negli ultimi decenni ha quindi suscitato particolare interesse lo studio di metodi atti a stimare rischi individuali di insorgenza della malattia entro un certo lasso di tempo a partire da informazioni anamnestiche, di stile di vita e genetiche dell'individuo. Ciò permetterebbe di meglio definire un sottoinsieme di donne ad alto rischio da indirizzare a pratiche preventive o diagnostiche, con un potenziale incremento in termini di costi/efficacia. Il presente lavoro si prefigge lo scopo di validare nella popolazione Italiana uno dei modelli di predizione più diffusi, proposto da Gail (Gail et al., 1989) nonché di valutare la possibilità di una sua estensione al fine di tener conto di variabili di tipo modificabile, ovvero legate allo stile di vita. A tal fine il modello di Gail è stato implementato in una libreria di R (R Development Core Team, 2004).

## Metodi

Lo scopo dei seguenti modelli è quello di stimare un rischio individuale assoluto, ovvero non relativo ad un qualsivoglia riferimento (diverso quindi da un RR). Il RA misura quindi la probabilità che un soggetto con determinati fattori di rischio possa andare incontro alla malattia nell'arco di un predefinito intervallo di tempo. La proiezione secondo il metodo Gail si compone di 3 parti:

1. Definizione dei fattori di rischio e dei relativi RR
2. Stima degli Hazard Rate età specifici di base, ovvero per una donna senza fattori di rischio
3. Proiezione della probabilità a lungo termine di sviluppare un tumore al seno sulla base dei RR e degli hazard di base, assumendo una situazione a rischi competitivi

La maggior difficoltà nel modello consta nella stima del tasso tumore al seno età specifico di base, ovvero al netto dei fattori di rischio. I dati disponibili dal SEER rappresentano dei tassi d'incidenza "compositi"  $h^*_i(t)$  ovvero derivanti da una popolazione composta in termini di esposizione ai fattori di rischi del modello. Sia  $r_i(t)$  il rischio relativo per un soggetto d'età  $t$  appartenente al livello di rischio  $i$ -mo, con  $r_1(t) = 1$ , laddove con livello di rischio intendiamo la presenza di una ben definita composizione dei fattori di rischio succitati. Se indichiamo con  $h_i(t)$  la proporzione di soggetti di età  $t$  esposti a rischio nel livello  $i$ , otteniamo

$$h_1(t) = h^*_1(t) \sum_{i=1}^l \left\{ \frac{\rho_i(t)}{r_i(t)} \right\} \equiv h^*_1(t) F(t)$$

dove  $F(t) = 1 - AR(t)$ , con AR ad indicare il rischio attribuibile, secondo Bruzzi, associato ai diversi fattori del modello.

La proiezione del rischio avviene assumendo che gli hazard competitivi  $h_2(t)$  età specifici siano uguali per tutti i soggetti. La probabilità di sviluppare un tumore tra l'età  $a$  e  $a + \tau$  per una donna appartenente al livello di rischio  $i$  si ottiene quindi dalla formulazione seguente:

$$\begin{aligned} P(a, \tau, r_i(t)) &= \int_a^{a+\tau} h_1(t) r_i(t) \exp \left\{ - \int_a^t h_1(u) r_i(u) du \right\} \frac{S_2(t)}{S_2(a)} dt \\ &= \int_a^{a+\tau} h_1(t) r_i(t) \exp \left\{ - \int_a^t [h_1(u) r_i(u) + h_2(u)] du \right\} dt \end{aligned}$$

I fattori di rischio impiegati nel modello di Gail originario sono stati ricavati dai dati del Breast Cancer Detection Demonstration Project (BCDP), un esteso studio di screening condotto tra il 1973 e il 1980 in 28 centri, nel quale furono raccolte informazioni epidemiologiche dettagliate su un campione complessivo di 284780 donne (BCDP 1982). I casi sono rappresentati da donne con carcinoma in situ ed invasivo.

Complessivamente il campionamento ha portato a 2852 casi e 3146 controlli per i quali si disponeva di tutte le informazioni d'interesse. La definizione dei rischi è stata condotta utilizzando una regressione logistica "unconditional".

Le variabili introdotte nel modello sono le seguenti:

- Età ( $< 50, \geq 50$  anni)
- Età al menarca ( $> 14$  anni, 12-13 anni,  $< 12$  anni)
- Età al primo figlio nato vivo ( $< 20$  anni, 20-24 anni, 25-29 anni o nullipare,  $\geq 30$  anni)
- Numero di familiari di I grado con tumore al seno (nessuno, 1,  $> 1$ )
- Numero di Biopsie con esito negativo (nessuna, 1,  $> 1$ )

Nel 1992 il modello di Gail è stato sottoposto ad una revisione (Anderson et al. 1992) per limitarne l'utilizzo alla proiezione di tumori invasivi (non *in situ*) introducendo inoltre la possibilità del calcolo di una proiezione per donne di colore. Nel proseguo del lavoro si farà riferimento a tale modello come USA Gail.

## **I Dati**

### ***Studio Multicentrico***

Al fine di stimare i rischi relativi ( $r_i(t)$ ) e il rischio attribuibile ( $AR(t)$ ) sono stati utilizzati dati raccolti nel contesto di uno studio caso-controllo multicentrico condotto in sei regioni italiane tra il giugno 1991 e aprile 1994 (Franceschi et al., 1996). I casi sono costituiti da 2569 donne di età compresa tra 23-74 anni (età mediana 55), ricoverate nei maggiori ospedali delle aree di studio per diagnosi di tumore al seno confermata istologicamente, entrò un anno precedente all'intervista, e senza alcuna storia pregressa di tumore. I 2588 controlli sono donne di età 20-74 (età mediana 56) ricoverati in ospedali delle medesime aree geografiche, per patologie acute.

Il data-set include informazioni di tipo sociodemografico, quali età, livello d'istruzione, attività lavorativa e indicatori socio-economici, abitudine al fumo, attività fisicap, misure antropometriche, consumo di alcool, abitudini dietetiche, storia medica personale e un insieme di informazioni selezionate sulla storia familiare di tumore al seno. Un questionario sulle frequenze dei consumi (FFQ), validato, è stato utilizzato per determinare le abitudini dietetiche nei due anni antecedenti la diagnosi, al fine di poter stimare l'apporto calorico giornaliero e il consumo di alcuni nutrienti selezionati.

### ***Coorte Epic-Firenze***

I dati EPIC Firenze includono informazioni su donne reclutate nell'area di Firenze, coperta dal Registro dei Tumori, nel contesto dello studio prospettico EPIC-Italia sulla relazione tra dieta e tumori (Calza et al., 2003). Il reclutamento è stato condotto tra gennaio 1993 e marzo 1998 con l'adesione su base volontaria di 10083 donne di età compresa tra i 35 e i 64 anni, residenti nelle provincie di Firenze e Prato.

Da ciascuna donna arruolata nello studio sono state raccolte informazioni dettagliate sulle abitudini alimentari e lo stile di vita, la storia riproduttiva e la storia familiare di tumore al seno. A causa del reclutamento prevalentemente su base volontaria, al fine di evitare possibili conseguenti bias, dall'analisi sono stati esclusi non solo i casi prevalenti ( $n = 30$ ), ma anche le donne a cui fosse stata diagnosticato un tumore entro 6 mesi dall'arruolamento nello studio ( $n = 12$ ) o il cui followup fosse minore di 6 mesi ( $n = 10$ ). Di conseguenza il dato persone-anno è stato calcolato ritardando di 6 mesi l'inizio del follow-up per ciascuna donna. Sul totale di 10031 donne considerate nell'analisi, si sono osservati 136 casi con diagnosi di tumore al seno invasivo entro 5 anni di follow-up (166 in totale).

## **Risultati**

### ***Validazione del Modello***

I rischi individuali di insorgenza di tumore al seno per le donne nella coorte Epic-FI sono stati stimati utilizzando 2 versioni del modello di Gail:

- 1) US-Gail: i tassi d'incidenza età specifici, i rischi relativi ( $r_i(t)$ ) e il rischio attribuibile (AR(t)) utilizzati nel modello sono i medesimi utilizzati nel modello riveduto dal NSABP (Anderson et al., 1992)
- 2) It-Gail: i rischi relativi ( $r_i(t)$ ) sono stati ricavati dallo studio caso controllo multicentrico su dieta e tumore al seno condotto in Italia mentre i tassi d'incidenza  $h_1(t)$  e  $h_2(t)$  sono stati ottenuti dal registro dei tumori di Firenze. Le stime di  $r_i(t)$  sono state ottenute codificando come ordinali le variabili inserite nel modello logistico, così come adottato da Gail et al. (1989)

La tabella 1 mostra la distribuzione dei fattori di rischio originariamente inclusi nel modello di Gail, sia per il modello USA, sia per la versione Italiana. Le stime degli OR risultano essere sostanzialmente coerenti nei due modelli.

Ciò si riflette in Tabella 2, che riporta le stime degli attesi e degli osservati, dove si può osservare una buona concordanza tra i due modelli, sebbene complessivamente il modello italiano dia una predizione, in termini di rapporto Attesi vs Osservati migliore rispetto al modello USA.

Entrambi modelli tendono a sottostimare in modo significativo nella classe d'età 40-49. Tuttavia questo sembra essere riconducibile alla struttura della popolazione in oggetto, dato che l'incidenza osservata in tale classe d'età è sensibilmente maggiore rispetto a quella della popolazione del registro.

Tabella 1: Distribuzione dei fattori irreversibili inclusi nel modello di Gail originario e OR stimati. Studio caso controllo multicentrico, Italia.

	Controlli		Casi		USA			Italia		
	N	%	N	%	OR	Lower .95	Upper .95	OR	Lower .95	Upper .95
<b>Età al Menarca</b>										
≥ 14 o Ignoto	953	36.82	852	33.16	1.00			1.00		
12-13	1147	44.32	1220	47.40	1.10	1.02	1.19	1.08	1.00	1.17
7-11	488	18.86	497	19.35	1.21	1.12	1.31	1.17	1.09	1.27
<b>Età nascita primo figlio</b>										
< 20 o Ignoto	218	8.42	121	4.71	1.00			1.00		
20-24	962	37.17	782	30.44	1.24	1.16	1.33	1.39	1.29	1.49
25-29 o Nullipare	1111	42.03	1255	48.85	1.55	1.41	1.66	1.92	1.78	2.07
≥ 30	297	11.48	411	16.00	1.93	1.80	2.07	2.66	2.47	2.86
<b>Familiarità</b>										
0	2467	95.32	2311	89.96	1.00			1.00		
1	110	4.25	238	9.26	2.61	1.99	3.42	2.28	1.43	3.63
>1	11	0.43	20	0.78	6.80	5.10	8.91	5.21	3.27	8.20
<b>Biopsie</b>										
No o Ignoto	2558	98.84	2524	98.25	1.00			1.00		
Sì	30	1.16	45	1.75	1.70	1.40	2.06	1.40	0.67	2.96

Tabella 2: Stime del Rischio Assoluto secondo il modello US-Gail e il modello It-Gail. Coorte Epic di Firenze

	Pop	Pop-Anno	Osservati	US-Gail		It-Gail	
				Attesi	Att/Oss	Attesi	Att/Oss
<b>Età</b>							
30-39	846	4187	4	4.33	1.08	5.50	1.38
40-49	2423	12027	47	23.08	0.49	28.24	0.60
50-59	5020	24804	64	70.14	1.10	71.77	1.12
≥ 60	1742	8585	27	31.64	1.17	29.04	1.08
<b>Età al Menarca</b>							
7-11	2530	12555	35	33.84	0.97	35.03	1.00
12-13	5104	25240	72	65.20	0.91	68.41	0.95
≥ 14 o Ignoto	2388	11809	35	30.16	0.86	31.10	0.89
<b>Età nascita primo figlio</b>							
< 20 o Ignoto	263	1300	1	2.09	2.09	1.72	1.72
20-24	2547	12622	40	28.01	0.70	25.76	0.64
25-29 o Nullipare	5318	26278	82	69.09	0.85	72.29	0.88
≥ 30	1903	9403	19	29.40	1.55	34.77	1.83
<b>Familiarità</b>							
No	9033	44662	119	107.55	0.90	109.61	0.92
Si	998	4943	23	21.64	0.94	24.93	1.08
<b>Biopsie</b>							
No o Ignoto	9769	48306	133	124.29	0.93	128.56	0.97
Si	262	1299	9	4.90	0.54	5.98	0.66
<b>Totale</b>			142	129.19	0.91	134.54	0.95

### *Estensione del Modello*

Il modello di Gail nella forma originaria prende in considerazione esclusivamente fattori di rischio non modificabili, ovvero laddove conseguenza di stili di vita o culturali non più modificabili. Sono tuttavia noti dalla letteratura (Mezzetti et al., 1998) fattori di rischio legati ad abitudini di vita ed all'alimentazione, ovvero a fattori modificabili. E' quindi plausibile ipotizzare un possibile contributo di tali fattori nella definizione del rischio di insorgenza della patologia.

Lo scopo consta nel quantificare la rilevanza proporzionale dei diversi fattori di rischio nella definizione del rischio assoluto di tumore al seno.

Due fattori sono stati presi in considerazione nel presente lavoro, alla luce di precedenti informazioni di tipo epidemiologico: consumo di alcool etilico e apporto di grassi monoinsaturi.

Le tabelle 3 e 4 mostrano un confronto tra il modello Italiano base e rispettivamente i modelli con l'aggiunta del consumo di alcool e di grassi monoinsaturi.

Il modello contenente il consumo di Alcool, in particolare, porta a delle stime degli attesi sensibilmente più vicine agli osservati.

Tabella 3: Stime del Rischio Assoluto secondo il modello It-Gail e il modello It-Gail con incluso il consumo di Alcool (It-Alc). Coorte Epic di Firenze

	Pop	Pop-Anno	Osservati	It-Gail		It-Alc	
				Attesi	Att/Oss	Attesi	Att/Oss
<b>Età</b>							
30-39	846	4187	4	5.50	1.38	5.68	1.42
40-49	2423	12027	47	28.24	0.60	29.24	0.62
50-59	5020	24804	64	71.77	1.12	74.35	1.16
≥ 60	1742	8585	27	29.04	1.08	29.01	1.10
<b>Età al Menarca</b>							
7-11	2539	12555	35	35.03	1.00	36.60	1.05
12-13	5104	25240	72	68.41	0.95	70.88	0.98
≥ 14 or Unknown	2388	11809	35	31.10	0.89	31.39	0.90
<b>Età nascita primo figlio</b>							
< 20 o Ignoto	263	1300	1	1.72	1.72	1.77	1.77
20-24	2547	12622	40	25.76	0.64	26.76	0.67
25-29 o Nullipare	5318	26278	82	72.29	0.88	74.33	0.91
≥ 30	1903	9403	19	34.77	1.83	36.02	1.90
<b>Familiarità</b>							
No	9033	44662	119	109.61	0.92	113.13	0.95
Sì	998	4943	23	24.93	1.08	25.75	1.12
<b>Biopsie</b>							
No o Ignoto	9769	48306	133	128.56	0.97	132.73	1.00
Sì	262	1299	9	5.98	0.66	6.15	0.68
<b>Totale</b>			142	134.54	0.95	138.88	0.98

Tabella 4: Stime del Rischio Assoluto secondo il modello It-Gail e il modello It-Gail con incluso il consumo di grassi Monoinsaturi (It-Mono). Coorte Epic di Firenze

	Pop	Pop-Anno	Osservati	It-Gail		It-Mono	
				Attesi	Att/Oss	Attesi	Att/Oss
<b>Età</b>							
30-39	846	4187	4	5.50	1.38	5.40	1.37
40-49	2423	12027	47	28.24	0.60	27.95	0.59
50-59	5020	24804	64	71.77	1.12	70.33	1.10
≥ 60	1742	8585	27	29.04	1.08	27.96	1.04
<b>Età al Menarca</b>							
7-11	2539	12555	35	35.03	1.00	34.52	0.99
12-13	5104	25240	72	68.41	0.95	67.25	0.93
≥ 14 or Unknown	2388	11809	35	31.10	0.89	29.94	0.86
<b>Età nascita primo figlio</b>							
< 20 o Ignoto	263	1300	1	1.72	1.72	1.67	1.67
20-24	2547	12622	40	25.76	0.64	25.38	0.63
25-29 o Nullipare	5318	26278	82	72.29	0.88	70.33	0.86
≥ 30	1903	9403	19	34.77	1.83	34.33	1.81
<b>Familiarità</b>							
No	9033	44662	119	109.61	0.92	107.21	0.90
Sì	998	4943	23	24.93	1.08	24.50	1.07
<b>Biopsie</b>							
No o Ignoto	9769	48306	133	128.56	0.97	125.75	0.95
Sì	262	1299	9	5.98	0.66	5.96	0.66
<b>Totale</b>			142	134.54	0.95	131.71	0.93

## Discussione

I metodi di previsione individuale del rischio di insorgenza di tumore al seno basati su modelli multivariati hanno conosciuto un notevole sviluppo nelle ultime decadi. Tali metodi sono finalizzati a fornire uno strumento utile nel processo di decisione nel caso di programmi di sorveglianza o preventivi, e possono essere di utile supporto all'esperienza clinica. Il metodo di Gail (Gail et al. 1989, Anderson et al. 1992) è ampiamente utilizzato nell'ottimizzazione di pratiche di intervento negli Stati Uniti ed è uno dei più promettenti nello specifico campo di ricerca. Tuttavia la letteratura relativa alla validità del modello in altre paesi è chiaramente carente, laddove i maggiori contributi alla validazione del modello sono stati rivolti alla verifica nel contesto di diversi gruppi etnici negli USA. A tutt'oggi esiste un solo lavoro che abbia preso in considerazione l'applicabilità del modello nella popolazione italiana (Boyle et al., 2004), tuttavia il campione su cui è stato applicato non è rappresentativo della popolazione generale, essendo composto da soggetti isterectomizzati. Nel presente lavoro il modello è stato applicato a dati relativi ad una coorte di donne reclutate in uno studio sulla relazione tra dieta e cancro, nel contesto del progetto EPIC (coorte di Firenze). I partecipanti sono soggetti pressoché normali, e soprattutto, sottoposte ad un programma di screening mammografico, situazione nella quale è stato sviluppato il modello di Gail stesso. L'analisi ha mostrato come, sebbene una certa tendenza alla sottostima del numero di casi totale prevedibile nell'arco di 5 anni, il modello sia ragionevolmente valido. Il confronto con una versione modificata del modello con l'aggiunta di stime di rischi e di tassi di incidenza relativi alla popolazione italiana, ha sostanzialmente mostrato una buona concordanza tra i modelli. Ciò non stupisce, in quanto i rischi relativi per le variabili inserite nel modello sono, in letteratura, sostanzialmente simili nel caso della popolazione statunitense quanto di quella italiana. Le maggiori differenze si riscontrano nel tasso di incidenza, di cui tuttavia il modello tiene conto. Questo farebbe supporre di poter applicare con successo il modello anche per popolazioni mediterranee, senza drastici cambiamenti. Un secondo aspetto del lavoro era rivolto a valutare la possibilità di inserire un maggior numero di variabili nel modello stesso, in particolare variabili relative a fattori di rischio (o protettivi) modificabili. Data la natura di strumento di supporto al decision-making, ciò potrebbe permettere di suggerire interventi preventivi sia legati all'utilizzo di farmaci (es. Tamoxifen), sia a pratiche di diagnosi precoce, sia a possibili cambiamenti nelle abitudini di vita. A tal scopo sono stati valutati due noti fattori di rischio quali il consumo di alcool etilico e l'apporto di grassi monoinstaurati, quale proxy di abitudini alimentari tipicamente mediterranee (consumo di olio d'oliva). I risultati sembrano suggerire un miglioramento nella capacità predittiva del modello e una possibile utilità nell'identificare soggetti ad incrementato rischio, oltre i fattori fisiologici, e soprattutto nel valutare l'effetto sul rischio di una modifica delle abitudini stesse.

## Bibliografia

- Anderson, S. J., S. Ahn, and K. Du : 1992, NSABP Breast Cancer Prevention Trial risk assessment program, version 2. Technical report, NSABP Biostatistical Center.
- Boyle, P., M. Mezzetti, L. C. Vecchia, S. Franceschi, A. Decarli, and C. Robertson: 2004, Contribution of three components to individual cancer risk predicting breast cancer risk in Italy. . Eur J Cancer Prev 13(3), 183-91.

- Bruzzi, P., S. B. Green, D. P. Byar, L. A. Brinton, and C. Schairer: 1985, Estimating the population attributable risk for multiple risk factors using case-control data. *Am J Epidemiol* 122(5), 904-14.
- Calza, S., C. Specchia, G. Frasca, R. Tumino, C. Sacerdote, L. Fiorini, R. Galasso, A. V. Ciardullo, D. Palli, S. Salvini, V. Krogh, S. Sieri, and A. Decarli: 2003, EPIC-Italy cohorts and multipurpose national surveys. comparison of some socio-demographic and life-style characteristics. *Tumori* 89(6), 615-23.
- Franceschi, S., A. Favero, A. Decarli, E. Negri, L. C. Vecchia, M. Ferraroni, A. Russo, S. Salvini, D. Amadori, and E. Conti: 1996, Intake of macronutrients and risk of breast cancer. *Lancet* 347(9012), 1351-6.
- Gail, M., L. Brinton, D. Byar, D. Corle, S. Green, C. Schairer, and J. Mulvihill: 1989, Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 81(24), 1879-1886.
- H, B. L.: 1982, Breast cancer detection demonstration project: Five year summary report. *CA* 32, 194-225.
- Landis, S. H., T. Murray, S. Bolden, and P. A. Wingo: 1999, Cancer statistics, 1999. *CA Cancer J Clin* 49(1), 8-31, 1.
- Mezzetti, M., L. C. Vecchia, A. Decarli, P. Boyle, R. Talamini, and S. Franceschi: 1998, Population attributable risk for breast cancer: diet, nutrition, and physical exercise. *J Natl Cancer Inst* 90(5), 389-94.