Metodi statistici per lo studio dei gemelli

Corrado Fagnani, Sonia Brescianini, Emanuela Medda e Maria Antonietta Stazi

Centro Nazionale di Epidemiologia, Sorveglianza e Promozione della Salute, Istituto Superiore di Sanità, Roma

Riassunto. Vengono descritti i più importanti metodi statistici attualmente utilizzati per l'analisi di dati riferiti a coppie di gemelli. L'obiettivo principale di questi metodi è stimare il contributo dei fattori genetici ed ambientali alla variabilità dei caratteri umani, normali o patologici, attraverso l'informazione ottenuta da coppie di gemelli monozigoti e dizigoti. In questo contesto, assume rilievo il concetto di ereditabilità. Oltre al semplice confronto tra gemelli monozigoti e dizigoti, basato su misure quali la concordanza e la correlazione, sono presentati nuovi approcci più complessi, con un'enfasi particolare sui modelli di equazioni strutturali, ed un cenno alla *DF-analysis* e ai *correlated frailty models*. Vengono anche illustrati alcuni esempi di applicazioni ai dati del Registro Nazionale Gemelli.

Parole chiave: gemelli, concordanza, ereditabilità, modelli di equazioni strutturali.

Summary *(Statistical methods for the analysis of twin data)***.** The most important statistical methods currently used for the analysis of twin data are described. The main objective of these methods is to estimate the contribution of the genetic and environmental factors to the variability of normal or pathological human traits, by means of the information obtained from monozygotic and dizygotic twin pairs. In this context, the concept of heritability becomes relevant. Not only the simple comparison between monozygotic and dizygotic twins, based on measures such as the concordance and the correlation, but also new and more complex approaches are presented, with a special emphasis on the structural equation models, and a synthetic view on the DF-analysis and the correlated frailty models. Some examples of applications to the data of the Italian Twin Registry are also illustrated.

Key words: twins, concordance, heritability, structural equation models.

INTRODUZIONE

Il metodo gemellare consiste nel confronto tra gemelli monozigoti (MZ) e dizigoti (DZ), ed ha lo scopo di investigare l'influenza che geni ed ambiente esercitano su un certo carattere di interesse. Tale confronto riguarda il livello di somiglianza fenotipica, ed è basato sulla diversa correlazione genetica nelle coppie delle due zigosità. Dato che i gemelli MZ sono geneticamente identici, mentre i DZ condividono in media il 50% dei geni, al pari di normali fratelli, è chiaro che un'eventuale maggiore somiglianza osservata tra i primi può essere assunta come indice di influenze genetiche sul carattere in studio. Il ragionamento è valido solo se si assume che i gemelli MZ condividano le esperienze ambientali, rilevanti per la caratteristica in esame, nella stessa misura dei DZ (*equal environments assumption* [1]). Per molte variabili tale assunzione può essere ritenuta valida. Sia i MZ che i DZ condividono i fattori della vita intrauterina, nonché le esposizioni dell'ambiente familiare e domestico nell'infanzia. Se l'assunzione non è valida, ovvero se una più forte condivisione di esperienze ambientali contribuisce all'eccesso di somiglianza dei MZ rispetto ai DZ, allora il semplice confronto tra MZ e DZ porta ad una sovrastima degli effetti genetici sulla caratteristica in studio. Uno dei campi di ricerca in cui l'assunzione viene messa in dubbio è quello della genetica del comportamento. In questo ambito, le critiche al metodo consistono nell'osservare che i gemelli MZ, specie nell'infanzia, sono spesso trattati dai genitori in modo più simile dei DZ, e ciò può renderli più correlati per numerosi tratti comportamentali nella vita successiva.

Dunque, il problema primario del metodo gemellare è stimare la somiglianza tra i fenotipi dei gemelli all'interno delle coppie. In questa fase, gli strumenti sono puramente statistici, e le principali misure utilizzate sono la concordanza [2] e la correlazione. La concordanza esprime il grado con cui un certo carattere binario, ad esempio una malattia, tende a ritrovarsi in entrambi i gemelli di una stessa coppia. Per caratteristiche quantitative, la correlazione mostra se c'è un'associazione lineare tra i valori rilevati nei due membri delle varie coppie.

Successivamente, la questione diventa quella di dedurre, dalle somiglianze (o dalle differenze) osservate, il ruolo svolto dai geni e dall'ambiente nel determinare la variabilità del fenotipo. Ed è qui che trovano applicazione modelli più o meno complessi, le cui assunzioni spesso interessano il campo della genetica. In questi modelli, un parametro rilevante, nel quale si traduce

Indirizzo per la corrispondenza (Address for correspondence): Corrado Fagnani, Reparto di Epidemiologia Genetica, Centro Nazionale di Epidemiologia, Sorveglianza e Promozione della Salute, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Roma, Italia. E-mail: fagnani@iss.it.

il quesito di partenza dell'indagine con il metodo gemellare, è dato dall'ereditabilità [1]. In generale, essa è definita come proporzione della varianza fenotipica dovuta ai fattori genetici, e quindi misura il contributo dei geni alla variabilità inter-individuale osservata. Diversi indici, sostanzialmente basati sulle correlazioni nei gemelli MZ e DZ, sono stati proposti per stimare l'ereditabilità.

Nel corso degli anni, lo studio dei gemelli non è rimasto estraneo ai progressi delle metodologie statistiche, ed ha beneficiato dei nuovi strumenti che a mano a mano si sono resi disponibili. In particolare, i modelli di equazioni strutturali (SEM) [3], integrando tecniche di regressione multivariata con elementi di genetica quantitativa, hanno fornito un approccio ben più complesso per la stima dell'ereditabilità. Nei SEM, i geni e l'ambiente sono descritti tramite variabili latenti, il cui effetto è stimato col metodo della massima verosimiglianza a partire dalle varianze e covarianze dei fenotipi gemellari. Numerose ipotesi possono essere testate, quali una diversa ereditabilità nei due sessi o nel tempo, oppure l'esistenza di fattori genetici ed ambientali comuni a più caratteri, ad esempio nel caso di studi di comorbidità. Nell'ambito della genetica molecolare, i SEM possono essere utilizzati in studi di *linkage* [4] per stimare il contributo di certi polimorfismi alla variabilità di tratti quantitativi.

Una strategia diversa dai SEM, meno flessibile e potente, per quantificare l'influenza dei geni e dell'ambiente sull'espressione di un dato carattere, è quella sviluppata da DeFries e Fulker, e nota come *DF-analysis* [5]. Nella *DF-analysis*, stime dirette dell'ereditabilità e della proporzione di varianza attribuibile a fattori ambientali sono ottenute col metodo dei minimi quadrati, considerando la regressione lineare del fenotipo di uno dei due gemelli sul fenotipo del co-gemello e su altre variabili opportune.

Infine, i *correlated frailty models* [6] consentono di combinare metodi di analisi della sopravvivenza e di genetica quantitativa per stimare, a partire dall'informazione sull'età di insorgenza di una malattia o sulla durata della vita in coppie di gemelli MZ e DZ, la componente genetica e quella ambientale della predisposizione alla morbidità e alla mortalità. Un simile approccio ha chiare applicazioni in studi sui determinanti della salute e della longevità.

Per una maggiore chiarezza sull'interpretazione dei vari parametri stimati e sui possibili ambiti di applicazione dei diversi modelli, vengono riportati esempi concreti, riferiti, quando possibile, ai dati del Registro Nazionale Gemelli [7]. Per gli approcci non ancora utilizzati nell'analisi dei dati del Registro, si trae spunto dalla letteratura straniera più o meno recente.

METODI DI BASE: CONCORDANZA, CORRELAZIONE ED EREDITABILITÀ

Il problema di partenza del metodo gemellare è stimare il grado di somiglianza fenotipica tra i gemelli all'interno delle coppie, per poi inferire, sotto la *equal environments assumption*, l'effetto di fattori genetici ed ambientali dal confronto tra MZ e DZ.

Per un carattere binario, si tratta di misurare la tendenza di una certa condizione ad occorrere in entrambi i membri di una stessa coppia. In tale situazione, si ricorre al concetto di concordanza, e le due misure tipicamente utilizzate sono la concordanza *casewise* (P_c) e quella *pairwise* (P_p). Supponendo che la condizione in esame sia una malattia, P_c è la probabilità che un gemello di una coppia sia malato dato che il co-gemello lo è, mentre P_p è la probabilità che entrambi i gemelli di una coppia siano malati noto che almeno uno lo è. Quando i gemelli costituiscono un campione casuale della popolazione, oppure ogni coppia è selezionata in base alla presenza della malattia in almeno uno dei due membri e la probabilità che un gemello affetto sia incluso nel campione è del 100% (accertamento completo), stime di massima verosimiglianza per i due tipi di concordanza sono date da $P_c = 2n_{11}/(2n_{11} + n_d)$ e $P_p = n_{11}/(n_{11} + n_d)$, dove n_{11} e n_d indicano, rispettivamente, il numero di coppie concordanti malate (cioè con entrambi i membri affetti) e discordanti (cioè con un solo membro affetto) [2].

Dalla concordanza si può ricavare il *recurrence risk ratio*, che è una misura di rischio genetico. Essa è definita dal rapporto tra la concordanza *casewise* e la prevalenza della patologia nella popolazione generale, ed è interpretabile come rischio relativo di malattia in un co-gemello di un gemello affetto rispetto ad un individuo nella popolazione [8].

La concordanza *casewise* può essere utilizzata per predire la malattia nel co-gemello di un gemello affetto, e quindi può avere una qualche applicazione in termini di counselling. Tuttavia, in questo ambito, risulta assai più informativa la stima della probabilità di malattia nel co-gemello ad un certo tempo dall'insorgenza della malattia nel gemello indice (probando). Ciò richiede un'analisi di tipo "sopravvivenza", in cui l'origine della scala dei tempi corrisponde all'insorgenza nel probando. Tramite il metodo di Kaplan-Meyer [9], si può stimare l'incidenza cumulativa nei co-gemelli entro un opportuno periodo, ed anche effettuare un confronto tra MZ e DZ. Se si ricorre al metodo di Cox [9], il confronto tra MZ e DZ può tenere conto di possibili confondenti. Un rischio di malattia più elevato (e quindi una minore sopravvivenza) in co-gemelli di probandi MZ può essere assunto come indice di influenze genetiche.

In uno studio sulla celiachia in coppie di gemelli italiani [10], identificate tramite l'incrocio tra il Registro Nazionale Gemelli e le liste dell'Associazione Italiana Celiachia, le stime di concordanza sono notevolmente più alte nei MZ rispetto ai DZ *(Tab. 1)*, e ciò conferma l'esistenza di un'importante componente genetica per la patologia. In un altro studio sul diabete di tipo 1 in gemelli finlandesi [11], l'incidenza cumulativa a 10 anni risulta essere del 32,8% nei co-gemelli MZ e del 3,2% in quelli DZ, indicando che, per un co-gemello di un diabetico MZ, il rischio di diabete entro la prima decade dall'insorgenza della malattia nel gemello indice è circa 10 volte superiore a quello sperimentato da un co-gemello DZ.

Tabella 1 | *Concordanza per celiachia in gemelli italiani*

	Coppie monozigote			Coppie dizigote		
N coppie	Concordanti 15	Discordanti 5	Totale 20	Concordanti 3	Discordanti 24	Totale 27
Concordanza casewise (%)	85,7			20,0		
(IC 95%)	(73,3-98,1)			(0,8-39,2)		
Concordanza pairwise (%)	75,0			11,1		
(IC 95%)	(62,0-94,0)			(9,9-23,0)		

IC 95% = intervallo di confidenza al 95%.

Nel caso di tratti quantitativi, la somiglianza fenotipica tra i gemelli è stimata dalla correlazione intraclasse e di Pearson. Le variabili su cui questi coefficienti sono calcolati rappresentano lo stesso carattere rilevato sui due gemelli di ogni coppia. La correlazione intraclasse può essere derivata dall'analisi della varianza, considerando come gruppi a confronto le diverse coppie e calcolando il rapporto (F-1)/(F+1), dove F è la statistica di Fisher; la correlazione di Pearson è data dal rapporto tra la covarianza dei due gemelli ed il prodotto delle rispettive deviazioni standard.

Nel caso di un carattere categorico, dicotomico o con più di due categorie ordinate, quale ad esempio la presenza/assenza di una malattia o i suoi livelli di gravità, si può ricorrere alla correlazione tetracorica o policorica, sotto il modello *liability-threshold* [1]. In questo modello si assume che, a sottendere la variabile osservata sugli individui, vi sia una suscettibilità *(liability)* di fondo, avente distribuzione normale, con dei valori soglia *(threshold)* che discriminano le diverse categorie; in tale situazione, la correlazione tetracorica o policorica tra i gemelli è la correlazione tra le relative suscettibilità.

Valori di correlazione (intraclasse, di Pearson o policorica) più alti per i MZ rispetto ai DZ indicano effetti genetici sul carattere in questione. In uno studio sull'altezza corporea in gemelli di otto registri, incluso quello italiano [12], nell'ambito del progetto europeo GenomEUtwin [13], l'importante ruolo dei geni nel determinare la variabilità del carattere emerge dalla maggiore correlazione di Pearson per i MZ in ogni registro *(Tab. 2)*.

La correlazione può essere usata per ottenere stime grezze delle proporzioni di varianza totale attribuibili a fattori genetici (ereditabilità) ed ambientali. Diversi indici di ereditabilità sono stati costruiti a partire dalle correlazioni in coppie di gemelli MZ e DZ [14], e tutti si basano su assunzioni ben precise, tra cui la già citata equal environments assumption. La formula più comunemente utilizzata è $h^2 = 2(r_{MZ} - r_{DZ})$, dove r_{MZ} e r_{DZ} sono le correlazioni per i MZ e i DZ. Negli studi di genetica quantitativa, i fattori ambientali vengono tipicamente raggruppati in due classi principali: quelli condivisi e quelli non condivisi tra gli individui. L'ambiente condiviso *(common environment)* include tutti quei fattori, come le esposizioni familiari e domestiche, che concorrono all'aggregazione all'interno delle famiglie e, quindi, anche all'interno delle coppie gemellari, indipendentemente dalla zigosità; nell'ambiente non condiviso *(unique environment)* sono inglobati, invece, quei fattori (infezioni, esperienze traumatiche, rapporti sociali ecc.) che, agendo separatamente su ogni individuo, contribuiscono alla variabilità all'interno delle famiglie e, nel caso dei gemelli, spiegano livelli di somiglianza tra i MZ spesso inferiori al 100%. Una stima degli effetti dei fattori ambientali condivisi e non condivisi sulla variabilità di un carattere si può ottenere, rispettivamente, da $c^2 = r_{MZ} - h^2$ ed $e^2 = 1 - r_{MZ}$, dove h^2 è l'ereditabilità e r_{MZ} è la correlazione osservata per i MZ. Ad esempio, in un lavoro riguardante il ruolo dei geni e dell'ambiente nello sviluppo della malattia di Alzheimer, Bergem *et al.* [15] riportano correlazioni tetracoriche di 0,97 per i MZ e 0,69 per i DZ, da cui ottengono, con le formule descritte, $h^2 = 0,55$, $c^2 = 0,42$ ed $e^2 = 0,02$; l'interpretazione di queste stime è che i fattori genetici sono responsabili del 55% della variabilità nello sviluppo della malattia di Alzheimer, con i rimanenti 42% e 2% a carico, rispettivamente, di fattori ambientali condivisi e non condivisi.

Tabella 2 | *Correlazione per l'altezza corporea in gemelli di diversi Paesi*

	Australia	Danimarca	Finlandia	Italia	Olanda	Norvegia	Svezia	Regno Unito
MZm	0,87	0,89	0,92	0,94	0,89	0,87	0,89	nd
DZm	0,42	0,47	0,53	0,57	0,47	0,49	0,56	nd
MZf	0,84	0,89	0,87	0,94	0,90	0,89	0,89	0,88
DZf	0,49	0,55	0,53	0,49	0,49	0,49	0,49	0,56
DOS	0,46	0,50	0,49	0,30	0,43	0,44	nd	nd

MZm = monozigoti maschi, DZm = dizigoti maschi, MZf = monozigoti femmine, DZf = dizigoti femmine, DOS = dizigoti di sesso opposto
nd = non disponibile.

METODI AVANZATI: I MODELLI DI EQUAZIONI STRUTTURALI, LA *DF-ANALYSIS* E I *CORRELATED FRAILTY MODELS*

Un approccio statistico molto potente e flessibile, inizialmente applicato per lo più alle scienze sociali (econometria, sociologia, psicometria), è quello dei modelli di equazioni strutturali (SEM). Negli ultimi anni, tale approccio ha avuto un numero crescente di applicazioni nel campo dell'epidemiologia genetica e, in particolare, degli studi sui gemelli. Nei SEM, accanto alle variabili osservate, che descrivono i valori fenotipici di un dato carattere rilevato sui gemelli, si considerano variabili non misurabili (latenti), che servono a modellare l'effetto di fattori genetici ed ambientali sul fenotipo in studio. Il caso più semplice è illustrato in *Fig. 1*, dove P è il fenotipo osservato sui due gemelli, mentre A, C ed E sono variabili latenti, con varianza unitaria, che rappresentano, rispettivamente, fattori genetici, ambientali condivisi ed ambientali non condivisi dai gemelli nelle coppie. I parametri a, c, e quantificano l'influenza esercitata dai fattori latenti sul fenotipo. Le frecce unidirezionali indicano influenze causali, quelle bidirezionali la covariazione tra variabili. Il fatto che i gemelli MZ siano geneticamente identici, e che quelli DZ condividano, in media, il 50% del loro patrimonio genetico, si traduce in una correlazione tra i fattori latenti genetici per i due gemelli pari a 1 nel caso MZ e a 0,5 nel caso DZ. Per la equal environments assumption, la correlazione tra i fattori ambientali condivisi è posta uguale a 1 in entrambi i gruppi di zigosità. Inoltre, per definizione, nessuna correlazione esiste tra i fattori ambientali non condivisi. Le frecce unidirezionali implicano la regressione lineare del fenotipo sulle variabili latenti, ovvero l'equazione $P = aA + cC + eE$, valida per ciascun gemello. Tale equazione consente di decomporre la varianza fenotipica e le covarianze tra i fenotipi dei MZ e dei DZ in funzione dei coefficienti a, c, e. Sotto le ipotesi sopra descritte, la varianza e le covarianze predette dal modello sono date da $V_P = a^2 + c^2 + e^2$, $Cov(MZ) = a^2 + c^2$, $Cov(DZ) = 0{,}5a^2 + c^2$ [3]. La stima dei parametri a, c, e consiste nel minimizzare un'opportuna funzione di adattamento che fornisce una misura della distanza tra le varianze-covarianze osservate e quelle attese (cioè predette dal modello), nei MZ e nei DZ. Più precisamente, se F è la funzione di adattamento, $\sum_{MZ}(\theta)$ e $\sum_{DZ}(\theta)$ sono le matrici di varianza-covarianza attese nei MZ e nei DZ, e $F(\sum_{MZ}(\theta))$ e $F(\sum_{DZ}(\theta))$ sono le discrepanze tra queste matrici e quelle osservate S_{MZ} e S_{DZ}, il processo di stima consiste nel trovare il vettore θ che rende minima la quantità $F(\sum_{MZ}(\theta), \sum_{DZ}(\theta)) = F(\sum_{MZ}(\theta)) + F(\sum_{DZ}(\theta))$. Sono state proposte varie funzioni di adattamento, alle quali corrispondono diversi metodi di stima. In genere, si ricorre al metodo di massima verosimiglianza per le buone proprietà statistiche che caratterizzano gli stimatori. In tal caso, la significatività di uno o più parametri è verificata con dei test chi-quadro di rapporto di verosimiglianza, basati su modelli annidati. In sostanza, si va a confrontare un modello saturo con un sotto-modello avente un minor numero di parametri da stimare: il rapporto tra le verosimiglianze e la differenza tra il numero di parametri dei due modelli forniscono, rispettivamente, la statistica chi-quadro ed i relativi gradi di libertà. Le stime dei parametri sono esibite sotto il modello più parsimonioso, ovvero quello da cui sono esclusi i parametri non significativi. Ad esempio, si possono considerare le coppie di modelli annidati (ACE, CE) e (ACE, AE) per verificare, con test chi-quadro di rapporto di verosimiglianza ad un grado di libertà, la significatività delle influenze genetiche e di quelle ambientali condivise, rispettivamente. Questi confronti gerarchici non riguardano mai il fattore latente E: esso ingloba anche l'errore di misura che interviene nella rilevazione del fenotipo, e quindi un modello senza tale fattore non sarebbe plausibile. Dalle stime dei parametri a, c, e si possono derivare le proporzioni di varianza totale attribuibili a fattori genetici (ereditabilità) ed ambientali condivisi e non condivisi, tramite le relazioni a^2/V_P, c^2/V_P ed e^2/V_P [1, 3].

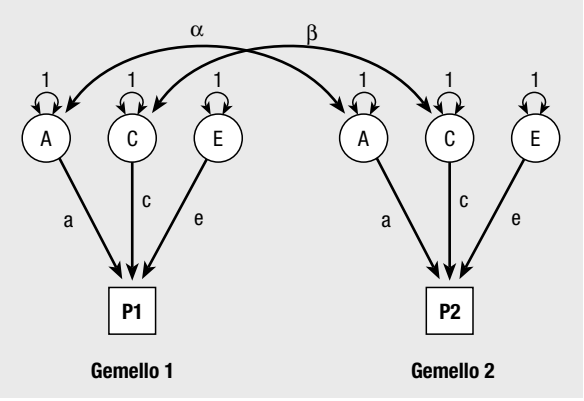

Fig. 1 | *Modello di equazioni strutturali per l'analisi univariata di dati su gemelli.*
P = fenotipo osservato su ogni gemello (i numeri 1 e 2 si riferiscono ai due gemelli di una stessa coppia);
A = fattori genetici; C = fattori ambientali condivisi dai gemelli nella coppia; E = fattori ambientali non condivisi dai gemelli nella coppia;
α = 1 (per i monozigoti) oppure 0,5 (per i dizigoti);
β = 1 (per i monozigoti e per i dizigoti).

Il semplice modello in *Fig. 1* può essere esteso ad una moltitudine di situazioni più complesse. Tramite la simultanea analisi di dati provenienti da gemelli MZ e DZ, maschi e femmine, i SEM permettono di investigare eventuali differenze tra i sessi nelle stime dei parametri, ovvero nella forza delle influenze genetiche ed ambientali. É anche possibile testare l'ipotesi che gli stessi geni siano espressi nei maschi e nelle femmine, includendo gemelli DZ di sesso opposto: se, per questi, la correlazione tra i fattori latenti genetici, stimata come parametro aggiuntivo nel modello, risulta significativamente più bassa di 0,5, che è il valore al quale viene fissata la stessa correlazione per i gemelli DZ dello stesso sesso, ciò indica che geni diversi potrebbero influenzare il carattere nei due sessi. Questioni del genere sono affrontate in uno studio sull'indice di massa corporea in gemelli di otto registri,

Tabella 3 | *Proporzioni (%) di varianza, sotto i modelli più parsimoniosi, dovute a fattori genetici (A) ed ambientali condivisi (C) e non condivisi (E) per l'indice di massa corporea in gemelli di diversi Paesi, e test chi-quadro per il confronto tra i modelli ACE ed AE*

	Maschi			Femmine			ACE vs AE
	A	C	E	A	C	E	$\chi^2_{(2)}$
Australia	67	---	33	72	---	28	0
Danimarca	77	---	23	73	---	27	0
Finlandia	74	---	26	78	---	22	0
Italia	81	---	19	85	---	15	4,59
Olanda	66	---	34	81	---	19	0
Norvegia	45	31	24	74	---	26	14,49*
Svezia	77	---	23	73	---	27	0
Regno Unito	nd	nd	nd	75	---	25	2,02**

$\chi^2_{(2)}$ = chi-quadro con 2 gradi di libertà per il test simultaneo sull'ambiente condiviso nei maschi e nelle femmine
--- = il relativo parametro è fissato a zero
** $p < 0,001$*
*** = 1 grado di libertà per il test sull'ambiente condiviso*
nd = non disponibile

tra cui quello italiano [16], nell'ambito del progetto europeo GenomEUtwin [13]. In questo studio, differenze tra maschi e femmine nelle stime dei parametri sono rilevate in tutti i Paesi, tranne l'Italia; inoltre, fatta eccezione per l'Olanda e la Norvegia, negli altri Paesi emergono indicazioni per geni diversi espressi nei due sessi. In *Tab. 3* sono riportate, per i vari Paesi, le proporzioni di varianza dell'indice di massa corporea, insieme ai test chi-quadro per il confronto tra i modelli ACE ed AE, riguardante gli effetti dell'ambiente condiviso. L'ereditabilità varia tra il 45% e l'81% nei maschi, e tra il 72% e l'85% nelle femmine, mentre i fattori ambientali condivisi risultano significativi solo per i maschi norvegesi.

Mediante i SEM si può anche rilevare un'interazione geni-ambiente, incorporando misure di esposizione ambientale, e verificando se la stima dell'ereditabilità differisce tra gemelli esposti e non esposti. Ad esempio, Heath *et al.* [17] trovano, per la depressione, un'ereditabilità più bassa nelle donne sposate rispetto a quelle non sposate, suggerendo che la relazione matrimoniale può attenuare l'espressione della suscettibilità genetica individuale alla patologia. Talvolta, ha senso considerare la reciproca influenza tra i fenotipi dei due gemelli. Essa si traduce in due relazioni causali aggiuntive nel diagramma in *Fig. 1*, e cioè quella dal primo gemello al secondo e viceversa, quantificabili attraverso uno stesso parametro. Una condizione di questo genere si verifica, ad esempio, quando l'uso di alcol o droga da parte di un gemello incoraggia lo stesso uso nell'altro, oppure quando l'estroversione di uno dei due gemelli ha un effetto inibitorio sull'altro. Nel caso di caratteri non stazionari ma dipendenti dal tempo, come la pressione sanguigna e la densità minerale delle ossa, che tendono rispettivamente ad aumentare e a diminuire con l'invecchiamento, i SEM consentono di includere l'età tra le sorgenti di variabilità del carattere in studio; in questo modo, si può distinguere l'effetto dell'età da quello dell'ambiente condiviso nel determinare la somiglianza all'interno delle coppie [3].

Oltre che per stimare le influenze genetiche ed ambientali esercitate su un certo carattere da sorgenti latenti, non direttamente misurabili, i SEM possono anche essere utilizzati per testare gli effetti di fattori genetici precisamente individuati. Un metodo fra i più noti è la cosiddetta analisi di *linkage* per la mappatura di QTL (quantitative trait loci), ovvero di geni che influenzano tratti quantitativi. Nel caso dello studio di gemelli, lo scopo del metodo è quello di stabilire se il grado di affinità genetica tra i gemelli nelle coppie, relativo ad un certo QTL, è importante per spiegare il livello di somiglianza fenotipica. Il modello usato è simile a quello presentato in *Fig. 1*, con un fattore latente (Q) ed un parametro (q) aggiuntivi per ciascun gemello, a descrivere rispettivamente il QTL ed il suo effetto sul carattere, e con il fattore latente A a rappresentare le influenze genetiche residue. Un concetto fondamentale alla base del metodo è lo stato IBD *(identical by descent)* di una coppia di individui, riferito al QTL. La regione cromosomica in cui un gene è localizzato è detta *locus*. Ad ogni *locus*, un individuo possiede due varianti del gene, chiamate alleli, una di origine paterna ed una di origine materna. Due individui, in particolare due gemelli, si dicono condividere n (n = 0, 1, 2) alleli IBD ad un certo *locus* se condividono l'origine genitoriale per n alleli. I gemelli MZ, geneticamente identici, condividono 2 alleli IBD in tutto il genoma, e quindi sono perfettamente correlati rispetto ad un qualunque *locus*; nel caso dei DZ, invece, la correlazione ad uno specifico *locus*, tipicamente indicata con π, viene misurata come proporzione attesa di alleli IBD, ovvero come probabilità che un allele scelto a caso da un gemello condivida l'origine genitoriale con uno dei due alleli del co-gemello. Allora, per la varianza e le covarianze predette dal modello valgono le equazioni $V_P = a^2 + c^2 + e^2 + q^2$, $Cov(MZ) = a^2 + c^2 + q^2$, $Cov(DZ) = 0,5a^2 + c^2 + \pi q^2$ [4]. Considerando la coppia di modelli annidati (ACEQ, ACE) si può verificare, con test chi-quadro di rapporto di verosimiglianza

ad un grado di libertà, la significatività dell'influenza esercitata dal QTL sulla varianza del tratto in studio. Mediante il rapporto q^2/V_P è anche possibile stimare l'ereditabilità specifica per il QTL. Con un approccio di questo tipo, applicato ad un campione di gemelli tedeschi MZ e DZ, Knoblauch *et al.* [18] trovano forti indicazioni di un effetto di un gene del cromosoma 13 sulla variabilità dei livelli di HDL, LDL e colesterolo totale nel sangue.

La flessibilità dei SEM rende agevole l'estensione all'analisi multivariata. Quando l'informazione su due o più caratteri è disponibile sui gemelli del campione, i SEM consentono di decomporre non solo le varianze dei singoli caratteri e le loro covarianze tra i gemelli nelle coppie, ma anche le covarianze tra i diversi caratteri in una componente genetica ed una ambientale. In termini pratici, l'obiettivo è capire se la correlazione fra tratti quantitativi o la co-morbidità tra condizioni patologiche siano dovute a comuni fattori genetici o ambientali. La decomposizione delle covarianze osservate tra più caratteri può essere effettuata tramite vari modelli, che assumono meccanismi latenti alternativi all'origine delle covarianze stesse. Un modello spesso utilizzato è quello di *Cholesky* [3], il cui diagramma è riportato in *Fig. 2* nel caso bivariato. Nel diagramma, X e Y sono due caratteri misurati sui gemelli, A_C, C_C ed E_C rappresentano, rispettivamente, fattori genetici ed ambientali (condivisi e non condivisi dai gemelli nelle coppie) che influenzano simultaneamente X e Y, mentre A_S, C_S ed E_S indicano analoghi fattori specifici per Y. Riguardo alle correlazioni tra le variabili latenti genetiche ed ambientali, valgono le stesse assunzioni descritte nel caso univariato. Questo modello implica per le varianze di X e Y, e per la covarianza tra X e Y le decomposizioni $V_X = a^2_{11} + c^2_{11} + e^2_{11}$, $V_Y = (a^2_{21} + a^2_{22}) + (c^2_{21} + c^2_{22}) + (e^2_{21} + e^2_{22})$, $Cov(X,Y) = a_{11}a_{21} + c_{11}c_{21} + e_{11}e_{21}$. In aggiunta all'ereditabilità di X e Y, altre quantità interessanti che si possono stimare sono la correlazione genetica e quelle ambientali tra X e Y, date da $r_A = a_{11}a_{21}/ [a^2_{11}(a^2_{21} + a^2_{22})]^{1/2}$, $r_C = c_{11}c_{21}/[c^2_{11}(c^2_{21} + c^2_{22})]^{1/2}$, $r_E = e_{11}e_{21}/[e^2_{11}(e^2_{21} + e^2_{22})]^{1/2}$; tali correlazioni misurano il grado con cui X e Y condividono gli stessi fattori genetici ed ambientali [3]. Inoltre, con dei test chi-quadro di rapporto di verosimiglianza ad un grado di libertà si può verificare la significatività dei singoli parametri a_{21}, c_{21}, e_{21}, e testare l'ipotesi che la covariazione tra i due tratti sia l'effetto di geni condivisi (pleiotropismo) o della sovrapposizione di influenze ambientali. L'indicazione di fattori genetici ed ambientali comuni a X e Y può essere derivata dalla correlazione tra X in un gemello e Y nel co-gemello, nei MZ e nei DZ; ad esempio, se questa correlazione è maggiore nei MZ rispetto ai DZ, ciò suggerisce l'esistenza di geni condivisi dai due caratteri. In uno studio longitudinale su coppie di gemelli inglesi, finalizzato a chiarire l'origine dell'associazione temporale tra insorgenza iniziale di ansia e sviluppo successivo di depressione, Rice *et al.* [19] utilizzano un modello di *Cholesky* bivariato, e trovano evidenze di fattori sia genetici che ambientali comuni ai due disturbi.

Dal punto di vista tecnico, l'implementazione dei SEM è alquanto laboriosa, e richiede l'utilizzo di software specifici, come ad esempio LISREL [20] e Mx [21]. Tali software permettono di gestire, attraverso gli strumenti dell'algebra matriciale, le complesse strutture di varianza-covarianza implicate dalle variabili osservate.

Una strategia alternativa ai SEM, meno flessibile e potente, per quantificare il peso relativo dei geni e dell'ambiente sull'espressione di caratteri umani, è quella introdotta da DeFries e Fulker, e nota come *DF-analysis*. Come per i SEM, si assume che il fenotipo in studio sia il risultato di influenze genetiche e di influenze esercitate da fattori ambientali condivisi *(common environment)* e non condivisi *(unique environment)* dai gemelli nelle coppie. In sostanza, la *DF-analysis* con-

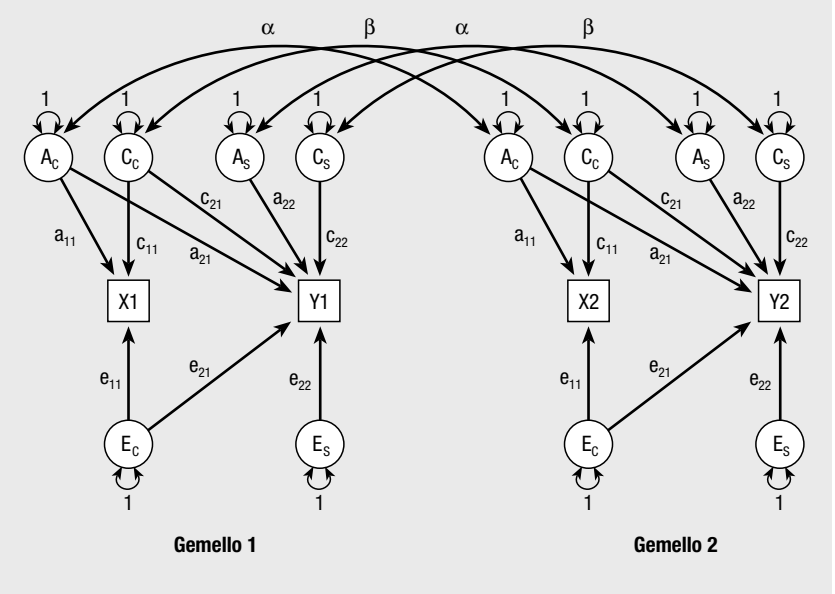

Fig. 2 | *Modello di equazioni strutturali per l'analisi bivariata di dati su gemelli. X, Y = due fenotipi osservati su ogni gemello (i numeri 1 e 2 si riferiscono ai due gemelli di una stessa coppia); A_C = fattori genetici comuni a X e Y; C_C = fattori ambientali condivisi dai gemelli nella coppia, comuni a X e Y; E_C = fattori ambientali non condivisi dai gemelli nella coppia, comuni a X e Y; A_S = fattori genetici specifici per Y; C_S = fattori ambientali condivisi dai gemelli nella coppia, specifici per Y; E_S = fattori ambientali non condivisi dai gemelli nella coppia, specifici per Y; α = 1 (per i monozigoti) oppure 0,5 (per i dizigoti); β = 1 (per i monozigoti e per i dizigoti).*

siste nella regressione lineare del fenotipo di uno dei due gemelli nella coppia sul fenotipo del co-gemello e su altre variabili. L'equazione base è data da $P_1 = A + B_1P_2 + B_2R + B_3(P_2 \times R) + E$, dove i P_i (i = 1, 2) sono misurazioni di uno stesso tratto quantitativo sui due gemelli nella coppia, R è la correlazione genetica tra i due gemelli nella coppia (R = 1 per i MZ, R = 0,5 per i DZ), E è il termine di errore, mentre i B_i (i = 1, 2, 3) sono i coefficienti di regressione, stimati col metodo dei minimi quadrati. Si dimostra che, sotto la equal environments assumption, B_1 fornisce una stima della proporzione di varianza del fenotipo spiegata dai fattori ambientali condivisi, mentre B_3 è una stima dell'ereditabilità [5, 22]. La significatività dei parametri B_1 e B_3 è verificata usando il test F di Fisher per confrontare i valori di R^2 relativi a modelli annidati (es., ACE e AE per il parametro B_1). Il semplice modello di regressione sopra presentato può essere esteso a situazioni più complesse, ad esempio per incorporare l'informazione sul sesso, e testare eventuali differenze nelle stime tra maschi e femmine.

La *DF-analysis* è stata inizialmente sviluppata per analizzare coppie di gemelli selezionate sulla base di probandi, ovvero gemelli in cui il carattere in esame assume valori estremi; in questo caso, si effettua la regressione lineare del co-gemello sul probando. Quando, invece, i gemelli sono un campione casuale della popolazione, l'ambiguità sull'ordine dei due membri nella coppia, e quindi sulla scelta delle variabili dipendente e indipendente per la regressione, è risolta considerando ogni coppia due volte, in un dato ordine ed in quello inverso *(double-entry)*, e tenendo conto della raddoppiata dimensione del campione per aggiustare gli errori standard delle stime.

Rodgers *et al.* [22], in uno studio riguardante gli effetti genetici ed ambientali sulla fertilità, utilizzano la *DF-analysis* per analizzare, in coppie di gemelli del registro danese, i due tratti quantitativi dati dal "numero di figli", come indicatore di successo riproduttivo, e dalla "età al primo tentativo di gravidanza", quale misura di motivazione alla riproduzione; le stime di ereditabilità che essi ottengono sono moderate, e variano dal 28% al 53% sotto il modello AE.

Infine, vi è una classe di modelli, i *correlated frailty models*, i quali permettono di combinare metodi di analisi della sopravvivenza e di genetica quantitativa per stimare, a partire dall'informazione sull'età di insorgenza di una malattia o sulla durata della vita *(life span)* in coppie di gemelli MZ e DZ, la componente genetica e quella ambientale della predisposizione *(frailty)* alla morbidità e alla mortalità. L'idea di fondo è che un'eventuale maggiore associazione tra i tempi di sopravvivenza in gemelli di coppie MZ rispetto a gemelli di coppie DZ può indicare un ruolo dei fattori genetici nel determinare la predisposizione all'evento di interesse. Attualmente, tra i numerosi modelli di tipo *frailty* [23], i *correlated frailty models* sono quelli più utilizzati per lo studio dei gemelli. La presenza di un parametro che rappresenta la correlazione tra le *frailty* dei due gemelli nella coppia consente di applicare in modo diretto il metodo gemellare (basato sulle correlazioni nei MZ e nei DZ) per effettuare l'analisi genetica della *frailty*, ed in particolare per stimare l'ereditabilità. Nel modello, il rischio istantaneo di un gemello nella coppia si scrive come $\mu_i(t) = Z_i\mu_{0i}(t)$ (i = 1, 2), dove $\mu_{0i}(t)$ è il rischio di base, e Z_i è la *frailty* individuale. Inoltre: I) $Z_i = A + C + E$ (i = 1, 2), con A, C ed E a rappresentare le influenze genetiche e quelle ambientali condivise e non condivise dai gemelli nella coppia; II) Z_1 e Z_2 sono correlate; III) per le correlazioni tra Z_1 e Z_2 nei MZ e nei DZ valgono le decomposizioni $\rho(MZ) = a^2 + c^2$ e $\rho(DZ) = 0{,}5a^2 + c^2$, dove a^2 e c^2 sono le proporzioni di varianza della *frailty* spiegate dai geni (ereditabilità) e dall'ambiente condiviso [6]. Con questo metodo, applicato all'analisi congiunta di dati sul *life span* in gemelli dei registri danese, svedese e finlandese, Iachine *et al.* [24] trovano una stima di ereditabilità per la *frailty* intorno al 50%, ovvero circa la metà della variabilità inter-individuale nella predisposizione alla mortalità è attribuibile a fattori genetici.

Tra i software che consentono l'implementazione dei modelli di tipo *frailty*, si usa spesso GAUSS [25] per la stima dei parametri col metodo della massima verosimiglianza.

In conclusione, sono stati illustrati i principali metodi statistici per l'analisi di dati su gemelli, con esempi di applicazioni tratti da studi pubblicati, alcuni dei quali riferiti ai dati del Registro Nazionale Gemelli. Si esprime l'augurio che il presente lavoro venga considerato dal lettore nella sua reale funzione, che è quella di fornire, senza pretese di completezza ma incoraggiando all'approfondimento, una panoramica generale ed unitaria su argomenti spesso descritti in modo separato nella letteratura.

Ricevuto il 27 luglio 2005.
Accettato il 2 novembre 2005.

Bibliografia

1. Sham P. *Statistics in human genetics*. New York: Wiley; 1998.
2. Witte JS, Carlin JB, Hopper JL. Likelihood-based approach to estimating twin concordance for dichotomous traits. *Genet Epidemiol* 1999;16:290-304.
3. Neale MC, Cardon LR. *Methodology for genetic studies of twins and families*. Dordrecht: Kluwer Academic Publisher; 1992.
4. Posthuma D, Beem AL, de Geus EJC, van Baal GCM, von Hjelmborg JB, Iachine I, Boomsma DI. Theory and practice in quantitative genetics. *Twin Research* 2003;6(5):361-76.
5. DeFries JC, Fulker DW. Multiple regression analysis of twin data. *Behav Genet* 1985;15(5):467-73.
6. Yashin AI, Iachine IA. Genetic analysis of durations: correlated frailty model applied to survival of Danish twins. *Genet Epidemiol* 1995;12(5):529-38.
7. Stazi MA, Cotichini R, Patriarca V, Brescianini S, Fagnani C, D'Ippolito C, Cannoni S, Ristori G, Salvetti M. The Italian twin project: from the personal identification number to a national twin registry. *Twin Research* 2002;5(5):382-6.

8. Spector TD, Snieder H, MacGregor AJ. *Advances in twin and sib-pair analysis*. Oxford University Press; 2000.
9. Armitage P, Berry G, Matthews JNS. *Statistical methods in medical research*. Oxford: Blackwell Science; 2002.
10. Greco L, Romino R, Coto I, Di Cosmo N, Percopo S, Maglio M, Paparo F, Gasperi V, Limongelli MG, Cotichini R, D'Agate C, Tinto N, Sacchetti L, Tosi R, Stazi MA. The first large population based twin study of coeliac disease. *Gut* 2002;50(5):624-8.
11. Hyttinen V, Kaprio J, Kinnunen L, Koskenvuo M, Tuomilehto J. Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. *Diabetes* 2003;52(4):1052-5.
12. Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, Dunkel L, de Lange M, Harris JR, Hjelmborg J, Luciano M, Martin NG, Mortensen J, Nisticò L, Pedersen NL, Skytthe A, Spector TD, Stazi MA, Willemsen G, Kaprio J. Heritability of adult body height: a comparative study of twin cohorts in eight Countries. *Twin Research* 2003;6(5):399-408.
13. Progetto europeo GenomEUtwin. Disponibile all'indirizzo: http://www.genomeutwin.org; ultima consultazione luglio 2005.
14. Jensen AR. The problem of genotype-environment correlation in the estimation of heritability from monozygotic and dizygotic twins. *Acta Genet Med Gemellol* 1976;25:86-99.
15. Bergem ALM, Engedal K, Kringlen E. The role of heredity in late-onset Alzheimer disease and vascular dementia. *Arch Gen Psychiatry* 1997;54:264-70.
16. Schousboe K, Willemsen G, Kyvik KO, Mortensen J, Boomsma DI, Cornes BK, Davis CJ, Fagnani C, Hjelmborg J, Kaprio J, de Lange M, Luciano M, Martin NG, Pedersen N, Pietilainen KH, Raissanen A, Saarni S, Sorensen TIA, van Baal GCM, Harris JR. Sex differences in heritability of BMI: A comparative study of results from twin studies in eight Countries. *Twin Research* 2003;6(5):409-21.
17. Heath AC, Eaves LJ, Martin NG. Interaction of marital status and genetic risk for symptoms of depression. *Twin Research* 1998;1:119-22.
18. Knoblauch H, Muller-Myhsok B, Busjahn A, Ben Avi L, Bahring S, Baron H, Heath SC, Uhlmann R, Faulhaber HD, Shpitzen S, Aydin A, Reshef A, Rosenthal M, Eliav O, Muhl A, Lowe A, Schurr D, Harats D, Jeschke E, Friedlander Y, Schuster H, Luft FC, Leitersdorf E. A cholesterol-lowering gene maps to chromosome 13q. *Am J Hum Genet* 2000;66:157-66.
19. Rice F, van den Bree MB, Thapar A. A population-based study of anxiety as a precursor for depression in childhood and adolescence. *BMC Psychiatry* 2004;4:43.
20. Joreskog KG, Sorbom D. LISREL 7: *a guide to the program and applications*. Chicago: SPSS Inc.; 1989.
21. Neale MC, Boker SM, Xie G, Maes H. Mx: *statistical modelling*. Richmond, VA: Department of Psychiatry, Virginia Commonwealth University; 2002.
22. Rodgers JL, Kohler HP, Kyvik KO, Christensen K. Behavior genetic modeling of human fertility: findings from a contemporary Danish Twin Study. *Demography* 2001;38(1):29-42.
23. Hougaard P. *Analysis of multivariate survival data*. New York: Springer-Verlag; 2000.
24. Iachine IA, Holm NV, Harris JR, Begun AZ, Iachina MK, Laitinen M, Kaprio J, Yashin AI. How heritable is individual susceptibility to death? The results of an analysis of survival data on Danish, Swedish and Finnish twins. *Twin Research* 1998;1(4):196-205.
25. *GAUSS*. Maple Valley, WA: Aptech Systems Inc.; 2002.