UNIVERSITA' DEGLI STUDI DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Pediatria

SCUOLA DI DOTTORATO DI RICERCA IN MEDICINA DELLO SVILUPPO E SCIENZE DELLA PROGRAMMAZIONE

INDIRIZZO: SCIENZE DELLA PROGRAMMAZIONE

XX CICLO

ASPETTI CRITICI NEGLI STUDI CLINICI: PROBLEMI DI METODO E APPLICAZIONE

Direttore della Scuola: Ch.mo Prof. Giuseppe Basso

Supervisore: Ch.ma Prof. Paola Facchin

Dottorando : Riccardo Pertile

DATA CONSEGNA TESI
Gennaio 2008

INDICE

Introduzione

La progettazione di uno studio clinico è al giorno d'oggi la forma base più comune per fare ricerca in qualsiasi campo medico ed epidemiologico. Può riguardare il lavoro di un singolo ricercatore o di un unico medico, come può essere seguito da una grande equipe di ricerca. Non sempre, però, i risultati ottenuti sono attendibili e veritieri perché in alcuni casi le premesse iniziali non sono solide oppure perchè in uno dei passaggi decisivi nel corso dello studio non si sono tenute in considerazione opportune regole per un corretto svolgimento della ricerca. In teoria il lavoro di verifica della qualità e dell'accuratezza delle metodologie statistiche utilizzate negli articoli oggetto di pubblicazione dovrebbe essere inutile, ma questo purtroppo non è sempre vero poiché anche ciò che viene pubblicato può non essere corretto (Glantz, 1997; Choi & Pak, 2006). Da parte di chiunque voglia cimentarsi nella lettura di bibliografia in campo clinico è utile saper verificare per conto proprio i metodi statistici usati dagli autori, mentre per chi abbia intenzione anche di intraprendere la strada della ricerca in questo campo, è indispensabile conoscerne gli ostacoli principali e sapere come affrontarli.

Esistono due principali tipologie di disegni di studio statistici, il disegno di studio sperimentale o *clinical trial* e l'indagine o *survey*. Il *clinical trial* richiede un'interferenza pianificata secondo il corso naturale degli eventi, tale da poterne osservare gli effetti. Nella maggior parte dei casi un *clinical trial* viene adottato in studi atti a testare un nuovo farmaco o una nuova terapia su un particolare gruppo di soggetti, vale a dire pazienti affetti da una specifica patologia. Nell'indagine, o *survey*, il ricercatore è un osservatore meno attivo, che interferisce il meno possibile con i fenomeni da registrare (Pocock, 1989; Breslow, 1996; Friedman, 1995). Il disegno e la conduzione di qualsiasi tipo di studio clinico richiedono sicuramente di esaminare ed indagare importanti questioni bio-mediche, ma richiedono anche di essere basati su una rigorosa metodologia che possa fornire una risposta corretta al quesito di ricerca. Un ultimo aspetto, sempre più di primaria importanza a livello internazionale, concerne le considerazioni etiche relative all'adesione dei pazienti partecipanti allo studio, i cui rischi di salute devono venire minimizzati (Sutherland et al., 1994; Glantz, 2007). In uno studio clinico, infatti, si prenderà in considerazione una specifica tipologia di unità statistiche, vale a dire le casistiche cliniche. Si tratta di gruppi di individui (pazienti o volontari sani) che prendono parte allo studio allo scopo di generare conoscenza in campo biomedico, ma la loro salute deve avere la priorità su qualsiasi fine scientifico.

Il lavoro oggetto della presente tesi di dottorato è stato suddiviso in cinque sezioni principali seguendo uno schema che possa guidare il lettore nell'esaminare gli aspetti critici che si possono affrontare in primo luogo nella gestione dei dati, in secondo luogo nella

progettazione, nello sviluppo e nelle fasi finali di uno studio clinico. Le cinque sezioni possono contenere uno o più capitoli in base all'estensione del problema. La sezione *A*, *Problemi di gestione di banche dati*, propone un capitolo sul *Record Linkage* ed un secondo sui *Missing Values*. Diversi studi si centrano su database ottenuti abbinando i record di dataset originariamente separati. Le procedure e le tecniche necessarie per eseguire questa fusione vengono raggruppate sotto il nome di *Record Linkage*. Il problema principale consiste nella disponibilità di variabili, nei diversi dataset, che rappresentino attendibili chiavi di unione o identificatori, in modo che le informazioni riportate in differenti archivi e relative alle stesse unità vengano integrate attraverso la creazione di un singolo database (Fellegi & Sunter, 1969; Newcombe, 1988; Armstrong & Mayda, 1992; Quan et al., 2006). Tale database finale è generalmente "affetto" da valori mancanti su una o più variabili, ma esistono tecniche specifiche che, attraverso programmi statistici, riescono a calcolare un opportuno valore per l'unità priva d'informazione. In letteratura si trovano quattro gruppi di metodi per il trattamento di dati mancanti: metodi basati sulle sole unità osservate, procedure di ponderazione, metodi basati sui modelli e metodi di imputazione, quest'ultimi i maggiormente adottati (Kalton & Kasprzyk, 1982; Little & Rubin, 1987; Barzi & Woodward, 2004; Acock, 2005).

La sezione *B*, *Disegni di studio*, rappresenta il nucleo del presente lavoro. Attraverso il terzo capitolo *Progettazione di uno studio* si analizzano in un primo tempo i *clinical trial*, in un secondo tempo le indagini statistiche, scorporando quest'ultime in studi coorte e studi caso-controllo. In uno studio coorte (altrimenti detto studio *follow-up* o studio prospettico) uno o più gruppi di individui vengono definiti secondo la presenza o l'assenza d'esposizione a uno o più fattori considerati di rischio per una patologia o una gamma di patologie. Gli individui oggetto di studio vengono seguiti prospetticamente nel tempo, in modo da poterne osservare l'incidenza di malattia/e e correlarla alla classificazione per fattori eziologici (Greenland, 1977; Hennekens & Buring, 1987; Adams et al., 2007). Uno studio caso-controllo (o studio retrospettivo) fornisce una strategia di ricerca per investigare fattori che possono prevenire o causare una particolare malattia. Il metodo comporta un confronto tra pazienti affetti dalla patologia (casi) con un gruppo di controllo (persone sane). Il confronto ha l'obiettivo di scoprire i fattori che possono differire nei due gruppi spiegando il manifestarsi della patologia nei pazienti (Schlesselman, 1982; Armitage & Berry, 1996).

La sezione *C* concerne il *Trattamento delle variabili* (capitolo 4) e le *Possibili classificazioni della variabile di outcome* (capitolo 5). Il quarto capitolo si concentra sulle tipologie di variabili che si possono incontrare in uno studio clinico, sulle loro scale di misura, sulla

sintetizzazione attraverso indicatori e sulle possibili trasformazioni. Quest'ultime possono essere necessarie per la linearizzazione delle relazioni tra variabili, per la normalizzazione della distribuzione degli errori e delle osservazioni e per la stabilizzazione delle varianze (Vajani,1997; Soliani, 2005). Il quinto capitolo si occupa invece delle classificazioni della variabile di outcome (o variabile dipendente) in uno studio clinico, che possono riguardare una specifica patologia, i sintomi riportati dal paziente oppure una conseguenza di malattia, vale a dire una menomazione, una disabilità o un handicap. In base alla scelta dello studio esistono numerose classificazioni dell'Organizzazione Mondiale delle Sanità (OMS), o di altri Enti, indicate non solo per il medico che può essere indirizzato nell'individuare la specifica patologia dai sintomi del paziente, ma anche al ricercatore che può ottenere un codice della patologia, archiviandone i dati medici e clinici più facilmente in un apposito registro informatizzato o database.

Le ultime due sezioni del presente lavoro affrontano i problemi nella scelta del metodo d'analisi multivariata. La sezione *D* (capitolo 6) mira a dare una sintesi dei principali metodi d'analisi, suddividendoli in base alla simmetria o meno delle variabili. Se esiste una distinzione tra variabile dipendente e variabili esplicative (modello d'analisi asimmetrico) si sceglie un metodo d'analisi di regressione multipla (*stepwise* o *logistica*) o l'analisi discriminatoria, sulla base del tipo (e della quantità) di variabile/i di outcome; se invece il modello d'analisi è simmetrico e le variabili in analisi sono tutte sullo stesso piano si sceglie tra l'*analisi fattoriale*, il metodo delle *componenti principali*, il metodo d'*analisi delle corrispondenze* e la *cluster analysis* sulla base degli obiettivi specifici dello studio (Kendall, 1980; Chatfield & Collins, 1980; Krzanowski, 1988; Mardia et al., 1979; Fabbris, 1997).

Il settimo ed ultimo capitolo (sezione *E*) si occupa di uno specifico metodo d'analisi, vale a dire l'analisi di regressione logistica, cogliendone gli aspetti centrali ed esaminando il caso dell'analisi di regressione logistica condizionata per il disegno *Matched Case-Control Study*. Nel disegno *Matched Case-Control Study* i casi vengono accoppiati ad uno o più controlli sulla base di variabili che si presumono associate con l'outcome. Dato che casi e controlli sono simili rispetto alle variabili d'accoppiamento, una loro differenza nei confronti della patologia è dovuta ad altri fattori non considerati per l'abbinamento (Breslow & Day, 1980; Schlesselman, 1982; Kelsey et al., 1986; Rothman & Greenland, 1998).

Ogni capitolo è strutturato in modo tale che nella prima parte sia riportata una sintesi dei metodi teorici e pratici proposti dalla letteratura per risolvere il problema in questione, mentre alla fine di ogni capitolo è stata inserita una sezione che spiega come il problema trattato è stato affrontato ed applicato in uno specifico studio clinico-genetico al Children's Hospital

dell'Università di Oulu – Finlandia. In tale Centro è stato svolto un lavoro sostanziale sulla creazione di un database inerente a dati clinici e genetici di nati prematuri nei tre principali ospedali della Finlandia centro-settentrionale, Oulu, Tampere e Seinäjoki. Attraverso l'utilizzo del database totale costituito dai pazienti dei tre nosocomi, è stato condotto uno studio caso-controllo (*match* 1-1) clinico-genetico sull'associazione tra Sindrome da Distress Respiratorio (in inglese Respiratory Distress Syndrome, RDS) ed i polimorfismi dei singoli nucleotidi dei geni SP-A e SP-B.

La pubblicazione sui lavori, relativi alle possibili complicanze legate alla prematurità che comprendono anche il contributo genetico all'RDS, ed eseguiti dal Children's Hospital dell'Università di Oulu in collaborazione col Biocenter di Oulu sotto la guida del Prof. Hallman, è stata inserita nella sezione *G* della presenti tesi di dottorato.

Inoltre sono stati allegati gli abstract di due ulteriori studi clinici svolti nel campo della Neuropsichiatria Infantile presso l'Università di Padova.

Sezione A

Problemi di gestione di banche dati

CAPITOLO 1 *Record Linkage*

1.1 Premessa

Negli ultimi decenni i processi computerizzati sono aumentati fortemente e i sistemi informatizzati avanzati sono diventati di uso comune nella gestione di informazioni archiviate portando alla proliferazione di differenti e numerosi dataset creati per diversi scopi e in ambienti in cui spesso si riscontra una forte incomunicabilità. Nel settore della sanità, in particolare, si trovano frequentemente settori in cui l'informazione è frammentata.

Molti studi di rilievo si centrano sull'individuo ed utilizzano dataset in cui l'unità statistica, che rappresenta il singolo record, possa in alcuni casi corrispondere ad un singolo evento nella vita del soggetto (nascita, decesso) e in altri ad un evento, un episodio o un momento dell'esistenza del soggetto che possa ripetersi, possa cambiare o possa occorrere nel tempo (un adempimento, un servizio, una malattia, una visita). Diversi studi combinano i record registrati in dataset separati, che si riferiscono ad uno stesso individuo, attraverso procedure di *Record Linkage*. In questo modo le informazioni riportate in differenti archivi e relative agli stessi soggetti vengono integrate attraverso la creazione di un singolo database. La possibile unione di informazioni concernenti variabili demografiche, socio-economiche e psicosociali con dati sulla salute e la storia clinica e diagnostica di un individuo rappresenta un prezioso aumento di informazioni fornito da studi che appunto usufruiscono di tecniche di *Record Linkage* (Kazanjian, 1998). Questa creazione di dataset concatenati sulla salute della popolazione ha una grande utilità per ricercatori nel campo sanitario e per chi deve prendere decisioni politiche (Chamberlayne et al., 1998). Per una pianificazione sanitaria adeguata, infatti, il riferimento è spesso fornito da studi che integrino differenti fonti di dati: questa procedura di unione di più dataset arricchisce il potenziale informativo in poco tempo e senza costi additivi. Il problema principale consiste nella disponibilità di variabili, nei diversi dataset, che rappresentino attendibili chiavi di unione o identificatori. Nei casi in cui record provenienti da fonti diverse condividono una stessa chiave identificativa, il *Record Linkage* è semplice da affrontare. Quando invece non è disponibile una stessa chiave identificativa la procedura di *Record Linkage* è più complicata. L'identificazione dei soggetti in un dataset è tuttora un problema irrisolto. Solo raramente un dataset include un'unica chiave identificativa (Bohmer et al., 2002), generalmente è disponibile più di una chiave identificativa, ma spesso nessuna di esse è di alta qualità perché a volte non è presente per ogni record, o perché si riscontrano errori di trascrizione dei valori relativi alla variabile identificativa. Molti studi si

prefiggono l'obiettivo di creare un'adeguata chiave identificativa che può essere rappresentata da un codice specifico o da una combinazione di dati personali contenuti nello stesso record (Grannis et al., 2002; Pates et al., 2001). Si trovano molteplici situazioni in cui si utilizzano specifici codici in base alle norme vigenti nel Paese d'interesse: si vedrà nell'applicazione considerata in questo capitolo (paragrafo 1.5) come la Finlandia, ed anche gli Stati Uniti (Weiner et al., 2003) e molti altri Paesi, utilizza il "*Social Security Number*".

Nella prima parte di questo capitolo si darà un orientamento generale dell'argomento di *Record Linkage*. Si affronteranno i metodi specifici di *Record Linkage* (paragrafo 1.2) soffermandosi sulle due principali possibili tecniche (probabilistica e deterministica) di cui si presenterà la teoria (paragrafi 1.3 e 1.4). La seconda parte del capitolo mira invece a fornire un'applicazione del *Record Linkage* su un lavoro svolto presso il Children's Hospital dell'Università di Oulu – Finlandia, attraverso l'utilizzo del programma Microsoft Access. Tale applicazione punterà ad effettuare due metodi di *Record Linkage* deterministico centrati su chiavi identificative differenti: nel primo caso sul *Social Security Number* (paragrafi 1.7 – 1.10), nel secondo caso su un set di variabili, quali nome, cognome, data di nascita e sesso (paragrafo 1.11). Infine si confronteranno i risultati dei due metodi adottati.

1.2 Metodi di *Record Linkage*

I metodi di *Record Linkage* possono essere riassunti in tre ampie categorie: manuali, deterministici e probabilistici. L'abbinamento manuale di record è il metodo più vecchio, quello che richiede naturalmente un consumo di tempo maggiore e anche quello più costoso in termini di risorse impiegate. Anche se rimane in molti casi il metodo standard, non è un'opzione fattibile quando si trattano database con grandi quantità di record (Liu & Wen, 1999).

Il *Record Linkage* deterministico abbina record da due dataset sulla base della completa concordanza di un'unica variabile identificativa (es. codice fiscale, codice sanitario o "Social security number") o attraverso la totale corrispondenza di un set di variabili comuni (es. nome, cognome, data di nascita, sesso, ecc.). Questo approccio deterministico minimizza le incertezze nell'unione di due database poiché solo un abbinamento di codici identici, o un abbinamento completo di un set di variabili personali, viene accettato, a discapito di una lieve riduzione del tasso di linkage (numero di record abbinati correttamente sul totale dei record del database in percentuale). È sufficiente, infatti, che sia stato commesso un minimo errore di trascrizione o d'imputazione del codice identificativo o di una delle variabili personali

utilizzata nel set per l'abbinamento, a portare ad un mancato abbinamento. In questo caso due record che in realtà rappresentano la stessa unità statistica non verranno agganciati. Per evitare un errore di questa specie è consigliabile non basarsi su un solo metodo di linkage, ma affiancare un procedimento centrato su un codice identificativo, o su una combinazione di variabili identificative, con un secondo sistema che prenda una diversa serie di variabili o un diverso identificativo come chiave per il *Record Linkage* deterministico.

Il *Record Linkage* probabilistico è utilizzato per identificare e agganciare i record di un dataset con i record corrispondenti in un altro dataset sulla base di una probabilità statistica, calcolata per un set di variabili identificative rilevanti (es. nome, sesso, data di nascita, ecc.). La probabilità è utilizzata per determinare se una coppia di record si riferisce approssimativamente alla stessa unità statistica (Li et al., 2006). Il linkage probabilistico teoricamente massimizza l'abbinamento ed è consigliabile in quanto il calcolo della probabilità può essere raffinato in vari punti per correggere pesi associati con valori identificativi e per aggiustare eventuali errori di codifica, in modo da massimizzare le informazioni disponibili nei dati (Newcombe et al., 1959; Newcombe,1988; Waien, 1997; Howe, 1998). Può tuttavia portare a distorsioni su alcuni link potenziali e inoltre richiede dettagliate conoscenze a priori su alcune importanti misure relative a specifici valori identificativi – per esempio la frequenza – in entrambi i file che devono essere agganciati. I ricercatori spesso non possiedono questo grado di conoscenza a priori (Van Den Brandt et al., 1990).

In conclusione, quando sono disponibili un numero sufficiente di variabili identificative, o ancora meglio un codice univoco per ciascun record, l'uso del linkage deterministico dovrebbe aumentare la frequenza di abbinamenti corretti con un minimo sacrificio nella diminuzione del tasso di linkage, problema in parte ovviabile, come si è visto, con l'interazione di due *Record Linkage* basati su identificativi diversi. D'altra parte quando non sono disponibili identificatori personali univoci il metodo di linkage dipende fortemente sull'univocità del set di variabili disponibili e spesso si predilige il linkage probabilistico.

1.3 La teoria del *Record Linkage*

Si supponga che dalla popolazione di dati A si prenda un campione casuale semplice (A_s) di n_a record e dalla popolazione B si prenda un secondo campione (B_s) con numerosità n_b (non si esclude tuttavia la possibilità che $A_s = A$ e $B_s = B$). Si denotino gli elementi dei campioni A_s e B_s rispettivamente con a e b. Ciascuno degli n_b record rappresenta un potenziale abbinamento

per ciascuno degli n_a record. In questo modo ci sono $n_a * n_b$ coppie di record per cui si deve determinare lo stato di concordanza o meno (Gomatam et al., 2002). Dal prodotto crociato $A_s * B_s$ si possono definire due dataset M e U, dove una coppia di record è parte del dataset M se quella coppia rappresenta un vero abbinamento, altrimenti è parte del dataset U. In altri termini:

$$M = \{(a,b); a = b, a \in A, b \in B\}$$

$$U = \{(a,b); a \neq b, a \in A, b \in B\}$$

Si assuma ora che ogni campione abbia la sua procedura di generazione di record: il risultato di un qualsiasi processo di generazione di record è la realizzazione di un record per ogni elemento del campione contenente delle caratteristiche selezionate (es. età in una certa data, indirizzo in una certa data, ecc.). La procedura di generazione dei record introduce anche degli errori e delle incompletezze nei record finali (es. errori di trascrizione o esiti negativi nel riporto di informazioni, errori di codifica, ecc.). Come risultato finale può succedere che due elementi (uno di A_s e uno di B_s) nella realtà diversi e non accoppiati possano dare luogo ad un identico record (sia per errori o per il fatto che un numero insufficiente di caratteristiche è stato incluso nel record) e, viceversa, due elementi accoppiati (identici) di A_s e B_s diano luogo a differenti record. Si denotino i record corrispondenti agli elementi di A_s e B_s con $\alpha(a)$ e $\beta(b)$ rispettivamente.

Infine si considerino i due file, L_A e L_B, come risultato dell'applicazione del processo di generazione dei record di A_s e B_s rispettivamente.

Il primo passo verso il raggiungimento del *link* tra i record dei due file (cioè identificare i record che corrispondono a elementi accoppiati di A_s e B_s) consiste nel confronto dei record. Formalmente si definisce il *vettore di confronto* come una funzione vettoriale dei record $\alpha(a)$ e $\beta(b)$:

$$\gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], ..., \gamma^K[\alpha(a), \beta(b)]\}$$

Si noti che γ è una funzione su $A * B$. La serie di tutte le possibili realizzazioni di γ è chiamata *spazio di confronto* ed è denotato con Γ (Newcombe et al., 1959).

Nel corso dell'operazione di *linkage* si osserva $\gamma(a,b)$ e si vuol decidere se (a,b) sia una coppia realmente appaiata, $(a,b) \in M$ (questa decisione prende il nome di *link positivo*, denotato con A_1), o se (a,b) sia una coppia non appaiata $(a,b) \in U$ (questa decisione prende il nome di *non link positivo*, denotato con A_3). Può accadere che nessuno dei due casi sopra elencati si verifichi, ma che ci si trovi nella situazione di una terza decisione, denotata con A_2 e chiamata *link possibile*.

Si può ora stilare una sorta di *regola di linkage L* sulla base di Γ, lo spazio di confronto, attraverso una serie di funzioni di decisione casuale $D = \{d(\gamma)\}$ dove

$$d(\gamma) = \{P(A_1|\gamma), P(A_2|\gamma), P(A_3|\gamma)\};\ \gamma \in \Gamma$$

e

$$\sum_{i=1}^{3} P(A_i|\gamma) = 1.$$

In altre parole, in corrispondenza di ogni valore osservato di γ la regola di linkage assegna le probabilità di ottenere ognuna delle tre possibili azioni (A_1, A_2 e A_3). Per alcuni o anche tutti i valori possibili di γ la funzione di decisione può assegnare una delle azioni con probabilità pari a 1.

Si devono però considerare anche i livelli d'errore associati alla regola di linkage. Si assuma che una coppia di record $[\alpha(a), \beta(b)]$ sia selezionata per il confronto in accordo con il processo probabilistico da L_A * L_B (è equivalente a selezionare una coppia di elementi (a,b) a caso da A_s * B_s, per conseguenza della costruzione dei file L_A e L_B). Il vettore di confronto risultante $\gamma[\alpha(a), \beta(b)]$ è una variabile casuale. Quando $(a,b) \in M$ si denoti la probabilità condizionata di γ con $m(\gamma)$:

$$m(\gamma) = P\{\gamma[\alpha(a), \beta(b)] | (a,b) \in M\} = \sum_{(a,b)\in M} P\{\gamma[\alpha(a), \beta(b)]\} \cdot P[(a,b)|M].$$

Similmente si denoti la probabilità condizionata di γ, nel caso in cui $(a,b) \in U$, con $u(\gamma)$:

$$u(\gamma) = P\{\gamma[\alpha(a), \beta(b)] | (a,b) \in U\} = \sum_{(a,b)\in U} P\{\gamma[\alpha(a), \beta(b)]\} \cdot P[(a,b)|U].$$

Ci sono due tipi di errore associati alla regola di linkage. Il primo occorre quando una coppia di elementi realmente diversi viene abbinata, ed ha la seguente probabilità (probabilità dei falsi accoppiati):

$$P(A_1|U) = \sum_{\gamma\in\Gamma} u(\gamma) \cdot P(A_1|\gamma).$$

Il secondo avviene quando una coppia di elementi realmente identici non viene abbinata, ed ha la seguente probabilità (probabilità dei falsi non-accoppiati):

$$P(A_3|M) = \sum_{\gamma\in\Gamma} m(\gamma) \cdot P(A_3|\gamma).$$

Una regola di linkage sullo spazio Γ si definirà regola di linkage ai livelli μ, λ ($0 < \mu < 1$ e $0 < \lambda < 1$) e verrà indicata con $L(\mu, \lambda, \Gamma)$ se $P(A_1|U) = \mu$ e $P(A_3|M) = \lambda$.

Tra la classe di regole di linkage su Γ che soddisfano i principi $P(A_1|U) = \mu$ e $P(A_3|M) = \lambda$, la legge di linkage $L(\mu, \lambda, \Gamma)$ verrà detta *regola ottimale di linkage* se la relazione

$$P(A_2|L) \leq P(A_2|L^I)$$

è valida per ogni $L^I(\mu, \lambda, \Gamma)$ nella classe.

Per spiegare quest'ultima definizione, si sottolinea che la regola ottimale di linkage massimizza le probabilità di abbinamento positive (vale a dire le decisioni A_1 e A_3) soggette ai livelli fissati di errore in $P(A_1|U) = \mu$ e $P(A_3|M) = \lambda$.

Non è difficile notare che per certe combinazioni di μ e λ la classe di regole di linkage su Γ che soddisfino i principi $P(A_1|U) = \mu$ e $P(A_3|M) = \lambda$, è vuota. Sono ammesse solo quelle combinazioni di μ e λ per le quali vengono soddisfatti i principi $P(A_1|U) = \mu$ e $P(A_3|M) = \lambda$ e contemporaneamente anche le funzioni di decisione $d(\gamma) = \{P(A_1|\gamma), P(A_2|\gamma), P(A_3|\gamma)\}$; $\gamma \in \Gamma$ e $\sum_{i=1}^{3} P(A_i|\gamma) = 1$. Una coppia di valori (μ, λ) sarà dunque non accettabile solo se una o entrambe le componenti sono troppo grandi. In questo caso si dovranno ridurre i livelli di errore (Fellegi & Sunter, 1969).

1.4 Teorema fondamentale

Si definisca innanzitutto una regola di linkage L_0 su Γ. Poi si definisca un unico set ordinato di possibili realizzazioni di γ. Se nessun valore di γ è tale che entrambe $m(\gamma)$ e $u(\gamma)$ siano uguali a zero, allora la probabilità (incondizionata) che si verifichi quello specifico valore di γ è pari a zero. Ora si assegni un ordine arbitrario a tutte le γ per cui $m(\gamma) > 0$, ma $u(\gamma) = 0$. In seguito vengano ordinate tutte le rimanenti γ in modo che la corrispondente sequenza $\frac{m(\gamma)}{u(\gamma)}$ risulti monotonicamente decrescente (quando il valore di $\frac{m(\gamma)}{u(\gamma)}$ è lo stesso per più di un γ si ordinino questi γ arbitrariamente). Successivamente si indicizzi il set ordinato $\{\gamma\}$ attraverso l'indice i (i = 1, 2, …, N_Γ) e si scriva $u_i = u(\gamma_i); m_i = m(\gamma_i)$.

Sia (μ, λ) una coppia ammissibile di livelli d'errore e si scelgano n e n' tali che

$$\sum_{i=1}^{n-1} u_i < \mu \leq \sum_{i=1}^{n} u_i$$

$$\sum_{i=n'}^{N_\Gamma} m_i < \lambda \leq \sum_{i=n'+1}^{N_\Gamma} m_i$$

dove N_Γ è il numero di punti in Γ e $1 < n < n'-1 < N_\Gamma$ in modo che i livelli (μ, λ) siano ammissibili. Avendo osservato un vettore di confronto γ_i, sia $L_0(\mu, \lambda, \Gamma)$ la regola di linkage definita come segue: si assume l'azione A_1 (link positivi) se $i \le n-1$, azione A_2 quando $n < i \le n'-1$ e azione A_3 (non-link positivo) quando $i \ge n'+1$. Quando $i = n$ o $i = n'$ allora è richiesta una decisione casuale per ottenere esattamente i livelli di errore μ e λ. Formalmente:

$$d(\gamma_i) = \begin{cases} (1,0,0) & i \le n-1 \\ (P_\mu, 1-P_\mu, 0) & i = n \\ (0,1,0) & n < i \le n'-1 \\ (0, 1-P_\lambda, P_\lambda) & i = n' \\ (0,0,1) & i \ge n'+1 \end{cases}$$

dove P_μ e P_λ sono definite come le soluzioni delle equazioni

$$u_n \cdot P_\mu = \mu - \sum_{i=1}^{n=1} u_i$$

$$m_{n'} \cdot P_\lambda = \lambda - \sum_{i=n'+1}^{N_\Gamma} m_i \ .$$

Il teorema afferma che se $L_0(\mu, \lambda, \Gamma)$ è la regola di linkage definita sopra, allora L è la miglior regola di linkage su Γ ai livelli (μ, λ).

Da questo teorema derivano due corollari molto importanti nella pratica.

Corollario 1:

se

$$\mu = \sum_{i=1}^{n} u_i, \ \ \lambda = \sum_{i=n}^{N_\Gamma} m_i, \ \ n < n',$$

la $L_0(\mu, \lambda, \Gamma)$, la miglior regola di linkage su Γ ai livelli (μ, λ), diventa

$$d(\gamma_i) = \begin{cases} (1,0,0) & se \quad 1 \le i \le n \\ (0,1,0) & se \quad n < i \le n' \\ (0,0,1) & se \quad n' \le i \le N_\Gamma \end{cases} .$$

Se si definiscono

$$T_\mu = \frac{m(\gamma_n)}{u(\gamma_n)}$$

$$T_\lambda = \frac{m(\gamma_{n'})}{u(\gamma_{n'})}$$

allora la regola di linkage $L_0(\mu, \lambda, \Gamma)$ espressa sopra, può essere scritta equivalente a

$$d(\gamma_i) = \begin{cases} (1,0,0) & se \quad T_\mu \le \dfrac{m(\gamma)}{u(\gamma)} \\ (0,1,0) & se \quad T_\lambda < \dfrac{m(\gamma)}{u(\gamma)} < T_\mu \\ (0,0,1) & se \quad \dfrac{m(\gamma)}{u(\gamma)} \le T_\lambda \end{cases}$$

Corollario 2:

siano T_μ e T_λ due numeri positivi qualsiasi tali che $T_\mu > T_\lambda$.

Esiste quindi una coppia ammissibile di livelli di errore (μ, λ) corrispondente a T_μ e T_λ tale che la regola di linkage

$$d(\gamma_i) = \begin{cases} (1,0,0) & se \quad T_\mu \le \dfrac{m(\gamma)}{u(\gamma)} \\ (0,1,0) & se \quad T_\lambda < \dfrac{m(\gamma)}{u(\gamma)} < T_\mu \\ (0,0,1) & se \quad \dfrac{m(\gamma)}{u(\gamma)} \le T_\lambda \end{cases}$$

sia la migliore con tali livelli d'errore. Questa coppia di livelli d'errore (μ, λ) è data da:

$$\mu = \sum_{\gamma \in \Gamma_\mu} u(\gamma)$$

$$\lambda = \sum_{\gamma \in \Gamma_\lambda} m(\gamma)$$

dove

$$\Gamma_\mu = \left\{ \gamma : T_\mu \le {m(\gamma)}/{u(\gamma)} \right\}$$

$$\Gamma_\lambda = \left\{ \gamma : {m(\gamma)}/{u(\gamma)} \le T_\lambda \right\}.$$

In molte applicazioni si possono tollerare livelli d'errore sufficientemente alti per precludere l'azione A_2. In questo caso si scelgono *n* e *n'* o, alternativamente, T_μ e T_λ tali che ogni (*a, b*) venga allocata o in *M* o in *U* (Fellegi & Sunter, 1969).

Jaro (1989, 1995) propone l'utilizzo dell'algoritmo EM (Dempster, 1977) per stimare $\{m(\gamma), \gamma \in \Gamma\}$. Nella sua formulazione del problema, il vettore di dati completo è dato da (γ, g), dove γ è definita come sopra, mentre g indica lo stato corrente (abbinata o non-abbinata) della coppia di record. Jaro considera le componenti di γ, nominalmente γ^j, come ristrette in valori compresi tra 0 e 1, e assume l'indipendenza condizionata (dato lo status di abbinato o non- abbinato della coppia di record) sui valori di j. g assume un valore corrispondente a un abbinamento con probabilità p e uno corrispondente a un non-abbinamento con probabilità 1-p. La verosimiglianza è scritta in termini di g, $m(\gamma)$ e $u(\gamma)$. Siccome g non può essere osservata, l'algoritmo EM è usato per ottenere le stime di massima verosimiglianza di $m(\gamma)$, $u(\gamma)$ e p sotto l'assunzione di indipendenza condizionale delle componenti del vettore di concordanza (Larsen & Rubin, 2001; Winkler, 1988; Armstrong & Mayda, 1992).

1.5 Applicazione del *Record Linkage* allo studio svolto al Children's Hospital dell'Università di Oulu – Finlandia

Il primo passaggio affrontato nel lavoro in campo genetico svolto al Children's Hospital dell'Università di Oulu – Finlandia è consistito nella creazione di un unico database da cui poi prelevare i dati per l'analisi sull'associazione tra Sindrome di distress respiratorio e covariate genetiche e cliniche. I dati erano già stati precedentemente raccolti nel corso del decennio 1996-2006 in appositi archivi di ciascuno degli ospedali partecipanti alla ricerca sulle patologie polmonari neonatali, vale a dire gli ospedali di Oulu, Tampere e Seinäjoki. Per ognuno dei tre nosocomi si dispone di due database: il primo contiene variabili individuali e cliniche raccolte al momento della nascita per i bambini i cui genitori hanno acconsentito di partecipare allo studio; il secondo contiene variabili genetiche provenienti da esami sul DNA e successive analisi di laboratorio, realizzate fino a un anno dalla nascita. L'obiettivo è stato quello di creare una base informativa statistica, costituita da diverse fonti d'informazione (dati genetici e clinici da tre ospedali della Finlandia centro-settentrionale) strutturate e rese disponibili, ai Centri d'interesse, in funzione del loro utilizzo per lo studio statistico di particolari fenomeni. L'integrazione dei sei dataset disponibili in file Microsoft Access è stata effettuata tramite *Record Linkage*. In un primo momento è stata utilizzata come chiave identificativa un particolare codice fornito ad ogni nato in Finlandia (*Social Security Number*) o un codice creato *ad hoc* nel caso non si disponesse del *Social Security Number* (la logica, simile a quella del codice identificativo, utilizzata per la realizzazione dei codici *ad hoc* verrà descritta nel paragrafo 1.6). In seguito si è eseguito un secondo *Record Linkage* deterministico utilizzando come identificativo un set di quattro variabili, quali nome, cognome, data di nascita e sesso per recuperare eventuali abbinamenti non avvenuti a causa di errori nella stringa del "*Social Security Number*" o del codice creato *ad hoc*.

Il *Social Security Number* adottato in Finlandia consiste in un codice identificativo di undici caratteri che viene fornito ad ogni nato e ai cittadini stranieri che ottengono un permesso di residenza. La composizione di questa stringa è la seguente: i primi 6 caratteri rappresentano la data di nascita, due cifre per il giorno, due per il mese e due per l'anno (solamente le ultime due cifre dell'anno vengono inserite; per es. 2 aprile 1998 sarà scritto come '020498'). Il settimo carattere è un trattino, '-', solamente per i nati fino all'anno 1999, mentre per i nati in un anno dopo il 1999 viene rappresentato da una 'A'. Gli ultimi quattro caratteri vengono assegnati all'individuo secondo un sistema computerizzato centrale dall'Ufficio Anagrafico di Stato e si suddividono in tre numeri più un ultimo simbolo che può essere una lettera come un ennesimo numero.

Le unità statistiche considerate in questo studio sono rappresentate da neonati con età gestazionale inferiore alle 37 settimane complete e sono nati vivi in uno dei tre ospedali della Finlandia centro-settentrionale (Oulu, Tampere e Seinäjoki) oppure ivi trasferiti subito dopo il parto.

I tre dataset relativi alle variabili individuali e cliniche provenienti dai tre ospedali sono stati uniti verticalmente, dato che le variabili raccolte erano le stesse per ogni ospedale e le colonne dei dataset coincidevano. Lo stesso procedimento è stato seguito anche per i tre dataset di dati genetici, dando così origine a due file totali, uno contenente tutti i dati clinici e individuali per i pazienti di tutti i nosocomi, l'altro relativo alle informazioni genetiche. Lo scopo è stato quello di ottenere un database unico per i pazienti dei tre ospedali entrati nello studio. In questo database le variabili individuali e cliniche dovevano essere associate a quelle genetiche, di conseguenza il secondo passaggio è consistito nell'unione orizzontale del file di dati clinici/individuali con quello di dati genetici attraverso il *Record Linkage* deterministico centrato sulla chiave identificativa "*Social Security Number*" o sul codice creato *ad hoc*.

Il problema principale del *Record Linkage* nello studio del Children's Hospital dell'Università di Oulu è consistito nel fatto che non per tutti i pazienti si disponeva sia di informazioni cliniche individuali, sia di variabili genetiche, vale a dire che vi sono stati alcuni pazienti per cui non si è riscontrato alcun abbinamento. I casi possibili sono stati tre:

1. bambini presenti in entrambi i dataset dell'ospedale di nascita, i cui genitori quindi hanno aderito allo studio subito dopo il parto e hanno continuato a far partecipare il figlio;
2. bambini per cui si dispone solo dei dati individuali e clinici: si tratta di nati pretermine che sono usciti dallo studio e che quindi non hanno eseguito gli esami sul DNA o altre analisi di laboratorio. Le motivazioni dell'uscita dallo studio possono essere: decesso del paziente, abbandono voluto dai genitori, trasferimento della famiglia in altra città o irreperibilità della famiglia;
3. bambini per cui si dispone delle sole variabili genetiche: si tratta di una situazione dovuta all'entrata nello studio da parte del paziente postuma alla raccolta dei dati clinici e individuali, oppure si tratta di neonati per cui è stato necessario un immediato trasferimento dopo il parto in un altro ospedale.

Naturalmente sia il database di dati clinici e individuali, sia quello di dati genetici contengono le colonne riferite alle variabili identificative del paziente, vale a dire il "*Social Security*

Number" nel primo caso di *Record Linkage*, il nome, il cognome, la data di nascita e il sesso per la seconda procedura. In entrambi i database è presente un'altra colonna relativa alla variabile "internal code", un codice identificativo del neonato, interno a ciascun ospedale, che però è stata considerata non attendibile per la procedura di *Record Linkage* data la logica incerta della sua struttura. Questo codice è stato però utilizzato, come si vedrà nell'applicazione del prossimo capitolo, per il recupero di alcuni dati mancanti.

Dalla figura 1.1 si osserva che il database finale contiene record provenienti dai tre diversi ospedali ed ogni riga rappresenta un paziente identificato dal "*Social Security Number*" o dal codice creato *ad hoc*: per la maggior parte dei record si possiedono sia le informazioni individuali/cliniche che quelle genetiche e si tratta dei record per cui è stato effettuato il link con successo, per alcuni casi si hanno solo le informazioni individuali/cliniche e si avranno quindi delle celle vuote in corrispondenza delle variabili genetiche, infine per altri casi le celle vuote saranno in corrispondenza delle colonne relative alle variabili individuali/cliniche.

Figura 1.1 Schema esemplificativo del database finale che si vuole creare. La "X" rappresenta la presenza del dato, la cella vuota indica invece che il dato non è disponibile per quel record.

Ospedale	Identificativo Paziente	Dato clinico	Dato clinico	Dato clinico	Dato genetico	Dato genetico	Dato genetico	Dato genetico
Oulu	150599-….	X	X	X	X	X	X	X
Oulu	200498-….	X	X	X				
Oulu	030701A….				X	X	X	X
Tampere	281101A….	X	X	X	X	X	X	X
Tampere	050997-….	X	X	X				
Seinäjoki	301196-….	X	X	X	X	X	X	X
Seinäjoki	081204A….	X	X	X				

Nello studio portato avanti dal Dipartimento di Pediatria dell'Università di Oulu – Finlandia è stato scelto un metodo di *Record linkage* deterministico sia perché non si possedevano informazioni a priori sui valori identificativi, sia perché la disponibilità del "*Social Security Number*" ha permesso un facile link laddove questa variabile era disponibile e corretta, mentre nei casi per cui non si possedeva si è creato un codice *ad hoc*. Si può quindi affermare che il linkage deterministico è stato affiancato da un linkage manuale nei casi in cui il valore della variabile identificativa era errata o mancante. Nello specifico, record sui nati pretermine, contenenti informazioni demografiche e cliniche, sono stati agganciati con dati genetici relativi ad analisi di laboratorio avvenute dopo la nascita, attraverso appunto la chiave

identificativa "*Social Security Number*" ove disponibile. In seguito si è eseguito un secondo *Record Linkage* deterministico utilizzando come identificativo un set di quattro variabili, quali nome, cognome, data di nascita e sesso per recuperare eventuali abbinamenti non avvenuti a causa di errori nella stringa del "*Social Security Number*" o del codice creato *ad hoc*.

Si sono considerati per il presente lavoro metodi di abbinamento uno a uno (in inglese *matching one-to-one* oppure *exact matching*), vale a dire che si sono presi in considerazione codici degli stessi individui per l'unione dei file esaminati. Questo è l'opposto dell'abbinamento statistico (*statistical matching*), dove l'interesse è agganciare informazioni su individui simili (quest'ultimo è una tipologia di *Record Linkage* molto usata, per esempio, nelle compagnie di marketing).

1.6 Pulizia dei dataset prima del *Record Linkage*

Prima di effettuare il *Record Linkage* per i dataset degli ospedali di Oulu, Tampere e Seinäjoki, è stata eseguita una verifica sull'attendibilità della variabile utilizzata per l'abbinamento dei pazienti e cioè il "*Social Security Number*". Il primo passo è stato controllare se c'erano codici doppi e si è riscontrato che sette pazienti erano stati inseriti due volte nello stesso database: cinque inseriti nel file dell'ospedale di Oulu, uno nel file dell'ospedale di Tampere e l'ultimo in quello di Seinäjoki. Questo controllo è stato necessario perché la variabile identificativa deve essere univoca.

Affinché la chiave identificativa non risultasse mancante per alcun record, successivamente si sono contati i casi per cui il "*Social Security Number*" non era presente e per cui si è iniziato un efficiente lavoro di recupero con l'aiuto del personale ostetrico. Quest'ultimo ha recuperato buona parte dei "*Social Security Number*" attraverso il database sulle nascite del corrispondente nosocomio. Con l'utilizzo di informazioni individuali, quali nome, cognome, data di nascita e in alcuni casi anche il cognome materno, è stato individuato il nato d'interesse e, se disponibile, è stato estrapolato il "*Social Security Number*". Per ciascuno dei sei file analizzati si riportano di seguito i numeri di pazienti per cui non è stato possibile recuperare l'identificativo.

- Oulu: 71 pazienti senza ID su 519 (13,7%) provenivano dal file di dati individuali/clinici, 33 su 592 (5,6%) dal file di dati genetici;
- Tampere: 178 su 246 (72,3%) non possedevano l'ID per il file di dati individuali/clinici e 116 su 159 (72,9%) per il file di dati genetici.

- Seinäjoki: 27 su 155 pazienti (17,5%) non riportavano il "*Social Security Number*" nel file di dati individuali/clinici, mentre per il file dei dati genetici 21 pazienti su 129 (16,4%) erano privi di chiave identificativa.

L'alta percentuale di casi senza identificativo riscontrati per l'ospedale di Tampere è dovuta al fatto che nel periodo dell'analisi dei dati non era presente alcun contatto di tale nosocomio che collaborasse al recupero dei "*Social Security Number*" mancanti, quindi si è dovuti ricorrere ad un ampio utilizzo dei codici creati *ad hoc*.

Altra spiegazione che è opportuno fornire concerne l'alto numero di record genetici dell'ospedale di Oulu, rispetto ai record provenienti dal file di dati individuali/clinici. Diversi esami sul DNA sono stati realizzati nell'ospedale di Oulu, anche per nati a Tampere e Seinäjoki, quindi il record genetico corrispondente a quello individuale/clinico di questi nati è stato trovato nel file di Oulu: il tentativo di *Record Linkage* deterministico è stato tentato su tutte le potenziali coppie di file, in modo da verificare tutte le situazioni possibili di spostamento dei pazienti.

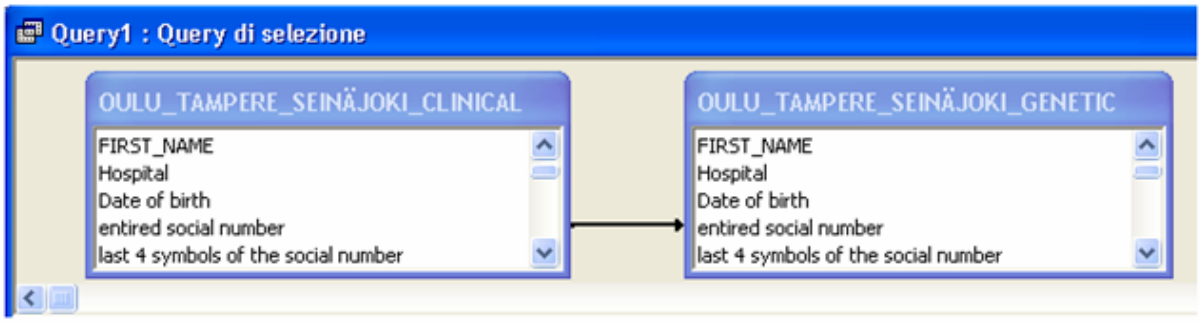
Per quei pazienti per cui non è proprio stato rinvenuto il "*Social Security Number*", si è dovuto creare un identificativo (ID) *ad hoc*. Per realizzare tali codici si è cercato di seguire una logica simile a quella dei "*Social Security Number*". Siccome la data di nascita è un'informazione presente per ogni record si è partiti da queste sei cifre, aggiungendovi altri cinque caratteri in modo da ottenere una stringa della stessa lunghezza del "*Social Security Number*". Il primo di questi ultimi cinque elementi è, come nel "*Social Security Number*", un trattino, '-', per i nati fino all'anno 1999, mentre per i nati in un anno dopo il 1999 è stata inserita la lettera maiuscola 'A'. La caratteristica del codice *ad hoc* rispetto al "*Social Security Number*" consiste nelle ultime quattro cifre della stringa: l'ottava e la nona sono rappresentate dalle prime due lettere del cognome (sostituendo le eventuali vocali Å, Ä e Ö con semplici vocali prive di dieresi o pallino). Nel caso di gemelli, o di nati con lo stesso cognome (o con cognomi inizianti con le stesse due lettere) nel medesimo giorno, si è lasciato che un codice seguisse la logica appena descritta, mentre per l'altro (o gli altri) sono state selezionate la prima lettera del cognome e la prima del nome; se anch'esse erano identiche alle lettere del primo codice, dopo l'iniziale del cognome è stata scelta la seconda lettera del nome, la terza, la quarta e così via fino alla selezione di un codice univoco.

Il decimo e l'undicesimo carattere indicano entrambi il sesso del nato: il decimo è una "T" per *tyttö*, che in finlandese significa femmina, o una "P" per *poika* (maschio in finlandese), mentre l'undicesimo carattere è rappresentato da un "2" per paziente femmina e da un "1" per paziente maschio.

1.7 Metodo di *Record Linkage* deterministico attraverso il *Social Security Number* con Microsoft Access

Possedendo i database in file Microsoft Access si è cercato un modo sicuro per la loro unione utilizzando questo programma di progettazione e gestione di uno o più database. Più specificatamente, il *Record Linkage* è stato reso operante attraverso l'utilizzo appropriato delle *Query*. Tramite una *Query* è possibile porre una domanda sui dati archiviati nelle tabelle, determinando con esattezza quali dati reperire attraverso la combinazione di dati provenienti da più tabelle (Microsoft Access, 1992).

Figura 1.2 Record Linkage nel programma Microsoft Access. La freccia indica la direzione verso cui va il link tra la chiave identificativa della prima tabella e quella della seconda.



Si sono agganciati il database totale dei dati individuali e clinici con quello totale dei dati genetici, dove per totale si intende l'unione dei tre ospedali. Nel processo di linkage si è utilizzato il database dei dati individuali/clinici come principale e ad esso si sono agganciati i record genetici con due passaggi intermedi:

1. il primo passaggio è consistito nel creare un database che comprendesse tutti i record individuali e clinici con o senza una corrispondenza nel file di dati genetici, ma laddove tale corrispondenza era presente si è richiesto di effettuare il link attraverso la variabile "*Social Security Number*", che in figura 1.2 è rappresentata da "entired social number". La freccia che evidenzia la direzione verso cui va il link tra i due database, indica che il dataset dei dati individuali e clinici è considerato il principale;
2. il secondo passaggio riguarda il recupero di tutti quei record presenti solamente nel dataset di dati genetici, che verranno aggiunti (come singole righe) al

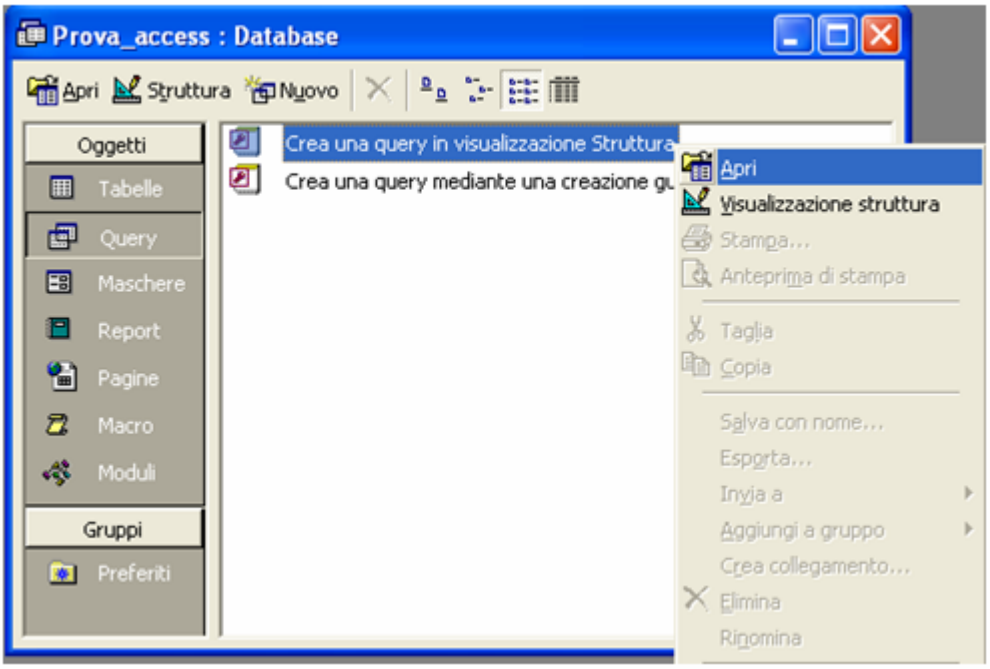
database totale e che avranno delle celle vuote in corrispondenza delle colonne relative alle variabili individuali e cliniche.

Si vedrà ora nel dettaglio come è stata eseguita questa operazione di link.

1.8 Creare una *Query*

Il punto centrale per effettuare un *Record Linkage* con Microsoft Access è sapere come creare una *Query*. Una volta aperto il programma ed aver importato i dataset d'interesse (se non fossero già stati progettati in Access), si deve entrare nella schermata delle *Query* e scegliere l'opzione "Crea una query in visualizzazione Struttura" come in figura 1.3.

Figura 1.3 Creazione di una Query in visualizzazione struttura.



Microsoft Access automaticamente aprirà una schermata in cui è possibile selezionare le tabelle dei dataset che si vogliono agganciare (figura 1.4).

Figura 1.4 Scelta delle tabelle da utilizzare per la *Query* nel Programma Microsoft Access.

Si selezioni prima la tabella relativa al dataset sui dati individuali e clinici dei nati pretermine (nello specifico è stata denominata OULU_TAMPERE_SEINÄJOKI_CLINICAL) scegliendo l'opzione "Aggiungi" e successivamente si segua lo stesso procedimento per la tabella dei dati genetici (nel presente caso OULU_TAMPERE_SEINÄJOKI_GENETIC).
Una volta scelte le due tabelle è possibile chiudere la maschera "Mostra Tabella" e nella parte superiore della schermata delle *Query* si osserva che le due tabelle sono già state agganciate automaticamente dal programma attraverso una linea che unisce due precise variabili. Dal momento che si era specificata in entrambi i dataset la chiave primaria, rappresentata dal "*Social Security Number*", Microsoft Access li unisce attraverso questa variabile.

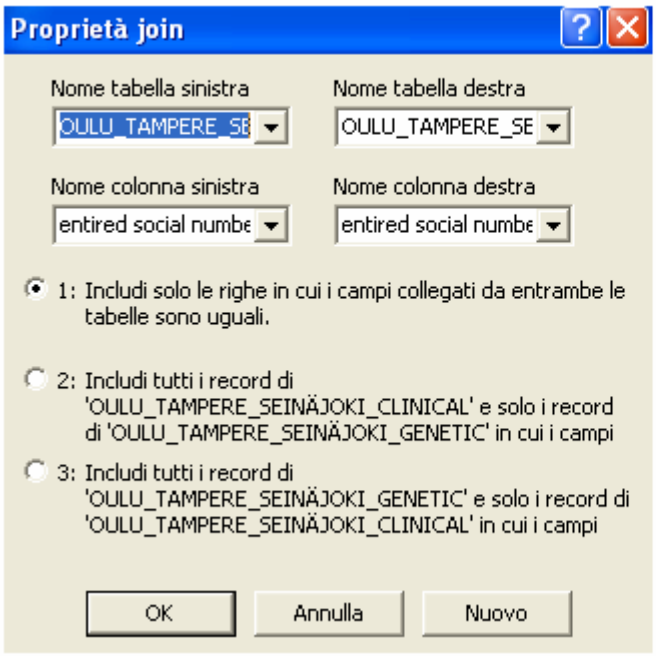
1.9 Proprietà della relazione tra i due dataset e loro unione finale

Uno dei passaggi più importanti per l'unione dei due dataset è la specificazione delle proprietà della relazione (in inglese "*join*"): cliccando col tasto destro del mouse sulla linea che unisce le due tabelle, apparirà una piccola finestra su cui si potrà selezionare l'opzione "Proprietà join" (figura 1.5).

Figura 1.5 Proprietà della relazione (*join*) tra le due chiavi identificative delle tabelle.

Si aprirà la seguente schermata in cui si possono scegliere tre tipologie di "join" tra i due dataset:

Figura 1.6 Dettagli delle proprietà della relazione (*join*) tra le due chiavi identificative delle tabelle.



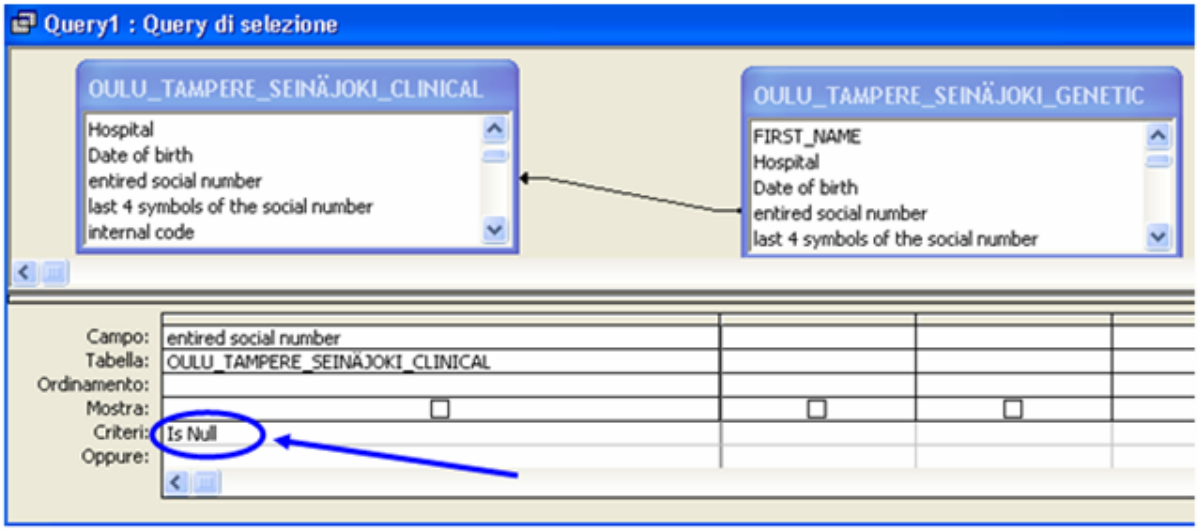
L'obiettivo è quello di non perdere alcun record dai due dataset e ottenere quindi un database finale in cui ogni riga rappresenti un paziente. Il record potrà contenere sia le informazioni individuali/cliniche che quelle genetiche (record accoppiati); potrà contenere solo le informazioni individuali/cliniche, oppure potrà contenere solo quelle genetiche. Per arrivare a questa situazione servono i due passaggi elencati in precedenza e che ora si presenteranno analiticamente.

1. Per includere tutti i record del dataset OULU_TAMPERE_SEINÄJOKI_CLINICAL e quelli del dataset OULU_TAMPERE_SEINÄJOKI_GENETIC solo nel caso in cui la chiave identificativa è identica, si sceglie l'opzione numero 2 nella maschera relativa alle proprietà del legame ("Proprietà join") in modo che la direzione del join vada dal

dataset di dati individuali e clinici verso quello di dati genetici, come mostrato in figura 1.6. Si ottiene in questo modo un primo database con tutti i dati individuali e clinici dei pazienti, a cui sono stati agganciati gli eventuali dati genetici che trovavano un abbinamento.

2. Il secondo punto riguarda la selezione di tutti quei record provenienti dal dataset di dati genetici per i quali non è presente un corrispettivo nel dataset di dati individuali e clinici per aggiungerli al database creato in precedenza. È necessario che la direzione dell'unione vada dal dataset di dati genetici verso quello di dati clinici, selezionando l'opzione 3 all'interno delle "join properties", ma con un accorgimento aggiuntivo. Si devono escludere, infatti, tutti i record del dataset OULU_TAMPERE_SEINÄJOKI_GENETIC per cui la chiave identificativa ("Social security number") è uguale a quella del dataset OULU_TAMPERE_SEINÄJOKI_CLINIC, perché tali soggetti sono già stati inclusi con il procedimento della *Query* precedente. Per raggiungere tale obiettivo è necessario inserire la variabile "Social security number" dalla tabella dei dati clinici e specificare nella riga dei "Criteri" la stringa "Is Null": questo significa che si stanno ricercando solo quei record dalla tabella di dati genetici per cui non esiste una corrispondenza (si impone che non ci sia lo stesso "Social security number") nella tabella di dati clinici (figura 1.7).

Figura 1.7 Passaggio relativo alla selezione dei record genetici per i quali non è presente un corrispettivo clinico individuale. Inserimento della stringa "*Is Null*" nei criteri del campo "entired social number" relativo alla tabella di dati clinici in modo da escludere i record con corrispondenza già inclusi in precedenza.



L'ultimo punto concerne l'unione verticale dei due database creati dai procedimenti di *Query* semplicemente copiando ed incollando i record della seconda operazione di *Query* sotto al primo database (questo è possibile perché le variabili sono le stesse in entrambi i dataset).

1.10 Risultati del *Record Linkage* deterministico attraverso il *Social Security Number*

Alla fine dei passaggi sopra elencati si è ottenuto un database conclusivo comprendente 1227 record, dei quali 573 hanno le informazioni sia individuali/cliniche sia genetiche, per i restanti non si è trovata alcuna corrispondenza tra i dataset. Come si nota dalla tabella 1.1 tra i pazienti appaiati, 314 provengono dall'ospedale di Oulu, 137 dall'ospedale di Tampere e 122 da quello di Seinäjoki. Le 573 coppie di record con corrispondenza rappresentano il 63,7% di tutti i record, clinici e genetici, che formavano i sei dataset iniziali per i tre ospedali; infatti trattandosi di coppie, si tratta di 1146 record tra clinici/individuali e genetici.

Se si vuole studiare una qualsivoglia relazione tra variabili cliniche/individuali e variabili genetiche si dovranno utilizzare solamente i 573 record che possiedono tutte le informazioni; se invece uno studio successivo avrà il fine di indagare relazioni tra variabili della stessa macro-area (clinica o genetica), si potranno estrapolare sia i record appaiati, sia quelli specifici della macro-area d'interesse.

Tabella 1.1 Tabella riassuntiva per ospedale delle coppie di dati e delle non corrispondenze trovate usando il "*Social Security number*" come chiave identificativa per il *Record Linkage*

	OULU	TAMPERE	SEINÄJOKI	Totale
Coppie di record con corrispondenza	314	137	122	573
Record individuali/clinici senza corrispondenza	205	109	34	348
Record genetici senza corrispondenza	278	22	6	306
Totale	797	268	162	1227

L'alta frequenza di record genetici senza una corrispondenza si spiega con il diverso momento d'entrata nello studio dei pazienti. Molti bambini, seppur nati prematuramente, sono stati arruolati nel campione di studio in un momento posteriore di qualche mese alla nascita, direttamente per le analisi sul DNA, quindi non si possiedono i dati clinici e demografici.

Invece, per quanto concerne le non corrispondenze nei dati clinici e individuali, si è scritto in precedenza che sono dovute ad abbandoni dello studio da parte di alcuni pazienti.

1.11 Metodo di *Record Linkage* deterministico attraverso un set di variabili con Microsoft Access

Per avere un riscontro sull'esito del *Record Linkage* effettuato attraverso la chiave identificativa del "Social security number" o del codice *ad hoc*, è susseguita una seconda procedura di *Record Linkage* deterministico attraverso quattro variabili: nome, cognome, data di nascita e sesso. Bisogna subito sottolineare che un *Record Linkage* di questo tipo è molto più inaffidabile rispetto a quello sopra analizzato, data la più alta probabilità di incontrare errori nella trascrizione di almeno una delle variabili prese in esame. Tuttavia si è intrapresa questa seconda procedura per tentare il recupero di eventuali accoppiamenti di record non avvenuti, utilizzando il "*Social Security Number*" come identificativo, per via di errori nel codice che hanno poi portato ad un'apparente non concordanza.

Per eseguire il *Record Linkage*, gli identificativi comuni sono stati formattati allo stesso modo nei due dataset totali (quello sui dati clinici e quello sui dati genetici dei tre ospedali uniti): le lettere sono state rese tutte maiuscole e sono stati rimossi eventuali spazi o trattini tra nomi e/o cognomi formati da più parole. Un problema frequente nell'utilizzo del nome come una delle variabili identificative per il *Record Linkage* concerne la possibilità di trovare il nome di un individuo rappresentato in differenti modi, con alterne compitazioni, con diverse iniziali, con abbreviazioni o soprannomi. Per affrontare questi problemi si è utilizzato un metodo di codifica (Soundex) in modo da identificare abbinamenti tra nomi che falliscono la concordanza per via di spelling varianti dei nomi nei due database (Knuth, 1973). L'algoritmo di Soundex associa dei numeri a diversi gruppi di consonanti, producendo un codice numerico basato sulla lettera iniziale considerata la più robusta rispetto alle variazioni nei nomi che suonano simili, e sulle consonanti successive.

Il *Record Linkage* tra il file di dati clinici/individuali e quello di dati genetici è stato portato a termine seguendo le stesse modalità analizzate in precedenza per il link sulla base del "*Social Security Number*". La differenza è consistita nella scelta della chiave identificativa: se prima la chiave era rappresentata solamente da una variabile, in questo caso si è considerato il set di variabili nome, cognome, data di nascita e sesso.

Tabella 1.2 Tabella riassuntiva per ospedale delle coppie di dati e delle non corrispondenze trovate usando il set di variabili nome, cognome, data di nascita e sesso come chiave identificativa per il Record Linkage

	OULU	TAMPERE	SEINÄJOKI	Totale
Coppie di record con corrispondenza	275	106	94	475
Record individuali/clinici senza corrispondenza	248	124	68	440
Record genetici senza corrispondenza	313	69	28	410
Totale	836	299	190	1325

Nella tabella 1.2 si osservano i valori degli accoppiamenti e dei record per cui non si è trovato un corrispettivo nell'altro dataset. È evidente come la procedura di *Record Linkage* attraverso la chiave identificativa rappresentata dal set di variabili nome, cognome, data di nascita e sesso, risulti meno efficace, portando ad un numero nettamente inferiore di abbinamenti: 475 rispetto ai 573 ottenuti con il "*Social Security Number*" (in termini di percentuali si riscontra un 52,8% rispetto al 63,7% precedentemente ottenuto). Si era però anticipato che l'obiettivo di questo secondo procedimento è stato quello di recuperare eventuali accoppiamenti di record non avvenuti nel primo *Record Linkage* deterministico. Andando perciò a confrontare i record con corrispondenza ottenuti dai due metodi di *Record Linkage* si è riscontrato che 7 dei 475 accoppiamenti riportati dal secondo procedimento non erano stati captati dal primo. Dopo un'operazione di controllo si è constatato che solo 5 di queste 7 coppie erano reali, le rimanenti due riguardavano un errore nella corrispondenza del nome per una coppia di gemelli: tale problema aveva dato luogo ad una duplicazione dei reali appaiamenti. Per quanto concerne invece le 5 nuove coppie recuperate, si è visto che nella procedura di *Record Linkage* attraverso il "*Social Security Number*" i record non erano stati abbinati a causa di errori in uno dei due codici identificativi, nel dataset dei dati clinici/individuali o in quello di dati genetici. L'ultimo passaggio è stato verificare (solamente se il paziente risultava nel dataset di Oulu o Seinäjoki) quale dei due codici era corretto attraverso il database sulle nascite del corrispondente nosocomio, correggere il codice e rilanciare la *Query* del *Record Linkage* deterministico basato sul "*Social Security Number*" di cui si presenta la tabella conclusiva (tabella 1.3).

Tabella 1.3 Tabella riassuntiva corretta per ospedale delle coppie di dati e delle non corrispondenze trovate usando il "*Social security number*" come chiave identificativa per il *Record Linkage*

	OULU	TAMPERE	SEINÄJOKI	Totale
Coppie di record con corrispondenza	318	137	123	578
Record individuali/clinici senza corrispondenza	201	109	33	343
Record genetici senza corrispondenza	274	22	5	301
Totale	793	268	161	1222

In conclusione il 64,2% del totale dei record iniziali, comprensivi di record clinici/individuali e genetici, ha trovato un abbinamento, mentre il restante 35,8% è rappresentato da record per cui si riscontra un settore di celle vuote nel database definitivo.

1.12 Conclusioni

È stata confermata l'ipotesi relativa alla migliore efficacia del *Record Linkage* deterministico effettuato utilizzando il "*Social Security Number*" come chiave identificativa rispetto al metodo attraverso il set di variabili identificative. Questa conclusione non può tuttavia essere generalizzata poiché nel caso dello studio al Children's Hospital dell'Università di Oulu era stato eseguito un ampio lavoro di pulizia della variabile "*Social Security Number*", pensata essere l'identificativo principale. Non è stato, invece, possibile avere alcun riscontro sull'attendibilità delle variabili utilizzate nel set preso come seconda chiave identificativa. Anche se è stato utilizzato il metodo di codifica Soundex per evitare che nomi trascritti con spelling differenti potessero portare a mancati appaiamenti, la digitazione manuale porta inevitabilmente ad una percentuale d'errore elevata. Sicuramente quindi la più alta percentuale di concordanze nel *Record Linkage* con il "*Social Security Number*" è dovuta al più alto rischio di errori su un insieme di quattro variabili rispetto ad una unica. È fortemente consigliabile, quindi, avere la possibilità di portare a termine due metodi di *Record Linkage* centrati su differenti identificativi, in modo da avere un riscontro. Nel caso si dovesse progettare dall'inizio una base informativa è indispensabile scegliere un identificativo utilizzato a livello nazionale in modo da poter permettere in un secondo momento l'integrazioni tra più fonti; inoltre il codice identificativo scelto deve puntare su un ottima qualità (nella coerenza e nella codifica) e su una totale completezza.

Il database ottenuto attraverso il lavoro di Record Linkage esaminato nel lavoro al Children's Hospital dell'Università di Oulu, sarà utile per tre diversi scopi:

1. portare a termine uno studio descrittivo o analitico con il fine di indagare relazioni tra variabili della stessa macro-area (clinica o genetica) nei nati pretermine negli ospedali di Oulu, Tampere e Seinäjoki. In questo caso si potranno estrapolare sia i record appaiati, sia quelli specifici della macro-area d'interesse;
2. studiare una qualsivoglia relazione tra variabili cliniche/individuali e variabili genetiche: si dovranno utilizzare solamente i 578 record (o un campione di essi) che possiedono tutte le informazioni.;
3. integrare verticalmente la fonte di dati, aggiungendo nuovi record di pazienti nati dopo il 2006 in modo da avere un quadro aggiornato della situazione e con l'obiettivo di seguire il trend delle nascite pretermine nei tre ospedali, continuando gli studi eziologici sulle patologie neonatali polmonari con un numero sempre maggiore di dati.

CAPITOLO 2 *Missing Values*

2.1 Premessa

Come si è visto anche nel precedente capitolo a proposito delle chiavi identificative, la completezza dei dati è di fondamentale importanza in un database poiché più alto è il numero di dati mancanti più le analisi produrranno stime distorte. Qualora non fosse possibile recuperare le informazioni dalle fonti di dati originarie, si possono utilizzare dei sistemi per l'analisi di valori mancanti. Questo capitolo presenta le tipologie di dati mancanti che si possono affrontare (paragrafo 2.2), soffermandosi poi sulla specifica trattazione delle mancate risposte parziali, o dati mancanti parziali (paragrafo 2.4). In letteratura si trovano quattro gruppi di metodi per il trattamento di dati mancanti (paragrafo 2.5): metodi basati sulle sole unità osservate, procedure di ponderazione, metodi basati sui modelli e metodi di imputazione. Dopo una breve presentazione di ciascun gruppo metodologico (paragrafi 2.5.1 – 2.5.4), questo capitolo si pone l'obiettivo di investigare il vasto campo dei metodi d'imputazione che vengono usati prevalentemente nel caso di mancate risposte parziali e consistono nel sostituire i valori mancanti con valori opportunamente calcolati (paragrafi 2.5.4.1 e 2.5.4.2). Si presenteranno in un primo momento i metodi d'imputazione deduttivi, ma l'interesse sarà orientato verso i metodi d'imputazione deterministici e stocastici dei quali verrà fornito un elenco dettagliato (paragrafi 2.5.4.3 e 2.5.4.4). Successivamente si sposterà l'attenzione sulla diversificazione dei metodi d'imputazione in approccio singolo (paragrafo 2.6) ed approccio multiplo (paragrafo 2.7) fornendone pure le basi teoriche. Infine si presenterà l'applicazione del trattamento di dati mancanti parziali nello studio condotto al Children's Hospital dell'Università di Oulu (paragrafo 2.9): è stato scelto un metodo d'imputazione deduttivo per i valori mancanti relativi alle variabili data di nascita, sesso ed età gestazionale, mentre per recuperare informazioni sul peso alla nascita è stato adottato un metodo d'imputazione multipla, e più in particolare si è ricorsi alla procedura MI del programma statistico Sas System.

2.2 Tipologie di valori mancanti (*missing values*)

Il problema delle mancate risposte su una o più variabili complica molto frequentemente le analisi dei dati negli studi scientifici. Questo è di particolare rilevanza nell'area medico-sanitaria e nelle scienze sociali, campi in cui i valori mancanti sono endemici (Juster & Smith, 1998; Horton & Lipsitz, 2001; Barzi & Woodward, 2004; Durrant, 2005). Un ricercatore è messo di fronte alla questione dei dati mancanti nel particolare momento in cui andrà a trattare i dati raccolti, ma può non essere familiare con i metodi di analisi statistica che si rivolgono a risolvere adeguatamente tale problema. È stata proposta una gran varietà di approcci per l'analisi dei dati mancanti e tra questi solo una piccola parte è diventata di uso stabile e comune (Ake, 2005). Gli approcci tradizionali possono portare a stime distorte e di conseguenza a conclusioni non valide (Acock, 2005), quindi prima di iniziare a trattare con i dati mancanti è indispensabile avere una buona conoscenza dei diversi metodi di analisi esistenti in letteratura.

Le cause che conducono all'incompletezza dell'informazione sono numerose e diverse, ma si possono riassumere in tre grandi categorie (Shill et al., 1993):

1. mancata copertura
2. mancate risposte totali
3. mancate risposte parziali

Per "***mancata copertura***" si intende l'esclusione dalla lista di campionamento di alcune unità appartenenti alla popolazione obiettivo. Dato che queste unità hanno probabilità nulla di essere selezionate, rimangono escluse dai risultati dell'indagine. Le cause possono riguardare omissioni nel preparare le liste della popolazione, esecuzioni difettose sul campo oppure la cattiva qualità delle liste di campionamento causata da mancate denunce o ritardi di aggiornamento. Questo tipo di incompletezza è abbastanza difficile da individuare e da trattare e la compensazione della mancata copertura avviene utilizzando informazioni provenienti da fonti esterne.

Il caso di "***mancate risposte totali***" si riscontra quando l'individuo oggetto dell'indagine ha rifiutato di collaborare (non rispondendo al questionario o non fornendo alcuna informazione richiesta), o ancora nel caso in cui il rilevatore o l'addetto alla raccolta dei dati non è riuscito a rintracciare o contattare la persona d'interesse. Altra possibilità concerne l'impossibilità da parte dell'intervistatore a comunicare con gli individui, ad esempio per problemi di lingua o malattia. L'effetto della non risposta totale può essere grave soprattutto quando le persone non

intervistate sono in qualche modo diverse da quelle intervistate. Ciò può infatti causare distorsioni nelle stime dei parametri delle quantità di interesse.

Infine il caso delle "***mancate risposte parziali***" riguarda situazioni in cui l'informazione dell'intervistato manca solo in uno o pochi campi (Grande & Luzi, 2002). Tale situazione si riscontra se:

- l'intervistato si rifiuta o non è in grado di rispondere ad una determinata domanda;
- l'intervistatore si dimentica di porre una domanda o registrare una risposta;
- la risposta rilevata risulta incongruente e quindi cancellata in fase di revisione dei dati;
- vi sono stati degli errori nella registrazione del supporto;
- nel caso di studi epidemiologici o clinici, non è stato possibile avere l'informazione d'interesse a causa di un cattivo funzionamento di apparecchiature, macchinari o terapie.

Tale tipo di incompletezza risulta la più semplice da gestire in quanto si dispone di una serie di informazioni ausiliarie sull'individuo in questione.

La conseguenza immediata della non risposta parziale è che non si dispone di un dataset rettangolare, cioè completo di tutte le informazioni, e dunque le analisi statistiche tradizionali non sono più direttamente applicabili, inoltre diminuisce l'efficienza delle stime poiché la numerosità campionaria viene ridotta.

2.3 Motivi per cui l'intervistato può non voler rispondere

Per quanto riguarda il rifiuto a collaborare da parte dell'intervistato, le variabili che hanno il potere di influenzare questo fenomeno della mancata collaborazione sono (Bosio, 1997):

- Caratteristiche del contesto sociale. Incidono sulla disponibilità a rispondere caratteristiche generali del contesto sociale, quali il livello di urbanizzazione, l'adesione al valore della privacy, la percezione di legittimazione delle istituzioni e il grado di coesione sociale.

- Caratteristiche del proponente, degli scopi della ricerca, dell'oggetto. Appare ampiamente verificata la maggiore propensione a collaborare a fronte di proponenti autorevoli o di ambito pubblico (università, organismi di emanazione pubblica). La conoscenza degli scopi di un'indagine (finalità, modalità di impiego) favorisce la partecipazione più che la semplice informazione sui contenuti.

Il tema dell'indagine infine influenza l'orientamento a cooperare; sembrano qui in gioco dimensioni di valutazione non solo individuali ma anche collettive. Di solito quesiti più delicati, che riguardano cioè la sfera personale e sensibile dell'individuo, vanno incontro a maggiori rifiuti. Soprattutto in indagini che puntano ad indagare aspetti come il reddito, le abitudini sessuali, metodi di fecondazione, gravi patologie o aspetti giudiziari, la persona si chiude a riccio e sente la propria intimità e privacy invase, autocensurandosi. Per ovviare a questo problema, già nella pianificazione dello studio è necessario avere la consapevolezza del fatto che si andrà incontro a possibili rifiuti da parte dei rispondenti o delle loro famiglie (nel caso di minorenni), quindi si cercherà di trovare punti di forza che facciano sentire il rispondente tutelato, come l'anonimato. È utile anche l'aiuto di particolari tecniche preventive, in fase di pianificazione del disegno d'indagine, miranti al controllo dell'atteggiamento dei rispondenti.

- Caratteristiche del disegno di ricerca. Tecniche specifiche di raccolta dei dati (indagine postale, telefonica, personale) si dimostrano diversamente efficaci nell'arruolare i rispondenti. Le proprietà dello strumento di rilevazione svolgono anch'esse un ruolo non trascurabile: la lunghezza del questionario, il formato e il colore, il "*wording*" e la diramazione delle domande.

- Caratteristiche delle persone del campione. E' questa senza dubbio l'area più esplorata e più ricca di riferimenti; sono state identificate connessioni fra propensione a rispondere e caratteristiche del rispondente, quali: età, istruzione, livello socioeconomico, sesso e razza.

- Caratteristiche dell'intervistatore. Per quanto sia diffusa la consapevolezza circa l'importanza della variabile intervistatore in rapporto al fenomeno dei non rispondenti, le evidenze di ricerca sono scarse e spesso sono sostituite da considerazioni di buon senso circa i tratti socioculturali, psicologici e professionali desiderabili per un intervistatore.

2.4 Mancate risposte parziali

La categoria delle mancate risposte parziali sarà la maggiormente trattata in questo studio in quanto rappresenta il problema più diffuso negli studi clinici. Presupponendo dunque questa situazione di dati mancanti, è possibile specificarne tre diversi tipi di struttura da un punto di vista teorico (Little & Rubin, 1987):

se Y è la variabile di studio e X una sua covariata possiamo distinguere tre casi:

1. la probabilità di risposta per Y è indipendente da X e da Y;
2. la probabilità di risposta per Y dipende da X ma non da Y;
3. la probabilità di risposta per Y dipende da Y ed eventualmente da X.

Se si verifica il caso 1, i dati sono mancanti completamente a caso, in inglese ***missing completely at random*** (**MCAR**). In questo caso i valori osservati di Y formano un campione casuale dei valori di Y. Se si verifica il caso 2, i dati sono mancanti a caso, ***missing at random*** (**MAR**). In questa situazione i valori osservati di Y formano un campione casuale dei valori di Y all'interno di classi definite sulla base dei valori di X. Nel caso 3 il meccanismo che genera i dati mancanti viene detto mancante non a caso, ***not missing at random*** (**NMAR**) o *nonignorable*.

2.4.1 Dati mancanti completamente a caso (missing completely at random − MCAR)

È l'ipotesi più facile da trattare, ma è anche quella più difficilmente riscontrabile in situazioni concrete: la probabilità di osservare una risposta mancante è indipendente sia dalla parte osservata che da quella non osservata dell'insieme di dati completo.

Il termine ha un preciso significato (Little & Rubin, 1987; Rubin, 1977): pensando al dataset come un'ampia matrice di dati, i valori mancanti sono casualmente distribuiti attraverso la matrice. Negli studi sulle famiglie questo accade raramente dato che è ben noto che individui in gruppi di minoranze, persone con alti redditi, soggetti con un basso livello d'istruzione e persone soggette a sindromi di depressione o di ansia, sono meno propense a rispondere a tutti gli item di un questionario rispetto alle loro controparti. Anche negli studi clinici è raro osservare dati mancanti completamente a caso, perché gli addetti all'inserimento dei dati

solitamente trascurano alcuni campi in modo sistematico, o perché i pazienti con particolari condizioni piuttosto che altre si rifiutano di fornire particolari informazioni.

2.4.2 Dati mancanti a caso (missing at random − MAR)

Si parla di dati MAR o non risposta ignorabile quando la probabilità di osservare una risposta mancante dipende soltanto dalla parte osservata dell'insieme di dati. I dati mancanti su una variabile specifica si definiscono "mancanti a caso" se la verosimiglianza dei dati mancanti sulla variabile d'interesse non è correlata con il punteggio individuale su quella variabile, dopo aver controllato le altre variabili nello studio. In uno studio sulla depressione materna (Acock, 2005), il 10% o più delle madri può rifiutare di rispondere alle domande sul loro livello di depressione. Si supponga che uno studio includa lo stato di povertà codificato come 1="in stato di povertà" e 0="non in stato di povertà". Un punteggio sulla depressione della madre è "mancante a caso" se i suoi valori mancanti sulla depressione non dipendono dalla variabile "livello di depressione", tenendo sotto controllo la variabile "stato di povertà". Se la verosimiglianza del rifiuto a rispondere al quesito sul livello di depressione è correlato con lo stato di povertà, ma non è correlato con il livello di depressione entro ogni classe di stato di povertà, allora i valori mancanti sono detti di tipo MAR. Per i dati MAR, in questo esempio, il punto focale non è se lo stato di povertà possa predire la depressione materna, ma se lo stato di povertà rappresenti un "elemento esplicativo" della presenza o meno dell'informazione "livello di depressione materna".

Una variabile è considerata "elemento esplicativo" quando aiuta a spiegare se un soggetto risponderà o meno ad un quesito (Raghunathan, 2004; Schafer, 1997). Molti "elementi esplicativi" vengono inclusi negli studi sulle famiglie di grande scala: i più comuni "elementi esplicativi" inclusi sono il livello d'istruzione, la razza, l'età, il sesso ed indicatori di benessere psico-sociale.

L'assunto per i valori MAR è valido solo se il modello dei dati mancanti è condizionatamente casuale, dati i valori osservati nelle variabili considerate "elementi esplicativi".

2.4.3 Dati mancanti non a caso (missing not at random − MNAR)

Si parla di dati MNAR o non risposta non ignorabile quando la probabilità di risposta dipende sia dai dati osservati che da quelli non osservati: in questo caso il meccanismo della non risposta deve essere tenuto esplicitamente in considerazione nel modello di analisi del

fenomeno. Quando si verificano situazioni di questo tipo l'applicazione delle tecniche di imputazione diventa assai più problematica, ragion per cui non si tratteranno i casi in cui i dati sono di questo terzo tipo.

2.5 Metodi per il trattamento di dati mancanti

Il fattore di base che fa propendere verso un particolare metodo per il trattamento di valori mancanti piuttosto che un altro, riguarda la specifica causa che conduce al problema, vale a dire se si è nel caso di mancata copertura, di non risposte totali o di non risposte parziali. Altro elemento che deve essere preso in considerazione nella scelta del metodo è, come si vedrà in seguito, il tipo di struttura dei dati mancanti (MCAR, MAR, MNAR).
In generale i metodi proposti in letteratura per l'analisi dei dati in presenza di osservazioni mancanti possono essere classificati in quattro gruppi:

1. metodi basati sulle sole unità osservate;
2. procedure di ponderazione;
3. metodi basati sui modelli;
4. metodi di imputazione.

2.5.1 Metodi basati sulle sole unità osservate

L'analisi sulle sole unità osservate concentra l'attenzione sui record per cui tutte le variabili sono presenti eliminando, completamente o parzialmente, quei record per cui una o più informazioni relative alle variabili d'interesse sono mancanti. I vantaggi di questo approccio sono la semplicità, dal momento che possono essere applicate analisi statistiche su dati completi senza trasformazioni dei dati, e la comparabilità di statistiche univariate, dato che quest'ultime sono tutte calcolate su un campione di casi comune. Gli svantaggi derivano dalla perdita potenziale di informazioni nel scartare casi incompleti. Tale perdita di informazione comporta una perdita di precisione ed una distorsione quando i dati non sono MCAR e quando i casi completi non sono un campione casuale di tutti i casi (Little & Rubin, 2002).
Esistono due diversi tipi di approccio:

- escludere completamente e in maniera definitiva in tutte le analisi (in inglese *listwise*) l'unità di cui manca il dato su una o più variabili;

- escludere per un solo confronto, o parzialmente (in inglese *pairwise*) l'unità di cui manca il dato, ovvero eliminarla solo nei confronti tra coppie di variabili su almeno una delle quali manca l'informazione, mentre l'unità rimane attiva per il computo degli indici costruiti con variabili sulle quali l'informazione è presente. Si tenga presente che ciò è possibile soltanto nel caso delle mancate risposte parziali.

Entrambi i metodi portano a risultati soddisfacenti, ma sono applicabili solo sotto ipotesi di dati MCAR.

Esclusione definitiva (*Listwise o Case Deletion*)

Si tratta di un metodo semplice che viene applicato agevolmente prima dell'analisi e dà risultati semplici da interpretare. È la soluzione più comune al problema di valori mancanti, così comune che molti pacchetti statistici standard la utilizzano per default. Molti ricercatori, però, considerano questo approccio conservativo e non vantaggioso data una perdita media del 20-50% dei dati. Se l'assunzione di dati MCAR è valida, il metodo *listwise* è considerato conservativo perché la numerosità del campione viene ridotta considerevolmente e questo problema gonfia gli errori standard e riduce il livello di significatività. Solo nel caso di campioni con numerosità molto elevata, dunque, può essere consigliato questo tipo di approccio. Inoltre se i dati non incontrano l'assunzione di MCAR, il metodo *listwise* può produrre stime distorte. Generalmente si potrebbe avere una distorsione perché le sole unità complete potrebbero non essere rappresentative dell'intera popolazione; per esempio, soggetti meno istruiti, con maggiori problemi di salute mentale e così via, potrebbero non essere rappresentati (Graham & Donaldson, 1993).

Se i valori mancanti sono MCAR, allora il metodo *listwise* fornirà stime non distorte e l'unico costo riguarda una riduzione della rilevanza statistica. In definitiva, se si ha un campione sufficientemente grande, se la rilevanza statistica non costituisce un problema e se i valori mancanti sono MCAR, allora il metodo *listwise* è una strategia ragionevole.

Esclusione parziale (*Pairwise Deletion*)

Ha un significato compiuto solo per le analisi di relazioni tra variabili o tra unità e può causare qualche impaccio nell'interpretazione dei risultati basati su numerosità statistiche che variano da indice a indice (Fabbris, 1997). Il metodo *pairwise* utilizza tutta l'informazione delle variabili, nel senso che tutti i rispondenti che hanno un'informazione completa per due variabili, vengono utilizzati per stimare la covarianza tra queste variabili, senza contare se hanno o meno un valore mancante per le altre variabili. Tuttavia c'è una ragione principale

che rende il metodo *pairwise* poco popolare, cioè il fatto che può produrre una matrice delle covarianze in cui ogni covarianza potrebbe essere basata su diversi sottocampioni di unità; infatti le covarianze non hanno i vincoli che avrebbero se le covarianze fossero basate sullo stesso dataset di unità. È possibile dunque che la matrice di correlazione ottenuta con il metodo *pairwise* non si riesca ad invertire, passaggio fondamentale per la stima dell'equazione di regressione. In alcuni programmi statistici può capitare che questo problema venga segnalato nella schermata di output come avviso del fatto che la matrice utilizzata non è definita positiva.
Un ultimo problema con il metodo *pairwise* riguarda la difficoltà nel calcolare i gradi di libertà perché diverse parti del modello hanno diversi campioni.

2.5.2 Metodi di ponderazione

Il trattamento dei dati mancanti nei casi di mancata copertura viene solitamente effettuato attraverso una procedura di ponderazione basata su informazioni provenienti da una fonte di dati esterna (Kalton, Kasprzyk, 1982).
Anche la compensazione per le mancate risposte totali avviene nella maggior parte dei casi tramite riponderazione. Infatti, nel caso in cui le informazioni relative ad alcune unità statistiche risultano completamente mancanti e non è possibile o non si ritiene opportuno procedere alla loro integrazione, è necessario tenere conto di questa assenza di informazione a livello di stima finale: ciò può essere fatto incrementando il valore dei pesi campionari di unità rispondenti considerate rappresentative di quelle non rispondenti (Kalton, Kasprzyk, 1986).
Le procedure di ponderazione consistono nel modificare i pesi assegnati alle unità effettivamente osservate a rilevazione avvenuta al fine di rappresentare anche quelle non registrate. Anche qui il vantaggio è la semplicità di applicazione, mentre gli svantaggi riguardano le difficoltà nel reperire le informazioni ausiliarie sui non rispondenti o sull'intera popolazione, per costruire i pesi; inoltre non sempre le quantità pesate sono facilmente interpretabili.
In genere si utilizza il metodo di ponderazione se:

- il campione non è autoponderante, ossia quando le unità statistiche sono state selezionate con probabilità variabili. Se in un campione di numerosità n, la probabilità di selezione dell'unità i-esima è p_i ($i = 1, \ldots, n$), il peso associato all'unità per il riporto all'universo è dato da: $w_i = 1/p_i$;

- la frequenza con la quale le unità statistiche hanno validamente partecipato alla rilevazione varia di categoria in categoria. Se, per esempio, in un'indagine sul reddito e sul risparmio degli italiani non si sono ottenuti per alcune unità i dati sull'impiego del risparmio, si possono surrogare le mancate risposte sul risparmio assegnando maggior peso a quelle dei rispondenti che appartengono alla stessa classe di reddito. Sia h la categoria di reddito cui appartiene il rispondente i (i = 1, …, n), il peso assegnato all'unità è inversamente proporzionale alla frequenza di risposta della categoria h: $w_i = 1/f_h$.

Ad un'unità selezionata con probabilità variabile e appartenente ad una categoria con mancate risposte si assocerà un peso pari a $(1/p_i)*(1/f_h)$.

La decisione di introdurre i pesi nell'analisi dipende dal rapporto tra costi (in termini di tempi di elaborazione, pesantezza dell'analisi) e rilevanza statistica e interpretativa del rischio di distorsione per mancata correzione dei dati. La ponderazione si applica correttamente se le unità mancanti sono mediamente simili a quelle che hanno collaborato (Fabbris, 1997).

2.5.3 Metodi basati su modelli

Questi metodi consistono nell'ipotizzare un modello parametrico sottostante ai dati e stimare i parametri attraverso metodo della massima verosimiglianza, sotto l'ipotesi che i dati siano MAR e che i parametri della funzione di densità dei dati siano distinti dai parametri del meccanismo generatore dei dati mancanti. Le stime risultanti sono corrette e, sotto ipotesi di normalità, ottime. Usualmente si usa l'algoritmo EM (*Expectation-Maximization*) per la massimizzazione della verosimiglianza (Dempster et al., 1977). Potenzialmente questi metodi si adattano bene nel trattamento di valori mancanti, ma spesso richiedono assunzioni di tipo distributivo che i dati non sono in grado di supportare.

Sono tuttavia metodi piuttosto flessibili che non necessitano di procedure *ad hoc* per aggiustare le stime e che dispongono di stime asintotiche della varianza (attraverso la derivata seconda della log-verosimiglianza), le quali tengono conto dell'incompletezza dei dati. D'altra parte, però, i calcoli sono piuttosto dispendiosi e, ad oggi, sono assai modeste le conoscenze sulle proprietà in campioni con bassa numerosità. Infine se il modello specificato non è corretto, le stime di massima verosimiglianza non sono consistenti.

Per una trattazione più approfondita si vedano Little e Rubin (1987), Little (1984).

2.5.4 Metodi di imputazione

I metodi d'imputazione rappresentano il nucleo d'interesse di questo capitolo poiché costituiscono il sistema per il trattamento dei valori mancanti adottato nella gran parte degli studi clinici. Per questa ragione si suddividerà in sotto paragrafi la loro trattazione.

2.5.4.1 *Caratteristiche generali*

I metodi di imputazione vengono usati prevalentemente nel caso di mancate risposte parziali e consistono nel sostituire i valori mancanti con valori opportunamente calcolati, producendo un dataset completo che possa poi essere analizzato con l'utilizzo di metodi d'inferenza per dati completi. Rientrano in questa categoria molti criteri effettivamente usati nella realtà, e nel caso particolare dell'applicazione che si vedrà in seguito è stato utilizzato proprio un metodo d'imputazione per risolvere il problema di dati parzialmente mancanti. Questi metodi risultano particolarmente attraenti perché semplici e intuitivi; permettono di ricondursi a situazioni di dati completi senza scartare nessuno dei dati osservati.
Numerosi sono i metodi di imputazione proposti in letteratura per predire valori sostitutivi per le mancate risposte parziali. In linea generale possiamo considerare tre classi di metodi:

- ***metodi deduttivi***, nei quali il valore imputato è dedotto da informazioni o relazioni note. Un semplice esempio può essere quello di un record che contiene una serie di cifre ed il loro totale, ma una delle cifre è mancante: quest'ultima viene dedotta per sottrazione;

- ***metodi deterministici***, nei quali imputazioni ripetute per unità aventi le stesse caratteristiche producono sempre gli stessi valori imputati;

- ***metodi stocastici***, nei quali imputazioni ripetute per unità aventi le stesse caratteristiche possono produrre differenti valori imputati; si caratterizzano per la presenza di una componente aleatoria, detta anche residuo, corrispondente ad uno schema probabilistico associato al particolare metodo d'imputazione prescelto.

Il trattamento di valori mancanti con tecniche d'imputazione richiede che le mancate risposte siano del tipo MCAR o MAR. Può però capitare che si debba lavorare con dati mancanti che non rispettino alcuna delle due ipotesi appena citate. Un modo per far fronte a questa situazione è quello di ricondursi a trattare sottoinsiemi omogenei di dati che presentino tali

caratteristiche di "mancanza casuale". Questo è possibile facendo ricorso alle cosiddette *classi di imputazione* per la variabile di risposta Y, ottenute operando all'interno del dataset con una serie di opportune stratificazioni in base alle diverse modalità delle covariate $X_1, X_2 ... X_n$. In questo modo si possono effettuare le imputazioni per la variabile di risposta Y utilizzando sottoinsiemi omogenei del dataset all'interno dei quali i rispondenti presentano caratteristiche per lo più simili. Uno dei principali obiettivi nella costruzione delle classi è, chiaramente, quello di fare in modo che esse siano in grado di spiegare la percentuale più elevata possibile di varianza della variabile sulla quale si effettua l'imputazione. Il numero delle classi deve essere determinato in modo da assicurare la presenza di un numero minimo di rispondenti in ogni classe al fine di ottenere stime affidabili dei valori mancanti.

Nel caso delle mancate risposte parziali solitamente è possibile esprimere il valore imputato come funzione di una o più variabili ausiliarie ritenute in grado di avere una buona capacità di rappresentare il dato mancante. I dati dovrebbero provenire da una distribuzione multivariata normale, altrimenti si può ricorrere a tecniche di normalizzazione dei dati (vedasi capitolo 4) oppure utilizzare un modello per distribuzioni non-normali (SAS Institute Inc., 1999).

Quasi tutti i metodi di imputazione deterministici e stocastici possono essere descritti, almeno in via approssimativa, come casi speciali del modello di regressione

$$y_{mi} = \beta_{r0} + \sum_j \beta_{rj} z_{mij} + e_{mi}$$

dove y_{mi} rappresenta il valore imputato per la i-esima unità con un valore mancante, z_{mij} è il valore delle variabili ausiliarie, β_{r0} e β_{rj} sono i coefficienti della regressione di *y* su *z* per i rispondenti, mentre e_{mi} costituisce un residuo corrispondente ad un determinato schema probabilistico.

La distinzione essenziale tra metodi deterministici e metodi stocastici dipende quindi dall'aver posto $e_{mi} = 0$ oppure no. In pratica ogni metodo deterministico ha la sua controparte stocastica e viceversa.

2.5.4.2 *Metodi d'imputazione messi a confronto*

I metodi deduttivi, che trovano ampia applicazione in campo amministrativo, sebbene adeguati e per certi versi "ideali" in quanto permettono di sostituire ai dati mancanti valori "reali", cioè il più possibile vicini a quelli che si sarebbero realmente osservati, risultano essere, da un punto di vista metodologico, specifici dei fenomeni investigati (ad esempio,

economici o demografici). Proprio per questa ragione, tali metodi presuppongono in genere la definizione di "modelli" di comportamento specifici del fenomeno in oggetto, sviluppati da esperti. Talvolta l'imputazione deduttiva è addirittura considerata come facente parte del processo di editing (Schulte Nordholt, 1998). Si è quindi ritenuto più opportuno privilegiare un'analisi comparativa dei soli metodi deterministici e stocastici, appartenenti alla classe più generale dei metodi "statistici".

La scelta tra un metodo di imputazione deterministico ed il suo corrispondente stocastico dipende principalmente dal tipo di analisi che si intende condurre.

Se, ad esempio, l'obiettivo dell'analisi è limitato alla stima della media della popolazione l'utilizzo dei metodi deterministici è da considerarsi preferibile. Infatti, sebbene entrambi gli approcci diano gli stessi risultati soddisfacenti in termini di distorsione delle stime, la presenza della componente casuale dei residui nei metodi stocastici genera comunque una certa perdita di precisione nella stima della media. Quando, invece, si intende effettuare analisi che richiedono la preservazione della varianza e della distribuzione di una data variabile, l'utilizzo dei metodi deterministici può condurre a risultati di qualità non elevata. Questi metodi, infatti, causano un'attenuazione della varianza della variabile per la quale è stata effettuata l'imputazione e generano delle distorsioni nella forma della sua distribuzione. Si consideri come esempio l'imputazione mediante valore medio (all'interno di classi di imputazione). Sostituendo in ogni classe a tutti i valori mancanti il valore medio dei rispondenti, la distribuzione delle risposte risulterà distorta a causa di una serie di picchi artificiali che si formeranno in corrispondenza della media di ciascuna classe. La variabilità della variabile oggetto di studio risulterà, quindi, attenuata in quanto i valori imputati riflettono soltanto la varianza *tra le classi* (*between*) e non quella *all'interno delle classi* (*within*). In questi casi l'uso di un metodo di imputazione stocastica permette di conseguire migliori risultati, grazie alla presenza della componente casuale dei residui che permette di catturare anche la parte di varianza all'interno delle classi.

Quando si decide di utilizzare un metodo di imputazione stocastica si pone il problema della scelta di un'opportuna distribuzione dalla quale estrarre la componente casuale. Nel caso di una imputazione mediante modello di regressione una scelta plausibile è rappresentata da residui con distribuzione normale, con media zero e varianza uguale alla varianza residua della regressione sui rispondenti. Possibili alternative sono rappresentate dalla scelta casuale della distribuzione empirica dei residui dei rispondenti o la scelta di un residuo a partire da unità rispondenti considerate "vicine" all'unità con valore mancante rispetto a prefissate variabili ausiliarie. Ciò, come si vedrà più avanti, è quello che ad esempio si verifica

effettuando una imputazione con metodi *hot deck* di tipo *nearest-neighbour*, in cui all'unità con dato mancante è assegnato il valore di un'unità rispondente estratta da un sottoinsieme di unità rispondenti considerate "vicine".

Un altro problema di fondamentale importanza è quello legato ai forti effetti che il processo di imputazione può avere sui legami tra due o più variabili, spesso con il risultato di attenuare le relazioni di associazione. A tal proposito Kalton e Kasprzyk (1986) cercano di valutare gli effetti dell'imputazione sulle relazioni tra la variabile di studio *y*, con dati incompleti caratterizzati da un meccanismo di mancata risposta *missing at random* (MAR), ed un'altra variabile *x*, che non presenta dati mancanti, attraverso gli effetti prodotti sulla loro covarianza. La conclusione a cui si giunge è che se lo studio della relazione tra *x* ed *y* rappresenta una componente importante delle analisi dei dati effettuate in un'indagine è necessario utilizzare *x* come variabile ausiliaria nel processo di imputazione dei valori mancanti di *y*. Questo infatti è l'unico modo per ottenere una stima non distorta dalla covarianza tra le due variabili. Se poi *x* ed *y* contengono entrambe mancate risposte, la covarianza può risultare attenuata per effetto dell'imputazione simultanea su entrambe le variabili. Un caso particolare si ha, infine, quando *x* ed *y* contengono mancate risposte in corrispondenza della stessa unità: in tal caso imputando congiuntamente, cioè utilizzando lo stesso "donatore" per l'imputazione sia di *x* che di *y*, la struttura della covarianza viene preservata.

2.5.4.3 *Elenco dei Metodi d'imputazione <u>Deterministici</u>*

a) **Imputazione con media (Mean imputation overall)**

Metodo

Con questo metodo si sostituiscono tutte le mancate risposte nella variabile y con un unico valore, la media calcolata sul totale dei rispondenti, cioè $\bar{y}_r$. E' un metodo che può essere utilizzato solo per le variabili quantitative (per le variabili qualitative al posto del valor medio si può imputare la moda).

Può anche essere interpretato come la trasformazione deterministica della funzione lineare (1) senza variabili ausiliarie:

$$y_{mi} = \beta_{r0} = \bar{y}_r$$

Campi di applicazione

E' consigliabile utilizzare questo metodo solo nei casi in cui: il numero dei dati mancanti per ciascuna variabile è esiguo; lo scopo dell'analisi è limitato alla stima di medie e totali; sembrano esistere poche relazioni tra le variabili; è richiesto un metodo di rapida applicazione.

Vantaggi

- Preserva la media dei rispondenti.
- Facile da applicare e da spiegare.

Svantaggi

- Introduce una seria distorsione nella distribuzione della variabile, creando un picco artificiale in corrispondenza del suo valor medio.
- Non dà buoni risultati nella stima della varianza.
- Provoca distorsioni nelle relazioni tra le variabili.

b) **Imputazione con media all'interno delle classi (Mean imputation within classes)**

Metodo

Si divide il campione totale in classi di imputazione in base ai valori assunti da prefissate variabili ausiliarie considerate esplicative di y e si calcola la media dei rispondenti della variabile y all'interno di ogni classe. Ciascuna media viene poi assegnata ai valori mancanti in unità appartenenti alla stessa classe: $y_{mhi} = \bar{y}_{rh}$, per l'i-esimo non rispondente della classe h (h = 1, 2…H).

Per individuare la migliore classificazione possibile possono essere utilizzati diversi metodi di analisi multivariata, come per esempio la *cluster analysis*.

Campi di applicazione

L'applicazione di questo metodo può essere utile nei casi in cui: l'obiettivo dell'analisi è rappresentato dalla stima di medie e aggregati; sembrano esistere poche relazioni tra le variabili; è richiesto un metodo di rapida applicazione.

Vantaggi

- Può ridurre le distorsioni generate dalle mancate risposte (se la scelta delle classi di imputazione è stata effettuata in modo appropriato).
- Semplice da applicare e da spiegare, una volta definite le classi di imputazione.

Svantaggi

- Introduce distorsioni (sebbene in maniera meno evidente del metodo precedente) nella distribuzione della variabile, creando una serie di picchi artificiali in corrispondenza della media di ciascuna classe.
- Provoca un'attenuazione della varianza della distribuzione dovuta al fatto che i valori imputati riflettono solo la parte di variabilità tra le classi (*between*) ma non quella all'interno delle classi (*within*).
- Provoca distorsioni nelle relazioni tra le variabili non considerate per la definizione delle classi di imputazione.

c) Imputazione con regressione (Predictive regression imputation)

Metodo

Con questo metodo si utilizzano i valori dei rispondenti per stimare i parametri della regressione per la variabile di studio y su prefissate variabili ausiliarie considerate esplicative di y. Le determinazioni della y sono, poi, imputate come valori stimati dell'equazione di regressione: $y_{mi} = \beta_{r0} + \sum_j \beta_{rj} z_{mij}$. Le variabili ausiliarie, nel modello di regressione, possono essere sia di natura quantitativa che qualitativa. Se la variabile y è quantitativa generalmente vengono utilizzati modelli di regressione lineare. Nel caso in cui, invece, la variabile y sia qualitativa, si possono adottare modelli log-lineari o logistici.

Un caso particolare del modello di regressione è costituito dal modello *ratio*: $y_{mi} = \beta_r z_i$ con una sola variabile ausiliaria ed intercetta zero, adoperato, ad esempio, nelle indagini *panel*, in cui z rappresenta la rilevazione di y in una precedente sessione dell'indagine.

Anche questo metodo può richiedere la *suddivisione in classi delle unità*. Infatti diversi modelli possono essere necessari in ogni classe, in quanto (soprattutto per variabili di tipo economico) le relazioni tra y e le covariate possono cambiare molto da strato a strato.

Campi di applicazione

Il metodo ben si adatta a situazioni in cui la variabile sulla quale effettuare l'imputazione è quantitativa oppure binaria oltre che naturalmente essere fortemente correlata con altre variabili. E' meno adatto, invece, a situazioni in cui le variabili qualitative presentano numerose modalità.

Vantaggi

- Si può fare uso di un numero elevato di variabili, sia quantitative che qualitative, in modo da ridurre, più che con altri metodi, le distorsioni generate dalle mancate risposte.
- Preserva bene le relazioni delle variabili usate nel modello.

Svantaggi

- Introduce distorsioni nella distribuzione della variabile (sebbene meno del metodo con donatore casuale all'interno delle classi).
- Essendo un metodo deterministico, non preserva sufficientemente la variabilità delle distribuzioni marginali.
- Provoca distorsioni nelle relazioni tra le variabili non utilizzate nel modello.
- E' necessario mettere a punto un modello diverso per ogni variabile sulla quale si intende effettuare imputazioni.
- Nel caso in cui si applica il metodo suddividendo in classi le unità, è necessario stimare molti modelli diversi tra loro, tanti quante sono le classi di imputazione.
- Può richiedere il possesso di conoscenze tecniche molto specifiche per la messa a punto di modelli appropriati.
- Metodo parametrico, richiede assunzioni sulle distribuzioni delle variabili.
- C'è il rischio che possano essere imputati valori non reali.
- È fortemente influenzato dalla presenza di dati anomali.

d) Imputazione con matching della media (Predictive mean matching)

Metodo

In questo metodo, il valore da imputare ottenuto mediante il metodo dell'imputazione con regressione è messo a confronto con i valori di y disponibili per i casi completi. Il valore di y che viene imputato è quello fornito dall'unità che presenta il valore più vicino al valore di riferimento ottenuto dalla regressione.

L'uso di records "donatori" rende questo metodo simile alle procedure di hot-deck anche se, rispetto a queste, consente l'utilizzo di un numero maggiore di variabili ausiliarie sia qualitative che quantitative.

Anche questo metodo può essere applicato su classi di imputazione distinte.

Campi di applicazione

Adatto per imputazioni su variabili quantitative ed indagini su larga scala.

Vantaggi

- Si può fare uso di un numero elevato di variabili, sia quantitative che qualitative, in modo da ridurre, più che con altri metodi, la distorsioni generate dalle mancate risposte.
- I valori imputati sono "reali".

Svantaggi

- Introduce distorsioni nella distribuzione della variabile (sebbene meno del metodo d'imputazione con regressione).
- Non preserva sufficientemente la variabilità delle distribuzioni marginali.
- Provoca distorsioni nelle relazioni tra le variabili non utilizzate nel modello.
- E' necessario mettere a punto un modello diverso per ogni variabile sulla quale si intende effettuare imputazioni.
- Nel caso si utilizzino classi di imputazione, deve essere stimato un modello diverso per ogni classe.
- Può richiedere il possesso di conoscenze tecniche molto specifiche per la messa a punto di modelli appropriati.

e) Imputazione dal più vicino donatore (Nearest-neighbour imputation)

Metodo

In queste tecniche si sostituisce ogni dato mancante con il valore del rispondente "più vicino". Quest'ultimo è determinato per mezzo di una funzione di distanza applicata alle variabili ausiliarie.

La procedura è la seguente:

1. Calcolare la distanza (considerando i valori assunti sulle variabili ausiliarie, poiché in genere i dati vengono stratificati) tra l'unità del campione con mancata risposta e tutte le altre unità senza dati mancanti usando un'appropriata *funzione di distanza*.
2. Determinare l'unità più vicina all'unità di interesse.
3. Utilizzare il valore dell'unità "più vicina" per effettuare l'imputazione.

Quando si usa una sola variabile ausiliaria si può ordinare il campione in base ai valori da essa assunti; in questo caso ogni donatore è selezionato calcolando la più piccola differenza

assoluta tra non rispondente ed altre unità. Quando, invece, sono disponibili molte variabili ausiliarie possono essere trasformate tutte nei loro ranghi.

Le varianti di questo metodo possono essere ricondotte all'uso di differenti funzioni di distanza. Le funzioni generalmente usate sono:

a) La *distanza Euclidea*.
b) La *distanza ponderata*, nella quale le variabili utilizzate nella funzione sono premoltiplicate per un peso rappresentativo della loro maggiore o minore importanza.
c) La *distanza di Mahalanobis*.
d) La *distanza Minmax*.

A seconda dell'utilizzo che viene fatto dei donatori selezionati, si possono distinguere *due versioni* del metodo:

- Ogni donatore viene usato per ogni valore mancante nel recipiente;
- Uno stesso donatore viene usato per tutti i valori mancanti nel recipiente.

Infine è possibile tenere sotto controllo l'*uso multiplo dei donatori* definendo la funzione di distanza come: $D(1+pd)$, dove D è la distanza di base, d è il numero di volte in cui il donatore è già stato utilizzato e p è una *penalty*, ossia un indicatore che viene incrementato ad ogni nuovo uso dello stesso donatore (College, Johnson, Pare, Sande 1978).

Campi di applicazione

Il metodo è particolarmente adatto nel caso di: indagini dove la percentuale delle mancate risposte è esigua (si limita l'uso multiplo dei donatori); indagini su larga scala in cui trovare un donatore per molte variabili simultaneamente sia più agevole, con notevoli vantaggi in termini di qualità dei risultati; indagini con informazioni di carattere quantitativo utilizzabili nelle funzioni di distanza; indagini in cui esistano relazioni fra variabili difficilmente esplicabili mediante "modelli" (statistici, economici etc.) e sia al contempo necessario preservare la variabilità delle distribuzioni marginali e congiunte.

Si sconsiglia, invece, l'utilizzo del metodo nel caso di: indagini con un numero elevato di mancate risposte (specialmente se una stessa risposta risulta mancante per una grossa percentuale di casi); indagini nelle quali si hanno solo informazioni di carattere quantitativo; indagini di piccole dimensioni.

Vantaggi

- Garantisce, in buona misura, il mantenimento delle relazioni tra variabili anche all'interno di dataset complessi, specialmente nei casi in cui uno stesso donatore è utilizzato per predire simultaneamente molte mancate risposte.

- Potenzialmente è in grado di gestire simultaneamente le informazioni relative ad un numero elevato di variabili.

Svantaggi

- Può provocare distorsioni di varia entità nella distribuzione delle variabili, sebbene i valori imputati includano una parte "residuale" implicitamente osservata nei donatori. In tal senso, la qualità delle imputazioni dipende dalla "ricchezza" del serbatoio dei donatori.
- Richiede una preparazione dei dati tale da assicurare che le variabili non abbiano effetti diseguali sulle misure di distanza.

f) Reti neurali (Neural networks)

Metodo

Le reti neurali sono sistemi di elaborazione realizzati per l'apprendimento di strutture complesse di dati, generalmente finalizzati alla ricostruzione di informazioni mancanti. Le reti neurali sono caratterizzate da due fasi principali:

1. *apprendimento* delle relazioni/associazioni fra i fenomeni osservati;
2. *integrazione* di insiemi di dati con valori mancanti, attraverso l'utilizzo delle relazioni apprese.

L'apprendimento/modellizzazione delle relazioni avviene mediante l'uso di strutture, funzioni e algoritmi di diversa complessità. In termini semplificativi, le reti neurali sono costituite da una serie di strati di unità chiamate *neuroni*. Ogni strato elabora le informazioni ricevute e le passa allo strato successivo allo scopo di ottenere, nell'ultimo strato, dei valori di *output*. Nel caso dell'imputazione, i valori di *input* ricevuti dal primo strato sono quelli delle variabili cosiddette ausiliarie, mentre i valori di *output* nell'ultimo strato sono quelli della variabile con valori mancanti e quindi da imputare. Questi ultimi vengono in pratica determinati attraverso modelli di regressione non lineare.

Campi di applicazione

Le reti neurali possono essere applicate con successo nei casi in cui si hanno a disposizione dati sufficientemente completi e rappresentativi sui quali identificare le relazioni tra le variabili (Scavalli, 2002)

Questo metodo è più adatto ad imputazioni di variabili qualitative piuttosto che quantitative.

Vantaggi

Da un punto di vista strettamente teorico, attraverso le reti neurali è possibile creare modelli migliori che con le tecniche statistiche tradizionali, nel caso in cui si elaborano insiemi di dati caratterizzati da relazioni non lineari molto complesse all'interno degli stessi.

Svantaggi

- Le reti neurali risultano difficili da capire e da spiegare.
- Potenzialmente più costose da sviluppare e da mantenere rispetto alle altre tecniche. I costi dipendono dal livello di complessità dell'applicazione che si intende effettuare.
- Possono verificarsi problemi quando il data set completo sul quale la rete identifica la struttura dei dati non fornisce informazioni "esaustive" sulle relazioni tra le variabili. In questi casi, infatti, si attenua "l'abilità" della rete di scoprire i legami all'interno dei dati e di fornire soluzioni.
- Essendo un metodo deterministico provoca distorsioni nella distribuzione della variabile. Ciò avviene soprattutto quando una determinata variabile o un set di variabili tra loro strettamente correlate presentano un elevato numero di valori mancanti.

2.5.4.4 *Elenco dei Metodi d'imputazione <u>Stocastici</u>*

a) Imputazione con donatore casuale (Random donor imputation overall)

Metodo

Con questa tecnica si attribuisce a ciascun non rispondente il valore y dato da un rispondente scelto a caso dal campione totale dei rispondenti. Alcune strategie di selezione del rispondente *donatore* possono dare risultati migliori di altre. Una delle più raccomandate è quella di selezionare i donatori attraverso un campionamento senza ripetizione per evitare che uno stesso donatore venga utilizzato troppe volte, con possibili effetti negativi su variabilità e relazioni.

Anche questo metodo costituisce la forma degenerata stocastica della funzione lineare $y_{mi} = \beta_{r0} + \sum_j \beta_{rj} z_{mij} + e_{mi}$ priva di variabili ausiliarie: $y_{mi} = \bar{y}_r + e_{mi}$, dove il fattore $e_{mi} = y_{rk} - \bar{y}_r$ riduce l'espressione alla seguente forma: $y_{mi} = y_{rk}$.

Campi di applicazione

Si tratta di un metodo molto semplice dal punto di vita metodologico, il cui utilizzo è consigliato solo nei casi in cui il data set è costituito da poche variabili possibilmente tra loro incorrelate e ci sono pochi dati mancanti.

Vantaggi

- Il valore sostituito al posto del dato mancate è un valore "reale".
- Uno stesso donatore viene utilizzato una sola volta (se selezionato senza ripetizione).
- I valori imputati non generano gravi distorsioni nella distribuzione della variabile.

Svantaggi

- Si basa sull'assunzione che tutte le unità abbiano uguale probabilità di risposta e che le unità con dati mancanti presentino caratteristiche simili a quelle con dati completi.
- Può provocare distorsioni nelle relazioni tra le variabili.

b) **Imputazione con donatore casuale all'interno delle classi (Random donor imputation within classes)**

Metodo

Questo metodo costituisce una forma più accurata dell'imputazione con donatore casuale. Infatti, a differenza del metodo visto precedentemente, si procede inizialmente alla creazione di classi di imputazione all'interno delle quali poi si sostituiscono i dati mancanti con quelli disponibili selezionati casualmente all'interno della medesima classe.

Rappresenta, inoltre, l'equivalente stocastico del metodo dell'imputazione della media all'interno delle classi: $y_{mhi} = \bar{y}_{rh} + e_{mhi}$, dove posto $e_{mhi} = y_{rhk} - \bar{y}_{rh}$, l'espressione si riduce alla forma: $y_{mhi} = \bar{y}_{rhk}$.

Come già detto per l'imputazione con donatore casuale, i migliori risultati si ottengono selezionando i donatori mediante un campionamento senza ripetizione all'interno delle classi.

Esistono diverse versioni del metodo a seconda che:

- le imputazioni tengono conto o meno dei vincoli;
- le imputazioni sono di tipo sequenziale (dato un record con più valori mancanti, viene utilizzato un donatore diverso per ogni mancata risposta) o congiunto (dato un record con più valori mancanti, viene utilizzato un solo donatore per integrarne simultaneamente le mancate risposte).

Campi di applicazione

Questo metodo va usato possibilmente nei casi in cui si lavora con data set di grosse dimensioni (in modo di avere molti donatori), ma con relativamente poche variabili (per ridurre l'entità delle distorsioni delle relazioni).

Vantaggi

- Il valore sostituito al posto del dato mancate è un valore "reale".
- In genere il donatore proviene da un'unità "simile", a differenza di quanto accade imputando senza classi di imputazione.
- Uno stesso donatore viene utilizzato una sola volta, con maggiore preservazione della variabilità delle distribuzioni marginali (se il selezionamento avviene senza ripetizione).
- Se uno stesso donatore è usato per imputare tutte le mancate risposte parziali di un record, vengono preservate le relazioni fra le variabili.

- Maggiore è il numero di classi, maggiori sono le possibilità di imputare un valore da un'unità vicina (come accade nell'hot-deck sequenziale).

Svantaggi

- Per ottenere un'imputazione da casi vicini è necessario un numero molto elevato di classi di imputazione, che comporta la messa a punto di complicate strategie di stratificazione.
- Possibile perdita di dettaglio nella formazione delle classi di imputazione dovuta alla eventuale conversione di dati continui in gruppi discreti di dati (quando si usano variabili continue per stratificare).

c) Imputazione tramite hot-deck sequenziale (Sequential hot-deck imputation)

Metodo

Il primo passo consiste nella determinazione delle classi di imputazione e nell'assegnazione, per ogni classe, di un valore per la variabile di studio y in modo da fissare un punto di partenza per la procedura. Questi "punti di partenza" possono essere ottenuti assegnando, ad esempio, un donatore selezionato (meglio se casualmente) tra i rispondenti di ciascuna classe, oppure un valore considerato rappresentativo della classe come il valor medio della classe ricavato da un precedente sviluppo dell'indagine. I records (unità) del data set sono trattati in maniera sequenziale. Se il record presenta una mancata risposta, questa viene sostituita dal valore iniziale assegnato alla sua classe di appartenenza. Se invece il record dispone della risposta, questa sostituisce il valore precedentemente memorizzato per la sua classe di imputazione.

Campi di applicazione

Di solito questo metodo è utilizzato in indagini su larga scala (ad esempio i censimenti) dove il numero di casi che presentano dati mancanti può essere anche molto elevato. Rispetto all'hot-deck gerarchico e alle tecniche di regressione, questo metodo fa uso solo di un numero limitato di variabili ausiliarie, così da risultare più appropriato nei casi in cui c'è un numero limitato di quesiti.

Vantaggi

- Il maggior pregio di questo metodo è rappresentato dalla sua efficienza computazionale (le imputazioni vengono eseguite effettuando una sola lettura del file di dati).

- Si tratta di una procedura relativamente semplice da comprendere.
- Il valore sostituito al posto del dato mancate è un valore "reale".

Svantaggi

- Programmare la procedura di imputazione può risultare complicato e dispendioso in termini di tempo.
- Problemi connessi all'utilizzo ripetuto di uno stesso donatore. Se all'interno di una classe di imputazione ad un record con una mancata risposta ne seguono altri, tutti con mancate risposte, a questi viene assegnato l'ultimo valore memorizzato, quello cioè relativo all'ultimo rispondente incontrato nella classe. Questo ha conseguenze negative sulla variabilità delle distribuzioni marginali.
- Se il numero delle classi di imputazione è elevato i valori da utilizzare come donatori potrebbero risultare non appropriati.
- Possibile perdita di dettaglio nella formazione delle classi di imputazione dovuta alla eventuale conversione di dati continui in gruppi discreti di dati (quando si usano variabili continue per stratificare).
- Non preserva le relazioni fra variabili, eccetto quelle usate per la stratificazione.

d) **Imputazione tramite hot-deck gerarchico (Hierarchical hot-deck imputation o Flexible matching imputation)**

Metodo

Metodo simile all'imputazione tramite hot-deck sequenziale. In questo caso tutte le unità sono raggruppate in un gran numero di classi di imputazione, costruite sulla base di set dettagliati di variabili ausiliarie. Quindi, rispondenti e non, sono collegati secondo una base gerarchica, nel senso che se non è possibile trovare un donatore adatto nell'iniziale classe di imputazione le classi sono collassate per consentire il *matching* ad un livello più basso.

Campi di applicazione

Metodo utilizzato nei censimenti e nelle indagini su larga scala, dove i data set sono abbastanza ampi da assicurare che la maggior parte delle classi d'imputazione contengano un numero sufficiente di valori da utilizzare come donatori.

Vantaggi

- Il valore sostituito al posto del dato mancate è un valore "reale".

- Le imputazioni possono essere effettuate da unità con le stesse caratteristiche (variabili ausiliarie usate per costruire le classi).
- Permette di considerare molte variabili ausiliarie senza che ciò comporti un complicato lavoro di modellizzazione come nel caso dell'utilizzo di tecniche d'imputazione mediante regressione (sebbene l'imputazione tramite regressione consente l'uso di un numero ancora maggiore di variabili ausiliarie).

Svantaggi

- Dal punto di vista computazionale risulta meno efficiente dell'hot-deck sequenziale.
- Possibile perdita di dettaglio nella formazione delle classi di imputazione dovuta alla conversione di dati continui in gruppi discreti di dati (quando si usano variabili continue per stratificare).
- Preserva le relazioni delle sole variabili usate per costruire le classi.
- Le imputazioni sono meno efficienti nel caso di donatori ricercati in classi "collassate".

e) Imputazione con regressione casuale (Random regression imputation)

Metodo

Questa tecnica costituisce la versione stocastica dell'imputazione con regressione esposta in precedenza, in cui i valori imputati sono sempre stimati con l'equazione di regressione nella quale si aggiunge, però, la componente residuale e_{mi}. In questo tipo di modello sono cruciali le assunzioni per la determinazione dei termini residui e_{mi}.

A tale proposito sono state proposte le seguenti soluzioni:

1) ipotizzare che i residui abbiano una distribuzione normale e rispettino il requisito di omoschedasticità e sceglierli, a caso, dalla distribuzione con media zero e varianza uguale a quella residua della regressione;
2) ipotizzare che i residui provengano dalla stessa distribuzione non specificata dei rispondenti e selezionarli casualmente dai residui di questi ultimi;
3) infine, se si hanno dubbi sulla linearità e sull'additività delle componenti del modello di regressione, si scelgono da quei rispondenti con valori simili nelle variabili ausiliarie.

Campi di applicazione

Vedi metodo di "Imputazione con regressione".

Vantaggi

- Si può fare uso di un numero elevato di variabili, sia quantitative che qualitative, in modo da ridurre, più che con altri metodi, le distorsioni generate dalle mancate risposte.
- I valori imputati non generano distorsioni nella distribuzione della variabile.
- Rispetto alla versione deterministica del metodo ("Imputazione con regressione") preserva meglio la variabilità della distribuzione.

Svantaggi

- Provoca distorsioni nelle relazioni tra le variabili non utilizzate nel modello.
- E' necessario mettere a punto un modello diverso per ogni variabile sulla quale si intende effettuare imputazioni.
- Nel caso si utilizzino classi di imputazione, deve essere stimato un modello diverso per ogni classe.
- Può richiedere il possesso di conoscenze tecniche molto specifiche per la messa a punto di modelli appropriati.
- C'è il rischio che possano essere imputati valori non reali.
- È fortemente influenzato dalla presenza di dati anomali.

2.6 Imputazione Singola…

L'imputazione singola è semplice da utilizzare dato che si tratta di un unico insieme di valori "plausibili" da imputare per ogni dato mancante. Tuttavia, ci sono alcuni svantaggi: l'imputazione singola non riflette infatti l'incertezza supplementare e non rileva la variazione dovuta ai dati mancanti; in altre parole le analisi che vengono effettuate sui dataset completati con il metodo d'imputazione singola, non tengono conto della mancata risposta come sorgente di incertezza, tendendo a considerare tutti i dati (anche quelli imputati) come se fossero stati effettivamente osservati. Inoltre le distribuzioni delle variabili di studio sono compresse e le relazioni tra le variabili possono essere distorte (Kalton, 1983; Lessler e Kalsbeek, 1992; Little e Rubin, 2002).
Tra i metodi elencati precedentemente, quelli che utilizzano l'imputazione singola sono:

Imputazione con media.
Imputazione con media all'interno delle classi.
Imputazione hot-deck sequenziale.
Imputazione tramite hot-deck gerarchico.
Imputazione con matching della media.
Imputazione dal più vicino donatore.
Imputazione con regressione.

Si può affermare che un metodo d'imputazione singola può venire trasformato in imputazione ripetuta, e più precisamente in imputazione frazionale; quando si assegnano M (M > 1) valori per ogni dato mancante, ripetendo un metodo d'imputazione singola più volte casualmente ("random"), si parla di imputazione frazionale (*fractional imputation*). Esempi di imputazione frazionale sono l'uso di un'imputazione ripetuta casualmente tramite hot-deck (*repeated random hot deck imputation*) e l'imputazione ripetuta con matching della media (*repeated predictive mean matching imputation*). Il vantaggio principale dell'imputazione frazionale è la riduzione della componente casuale della varianza dello stimatore derivante dall'imputazione, migliorandone l'efficienza. L'imputazione frazionale considera lo stimatore derivante come uno stimatore pesato con i pesi frazionali 1/M, per ogni valore imputato.
Per far fronte al problema della sottostima della varianza, dovuta al fatto che, nei metodi d'imputazione singola e pure in quello d'imputazione frazionale, si considerano i valori imputati come noti non riflettendo così la (eventuale) variabilità campionaria sotto il modello di non risposta, si opta per l'imputazione multipla.

2.7 …ed Imputazione Multipla

L'idea di base dell'imputazione multipla si può riassumere nei seguenti passaggi:

1. imputare i valori mancanti usando un appropriato modello d'imputazione che incorpori l'imputazione casuale (es. data augmentation), ripetendo l'operazione M volte;
2. produrre l'analisi d'interesse per ognuno degli M dataset risultanti utilizzando le analisi statistiche standard;
3. combinare le stime degli M dataset seguendo le regole di Rubin (Rubin, 1987) per produrre risultati inferenziali.

L'*imputazione multipla* è essenzialmente un metodo Monte Carlo che consente di effettuare un'ampia classe di analisi inferenziali in presenza di non-risposta mediante analisi standard su diversi data-set completi. In questo paragrafo descriveremo brevemente i principi su cui si fonda il metodo; una descrizione dettagliata si può trovare nel testo di Rubin (1987).

Attraverso l'imputazione multipla, ai valori mancanti $\boldsymbol{Y}_{mis}$ sono associati un numero prefissato M di insiemi di valori artificiali $\boldsymbol{Y}^{(1)}_{mis}$, $\boldsymbol{Y}^{(2)}_{mis}$, …, $\boldsymbol{Y}^{(M)}_{mis}$ così da ottenere M data-set completi che possano essere analizzati con metodi standard per dati completi. Il numero di imputazioni multiple è considerato soddisfacente in molte applicazioni quando si attesta tra 3 e 10 (Durrant, 2002). La variabilità dei risultati che si ottengono con le M analisi indipendenti effettuate, riflette l'incertezza associata alla mancata risposta e, combinata con la (eventuale) componente di variabilità di origine campionaria, può fornire una misura complessiva di incertezza nelle inferenze sui parametri di interesse.

Da un punto di vista Bayesiano l'imputazione multipla può essere vista come una procedura di generazione di M realizzazioni indipendenti dalla distribuzione predittiva a posteriori dei dati mancanti condizionatamente ai dati osservati: $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs})$. Le procedure più comuni sono basate sull'assunzione di un modello parametrico esplicito per i dati completi e di un'opportuna distribuzione a priori (generalmente scelta non informativa). In questo contesto, la distribuzione predittiva a posteriori dei dati mancanti può essere espressa come:

$$P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}) = \int P(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs}, \theta) P(\theta \mid \boldsymbol{Y}_{obs}) d\theta$$

cioè come la media della distribuzione predittiva condizionata $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \theta)$ di $\boldsymbol{Y}_{mis}$ dati i parametri θ, rispetto alla distribuzione a posteriori dei parametri "a dati osservati" $P(\theta \mid \boldsymbol{Y}_{obs})$. Per poter generare dalla distribuzione $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs})$ sarebbe pertanto sufficiente essere in grado di generare da ciascuna delle due distribuzioni di probabilità che compaiono nella formula

sotto il segno di integrale. Se la generazione dei dati mancanti $\boldsymbol{Y}_{mis}$ dalla $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \theta)$ non presenta particolari difficoltà, la distribuzione a posteriori a dati osservati $P(\theta \mid \boldsymbol{Y}_{obs})$ è in generale intrattabile. Si è soliti pertanto ricorrere a tecniche di tipo MCMC (Markov Chain Monte Carlo) che consentono di ricondurre il problema della generazione dalla distribuzione $P(\theta|\boldsymbol{Y}_{obs})$ a quello più semplice della generazione dalla distribuzione "a dati completi" $P(\theta|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$. Uno dei procedimenti più comuni consiste nel seguente schema iterativo (Schafer, 1997):

- dato un insieme di valori correnti $\theta^{(t)}$ per i parametri, generare dalla distribuzione predittiva condizionata $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \theta^{(t)})$ un insieme di valori dei dati mancanti $\boldsymbol{Y}_{mis}^{(t+1)}$ (***I-Step***);
- condizionatamente a $\boldsymbol{Y}_{mis}^{(t+1)}$, generare dalla distribuzioni a posteriori a dati completi $P(\theta \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(t+1)})$ un nuovo insieme di parametri $\theta^{(t+1)}$ (***P-step***).

Lo schema descritto è noto in letteratura come *Data Augmentation* (Schafer,1997). Esso fornisce, a partire da un insieme di parametri iniziali $\theta^{(0)}$, una successione aleatoria $\{\theta^{(t)}, \boldsymbol{Y}_{mis}^{(t)}\}_{t=1,2,\ldots}$ la cui distribuzione stazionaria è $P(\boldsymbol{Y}_{mis}, \theta| \boldsymbol{Y}_{obs})$ (Tanner e Wong, 1987). In particolare le sottosuccessioni $\{\theta^{(t)}\}_{t=1,2,\ldots}$ e $\{\boldsymbol{Y}_{mis}^{(t)}\}_{t=1,2,\ldots}$ hanno come distribuzioni stazionarie $P(\theta \mid \boldsymbol{Y}_{obs})$ e $P(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs})$ rispettivamente.

Come ultima osservazione sul metodo della *data augmentation* è utile sottolineare che, affinché si possa considerare ogni insieme di valori $\boldsymbol{Y}_{mis}^{(t)}$ come generato dalla distribuzione predittiva $P(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs})$ è necessario che t sia sufficientemente grande da garantire la stazionarietà; per la generazione di data-set multipli, si è soliti pertanto considerare solo valori di t maggiori di un valore prefissato t_0 (*burn-in period*) che si ritiene abbastanza elevato. Inoltre la richiesta di indipendenza per le imputazioni multiple effettuate suggerisce di non utilizzare data-set ottenuti da iterazioni consecutive dello schema descritto. Al contrario, è conveniente sottocampionare dalla successione $\{\theta^{(t)}, \boldsymbol{Y}_{mis}^{(t)}\}_{t= t_0, t_0+1,\ldots}$ estraendo da essa a passo costante K, dove K è abbastanza grande da poter considerare trascurabile la dipendenza tra gli insiemi estratti.

L'aspetto positivo dell'imputazione multipla è che le differenze ottenute negli M risultati, dagli M dataset completi, possono essere usate come misura d'incertezza dovuta ai dati mancanti.

Una volta ottenuto un insieme di M dataset completi, è semplice effettuare inferenze su un parametro di interesse G della popolazione combinando opportunamente i risultati delle M

analisi effettuate sui singoli dataset completati. Si supponga che $\hat{G} = \hat{G}(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ sia una stima puntuale a dati completi per G, cioè la stima che si otterrebbe se tutti i dati fossero osservati. Sia inoltre $U = U(Y_{obs}, Y_{mis})$ la varianza associata alla stima $\hat{G}$ e si supponga che possa considerarsi valida l'approssimazione:

$$\hat{G}(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) \approx E(G \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$$
$$U(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) \approx V(G \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$$

con $E(G \mid \bullet)$ e $V(G \mid \bullet)$ rispettivamente media e varianza a posteriori di G. Per il t-esimo dataset completato potranno essere calcolate le quantità $\hat{G}^{(t)} = \hat{G}(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(t)})$ e $U^{(t)} = U(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(t)})$. In accordo con le formule di Rubin (Rubin, 1987; Heitjan e Rubin, 1990; Schafer, 1997; Little and Rubin, 2002), per ottenere la stima puntuale di G basata sull'imputazione multipla viene presa la media delle stime puntuali dei dati completi:

$$\overline{G} = \frac{1}{M}\sum_{t=1}^{M} \hat{G}^{(t)}$$

e la varianza associata può essere calcolata come:

$$T = \overline{U} + (1 + \frac{1}{M})B$$

dove $\overline{U} = \frac{1}{M}\sum_{t=1}^{M} U^{(t)}$ è la *varianza all'interno dell'imputazione* (*within*) e $B = \frac{1}{M-1}\sum_{t=1}^{M} (\hat{G}^{(t)} - \overline{G})^2$ è la stima della varianza delle stime puntuali sui dati completi, definita come *varianza tra l'imputazione* (*between*). Il termine $\overline{U}$ approssima la stima della varianza standard per dati completi, mentre $(1 + \frac{1}{M})B$ riflette la correzione necessaria a catturare la crescente variabilità dovuta all'imputazione e alla non risposta.

Un modo semplice di definire un metodo di imputazione multipla è constatare che, sotto il metodo in questione, la formula sulla stima della varianza T sia veramente una valida formula, ricercando uno stimatore della varianza approssimativamente non distorto.

È utile riportare ora alcune considerazioni sulla definizione e sulla scelta del modello d'imputazione e sul rapporto col modello d'analisi. Generalmente, il modello d'imputazione dovrebbe essere scelto in linea con la successiva modalità di analisi, eseguita sull'insieme di dati osservati e imputati, come per esempio l'analisi di regressione. Il modello dovrà sicuramente preservare le associazioni e le relazioni tra le variabili che saranno di fondamentale importanza nell'analisi che verrà eseguita in seguito. Per esempio, variabili

esplicative ed interazioni che verranno incluse nel modello d'analisi sui dati completi, dovranno essere incluse nel modello d'imputazione (Schafer, 1997; Sinharay, Stern e Russel, 2001; Schafer e Olsen, 1998).

L'imputazione multipla ha il vantaggio di offrire una formula per la stima della varianza relativamente semplice e flessibile, nel senso che è applicabile, come principio, a qualsiasi tipo di stimatore imputato. L'imputazione multipla può anche essere utilizzata per completare valori mancanti in dataset multivariati di dati mancanti, ed è adatta sia per variabili numeriche che categoriali. In pratica esistono differenti metodi per realizzare un'imputazione multipla, alcuni dei quali non sono necessariamente diretti. Come si è visto in precedenza *MCMC* (*Markov Chain Monte Carlo*), e soprattutto algoritmi che utilizzano il *Data Augmentation*, possono essere usati per generare simulazioni su dati mancanti. In questo senso l'imputazione multipla è un approccio *MCMC* per dataset con dati incompleti (Rubin, 1996; Schafer, 1997; Lipsitz, Zhao and Molenberghs, 1998). Tuttavia un approccio di questo tipo è completamente parametrico e richiede forti assunzioni sulle distribuzioni sottostanti, quali la normalità multivariata che potrebbe non essere adeguata in alcune applicazioni. Potrebbe inoltre essere computazionalmente costoso e la convergenza potrebbe essere difficile da determinare (Horton and Lipsitz, 2001). Se quindi gli approcci standard d'imputazione multipla sono basati su assunzioni sulla distribuzione dei dati (come la normalità) ed inoltre sono parametrici, è necessario a volte ricercare delle vie alternative, puntando sui metodi d'imputazione semiparametrici o non parametrici. Infatti ci sono molte applicazioni, specialmente nelle scienze sociali, in cui approcci prettamente parametrici potrebbero non essere eseguibili. In circostanze in cui è probabile che le assunzioni distribuzionali non tengano (per es. l'assunzione di normalità è possibile che venga violata nelle variabili concernenti il reddito) è importante porre l'attenzione su metodi d'imputazione semiparametrici o non parametrici, che richiedono assunzioni distribuzionali più deboli sulla variabile da imputare o proprio non ne richiedono. Un modo per realizzare questo tipo d'imputazione multipla è utilizzare il metodo *ABB* (*Approximate Bayesian Bootstrap*) che viene considerato un approccio non parametrico per l'imputazione multipla (Rubin and Schenker, 1986). Si supponga che il campione originale di dati contenga delle classi d'imputazione definite, per esempio secondo variabili categoriali per cui tutti i valori sono osservati. Per ogni set d'imputazione i donatori all'interno di ciascuna classe d'imputazione vengono campionati (*bootstrapped*) con reinserimento basato sulla stessa proporzione con cui i rispondenti sono disponibili in ogni classe. Per ciascun non rispondente, in ogni classe un

donatore è selezionato con reinserimento dal set di rispondenti campionati casualmente per quella classe. Il metodo è ripetuto M volte.

2.8 Confronto tra approccio singolo ed approccio multiplo

Gli approcci d'imputazione non multipla sono stati definiti come sorpassati o convenzionali ed il loro uso è stato scoraggiato (Schafer and Graham, 2002; Allison, 2001). I metodi d'imputazione multipla invece sono stati promossi e vengono descritti come metodi moderni. È enfatizzato in letteratura che l'imputazione multipla non cerca di stimare ogni valore mancante attraverso valori forzati, ma piuttosto cerca di rappresentare un campione casuale di dati mancanti. Questo processo si risolve in valide inferenze statistiche che opportunamente riflettono l'incertezza dovuta ai valori mancanti e all'imputazione. L'imputazione multipla viene quindi considerata superiore ai metodi d'imputazione singola e a quelli cosiddetti d'imputazione multipla impropri, inclusi i metodi frazionali (Sinharay, Stern and Russell, 2001). Un altro vantaggio dell'imputazione multipla riguarda la possibilità di produrre file di micro-dati che possono essere utilizzati per diverse analisi successive. Questo è particolarmente utile quando la creazione di un dataset di uso pubblico può essere analizzato da un gran numero di ricercatori con differenti tipi di analisi in mente.

Nonostante sia raccomandato come metodo generale, l'imputazione multipla non può essere utilizzata senza le dovute considerazioni su assunzioni sottostanti i dati e sul modello scelto per una specifica applicazione. Fay (1996) raccomanda pure l'utilizzo di uno studio simulativo per verificare le performance dell'imputazione multipla prima di applicare tale metodo allo specifico problema.

2.9 Trattazione del problema dei dati mancanti nello studio al Children's Hospital dell'Università di Oulu - Finlandia

Negli studi medici è comune avere dati mancanti. Buchi nei dati possono comportare, come si è visto, stime distorte e interpretazioni fuorvianti dell'insieme dei dati. Anche nello studio genetico-clinico sulla sindrome di difficoltà respiratoria (*Respiratory Distress Syndrome*) nei nati pretermine condotto dal Dipartimento di Pediatria dell'Università di Oulu - Finlandia, si è dovuto affrontare tale inconveniente. Dopo la creazione di un database unico, il secondo passaggio del lavoro al Children's Hospital dell'Università di Oulu – Finlandia è consistito nel trattamento dei valori mancanti per alcune variabili cliniche ed individuali. Il database

definitivo, ottenuto dall'unione di più fonti provenienti dai 3 punti nascita principali della Finlandia centro-settentrionale (Oulu, Tampere e Seinajoki), presentava due macro-sezioni: la prima concernente le informazioni genetiche dei pazienti, la seconda inerente alle variabili cliniche degli stessi. I valori mancanti riferiti alle variabili genetiche sono stati impossibili da trattare. Dato che riguardano variazioni geniche (come SP-A e SP-B) o geni (SP-A1, SP-A2) provenienti da analisi effettuate in laboratorio e dunque dal DNA dei pazienti non è compito dello statistico o dell'epidemiologo cercare di recuperare informazioni in questa materia. Per un approfondimento su tecniche di recupero nelle espressioni geniche si rimanda ai lavori di Troyanskaya et al., 2001, Kim et al., 2006, Xian et al., 2006. L'analisi dei *missing value* trattata in questo capitolo ha dunque inglobato solamente i dati clinici dei pazienti. Come era stato anticipato, la tipologia di valori mancanti nel database creato al Dipartimento di Pediatria dell'Università di Oulu consiste esclusivamente in "non risposte parziali". Per il recupero di informazioni, quali la data di nascita, il sesso e l'età gestazionale in settimane, si sono utilizzati metodi d'imputazione deduttivi.

- Data di nascita: in alcuni casi non si disponeva dell'informazione, ma poteva essere recuperata, quindi dedotta, dalla variabile "*Social security number*", variabile considerata l'identificativo del paziente e dunque obbligatoriamente sempre presente. Il "*Social security number*" sarebbe un codice individuale per ciascun cittadino finlandese o per ogni straniero che ha ottenuto il permesso di soggiorno, corrispondente al nostro codice fiscale. Si è spiegata nel capitolo precedente la logica seguita dal codice *Social security number*, quindi è facile intuire come l'informazione sulla data di nascita sia recuperabile dalle prime 6 cifre di tale identificativo.
- Sesso del paziente: l'informazione è stata dedotta incrociando due variabili generalmente presenti per ogni record, cioè il nome (maschile o femminile) del paziente ed il codice interno. Quest'ultima variabile è un codice applicato dal personale ostetrico ad ogni nato ed è specifica di ciascun ospedale; anche se complessivamente simili, si possono incontrare piccole differenze tra i nosocomi. La struttura principale e comune a tutti e tre gli ospedali partecipanti allo studio comprende (come nel "*Social security number*") 6 prime cifre indicanti la data di nascita, seguite da un trattino, '-', o una 'A' rispettivamente se il paziente sia nato prima del 1° gennaio 2000 o da questa data compresa in poi. Le ultime 4 cifre invece possono seguire logiche diverse da punto nascita a punto nascita. All'ospedale di Oulu sono rappresentate dalle prime tre lettere del cognome (sostituendo le eventuali vocali Å, Ä e Ö con semplici vocali prive di dieresi o pallino) seguite da un numero indicante

il sesso del nato; "1" per maschio, "2" per femmina. Si deve specificare che nel caso di gemelli dello stesso sesso, tale codice potrebbe incorrere in duplicazioni; infatti si osserva la stessa data di nascita, lo stesso cognome ed un identico genere. In questi casi è stato pensato di cambiare semplicemente le tre lettere del cognome, sostituendole per ciascun gemello con quelle del rispettivo nome. Nei nosocomi di Tampere e Seinajoki, invece, l'ultima cifra del codice, identificativa del sesso del nato, viene specificata con "P" nel caso di maschio ("P" da *poika* che in finlandese significa appunto maschio) e "T" nel caso di femmina ("T" da *tyttö* che in finlandese significa femmina). È semplice quindi intuire come dall'ultima cifra del codice interno sia stato dedotto il sesso del nato, nel caso in cui fosse mancante, controllando anche il nome corrispondente.

- Età gestazionale in settimane: per i record in cui era disponibile l'informazione dell'età gestazionale in giorni si è calcolata la corrispettiva età in settimane, se mancante.

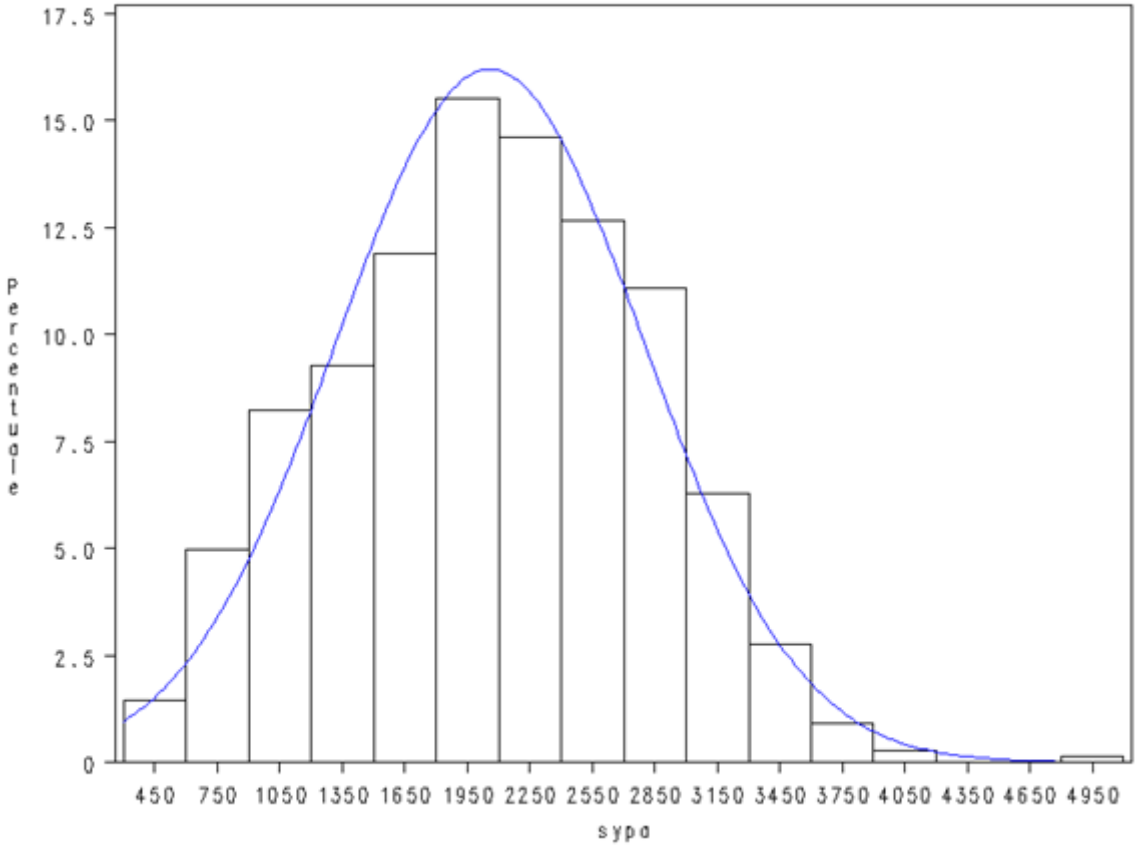
2.9.1 Imputazione multipla per trattare la variabile "peso alla nascita"

Per il recupero di dati mancanti relativi alla variabile peso alla nascita (*sypa* in lingua finlandese) si è ricorsi all'utilizzo dell'imputazione multipla, e più in particolare alla procedura MI del programma statistico Sas System. Il Sas System è un programma che comprende una successione integrata di software per l'erogazione di informazioni di grandi dimensioni e contiene appunto un ampio modulo per le analisi statistiche. La procedura MI è una procedura sperimentale del Sas, lanciata con la versione 8.1. Essa crea data-set d'imputazione multipla per dati multivariati incompleti, utilizzando metodi che incorporano un'appropriata variabilità attraverso le M imputazioni. I metodi d'imputazione disponibili nella procedura MI del Sas sono tre e dipendono dal tipo di modello che seguono i dati mancanti e dalla distribuzione delle variabili:

- per modelli di dati mancanti monotoni con l'assunto di normalità multivariata è appropriato il **metodo di Regressione**;
- per modelli di dati mancanti monotoni senza alcuna ipotesi di normalità, si consiglia il **metodo di *Propensity Score***;
- per modelli arbitrari di valori mancanti, ma con l'assunto di normalità multivariata, è appropriato il **metodo *Markov Chain Monte Carlo*** (**MCMC**).

Una volta che gli M data-set completi sono stati analizzati usando le procedure standard del Sas (nel caso oggetto di studio si ha optato per la procedura REG), la procedura MIANALYZE viene utilizzata per generare valide inferenze statistiche su specifici parametri combinando i risultati ottenuti dalle M analisi.

Figura 2.1 Distribuzione percentuale della variabile peso alla nascita (*sypa*). Grafico ottenuto con il programma Sas System.



Entrando ora nello specifico dello studio si riportano alcuni dati inerenti al database analizzato e per il quale è stato effettuato il processo d'imputazione. Su 1222 casi presenti nel database totale, i valori mancanti rappresentavano l'8,8%, vale a dire che 108 pazienti non riportavano il valore del peso alla nascita. Per imputare i valori mancanti del peso alla nascita si sono utilizzate come variabili esplicative l'età gestazionale in giorni e la misura della lunghezza del nato alla nascita. È stato verificato che i dati relativi al peso alla nascita provenissero da una distribuzione normale attraverso il test di Kolmogorov-Smirnov la cui significatività pari a 0.06 indica appunto questa distribuzione normale rappresentata anche dalla figura 2.1. Per le variabili età gestazionale e lunghezza è stata necessaria una trasformazione logaritmica (vedasi capitolo 4) per ricondursi ad una distribuzione normale dei dati. Altra condizione necessaria per l'applicazione del metodo d'imputazione multipla è che i dati mancanti sulla

variabile peso alla nascita si definiscano "mancanti a caso" MAR (le mancate risposte sul peso alla nascita non dipendono dal valore del peso alla nascita). Si è passati infine alla fase d'imputazione, attraverso la PROC MI.

Come è riassunto nella tabella 2.1 fornita dal Sas System, nel processo d'imputazione multipla è stato scelto il metodo MCMC, eseguibile perché i dati rispettavano l'assunto di normalità multivariata. Le successive specificazioni riguardano:

- la scelta della catena d'imputazione: si ha optato per una catena d'imputazione singola;
- le stime iniziali per MCMC: specificando nella procedura MI del Sas "INITIAL=EM", il programma usa le medie e le deviazioni standard dai casi completi come stime iniziali per l'algoritmo EM (Expectation-Maximization). Le correlazioni sono poste uguali a zero e le stime EM risultanti sono utilizzate per cominciare il processo MCMC;
- informazioni a priori: si è specificato che non si possedevano informazioni a priori per le medie e le covarianze attraverso l'istruzione "PRIOR=Jeffreys";
- il numero di imputazioni: si sono realizzate cinque imputazioni;
- il numero di iterazioni: si è lasciato un numero pari a 100 iterazioni per default;
- il numero di iterazioni Burn-in: 200 iterazioni.

Come si è visto in precedenza, il metodo MCMC utilizza l'inferenza Bayesiana per simulare realizzazioni indipendenti dalla distribuzione a posteriori dei valori mancanti, dati i valori osservati.

Tabella 2.1 Informazioni sul modello nel processo d'imputazione multipla (PROC MI del Programma Sas System).

Informazioni sul modello	
Metodo	MCMC
Catena di imputazione multipla	Catena singola
Stime iniziali per MCMC	EM
Inizio	Valore iniziale
A priori	Jeffreys
Numero di imputazioni	5
Numero di iterazioni Burn-in	200
Numero di iterazioni	100

La tabella 2.2 fornita dal Sas System elenca i distinti modelli di dati mancanti con le frequenze e le percentuali corrispondenti. In questa tabella "X" significa che la variabile è osservata nel gruppo corrispondente e "." significa che la variabile è mancante.

Tabella 2.2 Informazioni sulla distribuzione di frequenza dei dati mancanti relativa alle variabili età gestazionale, lunghezza e peso. "X" significa che la variabile è osservata nel gruppo corrispondente e "." significa che la variabile è mancante. Sas System.

Gruppo	Età gestazionale	Lunghezza	Peso	Frequenza	Percentuale
1	X	X	X	1117	91.3
2	X	X	.	106	8.6
3	.	X	.	2	0.1

La procedura MI crea dunque cinque dataset in ciascuno dei quali sono stati imputati dei valori plausibili al posto dei dati mancanti, il che rappresenta l'incertezza sul giusto valore da imputare. Inoltre viene aggiunta automaticamente dal programma Sas System una variabile, denominata "*Imputation"*, che indica il numero del dataset imputato. Nel caso studiato la variabile "*Imputation*" ha cinque valori (1, 2, 3, 4, 5) dato che le imputazioni sono state cinque.

Una volta che la PROC MI ha prodotto i cinque dataset con i valori imputati, l'uso di metodi e procedure per dati completi (che può essere PROC LOGISTIC come nel caso oggetto di studio, o PROC GLM o PROC PHREG) risulta facile ed intuitivo; l'unico fattore da sottolineare concerne la specificazione di un'istruzione "BY *Imputation*" per ripetere la procedura per ogni valore della variabile "*Imputation*".

Successivamente si possono richiedere delle inferenze statistiche per controllare la significatività delle variabili utilizzate per l'imputazione e si sono seguiti i due seguenti passaggi:

1. attraverso la procedura REG vengono calcolati i coefficienti di regressione per ciascuno dei cinque dataset imputati;
2. infine attraverso la procedura MIANALYZE, vengono combinati i cinque dataset e vengono generate valide inferenze statistiche.

Come si osserva dalla tabella 2.3, vengono riportati la media stimata ed un errore standard per ogni variabile (indicata con "parametro"). Le inferenze sono basate sulle distribuzioni *t*. La tabella riporta un intervallo di confidenza al 95% e un *t*-test con il *p*-value corrispondente per testare l'ipotesi che la media sia uguale al valore specificato con "Theta0", in questo caso zero

per default. Vengono fornite anche le stime parametriche minime e massime ottenute dai dataset imputati.

Tabella 2.3 Stime dei parametri dell'imputazione multipla

Parametro	Stima	Errore std	limiti di confidenza al 95%		DF	Minimo	Massimo	Theta0	t per H0: Parametro= Theta0	Pr > |t|
interc	-4350.8	184.4	-4739.9	-3961.7	16.9	-4480.8	-4210.9	0	-23.60	<.0001
gestp	23.5	1.8	19.8	27.2	23.2	22.2	24.8	0	13.11	<.0001
pituus	27.2	8.0	11.0	43.3	65.7	22.4	31.4	0	3.37	0.0013

Si osserva che i *p*-value per le variabili utilizzate nel modello sono minori di 0.05 e quindi significativi: questo indica che le variabili scelte per l'imputazione dei dati mancanti sul peso alla nascita sono corrette.

2.10 Conclusioni

Nella maggior parte degli studi biomedici ed epidemiologici si affrontano situazioni di dati mancanti parziali per le quali è utile ricorrere a metodi di trattamento basati sull'imputazione. Come si è sottolineato in precedenza, l'imputazione multipla è preferibile a quella singola per il principale motivo che la prima esprime l'incertezza relativa alla scelta del vero valore per i dati mancanti, mentre la seconda, non rispecchiando tale sistema, può portare a stime dei parametri distorte. L'imputazione multipla non cerca di stimare ciascun valore mancante attraverso valori simulati, ma cerca piuttosto di rappresentare un campione casuale di possibili valori mancanti.

Solitamente i programmi statistici offrono la possibilità di utilizzare l'imputazione multipla con alcuni accorgimenti relativi al metodo da scegliere che dipende dal modello con cui i dati mancanti si distribuiscono (monotono e/o con assunzione di normalità). Si è visto il caso d'imputazione multipla con il programma Sas System per il trattamento di dati mancanti della variabile peso alla nascita, attraverso le procedure MI e MIANALYZE. Questa operazione, aggiunta alle tecniche d'imputazione manuale per le variabili data di nascita, sesso ed età gestazionale, ha portato ad una maggior completezza del database da cui ora si potranno estrapolare le informazioni necessarie per studi di casistiche *ad hoc* al fine di indagare specifiche associazioni tra fattori clinici e genetici e patologie polmonari nei nati pretermine.

Sezione B

Disegni di studio

CAPITOLO 3 *Progettazione di uno studio*

3.1 Premessa

Uno studio statistico ha bisogno di una progettazione dettagliata affinché riesca a produrre informazioni utili nel campo di ricerca specifico. La progettazione consiste in un sistema di più tappe, tutte importanti e interdipendenti che conducono alla raccolta dei dati e alla loro elaborazione ideale. Gli elementi di base delle grandi quantità di dati raccolti sono rappresentati dalle unità statistiche che sono l'oggetto d'interesse della ricerca e che detengono l'informazione elementare che si vuole rilevare e analizzare (Armitage & Berry, 1996). In uno studio clinico si prenderà in considerazione una specifica tipologia di unità statistiche, vale a dire le casistiche cliniche. Si tratta di gruppi di individui (pazienti o volontari sani) che prendono parte a studi clinici, medici o genetici, allo scopo di generare conoscenza in campo biomedico. I pazienti che hanno preso parte allo studio presentato nell'applicazione della seconda parte di questo capitolo sono nati pretermine negli ospedali di Oulu, Tampere e Seinäjoki (o nati altrove, ma immediatamente trasferiti in uno di questi ospedali) i cui genitori hanno dato il consenso di adesione alla ricerca. Si tratta quindi di casi che si sono trovati in una particolare situazione e che, "a loro insaputa" sono entrati a far parte del campione di studio. Le informazioni raccolte su questi pazienti, oltre ai dati demografici e clinici individuali, riguardano il loro DNA.

Nel presente capitolo verranno presentate le due principali tipologie di disegni di studio statistici, il disegno di studio sperimentale o *clinical trial* (paragrafo 3.4) e l'indagine o *survey* (paragrafo 3.5), elencandone le caratteristiche e le modalità di programmazione. Soffermandosi poi sulle indagini si affronteranno i disegni di studio di indagini descrittive (paragrafo 3.5.1) ed i disegni di studio di indagini analitiche (paragrafo 3.5.2). Di quest'ultima categoria si presenteranno gli studi di coorte (paragrafi 3.6 e 3.6.1) e quelli caso-controllo (paragrafi 3.7, 3.7.1 e 3.7.2) riportandone anche un confronto relativo ai loro punti di forza e ai loro svantaggi (paragrafo 3.8). Successivamente si tratterà ampiamente la fase di progettazione dello studio al Children's Hospital dell'Università di Oulu (paragrafo 3.10). In principio verrà fornita un'introduzione alla patologia presa in esame nel lavoro, vale a dire la Sindrome di Difficoltà Respiratoria (in inglese *Respiratory Distress Sindrome*, RDS: d'ora in avanti si userà tale abbreviazione) (paragrafo 3.11) e le ipotesi iniziali che fanno supporre ad un'associazione tra RDS e fattori genetici (paragrafo 3.12), in seguito si passerà alla definizione della popolazione statistica (paragrafo 3.15), all'individuazione del disegno di

studio (paragrafo 3.16), alle modalità di raccolta dei dati (paragrafo 3.17), alla definizione di tempi e risorse (paragrafo 3.18) ed alla dimensione statistica dello studio (paragrafo 3.19).

3.2 La statistica ed il suo contributo negli studi programmati

La statistica può essere definita come la disciplina che concerne la trattazione e l'analisi di dati derivanti da gruppi di singole unità statistiche o di osservazioni. A volte per motivi amministrativi, possono essere richiesti dati statistici che si trovano in pubblicazioni ufficiali oppure che vengono forniti da Enti istituzionali di raccolta dati. Spesso, invece, si richiedono studi statistici programmati *ad hoc*. Prima di analizzare la prima grande differenziazione tra studi statistici, si presentano gli scopi principali per cui vengono utilizzati in generale. La statistica fornisce metodi specifici per:

1. **progettare**: pianificare ed effettuare studi di ricerca;
2. **descrivere**: riassumere e analizzare i dati;
3. **fare inferenza**: realizzare previsioni o generalizzare fenomeni specifici rappresentati dai dati.

La **progettazione** si riferisce ai metodi migliori per ottenere i dati richiesti. Gli aspetti del disegno di uno studio possono riguardare, per esempio, la prassi per condurre uno studio, includendo la fase di elaborazione di un questionario e la selezione di un campione di unità statistiche che parteciperanno allo studio. Fabbris (1997) propone una sottoclassificazione ulteriore per il passo di progettazione. Suddivide il momento del disegno in **fase d'astrazione** e **fase di rilevazione e manipolazione dei dati**. In particolare, nella fase d'astrazione solitamente viene richiesto di formulare precise ipotesi di ricerca; definire la popolazione statistica; individuare quali disegni di studio si desiderano progettare e perché si opta per l'uno anziché per l'altro; stilare le modalità di raccolta dei dati rispetto alle variabili d'interesse; definire tali variabili (scala di misura, se sono continue, dicotomiche o discretizzate) e capire come trattarle; prefissare quali saranno le scelte metodologiche per l'analisi dei dati; valutare i tempi e le risorse disponibili.

La fase di rilevazione e manipolazione dei dati riguarda, invece, la scelta di una rilevazione sull'intera popolazione o su un campione di essa; la scelta del procedimento per raccogliere l'informazione (rilevazione diretta o indiretta, attraverso una fonte o più fonti) e la

valutazione del trattamento degli errori di rilevazione, delle mancate risposte e degli eventuali errori campionari e/o extracampionari.

Le fasi di **rappresentazione dei dati** e d'**inferenza** sono i due elementi dell'analisi statistica, rappresentano cioè i modi per analizzare i dati ottenuti come risultato della fase di progettazione (Agresti & Finlay, 1999). Seguendo la distinzione di Fabbris (1997) si riportano la **fase d'analisi** e la **fase di descrizione e presentazione dei risultati** con le rispettive sottospecifiche. La fase d'analisi dei dati concerne la scelta del metodo d'analisi, cioè il procedimento logico ideato per soddisfare un dato obiettivo di analisi dei dati, e la scelta della tecnica d'analisi: una soluzione operativa quasi sempre caratterizzata da un algoritmo informatico, ideata per perseguire un obiettivo analitico. Si ha infine la **fase di descrizione e presentazione dei risultati** che punta a dare pubblica comunicazione dei risultati dell'analisi.

3.3 Studi statistici

È utile ora distinguere sul piano concettuale due diversi tipi di studio statistico:

l'esperimento (o clinical trial) e

l'indagine (o survey).

La sperimentazione richiede un'interferenza pianificata secondo il corso naturale degli eventi, tale da poterne osservare gli effetti. Nella maggior parte dei casi un clinical trial viene adottato in studi atti a testare un nuovo farmaco o una nuova terapia su un particolare gruppo di soggetti, vale a dire pazienti affetti da una specifica patologia. Nell'indagine, invece, il ricercatore è un osservatore meno attivo, che interferisce il meno possibile con i fenomeni da registrare. È facile pensare a esempi estremi che possano illustrare questa antitesi, ma nella pratica la linea di distinzione risulta a volte ardua da tracciare. Si vedano ora più nel dettaglio questi due tipi di studi medico-statistici.

3.4 Clinical Trial

Per clinical trial si intende una qualsiasi forma di esperimento pianificato che coinvolga pazienti e che sia designato a chiarire il trattamento più appropriato per pazienti futuri che hanno una specifica condizione di malattia. Si seleziona un campione di pazienti per fare inferenza sulla scelta del trattamento da utilizzare nell'intera popolazione.

I soggetti in uno studio sperimentale devono essere identificati solo nel caso abbiano la patologia in questione e devono essere ammessi nello studio per tempo, affinché venga

eseguita un'accurata diagnosi che permetta l'assegnazione del trattamento. Soggetti la cui patologia è troppo lieve o troppo seria per permettere la somministrazione del trattamento oggetto di studio o del trattamento alternativo, devono essere esclusi. L'assegnazione del trattamento dovrebbe essere eseguita in modo da minimizzare la variazione di fattori estranei che potrebbero falsificare il confronto (Pocock, 1989; Rothman, 1986).
Un clinical trial si può suddividere in quattro fasi principali che vengono riportate di seguito:

I. Farmacologia e tossicità cliniche: fase preliminare in cui viene testata la sicurezza del trattamento e non l'efficacia. Spesso in questa fase quelli che vengono sottoposti all'esperimento sono dei volontari sani (spesso gli stessi addetti delle case farmaceutiche) e viene stabilito un corretto dosaggio o una corretta modalità di utilizzo di un macchinario o di una terapia.

II. Indagine clinica iniziale per l'effetto del trattamento: si tratta di un piccolo studio di screening, destinato a selezionare provvedimenti terapeutici in grado di fornire le premesse per l'organizzazione della fase successiva. Viene testata l'efficacia del trattamento per cui si richiede uno stretto monitoraggio di ciascun paziente. Si selezionano quei farmaci o trattamenti che effettivamente hanno un autentico potenziale rispetto a tutti quelli che risultano inattivi o addirittura tossici.

III. Valutazione completa del trattamento: in questa fase si concepisce l'esperimento vero e proprio come se si trattasse di uno studio probante di ricerca (Hill & Hill, 1991). Dopo che un trattamento è stato dimostrato essere efficace, è essenziale compararlo con il trattamento fino ad allora standard per la stessa patologia in un trial esteso che coinvolga un numero elevato di pazienti.

IV. Controllo postmarketing: dopo che la ricerca ha condotto alla possibile commercializzazione del farmaco (o il possibile utilizzo del trattamento), devono essere intraprese altre ricerche sul monitoraggio di eventuali effetti avversi. Devono essere condotti studi addizionali di morbilità e mortalità a lungo termine e su larga scala.

Il lavoro statistico riguarda principalmente la terza fase: si tratta di comparare l'esperienza di un gruppo di pazienti a cui viene somministrato il nuovo trattamento, con un gruppo di

controllo di pazienti simili (per età, sesso, stadio di malattia, ecc.) che ricevono il trattamento standard (o placebo). Se non esisteva un trattamento standard, il gruppo di controllo sarà costituito da pazienti non trattati.

Un fattore importante è quello di assegnare casualmente ogni paziente al gruppo di controllo o a quello a cui viene somministrato il nuovo trattamento. Il metodo più affidabile per realizzare una ricerca clinica è denominato *esperimento randomizzato controllato* (randomized controlled trial RCT) (Cochrane, 1972) perché esso elimina tutte le forme di causalità spuria. Gli RCT clinici comportano un'assegnazione casuale dei trattamenti ai pazienti; questo assicura che diversi gruppi trattati siano statisticamente equivalenti.

Scopi e metodi di un RCT clinico vanno descritti in dettaglio in un documento, che viene chiamato *protocollo*, che contiene dettagli medici e amministrativi specifici del problema in esame. Il protocollo deve:

- includere condizioni chiare sul tipo di pazienti da ammettere,
- definire con precisione le misure terapeutiche da adottare,
- stabilire il numero di pazienti, la durata prevista per il loro reclutamento e, se è il caso, per il follow-up.

Si analizzano ora questi specifici punti generalmente presenti nel protocollo.

3.4.1 Definizione dei pazienti

È utile, di solito, affidarsi in parte alla flessibilità, non solo per aumentare il numero dei pazienti (posto che si disponga delle risorse economiche necessarie al loro reclutamento), ma anche perché ciò consente di effettuare separatamente i confronti tra trattamenti per diverse categorie di pazienti. L'ammissione di un vasto spettro di pazienti non impedisce in alcun modo la suddivisione in sottogruppi più omogenei ai fini dell'analisi dei risultati. Tuttavia, è meno facile che confronti basati su piccoli sottogruppi rivelino differenze reali tra effetti dei trattamenti rispetto a test basati sull'intero insieme dei dati. Inoltre, c'è il pericolo che, con troppi confronti in sottogruppi diversi, uno o più test diano risultati significativi solo per effetto del caso. Ogni sottogruppo considerato a priori va quindi definito nel protocollo e preso in considerazione nella pianificazione della numerosità campionaria.

3.4.2 Definizione dei trattamenti

I regimi terapeutici da confrontare sono solitamente ben noti sin dall'inizio, ma spesso non si è in grado di sapere se vadano definiti e standardizzati fin nei minimi dettagli oppure no. In un esperimento multicentrico può essere auspicabile adottare varianti minori dello stesso regime generale o adottare diverse terapie concomitanti. Spesso è meglio far rientrare tali varianti nello studio, in particolare quando ricorrono comunemente nella pratica medica, piuttosto che introdurre un grado di standardizzazione che non può essere accettato su vasta scala né durante l'esperimento, né dopo.

Provando più trattamenti terapeutici, si varia solitamente il programma dettagliato in base alle nuove condizioni del paziente. Per esempio, la dose di un farmaco può dipendere dalla risposta terapeutica o dagli effetti collaterali. Sperimentando su tali trattamenti, si raccomanda di mantenere sempre un certo grado di flessibilità.

3.4.3 Dimensione del campione

Il trial deve reclutare un numero sufficiente di pazienti affinché si ottenga una precisa stima di risposta su ciascun trattamento. L'approccio statistico che può fornire il numero di pazienti di cui si ha bisogno è denominata "previsione di potenza".

Per esiti di tipo qualitativo (per esempio il confronto tra un nuovo farmaco e un placebo):

$$n = \frac{p_1(100 - p_1) + p_2(100 - p_2)}{(p_2 - p_1)^2} f(\alpha, \beta) \qquad [3.1]$$

dove p_1 rappresenta la percentuale di successi attesi dal trattamento standard; p_2 indica la percentuale di successi che si desidera ottenere dal secondo trattamento (quello innovativo); α è il livello del test di significatività χ^2. α è comunemente chiamato errore di I tipo, cioè la probabilità del verificarsi di una differenza significativa tra i trattamenti quando però i trattamenti sono in realtà uguali come efficacia (rappresenta il rischio dei falsi positivi). β, comunemente chiamato errore di II tipo, indica la probabilità di non riscontrare una differenza significativa quando invece è presente una reale differenza di magnitudine (p_1- p_2). β rappresenta il rischio dei falsi negativi. I valori da attribuire alla funzione $f(\alpha, \beta)$ sono espressi nella tabella 3.1.

Tabella 3.1

		β (errore di II tipo)			
		0.05	**0.1**	**0.2**	**0.5**
α (errore di I tipo)	**0.1**	10.8	8.6	6.2	2.7
	0.05	13.0	10.5	7.9	3.8
	0.02	15.8	13.0	10.0	5.4
	0.01	17.8	14.9	11.7	6.6

Per esiti di tipo quantitativo (per esempio la misurazione dell'effetto nella somministrazione di una dose supplementare di vitamina D alle donne in gravidanza per prevenire l'ipocalchemia neonatale) la formula per stimare la numerosità del campione è la seguente:

$$n = \frac{2\delta^2}{(\mu_2 - \mu_1)^2} f(\alpha, \beta) \qquad [3.2]$$

dove μ_1 rappresenta la risposta media attesa; μ_2 rappresenta la risposta media che si vuole ottenere col nuovo trattamento e δ è la deviazione standard di μ_1. α e β vengono scelti con le stesse modalità spiegate precedentemente. Si aggiunge inoltre che la quantità μ_2 - μ_1 indica la differenza nella risposta media che si desidera realizzare con il nuovo trattamento.

3.4.4 Valutazione delle risposte

L'efficacia specifica di ogni trattamento viene valutata paragonando una o più risposte per ogni paziente, a certi intervalli di tempo dall'inizio del trattamento. Capita che le risposte possano essere influenzate dalla conoscenza del trattamento applicato da parte del paziente, del medico o del personale coinvolto nello studio. Se non viene presa l'opportuna precauzione, la distorsione può essere notevole. Possono emergere differenze spurie tra trattamenti che, in realtà, sono ugualmente efficaci o si possono ottenere risultati simili da trattamenti che differiscono di molto nei loro effetti.

In un semplice *esperimento cieco singolo*, l'identità del trattamento assegnato è tenuta nascosta al paziente. Nell'*esperimento in doppio cieco* anche il medico, insieme allo staff tecnico che valuta la risposta, non conosce l'identità del trattamento. Negli esperimenti farmacologici questa forma di mascheramento avviene con formulazioni speciali della sostanza da somministrare. Naturalmente, certi farmaci producono effetti collaterali caratteristici che, di fatto, non possono essere tenuti nascosti né al medico né al paziente

informato della sperimentazione in corso. Se l'esperimento è concepito per valutare l'efficacia di un certo farmaco in assenza di assunzione di altri farmaci, si può verificare una differenza di risposta specifica dovuta all'evidente intervento sperimentale e al conseguente atteggiamento del paziente che si attende un qualche beneficio; un controllo efficace di questa variabilità nelle risposte è spesso dato da un placebo, cioè da un preparato inerte apparentemente indistinguibile da quello del preparato supposto attivo.

3.5 Indagine (*Survey*)

Le indagini statistiche vanno distinte in *indagini descrittive*, programmate per fornire stime di certe caratteristiche semplici della popolazione, e *indagini analitiche*, programmate per studiare associazioni tra alcune variabili. Anche se la distinzione non è sempre così netta e definitiva, nelle indagini descrittive l'attenzione è posta soprattutto sulle modalità per ottenere stime attendibili delle caratteristiche di una popolazione ben definita. Nelle indagini analitiche, invece, il focus è orientato più sulla relazione tra specifiche variabili, piuttosto che sulla definizione della popolazione.

3.5.1 Indagini descrittive

Come implica il nome, studi clinici descrittivi riguardano la descrizione di caratteristiche generali della distribuzione di una patologia, in particolare relativamente a determinati pazienti, luoghi e periodi di tempo (Hennekens & Buring, 1987). Le variabili concernenti i pazienti includono sia fattori demografici di base, quali età, sesso, razza, stato coniugale, occupazione, sia variabili di stile di vita, quali l'alimentazione o l'uso di medicinali. Variabili relative al luogo si riferiscono alla distribuzione geografica della malattia, includendo variazioni tra Stati o all'interno di un singolo Paese, come per esempio tra aree urbane o rurali. Per quanto riguarda il periodo di tempo specifico, gli studi descrittivi possono esaminare modelli stagionali dell'inizio di una malattia oppure confrontare la frequenza di oggi con quella di 5, 10, 50 anni fa. Informazioni su diverse caratteristiche di pazienti, località geografiche e periodo sono disponibili su registri cartacei o su database informatizzati e quindi uno studio descrittivo può essere eseguito in poco tempo e senza impegnare troppe risorse. Dati descrittivi forniscono informazioni preziose per permettere alle istituzioni che forniscono servizi per la salute di allocare efficientemente le risorse, e di pianificare una buona prevenzione o programmi di educazione alla salute pubblica. In aggiunta, gli studi

descrittivi hanno spesso fornito le prime importanti indicazioni sulle possibili cause di una malattia. A causa delle limitazioni nel loro disegno, tuttavia, gli studi descrittivi in campo clinico ed epidemiologico sono principalmente utili per la formulazione di ipotesi iniziali che possono essere testate successivamente attraverso un disegno di studio analitico.

Esistono tre tipi principali di studi descrittivi clinici (Vajani, 1997). Il primo tipo, lo studio correlazionale, utilizza dati da intere popolazioni per confrontare le frequenze di una patologia tra differenti gruppi di pazienti durante lo stesso periodo di tempo o nella stessa popolazione di pazienti, ma in diversi momenti.

La seconda tipologia di studio descrittivo, forse la più diffusa, è il *Case Report*, che consiste in un rapporto dettagliato e accurato da parte di uno o più clinici sul profilo di un singolo paziente.

Il terzo tipo di studio descrittivo clinico è il *cross-sectional survey* nel quale lo stato di un individuo è accertato in un determinato momento controllando la presenza o l'assenza di esposizione a fattori di rischio rispetto alla presenza o assenza di una patologia.

3.5.2 Indagini analitiche

Ogni disegno di studio, implicitamente negli studi descrittivi ed esplicitamente in quelli analitici, comporta dei tipi di confronto tra esposizione e stato di malattia. In un *case report*, per esempio, dove un medico osserva una particolare caratteristica di un singolo caso, viene sicuramente formulata un'ipotesi basata su un implicito confronto con l'esperienza usuale o almeno attesa. Nei disegni di studio analitici il confronto è esplicito, dato che il ricercatore clinico raggruppa pazienti con lo specifico scopo di determinare sistematicamente se il rischio di manifestare una patologia è diverso o uguale tra individui esposti ed individui non esposti a fattori d'interesse.

Lo scopo dell'indagine statistica è quindi quello di stimare una o più caratteristiche (le variabili) della popolazione oggetto di studio, di accertare o meno ipotesi sulle stesse, di esplorare e verificare quali relazioni esistono tra fenomeni d'interesse. Una domanda comunemente posta nelle indagini epidemiologiche sull'eziologia di una malattia è se alcune manifestazioni di malessere si associno a qualche caratteristica o abitudine personale, a particolari aspetti dell'ambiente in cui la persona ha vissuto o a certe esperienze subite. Nello studio clinico-genetico che verrà presentato in seguito la domanda principale a cui si vuole dare una risposta è: *il rischio che un nato pretermine presenti gravi malattie croniche*

cardiopolmonari in periodo neonatale (dovute alla prematurità estrema) è connesso con uno o più polimorfismi di geni specifici?

Le indagini analitiche su popolazioni possono essere esaustive (censimenti) o campionarie. I vantaggi del campionamento risiedono nel risparmio di costi, di personale e di tempo. I vincoli imposti da uno qualsiasi di questi fattori sono sufficienti per consigliare il campionamento anziché la rilevazione esaustiva. Inoltre le ridotte dimensioni delle interviste campionarie permettono di eseguire osservazioni più accurate.

La definizione precisa della popolazione da esaminare non è di primaria importanza, dato che nelle indagini epidemiologiche è di solito amministrativamente impossibile studiare una popolazione nazionale o regionale, anche se solo su base campionaria. Il ricercatore può, tuttavia, essere facilitato nello studio di una specifica popolazione geograficamente collegata ad un particolare centro medico. Inoltre, sebbene la frequenza relativa o il valore medio di diverse variabili possano differire alquanto da una popolazione all'altra, entità e direzione delle associazioni tra variabili difficilmente variano di molto tra popolazioni geograficamente diverse (Armitage & Berry, 1996).

Per quanto riguarda le indagini analitiche eziologiche, i disegni principali sono due: lo studio *per coorti* e lo studio *caso-controllo*.

3.6 Studio per coorti

Nello studio per coorti si studia una popolazione di individui o un campione rappresentativo selezionato generalmente in base a criteri geografici. La popolazione viene classificata secondo uno o più fattori di interesse e seguita prospettivamente nel tempo, in modo da poterne osservare l'incidenza di varie manifestazioni morbose e correlarla alla classificazione per fattori eziologici. Gli studi coorte sono spesso chiamati *prospettici*, ma tale nomenclatura può essere fuorviante, dato che anche uno studio per coorti può basarsi interamente su documenti retrospettivi. Quest'ultimi studi sono talvolta chiamati *studi per coorte retrospettivi* o *studi prospettici storici* in cui l'evento che rappresenta l'entrata del paziente nello studio (sempre che acconsenta) è rappresentato dall'insorgere della malattia e dopo tale situazione si ricostruisce la storia del paziente attraverso dati retrospettivi.

Uno studio prospettico puro consiste invece nell'individuare due gruppi della popolazione obiettivo, il primo sottoposto a fattori di rischio noti, il secondo non esposto agli stessi. Seguendo i pazienti nel tempo si ha la possibilità di osservare l'eventuale manifestarsi di patologie. Nel cosiddetto *follow up di casistica*, invece, l'evento consiste nell'arrivo di un

paziente in un dato centro od ospedale e da quel momento inizia il suo monitoraggio, viene cioè seguita la sua storia clinica nel tempo analizzando i fattori a cui esso si espone e le patologie che eventualmente insorgono o si concludono.
Considerando la natura prospettica degli studi per coorti, si evidenzia che essi normalmente durano più a lungo nel tempo rispetto agli studi caso-controllo, oltre che essere anche più complessi dal punto di vista amministrativo. Il vantaggio è però poter studiare contemporaneamente diverse condizioni mediche e di poter ottenere informazioni dirette sullo stato di salute di ogni soggetto, durante un certo periodo di tempo.

3.6.1 Selezione dei soggetti

Studi di coorte che durano diversi anni presentano problemi logistici che possono influenzarne negativamente la validità. Il punto centrale è trovare i soggetti dello studio. Sia che lo studio sia retrospettivo o prospettico, è spesso difficile individuare le persone o i loro record molti anni dopo l'arruolamento nelle coorti dello studio. Negli studi prospettici potrebbe essere possibile mantenere periodici contatti con i partecipanti allo studio e nel frattempo recuperare informazioni sulla loro posizione e sui loro spostamenti. Questi passaggi nell'inseguire le tracce dei soggetti aumentano i costi dell'indagine, ma sono necessari perché una perdita sostanziale di individui può aumentare seriamente i dubbi sulla validità dello studio. Follow-up che recuperano meno del 60% dei soggetti iniziali sono generalmente guardati con scetticismo, ma anche studi che rintracciano solo il 70 o l'80% dei soggetti possono fornire insufficienti rassicurazioni contro la distorsione dei risultati se c'è ragione di pensare che perdite durante il follow-up possano essere correlate con uno dei fattori d'esposizione o con la malattia (Greenland, 1977).

3.7 Studio caso-controllo

Gli studi *caso-controllo*, spesso chiamati *studi retrospettivi*, forniscono un metodo di ricerca per analizzare fattori che possono prevenire o causare una patologia d'interesse (Schlesselman, 1982). Il metodo confronta un gruppo di individui affetti dalla patologia in esame (i casi) con un gruppo di individui senza tale patologia (i controlli) ed è diventato molto popolare negli anni '20 per lo studio del cancro (Breslow, 1996). Il confronto ha l'obiettivo di individuare fattori che possono differire nei due gruppi e spiegare il verificarsi della patologia nei pazienti. Si ottengono informazioni per ogni gruppo, di solito in modo

retrospettivo, sulla frequenza dei vari fattori personali o ambientali potenzialmente associati alla patologia. È conveniente usare questo tipo di indagine nello studio di condizioni morbose rare che si presenterebbero con troppa poca frequenza in un campione casuale della popolazione. In effetti, partendo da un gruppo di individui malati, si applica una frequenza di campionamento maggiore che nei controlli. Il metodo è appropriato anche quando la classificazione della patologia è semplice, in particolare nel caso di classificazioni dicotomiche di presenza o assenza di condizioni specifiche, ma in cui si devono studiare molti fattori eziologici potenziali. Ulteriore vantaggio del metodo è la relativa rapidità con cui, mediante l'inchiesta retrospettiva, si possono ottenere le informazioni rilevanti.

D'altro canto negli studi caso-controllo la modalità di selezione dei controlli risulta un problema fondamentale. Idealmente, dovrebbero essere in media simili ai casi sotto ogni profilo, tranne che per la condizione patologica in esame e per i fattori eziologici associati. Spesso i casi vengono selezionati da uno o più ospedali e condividono le caratteristiche della popolazione che viene curata nel medesimo ospedale, come condizioni sociali, ambientali e caratteristiche etniche. Normalmente è preferibile selezionare il gruppo di controllo dalla stessa area (o aree), addirittura dagli stessi ospedali, dei casi, purché sofferenti di patologie molto differenti, che difficilmente condividano gli stessi fattori eziologici. Inoltre, le frequenze con cui si osservano i diversi fattori variano con il sesso e l'età. I confronti tra i casi e i controlli devono perciò tener conto di ogni differenza di distribuzione per età e sesso dei due gruppi. Comunemente si evitano tali adattamenti, e si abbina ogni individuo malato a un individuo di controllo, deliberatamente scelto della stessa età e dello stesso sesso, che condivida con lui altre caratteristiche demografiche ritenute di rilevanza analoga.

3.7.1 Base per un disegno di studio caso-controllo

Si immagini una determinata popolazione dinamica di individui esposti e non esposti. I dati principali sull'incidenza di una malattia al tempo t, potrebbero essere riassunti come segue (Rothman, 1986):

$$I_1 = \frac{a}{P_1 t} \quad \text{e} \quad I_0 = \frac{b}{P_0 t}$$

dove I_1 e I_0 sono i tassi d'incidenza rispettivamente tra esposti e non esposti, a e b sono i corrispondenti numeri di individui che hanno sviluppato la malattia durante l'intervallo di tempo t, e P_1 e P_0 sono le rispettive numerosità delle due popolazioni. In uno studio coorte sia il numeratore che il denominatore di ciascun tasso vengono misurati, il che significa che è

necessario enumerare l'intera popolazione e tenerla sotto stretta sorveglianza. Uno studio caso-controllo è un tentativo di rendere le osservazioni rilevate sulla popolazione più efficienti. In uno studio caso-controllo i casi sono gli individui che hanno sviluppato la malattia durante un periodo di tempo, cioè un totale di (a + b). I controlli sono un campione delle coorti combinate che hanno fornito i casi. Se una proporzione delle coorti combinate di esposti e non esposti, k, è presa come controlli, ed il numero di tali controlli è c per gli esposti e d per i non esposti, allora i tassi d'incidenza tra esposti e non esposti possono essere stimati come:

$$I_1 = k\frac{a}{ct} \quad \text{e} \quad I_0 = k\frac{b}{dt}.$$

Se k, la frazione campionaria dei controlli, è nota, allora le stime dell'incidenza della malattia possono essere calcolate semplicemente per entrambi i gruppi di esposti e non esposti. Se invece k non è noto, si può calcolare l'incidenza relativa, o rischio relativo (RR), ottenuto come

$$RR = \frac{I_1}{I_0} = \frac{ad}{bc}.$$

Dal momento che la frazione campionaria k è uguale per gli esposti e i non esposti, essa viene eliminata nella divisione, come anche t. La quantità risultante, ad/bc, è l'odds ratio d'espisizione (rapporto dell'odds di esposizione tra i casi sull'odds di esposizione tra i controlli), spesso chiamato *rapporto crociato* (che chiameremo ψ). L'eliminazione della frazione di campionamento per i controlli nel rapporto crociato fornisce una stima non distorta del rapporto dei tassi d'incidenza da dati provenienti da uno studio caso-controllo (Sheehe, 1962; Miettinen, 1976). Il *rapporto crociato* è sempre positivo. Assume il valore 1 nell'ipotesi nulla di equilibrio nel rischio tra i due gruppi di unità posti a confronto; tende a valori tanto più grandi di 1 quanto più il rischio del gruppo di esposti è superiore al gruppo dei non esposti; è inferiore a 1 se il fattore di rischio è "protettivo" contro il rischio di malattia. Per i rapporti crociati ψ è possibile calcolare anche l'intervallo di confidenza e verificare se sono significativamente diversi da 1. Dati il $ln(\psi)$ e lo standard error di ψ, $SE(\psi) = \sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}$, si calcola l'intervallo di confidenza per $ln(\psi)$ con α pari a 0,05: $ln(\psi) \pm 1{,}96[SE(\psi)]$. Dato che il valore di ψ che indica indipendenza è 1, passando al $ln(\psi)$ per poter asserire che il valore del rapporto crociato d'interesse è significativamente superiore ad 1, al 95%, si dovrà ottenere un intervallo di confidenza che non contenga il valore 0.

La condizione centrale per condurre un valido studio caso-controllo è che i controlli siano selezionati indipendentemente dallo stato di esposizione in modo da garantire che la frazione di campionamento possa essere rimossa nel calcolo del rischio relativo.
Il rapporto crociato è solo una delle misure che può essere dedotta da una specifica tabella, denominata tabella tetracorica, nella quale sono presentate le frequenze del verificarsi congiunto della variabile dipendente *Y* (la patologia) e della variabile esplicativa *X* (il fattore di rischio) (tabella 3.2).

Tabella 3.2 Tabella tetracorica per lo studio della significatività di una variabile esplicativa dicotomica.

Variabile dipendente (Y)

Variabile esplicativa (X)	1 (presente)	0 (assente)	
1 (presente)	**a**	**b**	**a + b**
0 (assente)	**c**	**d**	**c + d**
	a + c	**b + d**	**n**

Oltre al *RR*, le misure della dipendenza tra le due variabili dicotomiche *X* e *Y* che si possono calcolare a partire dai dati della tabella 3.2 sono:

1 il coefficiente di correlazione di Bravais-Pearson: $\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$

il quale, misurando la relazione simmetrica fra Y e X, può essere assunto solo come prima approssimazione della dipendenza di cui si tratta;

2 la sensibilità, ossia la proporzione di soggetti malati correttamente diagnosticati dal test o dal fattore di rischio:
Sen = a / (a + c);

3 la specificità, ossia la proporzione di soggetti sani correttamente diagnosticati:
Spe = d / (b + d) (Fabbris, 1997).

3.7.2 Selezione di casi e controlli

I casi che vengono inclusi in uno studio caso-controllo devono rappresentare la totalità dei casi provenienti da un'ipotetica popolazione che ha prodotto i casi che sono stati selezionati (Rothman, 1986). Un caso non selezionato è quindi presunto provenire da una differente popolazione. I casi devono essere selezionati indipendentemente dall'esposizione al/i fattore/i di rischio. L'obiettivo nella selezione dei controlli è scegliere individui rappresentativi di coloro che sarebbero stati selezionati come casi se avessero sviluppato la malattia, e sicuramente scegliere questi controlli indipendentemente dall'esposizione al/i fattore/i di rischio, come si è osservato anche per i casi. In uno studio caso-controllo non è necessario che si includano tutti i casi che insorgono in una popolazione all'interno di confini geografici definiti, ma è sufficiente che i casi (identificati anche in una singola clinica o da un singolo medico) provengano da una popolazione limitata a quegli individui che sarebbero stati inclusi come casi se avessero sviluppato la malattia. I controlli devono essere selezionati per rappresentare tale popolazione.

La maggior parte degli studi caso-controllo sono basati su gruppi di soggetti constatati solo da specifici fattori di selezione avvenuti fuori dal controllo del ricercatore, quali, per esempio, la presenza regolare i uno o più centri clinici designati. Questi studi sono validi perché permettono che il ricercatore selezioni soggetti indipendentemente dall'esposizione e scelga come controlli coloro che avrebbero seguito lo stesso percorso di selezione forzata dei casi, se fossero stati malati. La sfida in studi di questo tipo è convertire questa definizione concettuale di soggetti di controllo in un protocollo di selezione dei controlli.

3.8 Punti di forza e debolezze dei disegni di studio caso-controllo e per coorti

Mentre gli studi per coorti sono utili nell'analisi di un serie di effetti legati ad un singolo fattore d'esposizione, gli studi caso-controllo possono solo fornire informazioni su un singolo effetto che colpisce i casi selezionati. È possibile, senza dubbio, selezionare molteplici serie di casi con diverse malattie, ma tale approccio comporta l'avvio di più studi casi-controllo simultanei (le differenti serie di casi potrebbero richiedere serie separate di controlli come no, ciò dipende dalle condizioni d'accertamento dei casi). D'altro canto uno studio caso-controllo può convenientemente fornire informazioni su una grande quantità di potenziali fattori eziologici d'esposizione che possono essere associati con una particolare patologia, mentre un tipico studio di coorti si focalizza su un unico fattore di rischio.

Punti di forza e debolezze degli studi per coorti hanno simmetrici svantaggi e vantaggi negli studi caso-controllo. È stata già citata la capacità degli studi caso-controllo di investigare malattie rare, procedimento problematico negli studi coorte. D'altra parte gli studi caso-controllo sono poco efficienti per la valutazione di effetti d'esposizione rari nella popolazione obiettivo dei casi. Un punto importante da tenere sotto controllo negli studi per coorte concerne la scelta dei soggetti in modo da evitare perdite durante il follow-up. Negli studi caso-controllo l'analogo concerne nel determinare i corretti fattori d'esposizione per tutti i soggetti, in modo da evitare la perdita o l'esclusione di soggetti a causa di fattori espositivi sconosciuti. Negli studi per coorti, dato che lo stato d'esposizione viene determinato prima del manifestarsi di una patologia ed è noto sia al soggetto sia al ricercatore, non c'è la possibilità che l'outcome relativo alla malattia influenzi la classificazione dell'esposizione. Negli studi caso-controllo invece, se l'informazione sull'esposizione proviene dal soggetto dopo l'inizio della patologia, la consapevolezza e le conoscenze sulla malattia potrebbero influenzare i dati sui fattori d'esposizione. A questo proposito vengono elencati tre rischi che il ricercatore deve tener presenti:

1. l'effetto memoria, vale a dire che individui malati cercano di ricordare qualsiasi dettaglio possa far pensare loro ad un rischio, mentre i controlli vanno meno in profondità;
2. l'effetto *telescoping*: i soggetti possono selezionare o avvicinare nella serie temporale gli eventi per loro più importanti e trascurarne altri;
3. l'effetto razionalizzazione: illustrare e spiegare con le ragioni e le giustificazioni di oggi, gli eventi del passato.

Gli svantaggi che invece si possono imputare alle indagini per coorti sono:

1. la perdita dell'anonimato per i soggetti;
2. il fenomeno dell'*attrition*: vengono perse unità col passare del tempo necessario per lo studio. Per evitare tale problema, come si è precisato nel paragrafo 3.6.1, a volte bisogna stimolare una partecipazione continuativa attraverso premi o agevolazioni, e dare una forte motivazione ai soggetti chiedendo loro informazioni sugli eventuali spostamenti di residenza o domicilio;
3. il costo elevato e la complessità dello studio.

Non si può nascondere che gli studi caso-controllo presentino più opportunità di distorsioni o errori d'inferenza rispetto ad altri tipi di studi (un esempio può essere come si è visto la distorsione nel ricordare i fattori d'esposizione), ma la maggiore causa di questi problemi non ha nulla a che vedere con la validità dell'informazione ottenibile da questi disegni di studio,

riguarda piuttosto l'eccessiva facilità con cui uno studio di questo tipo può essere eseguito. Siccome non richiede consumi di tempo e denaro estremamente elevati, può essere progettato anche da ricercatori poco esperti che mancano dei principi epidemiologici di base. Studi di questo tipo portano sì a risultati, ma tali risultati molto spesso sono errati perché alcune leggi di epidemiologia erano state violate in precedenza. È quindi ammirevole ed incoraggiante osservare l'aumento degli studi caso-controllo nel campo di ricerca internazionale, ma bisogna puntare al rispetto dei principi di base per un buono disegno di studio.

3.9 Dimensione di una ricerca statistica

A parità di ogni altra condizione, quanto maggiori sono le dimensioni del campione, più precise sono le stime dei parametri e delle loro differenze. La difficoltà sta nel decidere che grado di precisione si vuole ottenere. Incrementare le dimensioni di un'indagine costa più denaro e richiede più tempo. A volte i limiti della ricerca sono imposti dalle ridotte risorse finanziarie o dal limitato tempo disponibile, quindi il ricercatore vorrà eseguire tutte le osservazioni che le sue risorse organizzative e logistiche gli permettono, tenuto conto dei tempi e dei costi di elaborazione e di analisi dei dati. In altre situazioni non vi sono limiti così ovvi e il ricercatore dovrà fare il bilancio tra i benefici di livello più elevato di precisione e i maggiori costi di una raccolta dati.

A livello di pianificazione è importante mettere in relazione la dimensione campionaria con un grado di precisione predeterminato.

Si consideri innanzitutto il problema del confronto delle medie di due popolazioni μ_1 e μ_2 nell'ipotesi che la deviazione standard, σ, sia nota e uguale per le due popolazioni distribuite normalmente e si consideri inoltre che si estraggano due campioni casuali di uguali dimensioni n. Se le deviazioni standard sono note ma diverse, i risultati seguenti andrebbero considerati come approssimazioni (con σ^2 pari alla media delle due varianze). Per il confronto di due frequenze relative, π_1 e π_2, σ può essere considerata approssimativamente uguale al valore ponderato:

$$\sqrt{\left\{\frac{1}{2}\left[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)\right]\right\}}\,. \qquad [3.3]$$

Si considerano, adesso, tre modi in cui la precisione si può esprimere:

1. *errore standard dato*. Si supponga che la richiesta sia che l'errore standard della differenza tra le medie campionarie osservate, $\bar{x}_1 - \bar{x}_2$, sia inferiore a ε; equivalentemente, si potrebbe richiedere che l'ampiezza dell'intervallo di confidenza al 95% non superi ± 2 ε. Ciò implica che sia:

   $$\sigma\sqrt{(2/n)} < \varepsilon ,$$

   oppure

   $$n > 2\sigma^2 / \varepsilon^2 . \qquad [3.4]$$

   Se si richiede che l'errore standard della media di un campione sia minore di ε, la corrispondente disuguaglianza per n è:

   $$n > \sigma^2 / \varepsilon^2 . \qquad [3.5]$$

2. *Differenza significativa data*. Si richiede che, se la differenza $\bar{x}_1 - \bar{x}_2$ è maggiore in valore assoluto di un certo valore d_1, allora essa deve essere significativa a un certo livello (per esempio 2α bilaterale). Si indichi con $z_{2\alpha}$ il valore dello scarto normale standardizzato superato in valore assoluto con probabilità 2α (per α = 0.05, $z_{2\alpha}$ = 1.96), allora:

   $$d_1 > z_{2\alpha}\sigma\sqrt{(2/n)}$$

   oppure

   $$n > 2\left(\frac{z_{2\alpha}\sigma}{d_1}\right)^2 . \qquad [3.6]$$

3. *Potenza data per una determinata differenza*. Il criterio 2 è definito nei termini di una differenza osservata, d. La vera differenza, δ, può essere o più grande o più piccola di d e pertanto sembra preferibile basare le richieste di precisione sul valore δ. Si può specificare un valore di δ, per esempio δ_1, come il più piccolo dei valori che non si vogliano trascurare, nel senso che se $\delta > \delta_1$ si vorrebbe ottenere un risultato significativo al livello della significatività bilaterale 2α. Tuttavia, non si può mai garantire una differenza significativa. Le fluttuazioni campionarie possono portare a un valore di $|d|$ molto inferiore a $|\delta|$ e non significativamente diverso da zero. Si indichi con β la probabilità di questo evento ed è nota come *errore di II tipo*, ossia è la

probabilità di non riuscire a riscontrare una reale differenza (falso negativo). Il livello di significatività α è detto *errore di I tipo* ed è la probabilità di rifiutare, sbagliando, l'ipotesi nulla (falso positivo). Il suo complemento, 1-α, esprime quindi la probabilità di accettare l'ipotesi nulla quando è vera. Mentre l'errore di I tipo è fissato a un basso valore dalla scelta di un livello di significatività durante l'analisi, l'errore di II tipo può essere controllato soltanto durante la fase di programmazione. Si potrebbe sostenere che l'errore di II tipo non sia maggiore di un valore comunque basso o, equivalentemente, che la probabilità di riscontrare correttamente la differenza come significativa, 1-β, non sia minore di un certo valore elevato. Questo valore è la cosiddetta *potenza* di uno studio.

Valori positivi della distribuzione dei d con errore standard ES(d) sono significativi al livello stabilito se:

$$d > z_{2\alpha} ES(d) \qquad [3.7]$$

Per una potenza maggiore di 1-β:

$$z_{2\alpha} ES(d) < \delta_1 - z_{2\beta} ES(d)$$

oppure

$$\delta_1 > (z_{2\alpha} + z_{2\beta}) ES(d)\,. \qquad [3.8]$$

Se d rappresenta la differenza, $\bar{x}_1 - \bar{x}_2$, delle medie campionarie di una variabile continua tra due gruppi indipendenti di dimensioni n, allora $ES(d) = \sigma\sqrt{(2/n)}$ e la [3.8] diventa:

$$n > 2\left[\frac{(z_{2\alpha} + z_{2\beta})\sigma}{\delta_1}\right]^2. \qquad [3.9]$$

L'espressione [3.9] segue la stessa logica della formula [3.2] vista nel caso del clinical trial, solamente che ha una formulazione differente.

La distinzione tra i criteri 2 e 3 è importante. Per esempio con 2α = 0.05 e 1-β = 0.95, $z_{2\alpha} = 1.96$ e $z_{2\beta} = 1.64$. Ponendo d_1 uguale a δ_1, i valori di *n* dati dalla [3.9] e dalla [3.6] stanno nel rapporto $(1.96+1.64)^2 : (1.96)^2$ o circa 3.4 : 1.

Le formule riportate ammettono che si conosca la deviazione standard, σ. In pratica raramente si conosce σ prima dell'elaborazione dei dati, anche a volte il ricercatore è in grado di utilizzare una stima di σ sulla base di dati precedenti che si reputino ragionevolmente accurati. Tuttavia ci sono due modi per modificare gli approcci visti in precedenza. Il primo parte dalla considerazione che il valore richiesto di *n* può essere determinato solo in funzione del rapporto di una distanza critica (ε, d_1 o δ) rispetto a una deviazione standard stimata o vera (*s* o σ) e quindi viene proposto di assegnare multipli di *s* a ε o a d_1 (casi 1 e 2) oppure specificare la potenza (caso 3) in funzione di un dato rapporto di δ_1 rispetto a σ. Il secondo modo suggerisce di stimare σ tramite una ricerca pilota relativamente limitata e quindi utilizzare questo valore nelle formule [3.4]-[3.9] per stimare *n*, la dimensione campionaria totale.
Il confronto tra due frequenze relative indipendenti segue un approccio per il calcolo di *n* simile al criterio 3 visto per il confronto delle medie. La soluzione è data in termini di ampiezza da rilevare della differenza tra due frequenze relative vere, π_1 e π_2, testando l'ipotesi nulla che ogni frequenza relativa sia uguale al valore ponderato π. L'equazione corrispondente alla [3.9] è:

$$n > \left\{ \frac{z_{2\alpha}\sqrt{[2\pi(1-\pi)]} + z_{2\beta}\sqrt{[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}}{(\pi_1-\pi_2)} \right\}^2 .$$

3.10 Progettazione dello studio al Children's Hospital dell'Università di Oulu

Attraverso il record linkage e l'imputazione di eventuali valori mancanti si è conclusa la preparazione del database totale che ha raccolto tutti i casi che sono entrati a far parte della ricerca portata avanti dal Children's Hospital dell'Università di Oulu durante il periodo 1996-2006. Come si è già spiegato in precedenza, le unità statistiche del database sono rappresentate da nati pretermine (prima delle 37 settimane di gestazione) negli ospedali di Oulu, Tampere e Seinäjoki (o nati altrove, ma immediatamente trasferiti in uno di questi ospedali) i cui genitori hanno dato il consenso di adesione alla ricerca. Un database completo e sistemato in ogni campo ha permesso l'avvio di nuovi studi e il corretto proseguimento di ricerche già in corso. L'obiettivo a lungo termine dell'equipe guidata dal Prof. Hallman al Children's Hospital dell'Università di Oulu, in collaborazione con il Biocenter sempre dell'Università di Oulu, è quello di sviluppare terapie (anti-infiammatorie) efficaci nella prevenzione o nel trattamento di gravi malattie croniche cardiopolmonari in periodo neonatale dovute alla prematurità estrema, rendendo possibile il rafforzamento del sistema di difesa. Trovare terapie efficaci significherebbe ridurre la richiesta di cure intensive dopo il parto ed aumentare la sopravvivenza dei nati senza incorrere in handicap neurologici o cognitivi.

Durante gli ultimi quarant'anni la mortalità neonatale in Finlandia è diminuita da un due per cento a un due per mille, ma non è stato riscontrato un altrettanto netto decremento della morbilità a lungo termine. Nonostante gli sforzi, il tasso di prematurità non è sceso e l'attuale stile di vita (aumento dell'età media al primo parto, riproduzione con procreazione medico assistita ed aumento di fumo ed uso di alcool) tende ad incrementare il rischio di nascite estremamente pretermine.

Le evidenze correnti dimostrano che solo pochi nati estremamente pretermine sono malati già al momento del parto, la maggior parte dei problemi patologici insorgono subito dopo la nascita. Gli insuccessi nel campo delle ricerche cardio-vascolari subito dopo il parto rimangono il più significativo problema nei nati pretermine. Una comune e seria patologia dei polmoni che si manifesta nei nati pretermine è la RDS e sarà su questa malattia che si centrerà lo studio condotto al Children's Hospital dell'Università di Oulu e presentato in questo testo.

3.11 Introduzione alla Sindrome da Distress Respiratorio

A livello mondiale circa il 5-10% dei neonati sono nati prematuramente (Gomez et al., 1995). Negli Stati Uniti la mortalità dovuta all'RDS è scesa dal 2,6 allo 0,4 per mille nati vivi durante il periodo che va dal 1970 al 1995 (Lee et al., 1999). In Finlandia l'incidenza di RDS è circa 0,6% di tutti i neonati e circa il 12% dei nati prematuri (Koskinen et al., 1999).

L'RDS è un'acuta conseguenza del parto prematuro. È una patologia caratterizzata da insufficienza respiratoria e insufficiente scambio dei gas, entro le prime poche ore dopo la nascita. Senza trattamenti, l'RDS è una malattia fatale nel neonato. Il fattore principale che predispone all'RDS è la prematurità e, più specificatamente, un'insufficienza del surfattante polmonare, dovuta al ridotto sviluppo del polmone (Avery & Mead, 1959). Il surfattante polmonare consiste in una mistura di lipoproteine richiesta per ridurre la tensione superficiale dell'interfaccia aero-liquida degli alveoli ed è fondamentale per prevenire atelectasia generalizzata. Nel paragrafo 3.8 verrà spiegata la funzione e la composizione del surfattante polmonare.

Grazie allo sviluppo delle cure intensive prenatali e neonatali nelle ultime due decadi la sopravvivenza dei nati prematuri è cresciuta significativamente e con essa anche la prognosi di RDS. Il trattamento glucocorticoide di profilassi materna è solitamente adottato nel caso di minaccia di travaglio prematuro per accelerare la maturazione polmonare e per prevenire RDS (Crowley, 1995). Il trattamento glucocorticoide prenatale abbassa l'incidenza di RDS di circa il 50%. Un più alto progresso nel trattamento dell'RDS è stato raggiunto attraverso la ventilazione artificiale. La somministrazione di surfattante direttamente nelle vie respiratorie migliora rapidamente il problema d'insufficienza respiratoria, diminuendo lesioni serie ai polmoni e la mortalità. L'integrazione di surfattante è attualmente una pratica comune come trattamento di RDS o come terapia profilattica dopo una nascita prematura (Hallman et al., 1988). La combinazione di terapia con supplemento d'ossigeno, supporto ventilatorio e terapia con surfattante esogeno è un efficace trattamento nella maggior parte dei casi di RDS.

Tuttavia, nonostante l'efficienza della profilassi e dei trattamenti, l'RDS rimane ancora la maggiore causa di mortalità e morbilità tra i nati prematuri, e particolarmente tra i nati estremamente sotto peso (ELBW dall'inglese *extremely low birth weight*, <1000g). Inoltre la displasia broncopolmonare (in inglese *Bronchopulmonary Dysplasia*, BPD) continua ad essere un problema comune tra la porzione di nati con RDS, la cui patologia è aggravata da estese complicanze ai polmoni.

I bambini che sopravvivono alla prematurità e all'RDS sono a rischio di morbilità a breve e lungo termine, incluso BPD e malattie neurosensoriale (Hallman, 1999). I modi più efficienti

di ridurre l'incidenza di RDS sono prolungare la gestazione e portare a termine la somministrazione di glucocorticoidi per l'aumento della maturazione dei polmoni nel feto.
In accordo con la letteratura corrente, l'infezione intrauterina gioca un ruolo chiave nella patogenesi del travaglio e del parto spontaneo pretermine attraverso l'attivazione di citochine e altri mediatori infiammatori che provocano il parto prematuro con effetti a cascata (Gomez et al., 1995; Greci et al., 1998).
Più della metà delle nascite premature vengono da parto vaginale spontaneo o dalla rottura prematura delle membrane fetali (Goldenberg & Rouse, 1998), che sono spesso provocate da un'incipiente infezione. Altre cause comuni di parti prematuri sono la pre-eclampsia e la gravidanza multipla.
Sia il tasso di mortalità neonatale che il rischio di RDS differiscono tra maschi e femmine: i maschi presentano una più alta incidenza di RDS e una più elevata mortalità legata all'RDS rispetto alle femmine, probabilmente per via della più lenta maturazione dei polmoni (Farrell & Wood, 1976; Khoury et al., 1985). Altra variabile che influisce sulla probabilità di presentare RDS nei nati pretermine è l'etnia. Essere di razza nera sembra consistere in un fattore protettivo da RDS. Prematuri di razza nera hanno una più bassa incidenza di presentare RDS con conseguenze fatali rispetto ai prematuri di razza bianca (Farrell & Wood, 1976). Sembra, inoltre, che l'RDS si presenti meno frequentemente, in modo meno severo e accompagnato da minori complicazioni, nei nati neri pretermine che nei bianchi (Hulsey et al., 1993).

3.12 Ipotesi iniziali: evidenza del contributo genetico all'RDS

Sono stati condotti diversi studi epidemiologici che sostengono l'idea di una tendenza familiare ed ereditaria nella presenza di RDS. Le madri che partoriscono bambini sottopeso (<2500 g) possono essere suddivise in due gruppi, madri a basso rischio di RDS e madri ad alto rischio di RDS, sulla base della frequenza di RDS nei figli precedenti (Graven & Misenheimer, 1965). Siccome i due gruppi non differiscono nelle modalità di gravidanza, travaglio o parto, è stato esposto un fattore genetico sconosciuto nella spiegazione dello sviluppo della malattia. Il primo studio sull'associazione di alleli con l'RDS ha evidenziato la presenza di geni legati all'RDS, con un'evidente prevalenza degli alleli A3 e B14 dell'antigene HLA nei casi con RDS rispetto ai controlli (Hafez et al, 1989). Questi risultati non sono, però, stati confermati dagli ultimi studi.

Gli sforzi più recenti di identificare i fattori genetici influenti sull'RDS sono stati ristretti solo a pochi studi sui geni candidati SP-A e SP-B (Floros et al, 1995; Veletza et al., 1996; Kala et al., 1997; Kala et al., 1998). La codifica genetica per i SP-, specialmente per SP-A e SP-B, è stata stimata essere la più significativa tra i geni candidati nello spiegare il manifestarsi dell'RDS. Questo dato è correlato con l'importanza funzionale diretta di queste proteine nella biologia del surfattante e nello sviluppo della patologia. L'estensione della variazione dell'introne 4 di SP-B è stato il primo ad essere individuato come legato all'RDS, perché gli alleli della variante Δi4 si manifestano ad una più alta frequenza nella popolazione affetta da RDS rispetto alla popolazione di controllo (Floros et al., 1995). Anche la variazione allelica di SP-A è stata indicata come legata all'RDS, in particolare è stata riscontrata un'associazione positiva dello specifico allele SP-A2. Queste associazioni alleliche non sono state riscontrate ripetutamente e sono risultate evidenti in piccoli sottocampioni di nati. Inoltre gli studi sono stati ostacolati dall'eterogeneità delle popolazioni oggetto di studio, dalle differenze di razza nelle frequenze alleliche, dalle dimensioni dei campioni e da un inappropriato controllo dei fattori di confondimento, quale il grado di prematurità.

Il gene SP-C è polimorfico a diverse sedi, ma nessuna delle comuni variazioni è risultata associata con l'RDS o altre patologie (Warr et al., 1987; Hatzis et al., 1994; Nogee, 1998).

In luce dei risultati ottenuti dalla letteratura il quesito principale a cui vuole rispondere il presente studio è se i geni SP-A1, SP-A2 e SP-B abbiano un ruolo causale o proteggente nell'eziologia dell'RDS neonatale. Per rispondere a tale domanda si monitorano le varianti intrageniche dei geni candidati in una popolazione ben definita.

3.13 Funzione e composizione del surfattante polmonare

Il surfattante polmonare è un complesso di lipoproteine che riveste la superficie alveolare del polmone. Le funzioni principali del surfattante polmonare consistono nel mantenere bassa la tensione superficiale dell'interfaccia aero-liquida e nel prevenire il collasso alveolare nell'espirazione (Hawgood & Clements, 1990).

Il surfattante è un materiale eterogeneo che esiste in specifiche forme intracellulari ed extracellulari. La quantità di surfattante extracellulare comprende differenti complessi morfologici, quali lo strato reale dell'interfaccia aero-liquida e la mielina tubolare (in inglese *tubular myelin*, TM), un complesso di strutture a due strati entro la subfase liquida degli alveoli. La quantità di surfattante intracellulare risiede nei corpi lamellari (in inglese *lamellar*

body, LB) che sono densi organelli di accumulo a più strati specifici delle cellule di II tipo (Haagsman & Golde, 1991; Johansson & Curstedt, 1997).

Il surfattante consiste di circa il 90% di lipidi e il 10% di proteine (King, 1982). I lipidi sono principalmente fosfolipidi, e la maggior parte di essi consistono in fosfatidilcolini (in inglese *phosphatidylcholine*, PC) (Clements, 1977).

Il surfattante polmonare contiene proteine di siero e proteine associate al surfattante. Le proteine di siero hanno un ruolo funzionale sconosciuto nel surfattante e possono inibire la sua attività (Hallman et al., 1991). Delle quattro proteine del surfattante, SP-A, SP-B ed SP-C sono conosciute essere importanti determinanti della struttura dell'omeostasi e dell'attività superficiale del surfattante, mentre SP-D assieme a SP-A ha un ruolo nelle funzioni immunomodulatorie (Mason et al., 1998). Il componente proteico più abbondante è l'SP-A, che conta circa il 50% della proteina. L'SP-A e l'SP-D sono proteine idrofiliche, mentre l'SP-B e l'SP-C sono piccole proteine molto idrofobiche.

3.14 Metabolismo del surfattante

I componenti del surfattante vengono prodotti, assemblati, secreti e riciclati dalle cellule epiteliali di tipo II degli alveoli (Rooney et al., 1994).

Dopo la sintesi nel reticolo endoplasmatico (in inglese *endoplasmic reticulum*, ER) delle cellule di II tipo, i componenti del surfattante vengono modificati nel Golgi. I componenti sono processati e assemblati attraverso i corpi multivescicolari (in inglese *multivescicular body*, MVB) ed i corpi compositi risultanti vanno verso i corpi lamellari (in inglese *lamellar body*, LB) da cui i componenti del surfattante sono secreti con esocitosi. Dopo la secrezione, i contenuti dei corpi lamellari vengono trasformati in una struttura extracellulare chiamata mielina tubolare (in inglese *tubular myelin*, TM), da cui i lipidi sono inseriti nell'interfaccia aero-liquida in modo da formare lo strato del surfattante (Williams, 1977).

Durante la fase di espirazione, la tensione superficiale nell'interfaccia aero-liquida viene ridotta per prevenire il collasso alveolare ed il surfattante viene continuamente riciclato. Per raggiungere una bassa tensione superficiale su compressione, lo strato di surfattante diventa arricchito di DPPC (*dipalmitoyl phosphatidylcholine*). Durante la successiva inalazione ed espansione della superficie degli alveoli, i lipidi vengono di nuovo distesi con l'aiuto delle proteine idrofobiche del surfattante (Oosterlaken-Dijksterhuis et al., 1991). La maggior parte del surfattante extracellulare viene riciclato dalle cellule di II tipo (Wright & Clements, 1987). Il surfattante intracellulare, quindi, non è assemblato solo dai componenti del surfattante

nuovamente sintetizzati, ma anche dai componenti presi dalle cellule di II tipo attraverso endocitosi. Il surfattante viene catabolizzato dai macrofagi alveolari e dalle cellule di tipo II (Gurel et al., 2001).

3.15 Definizione della popolazione statistica

Il presente studio è stato condotto al Children's Hospital dell'Università di Oulu – Ospedale Centrale, in collaborazione con l'Ospedale Centrale dell'Ostrobotnia del Sud di Seinäjoki e con il Children's Hospital dell'Università di Tampere – Ospedale Centrale. I comitati etici di questi centri hanno approvato il protocollo dello studio. La popolazione dello studio è rappresentata da tutti i nati prematuramente (< 37 settimane di gestazione) in uno degli ospedali sopra elencati tra il 1996 ed il 2006. I genitori dei neonati che sono entrati a far parte dello studio hanno dato un consenso scritto relativo alla partecipazione dei loro figli alla ricerca. Sono stati tracciati due criteri di esclusione dallo studio: i neonati sono stati esclusi sia se era stata effettuata una trasfusione di sangue intrauterina, sia se uno o entrambi i genitori non erano di origine finlandese.

Nel caso dei nati pretermine negli ospedali di Oulu, Tampere e Seinäjoki o dei nati che vi vengono trasferiti il ricercatore è facilitato nello studio di una specifica popolazione geograficamente definita perché essa è collegata a particolari centri medici in cui giungono spontaneamente tutti i pazienti che hanno bisogno di cure specifiche. Questi centri sono infatti specializzati in cure intensive per patologie neonatali e quindi, anche se un bambino è nato in un nosocomio diverso dai tre sopra citati, verrà trasferito in uno di questi centri se presenta malattie croniche cardiopolmonari o altre patologie. Il presente studio ha quindi il vantaggio di utilizzare dati su un'ampia popolazione omogenea sotto il profilo etnico e di avere a disposizione dei record con molte informazioni cliniche.

3.16 Individuazione del disegno dello studio

Tra gli studi di suscettibilità genetica generalmente si preferiscono disegni caso-controllo quando i casi e i controlli presentano una riproduzione della popolazione quasi identica e le variabili indipendenti principali (almeno la durata gestazionale) possono essere considerate sullo stesso piano. D'altra parte coorti di popolazione con un errore di selezione minimo rendono possibile la valutazione dell'impatto generale dei fattori genetici (Hallman et al., 2007).

La presente ricerca si basa su pazienti che sono nati pretermine (prima delle 37 settimane di gestazione) in uno dei tre centri partecipanti allo studio ed è stato proposto loro, o meglio ai genitori, di aderire allo studio. Nella progettazione del disegno si è quindi partiti dal valutare queste casistiche iniziali (studio di casistica) per poi suddividerle in casi e controlli sulla base della presenza o assenza della patologia oggetto di studio. L'obiettivo della presente ricerca è, infatti, quello di individuare le eventuali varianti intrageniche che possano essere associate con il manifestarsi dell'RDS neonatale, quindi è appropriata la scelta di uno studio caso-controllo in cui i casi sono rappresentati da nati prematuri affetti da RDS e i controlli da nati prematuri sani. Si è visto che solitamente gli studi caso-controllo vengono identificati con gli studi retrospettivi, ed anche in questo contesto è corretto parlare di studio retrospettivo. Sono state infatti recuperate, sia per i casi (una volta diagnosticata l'RDS) che per i controlli, differenti informazioni cliniche del nato ed una serie di informazioni genetiche attraverso l'analisi del DNA ottenuto da un campione di sangue ombelicale.

A differenza degli studi coorte, la variabile dipendente binaria in uno studio caso-controllo clinico-epidemiologico, è stabilita dalla stratificazione dovuta alla presenza o assenza di una malattia, mentre le variabili indipendenti sono variabili d'esposizione a rischi o variabili genetiche che portano ad un rischio maggiore o inferiore nel manifestarsi della patologia d'interesse. In un disegno di studio di questo genere, vengono scelti due campioni di dimensione fissa dai due strati definiti dalla variabile di outcome. I valori delle variabili d'esposizione vengono misurati per ogni soggetto selezionato, assumendo che esse includano tutti i principali termini di esposizione, confondimento ed interazione.

Un importante caso speciale dello studio caso-controllo stratificato è il *Matched Case-Control Study*. Per contrastare il problema dei fattori di confondimento (variabili che si pensa siano associate con l'outcome), quali età gestazionale e sesso, ogni caso è stato accoppiato ad un controllo con identica durata gestazionale (calcolata in settimane) e dello stesso sesso. Inoltre le coppie di pazienti sono state scelte sulla base dell'avvenuto trattamento glucocorticoide prenatale o meno per la prevenzione di RDS. Le coppie di gemelli, infine, non sono state abbinate tra loro.

Anche se non è indispensabile che il numero di casi e controlli sia costante tra gli strati, il disegno di abbinamento previsto nel presente studio consiste in un controllo per ogni caso.

3.17 Modalità di raccolta dati

Dati clinici sul sesso, sull'età gestazionale e sulle storie cliniche materna e neonatale sono stati ottenuti da record medici con l'aiuto del personale ostetrico. Quest'ultimo ha fornito le informazioni attraverso il database sulle nascite del corrispondente nosocomio che è inserito in un flusso informativo nazionale sulle nascite.

L'RDS è stata diagnosticata nei neonati in accordo con i criteri clinici, radiografici e patologici indicati in letteratura. I sintomi clinici riguardano suoni respiratori del neonato definiti *grunting*, retrazioni, ostruzione respiratoria nasale e necessità di ossigeno aggiuntivo per più di 48 ore o necessità di terapia con surfattante esogeno; i criteri radiografici consistono in un diffuso sviluppo del reticolo granulare e nella presenza di broncogrammi aerei; infine i sintomi patologici sono diffusa atelectasia e membrane ialine. Nessun neonato è stato trattato con surfattante in profilassi. Una volta diagnosticata la patologia si è chiesto ai genitori se acconsentivano all'entrata del proprio figlio nello studio ed è stato individuato un neonato di controllo, di ugual sesso, settimana gestazionale e trattamento glucocorticoide prenatale. Anche per i pazienti sani si è naturalmente richiesto il consenso scritto dei genitori. Successivamente per ogni paziente è stato prelevato un campione di sangue (0.5-3 ml) dal cordone ombelicale ed è stato raccolto in una provetta EDTA e conservato a –70° C fino a quando non è stato inviato nel laboratorio del Dipartimento di Pediatria dell'Ospedale Centrale dell'Università di Oulu per le analisi genetiche. Da ogni campione di sangue è stata isolata la genomica del DNA attraverso l'uso del Kit d'Isolamento del Gene Puro dal DNA (Gentra Systems). Una parte della soluzione del DNA è stata diluita in 50 ng/µl per l'amplificazione PCR (reazione a catena della polimerasi dall'inglese *Polimerase Chain Reaction*). Quando i campioni di sangue dal cordone ombelicale non erano disponibili, si sono determinati i genotipi utilizzando una macchia di sangue assorbita in un filtro di carta. Un dischetto di 3 mm (corrispondente a circa 12.000 globuli bianchi) è stato estratto dalla macchia di sangue attraverso lo stampo di un particolare tipo di carta. Per decontaminare lo stampo tra i campioni, sono stati presi stampi multipli sul filtro di carta pulito. Il DNA è stato delimitato nel dischetto, e i contaminanti cellulari sono stati prelevati con tre incubazioni successive di 15 minuti con 50 µl di soluzione purificata di DNA (Gentra Systems), seguiti da tre lavaggi con etanolo al 100%. Dopo aver asciugato ad una temperatura di 55° C, il dischetto di carta purificato è stato direttamente utilizzato come modello per l'amplificazione PCR. Un nuovo e pulito dischetto di carta, trattato in modo simile, è stato incluso in ogni serie come controllo per la contaminazione crociata del DNA.

Ultimo passaggio è consistito nella genotipizzazione dei geni SP-A e SP-B attraverso PCR.

3.17.1 Genotipizzazione dei geni SP-A

La genotipizzazione di entrambi i geni SP-A è stata svolta come descritto in DiAngelo et al. (1999). In breve, i geni SP-A sono stati amplificati con l'uso di primer di geni specifici sotto le condizioni descritte in Floros et al. (1996).

I cloni genomici SP-A1 e SP-A2 sono stati usati come controlli della specificità genica in ogni set delle reazioni PCR. La metodologia basata sulla PCR-cRFLP[1] è stata usata per individuare i polimorfismi dei singoli nucleotidi per i codoni 19, 50, 62, 133 e 219 (nel caso del gene SP-A1) e per i codoni 9, 91, 140 e 223 (nel caso del gene SP-A2). Il codone 85 è stato analizzato per entrambi i geni, in modo da garantire ulteriormente la specificità genica delle amplificazioni PCR. Varie combinazioni di polimorfismi in questi siti distinguono differenti alleli che vengono denotati con $6A^n$ per SP-A1 e con $1A^n$ per SP-A2.

3.17.2 Genotipizzazione dei geni SP-B

Sono stati genotipizzati due polimorfismi del gene SP-B: la variazione del singolo nucleotide T/C nell'ultimo codone dell'esone 4, causante una variazione dell'amminoacido, è stata denotata con SP-B Ile 131 Thr e la variazione di lunghezza dell'introne 4 è stata denotata con SP-B Δi4.

Per la genotipizzazione di SP-B Ile 131 Thr è stata utilizzata un'amplificazione PCR attraverso il primer SPBTaaF1, 5'-TGGGGGATTAGGGGTCAGTC-3', seguita da una seconda amplificazione PCR con il primer SPBTaaF2, 5'-GGGGGATTAGGGGTCAGTCT-3' e con il primer inverso SPBTaaR2, 5'-CATGGGTGGGCACAGGGGC-3', in 10 µl di mistura reagente.

La genotipizzazione del gene SP-B, variante Δi4 è stata ottenuta attraverso un passo dell'amplificazione PCR con l'utilizzo del primer SPBi4F, 5'-CTGGTCATCGACTACTTCCA-3' e del primer inverso SPBi4R 5'-TGAAGGGCACGTAGTTTCCTA-3'. Il frammento dell'SP-B Δi4 più comune (510 bp) [2] è stato denotato come "allele invariante". Tutte le varianti di delezione (cinque diverse dimensioni) sono state raggruppate come alleli di delezione. La variante di delezione più comune (366 bp) risponde di circa l'80% di tutti gli alleli di delezione; gli altri (240, 330, 400 e 480 bp) sono meno frequenti.

[1] PCR-converted restriction fragment length polymorphism.

[2] bp = base pairs -coppia di basi. È l'unità di misura più piccola usata per quantificare il DNA che fa riferimento al numero delle basi. I simboli sono bp, kb, Mb (per esteso: base pairs, kilobasi, megabasi).

3.18 Tempi e risorse

Gli studi portati avanti dal Children's Hospital dell'Università di Oulu sull'RDS, a partire dal 1996 sono in parte finanziati dall'Università di Oulu – Dipartimento di Pediatria, in parte dal Biocenter di Oulu e per un'altra parte più ridotta sono sponsorizzati con i fondi provenienti da Fondazioni finlandesi per la ricerca pediatrica genetica. Nel caso dello studio caso-controllo presentato, il maggior impiego di risorse è stato richiesto nella parte relativa alle analisi genetiche di laboratorio, condotte dal team del Biocenter di Oulu diretto dalla dott.ssa Haataja.

La tempistica dipende sostanzialmente dal numero di soggetti che aderiscono alla ricerca in modo da raggiungere una dimensione adeguata per la significatività dell'analisi statistica. Per il presente studio sono stati utilizzati dati già raccolti precedentemente per studi di carattere diverso, come per esempio studi di coorte, più dati di pazienti entrati appositamente nel presente studio. Nel complesso è stato impiegato un periodo di un anno e cinque mesi per coprire la numerosità di pazienti richiesta e per effettuare tutte le analisi e gli esami necessari per il recupero dei dati genetici. Successivamente è iniziata la parte relativa alla creazione e pulizia delle banche dati e all'analisi dei dati.

3.19 Dimensione statistica

Fissando un errore standard che non superi lo 0.05% e ottenendo dalla letteratura (Floros et al., 1996; Rämet et al., 2000) le frequenze relative del fattore di rischio principale (l'allele $6A^2$ del gene SP-A1) nel gruppo dei casi e nel gruppo di controllo, si può calcolare la numerosità statistica per lo studio attraverso la formula [3.4] $n > 2\sigma^2 / \varepsilon^2$.

$\sigma = \sqrt{\left\{\frac{1}{2}\left[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)\right]\right\}}$ e π_i i = 1, 2 sono le frequenze relative del fattore di rischio, allele $6A^2$, rispettivamente del gruppo dei casi e di quello di controllo. Sostituendo $\pi_1 = 0.65$ e $\pi_2 = 0.50$, il calcolo della varianza diventa $\sigma^2 = \left\{\frac{1}{2}\left[0.65 \cdot (1-0.65) + 0.50 \cdot (1-0.50)\right]\right\} = 0.23875.$

Andando infine a calcolare la dimensione statistica $n > 2\sigma^2 / \varepsilon^2$, si ottiene $n > 2(0.23875)/(0.05)^2 = 96$. Sono stati quindi arruolati nello studio almeno 96 pazienti per gruppo, più precisamente 137 casi e 137 controlli.

3.20 Conclusioni

Si è visto come negli studi sperimentali l'investigatore agisca sul campo di ricerca assegnando il fattore d'esposizione ai soggetti, mentre nelle indagini il ricercatore diventa più un osservatore, descrivendo la situazione di una specifica condizione morbosa in relazione a soggetti, tempo e luogo (nelle indagini descrittive), registrando gli eventi (patologie) nel corso del tempo (negli studi per coorti) e investigando retrospettivamente possibili fattori di rischio che possono aver causato una malattia (negli studi caso-controllo).

Il lavoro presentato nell'applicazione al Children's Hospital dell'Università di Oulu è centrato su uno studio caso-controllo 1-1 (*Matched Case-Control Study*), che prende in considerazione 137 coppie di nati pretermine in cui i casi sono rappresentati da neonati affetti da RDS ed i controlli invece da neonati sani. Le variabili attraverso cui sono state formate le coppie di pazienti sono età gestazionale, sesso e terapia glucocorticoide prenatale.

Sezione C

Trattamento delle variabili e possibili classificazioni dell'outcome

CAPITOLO 4 *Variabili e loro trattamento*

4.1 Premessa

La materia prima di ogni indagine statistica è rappresentata da osservazioni individuali che devono sempre venire organizzate e sintetizzate prima di essere utilizzate in elaborazioni più complesse (Armitage & Berry, 1996). Ogni classe di misura o classificazione in cui vengono comprese le osservazioni individuali si chiama variabile. La presentazione e la sintetizzazione dei dati richiede innanzitutto la comprensione dei tipi di variabili che si possono incontrare in uno studio clinico e le modalità in cui esse vengono misurate (paragrafo 4.2). Una volta rilevate le variabili sulle osservazioni (paragrafo 4.3) si passa alla loro rappresentazione attraverso tabelle, grafici, distribuzioni di frequenza e indicatori. In questo capitolo si tratterà principalmente la sintetizzazione delle osservazioni attraverso indicatori (paragrafo 4.4). L'ultimo aspetto fondamentale nel trattamento delle variabili riguarda la loro eventuale trasformazione (paragrafo 4.5), a volte necessaria per ottenere distribuzioni normali dei dati per poter applicare test statistici *ad hoc* o altre analisi. Nell'ultima parte del capitolo verranno presentate le variabili analizzate nello studio clinico-genetico al Children's Hospital dell'Università di Oulu, indicandone il loro trattamento specifico (paragrafo 4.6).

4.2 Definizione di variabili e scala di misura

Una variabile può essere definita come una classe di attributi, X, associabile alle unità statistiche. Se gli attributi sono misure numeriche si parlerà di variabili quantitative, se gli attributi sono qualificatori non metrici (aggettivi), si parlerà di variabili qualitative o categoriali. Le modalità di una variabile rappresentano gli elementi x della classe di attributi X. In altre parole una variabile è definita dalla classe delle sue modalità (Guseo, 1997).

Le variabili si possono classificare in funzione delle relazioni che si possono instaurare tra modalità corrispondenti, cioè all'interno della stessa variabile. Tali relazioni informano quindi sul tipo di misurazione consentita (tabella 4.1). Le principali scale di misurazione sono:

- scala nominale (o sconnessa): le modalità x costituiscono un insieme privo di ordinamento e costituiscono nomi. Alcuni esempi sono: sesso, nazionalità, gruppo sanguigno. Le uniche relazioni possibili sono quelle di uguaglianza e disuguaglianza tra modalità.

- Scala ordinale: le modalità *x* costituiscono un insieme ordinato. A titolo d'esempio si considerino il titolo di studio, il grado militare o il livello del personale amministrativo. Oltre alle relazioni di uguaglianza e disuguaglianza tra modalità, sono possibili anche quelle di maggiore o minore grado.

- Scala intervallare: le modalità appartengono ai numeri reali. L'origine è convenzionale nel senso che lo zero non indica l'estremo inferiore dell'ordine di grandezza della variabile. Un esempio tipico è costituito dalla scala in gradi Celsius per la misura della temperatura. In aggiunta alle relazioni presenti per la scala ordinale può essere eseguita anche la differenza tra due modalità.

- Scala rapporto: le modalità appartengono ai numeri reali non negativi. Lo zero quindi indica l'estremo inferiore dell'ordine di grandezza della variabile.

Tabella 4.1 Relazioni tra le modalità dele scale di misurazione

	SCALE			
	QUALITATIVE		**QUANTITATIVE**	
Relazioni	Nominale	Ordinale	Intervallare	Rapporto
$x_i = x_j$	*	*	*	*
$x_i \neq x_j$	*	*	*	*
$x_i < x_j$		*	*	*
$x_i > x_j$		*	*	*
$x_i - x_j$			*	*
x_i / x_j				*

Un particolare tipo di variabile è la variabile dicotomica, vale a dire una variabile con sole due modalità. La variabile dicotomica si può concepire come l'espressione del possesso di un determinato attributo, oppure dell'appartenenza a una categoria di unità. Un classico esempio di variabile dicotomica è l'avere o meno una certa patologia in un determinato momento. Alle unità che possiedono l'attributo si associa il valore 1, a quelle che non lo possiedono il valore 0. Qualunque variabile, quantitativa o qualitativa, può essere trasformata in una o più variabili dicotomiche. Per esempio il peso alla nascita si può trasformare nelle tre variabili dicotomiche "peso < 1500 grammi", "1500 grammi ≤ peso ≤ 3500 grammi" e "peso > 3500 grammi".

L'operazione di trasformazione di una variabile a più modalità in variabili dicotomiche si dice dicotomizzazione.

Nell'ambito delle variabili quantitative si ricordano i due sottocasi che seguono:

1) Discreto. Le modalità distinguibili *x* costituiscono un insieme di cardinalità al più numerabile; in altre parole le modalità assumono un numero finito di valori (o infinito numerabile). Un esempio può essere il numero di ricoveri di un paziente negli ultimi cinque anni.
2) Continuo. Le modalità distinguibili *x* costituiscono un insieme di cardinalità, cioè quando assumono tutti valori di un intervallo [a, b]. Tipici esempi sono il peso o l'altezza e la pressione sanguigna.

4.3 Rilevazione delle variabili sulle unità statistiche

Sia *X* la variabile statistica che si intende studiare. Sia $I = \{i : i = 1,2,...,N\}$ l'insieme delle etichette che individuano le unità statistiche di una popolazione. La rilevazione associa a ciascuna unità statistica $i \in I$ una modalità di *X*, indicata con il simbolo minuscolo, x_i, $i = 1,2,...,N$. Se si osservano simultaneamente alcune variabili, *X*, *Y*, *Z*, sulla stessa popolazione, la rilevazione produrrà concettualmente la matrice $(x_i, y_i, z_i), i = 1,2,...,N$, detta anche matrice dei dati.

Le rilevazioni possono essere dirette, mediante l'osservazione delle unità statistiche della popolazione selezionata, o indirette, se si utilizzano annuari, fonti cartacee o informatizzate già esistenti. Il database rappresenta l'archivio elettronico per la gestione dei differenti livelli di aggregazione delle unità statistiche primarie.

Una volta rilevati i dati si passa all'operazione di riduzione statistica (lo spoglio) che consente di passare dalle osservazioni alla distribuzione di frequenza corrispondente.

4.4 Indicatori

A volte la sintesi delle osservazioni, mediante la distribuzione di frequenza, non è molto soddisfacente se la numerosità delle distinte modalità di una variabile è elevata, quindi è necessario ricorrere a indicatori ed indici del fenomeno oggetto di studio. In base al tipo di variabili si possono enumerare differenti indicatori.

Per le variabili qualitative esistono indici di localizzazione, che studiano cioè la forma della distribuzione dal punto di vista della posizione in relazione ad un ordinamento delle modalità intrinseco o convenzionale. Tali indici specifici sono:

- la *moda*: si tratta della modalità di una variabile maggiormente presente tra le osservazioni.
- la *mediana*: disponendo le osservazioni in ordine crescente o decrescente, l'osservazione posta al centro è la mediana. Se il numero di osservazioni è dispari, la mediana viene calcolata prendendo la somma delle due osservazioni centrali e dividendola per 2.

Esistono inoltre indicatori che analizzano la dispersione o la variabilità di una variabile quantitativa, indicano cioè l'attitudine della variabile di assumere valori diversi. I più comuni sono:

- l'*indice di Gini*: ad ogni coppia di osservazioni *r* ed *s* (r, s = 1,2,…,*N*) corrisponde una coppia di modalità x_r ed x_s cui può essere associato un confronto c_{rs}, funzione indicatrice che assume il valore 0 se $x_r = x_s$ ed il valore 1 se $x_r \neq x_s$. essendo *N* la numerosità della popolazione sono possibili N^2 confronti distinti. Se si effettuano tutti i confronti ammissibili, quante più unità statistiche sono diverse, tanti più termini c_{rs} assumono valore 1. L'indice *G* di Gini è definito mediante la formula: $G = \frac{1}{N^2}\sum_{r=1}^{N}\sum_{s=1}^{N} c_{rs}$.
- L'*entropia di Shannon*. Per la sua trattazione si rimanda a Shannon & Weaver, 1949.

Anche nel caso delle variabili quantitative si può procedere nell'individuazione di specifiche riduzioni dell'informazione. A tale proposito si può dare, con un solo numero, un'idea del livello generale dell'ordine di grandezza di una serie di misure quantitative. Tale numero si può chiamare misura di localizzazione, come si è visto per le variabili qualitative, e può essere espresso con:

- la *media aritmetica*, $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- la *media geometrica*: viene utilizzata quando le osservazioni crescono in base ad un fattore moltiplicativo
- la mediana o la moda.

Le misure di variabilità che si utilizzano per variabili quantitative sono:

- il *range* o intervallo, definito come differenza tra valore massimo e valore minimo. Dato che tale valore numerico è determinato solo da due osservazioni originali, questo indice può risentire molto di eventuali valori anomali.
- Lo *scarto interquartile*. Dopo aver ordinato le osservazioni in senso ascendente o discendente si possono individuare due valori che lasciano fuori una piccola frazione di osservazioni a ogni estremo. Il valore al di sotto del quale cade un quarto delle osservazioni si chiama quartile inferiore, mentre il valore che è superato da un quarto delle osservazioni si chiama quartile superiore. La distanza tra essi si chiama scarto o distanza interquartile.
- I *percentili*. Il valore sotto il quale cade il P% dei valori è chiamato P-esimo percentile. Così il quartile inferiore rappresenta il 25-esimo percentile e quello superiore il 75-esimo percentile.
- La *varianza*. La media delle deviazioni dalla media ($x_i - \bar{x}$) elevate al quadrato si chiama varianza: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$.
- La *deviazione standard* o *scarto quadratico medio*: consiste nella radice quadrata positiva della varianza.

4.5 Trasformazioni delle variabili

Talvolta per esigenze operative, come per esempio applicare nel rispetto pieno delle condizioni di validità i test di statistica o specifiche analisi multivariate, si ritiene conveniente trasformare la variabile qualitativa o quantitativa di riferimento, *X*, dotata di distribuzione relativa $(x_i, p_i), i = 1,2,...,K_X$, in una nuova variabile *Y*. Si noti che una qualsiasi trasformazione è una funzione che associa ad una modalità *x* di *X* una ed una sola modalità *y* di *Y*. Se la trasformazione induce una funzione univoca $g : X \longrightarrow Y$ si tratta di stabilire in quale modo si trasforma la frequenza relativa modalità *y*. È sufficiente far riferimento all'antimmagine di una modalità y_j di *Y* ed attribuire alla modalità stessa la frequenza relativa pari alla somma delle frequenze relative corrispondenti appartenenti alle modalità *x* della classe di equivalenza. Sia ad esempio, $y_j^{-1} = \{x_1, x_3, x_5\}$ la specifica antimmagine di y_j tale che $g(x_1) = g(x_3) = g(x_5) = y_j$. Si indichi con q_j la frequenza relativa alla modalità y_j, allora, $q_j = p_1 + p_3 + p_5$. Si noti pertanto che la nuova variabile *Y* sarà caratterizzata dalla

distribuzione $(y_j, q_j), j = 1,2,..., K_Y$ con $K_Y \leq K_X$. Se la trasformazione $g(\cdot)$ è biunivoca, allora si conseguirà un semplice mutamento delle denominazioni originarie: $(y_i, q_i = p_i), i = 1,2,..., K_Y = K_X$.

Le condizioni di carattere sostanziale che possono portare alla necessità di trasformazioni di dati su specifiche variabili sono fondamentalmente tre e riguardano (Soliani, 2005):

a) *la linearizzazione delle relazioni tra variabili*. Per relazione lineare tra le variabili *x* e *y* s'intende, che *y* è esprimibile in funzione di *x* e di un termine residuale ε: $y = \alpha + \beta x + \varepsilon$, con α e β costanti opportune. Tutte le relazioni che si presentano con una forma diversa da quella appena descritta sono non lineari;

b) *la normalizzazione della distribuzione degli errori e delle osservazioni*. I test parametrici sono validi se la distribuzione dei dati è normale e quindi quella degli errori è normale intorno alla media. La verifica avviene con il controllo della simmetria e della curtosi oppure attraverso i test di Kolmogorov-Smirnov e di Shapiro-Wilk;

c) *la stabilizzazione delle varianze.* Dati suddivisi in gruppi spesso hanno varianze eterogenee e necessitano di essere trasformati. L'omogeneità delle varianze o omoschedasticità viene verificata mediante i test per due o più campioni. Nella statistica parametrica, tutti i confronti tra le medie e la stima degli effetti aggiunti sono fondati sull'assunto che tutti i gruppi abbiano la stessa varianza naturale o varianza vera (σ^2).

Quando un ricercatore deve applicare un test a dati campionari, ma incorre in problemi derivanti dalla non normalità e dalla eterogeneità delle varianze, egli può scegliere tra tre soluzioni:

1. ricorrere a metodi non parametrici, anche se si determina una perdita nell'informazione della misura rilevata, poiché da una scala di rapporti o di intervalli si scende a una scala di rango o binaria;
2. utilizzare una trasformazione dei dati, che elimina i due problemi elencati in precedenza e offre il vantaggio di applicare ugualmente il test parametrico;
3. utilizzare ugualmente il test parametrico senza trasformare i dati, contando sulla robustezza del test; è una soluzione accettata soprattutto quando il

campione è grande ma, è una procedura da non raccomandare e che in questi ultimi anni è sempre più contestata.

In questo paragrafo ci si occuperà della seconda soluzione. Riassumendo, con la trasformazione dei dati si effettua un tentativo, che in varie situazioni raggiunge lo scopo, di ottenere stabilità delle varianze, distribuzioni normali e linearità tra le variabili.
Le trasformazioni riportate in letteratura e alle quali più frequentemente si ricorre sono: la lineare, la logaritmica, le potenze (che comprendono le radici e soprattutto la radice quadrata e cubica, la reciproca e la quadratica), le angolari e i probit, i logit, i normit.

La *trasformazione lineare*: consiste nel cambiamento di scala o dell'origine delle misure, per facilitare la loro comprensione delle caratteristiche dei dati o i calcoli da effettuare. Può essere moltiplicativa, additiva e una combinazione di queste due modalità. E' il caso della trasformazione di una lunghezza da pollici a centimetri (trasformazione moltiplicativa). In una trasformazione moltiplicativa, la variabile trasformata (X_T) è ottenuta con una semplice moltiplicazione della variabile originaria (X_0): $X_T = C \cdot X_0$, dove C è la costante di conversione.

La *trasformazione in ranghi*: è una tecnica molto semplice e sempre più frequentemente raccomandata. Quando i dati sono abbastanza numerosi, utilizzare i ranghi al posto dei valori originari permette di ricostruire le condizioni di validità e di applicare tutti i test parametrici. Quando il campione è abbastanza numeroso (n > 30), i ranghi sono sempre distribuiti in modo normale; inoltre questa trasformazione elimina immediatamente l'effetto dei valori anomali. E' utile soprattutto nel caso di disegni sperimentali complessi, analisi di causalità a tre o più fattori con eventuale interazione o analisi gerarchica, per i quali nella statistica non parametrica non esistono alternative ai test di statistica parametrica. L'unico inconveniente che comporta tale trasformazione può consistere nella perdita d'intensità della variabile. Se le misure originali erano su una scala ad intervalli o di rapporti, nelle quali la distanza tra i valori è una indicazione importante da permettere il calcolo della media e della varianza, nella trasformazione in ranghi si ha una perdita di informazione. Ma si ha un vantaggio complessivo, poiché la perdita è limitata.

La *trasformazione logaritmica*, $Y = log_a X$, di solito avviene con base 10 o con base naturale (*e*), anche se non sono infrequenti quelli con base 2. Si applica quando la distribuzione ha simmetria positiva, per ottenere una distribuzione normale. In variabili continue, è utile per rendere omogenee le varianze quando esse crescono all'aumentare della media. Nel caso di

effetti moltiplicativi tra variabili, come nell'interazione, è utile per ritornare agli effetti additivi richiesti dal modello statistico dell'ANOVA.
La trasformazione logaritmica può essere applicata solamente a valori positivi, in quanto non esistono i logaritmi di valori negativi. Quando si hanno valori nulli, poiché log 0 = - ∞ (meno infinito), la trasformazione richiede l'accorgimento di aggiungere una costante (con C = 1 oppure C = 0,5) a tutti i dati (non solo a quelli nulli): $Y = log_a(X + C)$.

La *trasformazione in radice quadrata*, $Y = \sqrt{X}$ è uno dei casi più frequenti di trasformazioni mediante potenze, in cui c = ½: $Y = X^c$. E' utile in particolare sia per normalizzare distribuzioni con asimmetria destra (ma meno accentuata rispetto alla trasformazione logaritmica) per omogeneizzare le varianze. Spesso è applicata a conteggi, quindi a valori sempre positivi o nulli, che seguono la distribuzione poissoniana. Quando si ha la presenza di almeno uno zero è consigliabile (per tutti i dati) la trasformazione $Y = \sqrt{X + 0.5}$ che risulta appropriata per valori piccoli, con medie inferiori ad 1, in cui la semplice trasformazione in radice quadrata determinerebbe un ampliamento delle distanze tra i valori minori.

La *trasformazione quadratica*, $Y = X^2$ è utile quando la varianza tende a decrescere all'aumentare della media e la distribuzione dei dati ha una forte asimmetria negativa.

La *trasformazione probit* (probits da *probability units*)

$$P = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{z-5} \exp\left[-\frac{(X-\mu)^2}{2\sigma^2} \right] dx$$

è definita come la deviata normale equivalente, aumentata di 5.
Nello studio della relazione dose-risposta, la percentuale di individui che rispondono all'effetto causato della dose viene di solito rappresentato con una curva cumulata. Essa ha forma sigmoide, se la curva della distribuzione originaria è normale, con la conseguenza che a parità di errore nella dose l'errore nella risposta non è costante, ma varia secondo il punto in cui incontra perpendicolarmente la sigmoide. Per un errore costante nella risposta, occorre trasformarla in una retta. La curva percentuale cumulata può essere linearizzata in vari modi. Uno dei più diffusi consiste appunto nei probit, ottenuti con due passaggi logici:
1- Sostituire ai valori di p dell'ordinata quelli corrispondenti all'ascissa della distribuzione normale standardizzata $Y' = (X - \mu)/\sigma$. A causa della simmetria della distribuzione normale, il 50% dei valori Y' è negativo e l'altro 50% è positivo.

2 - Successivamente a tutti i valori trasformati in Y' aggiungere la quantità 5: si eliminano tutti i valori negativi. Questi valori trasformati mediante la relazione

$$Y = 5 + Y' = 5 + (X - \mu)/\sigma$$

sono i probit.

La *trasformazione normit*

$$P = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{z} \exp\left[-\frac{\mu}{2} \right] du$$

è un'altra trasformazione di percentuali cumulative basate sull'integrale di probabilità della curva normale. Fornisce valori diversi dai probit.

La *trasformazione logit* viene anche essa applicata a osservazioni percentuali (*p*) ed è ottenuta con $Y = \log_e \frac{p}{1-p}$. L'effetto di questa trasformazione logistica o *logit* è simile a quella probit e può determinare analisi del tutto uguali, in particolare nello studio del dosaggio con risposte quantali. L'attuale diffusione dell'informatica, che ha superato le difficoltà derivanti dalla complessità dei calcoli e dal tempo richiesto nei calcoli manuali, ha annullato la necessità di linearizzare le distribuzioni. Di conseguenza, le trasformazioni probit e logit sono sempre meno usate.

4.6 Definizione delle variabili e loro trattamento nell'applicazione al Children's Hospital dell'Università di Oulu – Finlandia

Innanzi tutto bisogna sottolineare che la variabile "*Social Security Number*", oppure il codice creato *ad hoc* qualora il "*Social Security Number*" non fosse stata disponibile, è stata usata come identificativo di record.

Come si sottolineato in precedenza, si vuole studiare la variabile RDS che viene presa come dipendente. Si tratta di una variabile nominale dicotomica espressa nelle modalità "presenza di RDS" e "assenza di RDS" diagnosticata come descritto precedentemente. In caso di presenza di RDS la modalità della variabile è stata espressa pari a 1, mentre in sua assenza si è registrato il valore 0.

Le variabili esplicative (o indipendenti) possono essere suddivise in due gruppi, le variabili cliniche e le variabili genetiche. La tabella 4.2 riporta questa suddivisione specificando il nome delle variabili nel database utilizzato per l'analisi, una descrizione di tale variabile con corrispettivi tipo, scala di misura e le alternative possibili. Le variabili cliniche prese in considerazione nel presente lavoro sono ospedale di nascita, sesso, durata gestazionale in giorni (e durata gestazionale in settimane), trattamento glucocorticoide prenatale, peso alla nascita, apgar alla nascita, lunghezza, genere del parto e modalità del parto. Le variabili genetiche riguardano invece polimorfismi dei geni SP-A1 e SP-A2 ed il polimorfismo nell'ultimo codone dell'esone 4 del gene SP-B. Per chiarire la trattazione dei due geni considerati ai lettori meno competenti in questo campo, si forniscono ora alcune informazioni relative alla funzionalità genica di SP-A e SP-B.

Il locus umano SP-A nel cromosoma 10q22-q23 consiste in due geni funzionali, SP-A1 e SP-A2. Entrambi i geni SP-A hanno quattro esoni codice. I prodotti di SP-A1 e SP-A2 differiscono gli uni dagli altri per le posizioni di quattro amminoacidi nel dominio collageno. La differenza più significativa nelle regioni codice tra SP-A1 e SP-A2 si trova nel codone 85. Sia SP-A1 che SP-A2 sono fortemente polimorfici: polimorfismi del singolo nucleotide (in inglese *single-nucleotide polymorphism SNP*) si verificano in ogni parte della sequenza codice e sono dovuti a sostituzioni di amminoacidi o cambiamenti silenti sia nel gene SP-A1 che in SP-A2 (McCormick et al., 1994). Gli alleli del gene SP-A1 sono stati denotati con $6A^n$ e gli alleli del gene SP-A2 con $1A^n$. Gli alleli dentro ogni gene differiscono in uno o più SNP negli esoni codice che nella tabella 4.2 sono indicati con SP-A1aa19, SP-A1aa50, SP-A1aa62, SP-A1aa133,SP-A1aa219, SP-A2aa9, SP-A2aa91, SP-A2aa140 e SP-A2aa223.

Tabella 4.2 Variabili cliniche e genetiche considerate nel presente studio, loro descrizione, tipologia, scala di misura e alternative possibili.

	Nome della variabile	Descrizione della variabile	Tipo di variabile	Scala di misura	Alternative possibili
Variabili cliniche	Hospital	ospedale di nascita	categoriale	nominale	oys = ospedale di Oulu sj = ospedale di Seinajoki tays = ospedale di Tampere
	Sukup	sesso	categoriale	nominale dicotomica	t = tytto (femmina) p = poika (maschio)
	Gestp	durata gestazionale in giorni	quantitativa	continua	es. 250
	Gestv	durata gestazionale in settimane	quantitativa	continua	es. 36
	Gluc	trattamento glucocorticoide prenatale	categoriale	nominale dicotomica	kylla = trattamento ei = non trattamento
	Sypa	peso alla nascita	quantitativa	continua	es. 2025 (in grammi)
	Apgar	apgar alla nascita	quantitativa	continua	valori da 0 a 10
	Pituus	lunghezza	quantitativa	continua	es. 50 (in centimetri)
	Twin	genere del parto	categoriale	nominale dicotomica	semplice, plurimo
	Birth	modalità del parto	categoriale	nominale	vaginale, cesareo, forcipe, ventosa, altro
Variabili genetiche	SP-Baa131	gene SP-B, polimorfismo del singolo nucleotide (T con C) nell'ultimo codone dell'esone 4	categoriale	nominale	CC, TC, TT
	SP-A1aa19	gene SP-A1, polimorfismo del singolo nucleotide per il codone 19	categoriale	nominale	CC, TC, TT
	SP-A1aa50	gene SP-A1, polimorfismo del singolo nucleotide per il codone 50	categoriale	nominale	CC, GC, GG
	SP-A1aa62	gene SP-A1, polimorfismo del singolo nucleotide per il codone 62	categoriale	nominale	AA, AG, GG
	SP-A1aa133	gene SP-A1, polimorfismo del singolo nucleotide per il codone 133	categoriale	nominale	AA, AG, GG
	SP-A1aa219	gene SP-A1, polimorfismo del singolo nucleotide per il codone 219	categoriale	nominale	CC, CT, TT
	SP-A2aa9	gene SP-A2, polimorfismo del singolo nucleotide per il codone 9	categoriale	nominale	AA, AC, CC
	SP-A2aa91	gene SP-A2, polimorfismo del singolo nucleotide per il codone 91	categoriale	nominale	CC, GC, GG
	SP-A2aa140	gene SP-A2, polimorfismo del singolo nucleotide per il codone 140	categoriale	nominale	CC, CT, TT
	SP-A2aa223	gene SP-A2, polimorfismo del singolo nucleotide per il codone 223	categoriale	nominale	AA, CA, CC

Il locus umano SP-B è codificato da un singolo gene nel cromosoma 2p12-p11.2 (Emrie et al., 1988; Vamvakopoulos et al., 1995). Il gene SP-B consiste di 11 esoni ed è conosciuto essere polimorfico in almeno cinque posizioni. Di queste variazioni, il polimorfismo (T con C) del singolo nucleotide in posizione +1580 nell'ultimo codone dell'esone 4 ha la più alta conseguenza funzionale. Tale polimorfismo è stato preso in esame nel presente lavoro (tabella 4.2).

Tutte le variabili genetiche originali sono qualitative e su scala nominale, ma verranno dicotomizzate nell'analisi statistica al fine di individuare il fattore ipotizzato essere di rischio che per ogni variabile è rappresentato dalla coppia di nucleotidi eterozigotica (per esempio nello SNP di SP-A1aa19 il fattore di rischio considerato è la coppia TC, Timina e Citosina).

Le variabili cliniche si presentano su scale più eterogenee. Come si è scritto in precedenza, le variabili sesso (nominale dicotomica), durata gestazionale in settimane (quantitativa) e trattamento glucocorticoide prenatale (nominale dicotomica) sono state utilizzate per abbinare ciascun caso al rispettivo controllo e quindi non verranno inserite nell'analisi dei dati. Le variabili quantitative continue apgar, lunghezza e peso alla nascita verranno analizzate come tali per studiare la loro relazione con la patologia RDS. Si presume che le variabili peso e lunghezza siano correlate con l'età gestazionale e che quindi non risultino significative: un peso alla nascita ridotto, come una lunghezza inferiore sono caratteristiche specifiche di neonati molto pretermine (< 32 settimane di gestazione). Infine si è analizzata la variabile nominale ospedale di nascita per controllare che non vi fossero fattori ambientali di confondimento, tenendo come modalità di riferimento l'ospedale di Oulu.

4.7 Conclusioni

Sebbene certe variabili, data la loro naturale scala, possano essere classificate solo in un modo, si è visto che molte altre misure possono essere considerate in forme diverse. Ridurre o condensare i valori multipli di una variabile in un numero inferiore di categorie discrete nell'analisi di uno studio, oppure trasformarli attraverso una delle tecniche appena elencate, può sicuramente essere utile e necessario per valutare certe ipotesi o per applicare certi test statistici. Tuttavia è importante tener ben presente che, mentre la riduzione o la trasformazione dei dati spesso ne rendono più facile la presentazione e possono semplificare le successive analisi statistiche, c'è sempre una conseguente perdita di dettaglio dalle osservazioni originali che può mascherare importanti informazioni o trend. Inoltre, mentre i dati possono sempre essere ridotti in una forma più semplice, non è mai possibile ricostruire

informazioni più dettagliate da dati che in origine sono stati registrati in semplici categorie. È quindi basilare considerare attentamente tutte le ipotesi e le domande interessanti per lo studio che si sta operando prima di iniziare la raccolta dei dati, in modo che possa essere ottenuto il necessario livello di dettaglio. La riduzione o trasformazione dei dati, se desiderabile, può essere eseguita in un secondo momento in preparazione all'analisi dei dati.

CAPITOLO 5 *Classificazioni dell'outcome: sintomi, patologie e loro conseguenze (menomazioni, disabilità e handicap)*

5.1 Premessa

In uno studio di casistiche cliniche assume una gran importanza l'outcome (la variabile dipendente) che si vuole indagare. Esso può essere analizzato sotto forma di patologia (o traumatismo), sotto forma di menomazioni, disabilità o handicap, oppure prendendo in considerazione i sintomi che un individuo presenta. Nella maggior parte dei casi si considera l'analisi di una singola patologia soprattutto perché è più semplice definirne la presenza o l'assenza su un singolo paziente mediante test od esami clinici specifici, e perché ne risulta più semplice la trattazione dei dati. In questo caso si tratta la variabile di outcome come variabile dicotomica e il disegno di studio adottato, come si è visto, è spesso il caso-controllo. In alternativa si può prendere in esame l'intensità o il livello (di gravità) della patologia, trovandosi ad analizzare una variabile di outcome quantitativa (scala rapporto) nel caso il livello di gravità sia espresso da un numero; nel caso invece l'outcome venga espresso in categorie di gravità la variabile di sarà trattata su scala ordinale. Tutte le malattie, possibili oggetto di studio come outcome, vengono raggruppate in apposite classificazioni (ICD, *International Classification of Diseases)* di cui questo capitolo si pone l'obiettivo di fornire una descrizione dettagliata (paragrafi 5.3.1, 5.3.2 e 5.3.3). Anche i sintomi di patologie specifiche sono classificati in apposite sezioni delle due versioni ICD presentate in questo capitolo, ma raramente vengono utilizzati come variabili di outcome in uno studio. Variabili che al contrario sono sempre più spesso inserite in ambito di ricerca clinico-epidemiologico consistono nelle cause di mortalità, classificate nell'ICD X.

Nel caso in cui si considerino le conseguenze di una patologia, identificate da menomazioni, disabilità ed handicap, normalmente uno studio clinico affronta l'analisi di più variabili di outcome contemporaneamente attraverso studi di coorte. Tali outcome sono classificate nell'ICIDH (*International Classification of Impairments, Disabilities and Handicaps*) e nell'ICF (*International Classification of Functioning, Disabilities and Health*) la cui trattazione concerne i paragrafi 5.4.1 e 5.4.2. Nell'ambito delle capacità e dell'autonomia di una persona disabile esistono altre classificazioni o scale di misure oltre a ICIDH e ICF che sono presentate nei paragrafi 5.5 e 5.6. Si vedrà infine come viene classificata la Sindrome da

Distress Respiratorio, variabile di outcome nello studio al Children's Hospital dell'Università di Oulu.

5.2 Le classificazioni e la loro storia

Dal punto di vista medico i nomi dati alle diverse patologie e alle loro conseguenze sono importanti strumenti d'interpretazione, di riflessione e di comunicazione, ma devono venire considerati nella giusta prospettiva, poiché tendono a mascherare le differenze esistenti fra i pazienti. L'aumento delle conoscenze sulle cause di malattia ha portato a modificare nomi descrittivi in nomi che si rifanno alla causa. Diverse forme morbose identificate un tempo da molteplici termini descrittivi vengono ora classificate in base all'eziologia, ma può anche accadere che una descrizione includa alcune delle manifestazioni che caratterizzano la malattia. Il passaggio da nomi descrittivi di patologie a nomi inglobanti una o più specifiche cause di malattia può drasticamente alterare il modo di trattare un particolare problema da parte sia del medico che del paziente, dato che già nel nome della malattia vengono indicati con esplicita chiarezza i mezzi per la prevenzione o la terapia. Si pensi per esempio a come cambierebbero i comportamenti di una società se il "tumore al polmone" venisse chiamato "malattia da fumo". Tuttavia i nomi che si riferiscono alle cause di malattia a volte possono limitare le vedute del medico. Per quanto riguarda le malattie infettive, per esempio, l'attenzione si concentra sull'agente causale e spesso si tende a considerare il microrganismo come la sola causa dimenticando gli altri fattori. Per ovviare a questo problema, oltre a catalogare le patologie e le loro conseguenze, si classificano anche i sintomi peculiari.

Dal punto di vista statistico-informatico l'etichettare le singole malattie e le loro conseguenze, ma soprattutto il codificarle dando loro un codice numerico o alfanumerico, rappresenta un ottimo aiuto nell'archiviazione, nell'analisi e nell'elaborazione di dati medici. La creazione di registri o di data-base con grandi moli di dati medici è infatti possibile soprattutto grazie all'ideazione delle Classificazioni.

Sin dai primi tentativi di classificazione sistematica delle malattie, che si fanno risalire al secolo XVIII, gli operatori sanitari e statistici hanno sentita forte l'esigenza di una standardizzazione e corretta descrizione degli eventi al fine di poterli codificare in maniera soddisfacente ed esauriente. A buona ragione, quindi, Edgar Sydenstricker nel 1920 affermava: "Non è possibile sperare in uno sviluppo uniforme della epidemiologia sino a quando non saranno raccolte correntemente e dettagliatamente statistiche accurate e complete

sull'incidenza delle malattie in differenti gruppi di popolazione e in diverse condizioni ambientali".

I primi tentativi classificativi si riferiscono a singoli medici, scienziati o ricercatori che utilizzavano delle particolari catalogazioni ad uso personale ed interno al loro laboratorio, poi l'area d'interesse si è estesa inglobando intere città, regioni, nazioni fino ad arrivare oggi a livello internazionale. In quest'ultimo campo solo nel 1855 il Congresso Internazionale di Statistica di Parigi diede vita ad una Classificazione Internazionale delle Cause di Morte. E' stato necessario, però attendere il 1893 perché l'Istituto Internazionale di Statistica a Chicago proponesse uno schema di classificazione che incontrasse i favori e la piena accettazione da parte degli operatori interessati. Solo a partire dal 1948 la classificazione internazionale è stata pensata ed elaborata in modo da poter essere utilizzata sia per le cause di morte che per gli eventi morbosi in sé, anche se non seguiti dal decesso della persona affetta. Solo dal 1980 si sono cominciate a classificare anche le conseguenze delle malattie.

Un fattore basilare per riuscire ad ottenere buoni dati in una determinata area territoriale è la scelta e la diffusione di un unico strumento di classificazione e di definizione che venga accettato non solo in ambito clinico ed epidemiologico, ma anche in ambito amministrativo e legislativo.

Si vedranno ora nel dettaglio le Classificazioni Internazionali attualmente in vigore ed emesse dall'Organizzazione Mondiale della Sanità (OMS).

5.3 Classificazione Internazionale delle Malattie (ICD)

5.3.1 ICD-9-CM

Nel 1893, la Conferenza dell'Istituto Internazionale di Statistica, che ebbe luogo a Chicago, approvò la Classificazione internazionale delle cause di morte. L'Italia adottò tale Classificazione a partire dal 1924. Sottoposta periodicamente a revisione, la Classificazione internazionale, a partire dalla 6ª revisione (1948), fu adottata anche per la rilevazione delle cause di morbosità, oltre che di mortalità.

Nel 1975, a Ginevra, nel corso della 29ª Assemblea dell'Organizzazione Mondiale della Sanità è stata approvata la 9ª revisione della Classificazione (ICD-9). Negli Stati Uniti, un comitato in cui sono rappresentati sia le Associazioni professionali ed accademiche dei medici, sia le associazioni degli ospedali, sia l'ufficio regionale della Organizzazione Mondiale della Sanità (OMS), ha sviluppato la ICD-9-CM ("International Classification of

Diseases, 9th revision, Clinical Modification"), la quale è stata utilizzata di norma dal 1979. Il termine "clinical" è utilizzato per sottolineare le modificazioni introdotte: rispetto alla ICD-9, che è fortemente caratterizzata dall'orientamento a scopo di classificazione delle cause di mortalità, la ICD-9-CM è soprattutto orientata a classificare i dati di morbosità. Le principali modificazioni introdotte sono infatti finalizzate a consentire sia una classificazione più precisa ed analitica delle formulazioni diagnostiche, sia l'introduzione della classificazione delle procedure.

La nuova classificazione ICD-9-CM è stata tradotta in italiano nel 1997 e nel 2002 ne è uscito un aggiornamento in cui vengono anche forniti gli strumenti per la codifica e compilazione della Scheda di Dimissione Ospedaliera (SDO).

Nel 1994 l'OMS ha ultimato la pubblicazione dei tre volumi della International Statistical Classification of Deseases and Related Health Problems – 10th Revision, comunemente chiamata ICD-10 e l'Ufficio di Statistica del Ministero della Sanità, congiuntamente all'Istituto Nazionale di Statistica ne ha curato la traduzione in lingua italiana. Nonostante questa sia l'ultima versione ICD, l'ICD-9-CM è il sistema tuttora utilizzato per la codifica delle diagnosi e dei traumatismi, mentre la decima revisione è al momento prevalentemente fruibile ai fini dell'elaborazione statistica dei dati di mortalità.

L'ICD-9-CM è uno strumento che riporta in modo sistematico e secondo precise regole d'uso, la nomenclatura delle diagnosi, dei traumatismi, degli interventi chirurgici e delle procedure diagnostiche e terapeutiche, i quali sono ordinati per finalità statistiche in gruppi tra loro correlati. A ciascun termine è associato un codice numerico o alfanumerico per un totale di oltre undicimila codici finali di diagnosi e oltre tremila codici finali di procedure.

Come si è detto il Sistema ICD-9-CM contiene due classificazioni, una per le malattie e traumatismi e una per gli interventi chirurgici e le procedure diagnostiche e terapeutiche. La prima di queste due classificazioni è un elenco sistematico che riporta, in ordine progressivo, i codici delle malattie, dei traumatismi e di altre cause di ricorso ai servizi sanitari e la relativa descrizione. La Classificazione delle malattie e dei traumatismi comprende 17 capitoli o macrogruppi, dei quali 10 sono dedicati a specifici organi o apparati anatomici, mentre gli altri 7 descrivono specifiche tipologie di condizioni che interessano l'intero organismo. Si propone di seguito l'elenco dei 17 macrogruppi.

- Malattie infettive e parassitarie
- Tumori

- Malattie delle ghiandole endocrine, della nutrizione e del metabolismo e disturbi immunitari
- Malattie del sangue e degli organi emopoietici
- Disturbi psichici
- Malattie del sistema nervoso e degli organi di senso
- Malattie del sistema circolatorio
- Malattie dell'apparato respiratorio
- Malattie dell'apparato digerente
- Malattie dell'apparato genitourinario
- Complicazioni della gravidanza, del parto e del puerperio
- Malattie della pelle e del tessuto sottocutaneo
- Malattie del sistema osteomuscolare e del tessuto connettivo
- Malformazioni congenite
- Alcune condizioni morbose di origine perinatale
- Sintomi, segni e stati morbosi mal definiti
- Traumatismi e avvelenamenti

Ogni capitolo comprende un gruppo di codici che classificano le malattie attinenti uno stesso apparato anatomico oppure una stessa tipologia clinica. Ciascuno dei 17 capitoli è suddiviso a sua volta nelle seguenti parti:

- **Blocco:** insieme di condizioni tra loro strettamente correlate (es. malattie infettive intestinali, 001-009);
- **Categoria:** codici a tre caratteri, alcuni dei quali molto specifici e non ulteriormente suddivisibili (es. 462 faringite acuta), mentre altri sono ulteriormente suddivisi, con l'aggiunta di un quarto carattere dopo il punto decimale;
- **Sotto-categoria:** codici a quattro caratteri; il quarto carattere fornisce ulteriore specificità o informazione relativamente ad eziologia, localizzazione o manifestazione clinica.

La Classificazione supplementare dei fattori che influenzano lo stato di salute ed il ricorso alle strutture sanitarie è composta da codici alfanumerici che iniziano con la lettera V e sono usati per descrivere problemi clinici, servizi erogati oppure circostanze particolari. Questi codici sono suddivisi in due gruppi di rubriche: le rubriche comprese fra 01 e 86 comprendono

interventi chirurgici maggiori, endoscopie e biopsie. Queste rubriche sono raggruppate in 15 sezioni, identificate sulla base del criterio anatomico (sede dell'intervento della procedura). Le rubriche comprese fra 87 e 99 comprendono invece altre procedure diagnostiche e terapeutiche. Sono raggruppate in 11 sezioni, identificate sulla base della tipologia della procedura (specialità che di norma o nella maggior parte dei casi eroga la procedura stessa).
Il sistema di codici utilizzato nella classificazione degli interventi chirurgici e delle procedure della ICD-9-CM è articolato in quattro caratteri numerici, dei quali i primi due identificano generalmente un organo, mentre il terzo e il quarto specificano la sede e il tipo dell'intervento. In alcuni casi, i codici si limitano al terzo carattere, per designare interventi che non necessitino di ulteriori specificazioni.

5.3.2 ICD-10

La decima revisione dell'ICD è stata pubblicata dall'Organizzazione Mondiale della Sanità nel 1992 e dall'anno 2000 è disponibile anche in lingua italiana grazie ad una traduzione curata dal Ministero della Sanità, dall'Istituto Nazionale di Statistica e dalla Federazione delle Società Medico-Scientifiche Italiane. Rispetto all'ICD-9-CM, l'ICD-10 è più orientata alla descrizione della comorbidità.
Il nucleo centrale della decima revisione dell'ICD è la codifica in categorie a tre caratteri, ognuna delle quali può essere ulteriormente suddivisa fino a dieci sottocategorie costituenti il quarto carattere. Il minimo livello necessario per fornire dati per la banca dati di mortalità internazionale dell'OMS e per confronti internazionali e di carattere generale è il codice a tre caratteri. Le sottocategorie a quattro caratteri sono consigliate per altri fini e costituiscono una parte integrante dell'ICD-10.
Al posto del sistema di codifica puramente numerico delle precedenti revisioni, la decima revisione utilizza un codice alfanumerico con una lettera in prima posizione e delle cifre in seconda, terza e quarta posizione. Il quarto carattere segue il punto decimale. I codici possibili sono perciò compresi tra A00.0 e Z99.9. La lettera U (U00-U49) deve essere utilizzata per classificare provvisoriamente nuove malattie di dubbia eziologia.
L'ICD-10 è strutturato in tre volumi: il volume 1 contiene la lista tabulare delle malattie, il volume 2 fornisce una guida agli utenti, il volume 3 presenta l'indice alfabetico della classificazione.
La classificazione è divisa in 21 settori o capitoli, ciascuno dei quali contiene un numero sufficiente di categorie a tre caratteri per trattare il proprio contenuto. I codici disponibili non

sono tutti utilizzati, lasciando spazio a future revisioni ed espansioni. I primi 17 settori trattano delle malattie e di altre condizioni morbose; il settore XVIII tratta dei sintomi, segni e risultati anormali di esami clinici e di laboratorio, non classificati altrove; il settore XIX tratta dei traumatismi, avvelenamenti ed alcune altre conseguenze di cause esterne; il settore XX classifica cause esterne di morbosità e mortalità; infine il settore XXI (fattori influenzanti lo stato di salute ed il ricorso ai servizi sanitari), va utilizzato per la classificazione di dati che spiegano:

- i motivi di ricorso ai servizi sanitari di persone non con malattie in atto;
- le circostanze per le quali il paziente sta ricevendo delle cure in quel particolare momento;
- circostanze relative ad altra tipologia di cure che il paziente sta ricevendo in rapporto con i sevizi sanitari.

5.3.3 Altre Classificazioni di malattie

Relativamente alle malattie mentali, il principale riferimento dopo l'ICD è dato dal DSM (Manuale Diagnostico e Statistico dei Disturbi Mentali), anch'esso sottoposto a revisioni e giunto ora alla versione IV. Nel D.S.M. IV viene proposto un sistema multiassiale grazie al quale vengono presi in considerazione tipi di disturbi, aspetti dell'ambiente ed aree del funzionamento. Nella classificazione del D.S.M. IV gli assi sono cinque: l'Asse I comprende i disturbi clinici; l'Asse II viene utilizzato per descrivere i disturbi di personalità (paranoide, schizoide, narcisistico, ecc.) e il ritardo mentale; l'Asse III attiene alle condizioni mediche generali che possono influenzare i disturbi mentali; l'Asse IV elenca una serie di situazioni esterne al soggetto, sociali, ambientali e relazionali che in qualche modo possono avere un ruolo causale e concausale nell'insorgenza e nell'andamento della forma morbosa; l'Asse V serve per codificare il livello globale di "funzionamento sociale" di un soggetto, ossia la sua capacità di adattarsi all'ambiente socioculturale in cui vive e di adempiere ai ruoli ritenuti normali per un essere umano. A differenza dell'ICIDH (affrontato nel paragrafo 5.4.1), in cui il "funzionamento sociale" viene valutato attraverso un approccio sistematico, costituito da tre assi tra loro relazionati (una menomazione determina una disabilità, la quale a sua volta può provocare uno o più handicap), il D.S.M., nelle sue varie versioni, affronta il modo in cui una persona "funziona" socialmente solo all'interno dello specifico asse V, tramite la "Scala per la Valutazione Globale del Funzionamento" (VGF). Si tratta di una scala a 100 punti, nella quale

il punteggio più basso si riferisce ad un livello di funzionamento grossolanamente alterato, mentre il punteggio più alto indica un buon funzionamento in tutte le aree. Osservando la scala VGF è possibile constatare come il funzionamento di un individuo venga valutato facendo riferimento alle aree del lavoro o della scuola, delle relazioni interpersonali e della comunicazione. Nonostante l'Asse V fornisca informazioni utili per la pianificazione di un trattamento e per la valutazione del suo esito, è stato fortemente criticato, in quanto vengono combinate in un unica scala misure di funzionamento psicologico, sociale, occupazionale e gravità dei sintomi. Poiché si tratta di quattro dimensioni poco omogenee tra loro, può risultare problematico integrarle in un'unica valutazione, soprattutto perché il miglioramento nell'una non è necessariamente correlato al miglioramento nelle altre. È evidente che tale problema non si pone invece per l'ICIDH, nel quale le conseguenze dei fenomeni morbosi vengono mantenute ben distinte nei tre assi di cui la classificazione è costituita (menomazioni, disabilità, handicap).

Aspetto positivo del DSM IV è la sua congruenza con l'ICDX, cosicché quasi sempre è possibile ricondurre una diagnosi fatta mediante il DSM IV a una codifica secondo l'ICDX. Tale corrispondenza non era presente nelle precedenti versioni DSM.

Ancora con riferimento alle patologie mentali, vi sono altri sistemi classificatori, tra cui il sistema Zero To Three che, configurandosi come un'estensione dell'ICD X, è finalizzato alla classificazione delle patologie psichiche ad esordio precoce, ovvero fra 0 e 3 anni.

Infine, per quanto riguarda le cause esterne, va citata la "External Causes of Diseases", implementata dall'Organizzazione Mondiale della Sanità congiuntamente al Center of Disease Control di Atlanta.

5.4 Classificazioni Internazionali delle conseguenze di Malattie (ICIDH e ICF)

Le risorse destinate alla sanità vengono generalmente allocate proporzionalmente alla prevalenza e all'incidenza di malattie e traumatismi ed in base alla gravità delle loro conseguenze, che possono essere il decesso del paziente od una sua disabilità temporanea o permanente. L'invecchiamento della popolazione ed il crescente progresso delle tecniche di terapia e di cura, che riescono a tener in vita anche pazienti molto gravi, hanno portato ad un aumento delle patologie croniche e delle loro conseguenze. Il Ministero della Salute deve quindi tener sempre più presente nella sua pianificazione e valutazione dei programmi di prevenzione e riabilitazione che le conseguenze delle malattie croniche hanno assunto un peso preponderante. Se le statistiche sulla mortalità sono registrate in modo adeguato (almeno nei Paesi sviluppati) non è così per le statistiche sulla disabilità, dato che insiste un'alta sottonotifica dovuta a problemi di documentazione della loro distribuzione nel territorio, non solo nazionale, ma anche regionale. La prima classificazione volta a catalogare le conseguenze delle malattie è stata proposta dall'Organizzazione Mondiale della Sanità (OMS) nel 1980: il lavoro, realizzato da un gruppo di esperti guidato da P. Wood, venne pubblicato nella sua edizione definitiva con il titolo di "*International Classification of Impairments, Disabilities, and Handicaps"* (ICIDH, tradotta in italiano con il titolo di "Classificazione Internazionale delle Menomazioni, delle Disabilità e degli Svantaggi Esistenziali"). A partire dalla sua introduzione l'ICIDH è stato largamente utilizzato in diversi settori di ambito sanitario come mezzo di categorizzazione delle conseguenze di malattie e traumi di un individuo.

Nonostante l'ICIDH sia rapidamente divenuto uno standard internazionale (è stato infatti tradotto in 13 lingue), fin dal momento della sua pubblicazione è stato oggetto di un importante dibattito internazionale che ha condotto ad un processo di revisione dell'ICIDH, iniziato nel 1993. Il primo risultato del lavoro di revisione è la bozza alfa del 1996, seguita nel 1997 dalla bozza beta-1 e nel 99 dalla bozza beta-2. La stesura finale è stata presentata nel maggio 2001 ed è intitolata "*International Classification of Functioning, Disability and Health*" (ICF, tradotta in Italia nel 2002 con il titolo di "Classificazione Internazionale del Funzionamento, della Disabilità e della Salute").

Anche se l'ICF è stato presentato come una revisione e un aggiornamento dell'ICIDH, in realtà si struttura su principi e presupposti molto differenti. Pur mantenendo l'obiettivo originale dell'ICIDH, di fornire cioè un linguaggio comune e unificato che serva da modello di riferimento per la descrizione degli stati di salute, l'ICF non ha lo scopo di descrivere lo

stato di salute solo di soggetti con disabilità, ma di ogni individuo, sano o malato che sia. Per concretizzare questo concetto l'ICF presenta un cambiamento sostanziale nella struttura rispetto all'ICIDH: invece di elencare le conseguenze di malattie nei tre concetti base dell'ICIDH (Menomazioni, Disabilità ed Handicap), organizza le informazioni relative al funzionamento del corpo umano e alle sue restrizioni, in due parti: C*omponenti del Funzionamento e della Disabilità* e *Componenti dei Fattori Contestuali*. La prima parte descrive "*Funzioni e Strutture Corporee*" e "*Attività e Partecipazione*" di un individuo; la seconda invece si occupa di "*Fattori Contestuali*" e di "*Fattori Personali*". Tutte le categorie all'interno di queste quattro sezioni contengono termini neutri, cioè inerenti a qualsiasi individuo, e vengono resi esplicativi attraverso dei qualificatori numerici che identificano un'eventuale deviazione dalla norma. Dal punto di vista statistico-informatico l'ICF risulta meno agevole dell'ICIDH per la trattazione delle singole conseguenze di patologia. Attraverso l'ICIDH ogni menomazione, disabilità o handicap viene identificato con un solo codice specifico che ne facilita l'archiviazione in un database (Facchin et al., 2002). Con la classificazione ICF invece, per individuare una menomazione o una disabilità o un handicap di un singolo paziente è necessario ricorrere all'utilizzo di più codici di Funzioni e/o Strutture interessate.

Anche se l'ICF è applicabile in diversi contesti culturali (sanità, scuola, lavoro, ambito sociale, ecc.) e considera tutti i disturbi in relazione al loro livello di funzionamento fornendo strumenti per la valutazione degli outcome, non facilita l'archiviazione dei dati attraverso una normale struttura di record in un dataset e rende difficile l'analisi dei dati.

La Legge Quadro 104/92 aveva introdotto in Italia la terminologia dell'ICIDH, ancora oggi utilizzata in diverse strutture nonostante la forte azione di stimolo da parte del Ministero del Lavoro e delle Politiche Sociali Italiano attraverso il "Progetto ICF in Italia" ad introdurre l'ICF.

Si vedano ora nel dettaglio le due Classificazioni.

5.4.1 ICIDH (*International Classification of Impairment, Disability and Handicap*)

L'obiettivo principale dell'ICIDH è fornire un linguaggio il più possibile unitario e standardizzato nella valutazione delle problematiche relative allo stato di salute di un individuo.

La classificazione tiene conto di qualsiasi alterazione delle condizioni di salute di un individuo in termini di cambiamenti nella funzionalità, guardando all'individuo secondo tre piani di osservazione: quello corporeo, quello personale e quello sociale.
L'ICIDH, a differenza dell'ICD che si preoccupa di classificare malattie e/o sintomi, si occupa della condizione di salute di un individuo, intendendo con ciò qualsiasi alterazione o modificazione dello stato di salute che può interferire con le attività quotidiane e determinare nell'individuo la necessità di rivolgersi ai servizi sanitari. La condizione di salute così intesa può, perciò, essere una malattia (acuta o cronica), un sintomo, oppure può riflettere altri problemi collegati allo stato di salute quali la gravidanza, l'età, alterazioni genetiche e altro.
L'ICIDH si divide in tre parti indipendenti (Menomazioni, Disabilità e Handicap) classificate attraverso dei codici gerarchici (vedasi tabella 5.1) secondo una struttura ad albero che identifica con la prima cifra il macrogruppo d'appartenenza della menomazione, della disabilità o dell'handicap per poi specificare il dettaglio fino ad un massimo di 4 cifre.

Tabella 5.1 Esempio della gerarchia tra i codici nell'ICIDH

3	**Menomazioni del linguaggio**				
		31	Menomazioni della comprensione e dell'uso del linguaggio		
				31.0	Disordini centrali a carico della funzione visiva con incapacità di comunicare
				31.1	Altre forme di dislessia
				31.2	Altri disordini centrali a carico della funzione visiva
				31.3	Riduzione del vocabolario
				31.4	Disturbo della sintassi
				31.5	Disturbo della funzione semantica
				31.8	Altre
				31.9	Non specificate

La prima parte dell'ICIDH (individuata dalla lettera I = *Impairment*) riguarda la menomazione intesa come "qualsiasi perdita o anomalia a carico di strutture o funzioni psicologiche, fisiologiche o anatomiche" (World Health Organization, 1980). Le menomazioni riflettono, in linea di principio, i disturbi a livello d'organo e sono raggruppate nell'ICIDH in nove gruppi di categorie:

1. Menomazioni della capacità intellettiva;
2. Altre menomazioni psicologiche;

3. Menomazioni del linguaggio;
4. Menomazioni auricolari;
5. Menomazioni oculari;
6. Menomazioni viscerali;
7. Menomazioni scheletriche;
8. Menomazioni deturpanti;
9. Menomazioni generalizzate, sensoriali e di altri tipo.

Le menomazioni, in questa classificazione, ricordano molto da vicino la terminologia utilizzata nell'ICD, anche se la loro descrizione appare estremamente particolareggiata e ciò sia per precisare il contenuto delle nove grosse categorie (che apparirebbero altrimenti troppo onnicomprensive), sia per permettere il raggiungimento di una buona precisione classificativa a chi ne fosse interessato.
La seconda parte dell'ICIDH (individuato dalla lettera D = *Disability*) considera le disabilità intese "nel contesto delle conoscenze e delle esperienze sanitarie, come qualsiasi restrizione o carenza della capacità di svolgere una attività nel modo o nei limiti ritenuti normali per un essere umano" in relazione alla sua età e sesso (World Health Organization, 1980). Le disabilità possono essere conseguenza diretta di una menomazione o riflettere la reazione di una persona a una menomazione fisica, sensoriale o di altro genere: se la menomazione riflette disturbi a livello di organo, la disabilità riflette disturbi a livello di persona.
Anche le disabilità sono divise in nove categorie:

1. Disabilità nel comportamento
2. Disabilità nella comunicazione
3. Disabilità nella cura della propria persona
4. Disabilità locomotorie
5. Disabilità dovute all'assetto corporeo
6. Disabilità nella destrezza
7. Disabilità circostanziali
8. Disabilità in particolari attività
9. Altre restrizioni all'attività.

Rispetto alla classificazione delle menomazioni, le sottocategorie sono qui meno particolareggiate ma ancora una volta gli operatori vengono invitati ad usare lo strumento

classificativo come di una sorte di elenco valutando per ogni individuo la presenza o meno di ciascuna disabilità. Nell'estendere il capitolo riguardante la disabilità è sorta l'esigenza di considerare due ulteriori elementi: la gravità e la prospettiva prognostica.
La gravità della disabilità riflette il grado di limitazione nell'esecuzione di una attività da parte di un soggetto e viene graduata secondo diverse categorie che vanno dalla esclusione della disabilità fino alla inabilità completa.
Interessante appare il raggruppamento delle diverse categorie secondo un'ottica di intervento da porre in essere per eliminare o limitare la disabilità del soggetto secondo quattro livelli che descrivono una inabilità crescente:

I. prevenzione della disabilità: l'individuo esegue le attività autonomamente e senza difficoltà (categoria 0);
II. potenziamento: l'individuo esegue le attività autonomamente ma solo con difficoltà (categoria 1);
III. integrazione: l'individuo esegue le attività solo grazie ad un ausilio, compreso l'aiuto di altre persone (categoria 2 - 4);
IV. sostituzione: l'individuo non è in grado di svolgere attività anche se aiutato (categoria 5 - 6).

Per quanto riguarda la prospettiva prognostica essa esprime la probabile evoluzione futura dello stato di disabilità del soggetto e anche in questo caso è prevista una graduazione in più categorie.
Tale ulteriore specificazione trova la sua giustificazione nel tentativo di indicare la possibilità di intervento sia in termini di recupero che di limitazione della disabilità e perciò dovrebbe contenere anche indicazioni riguardo al potenziale residuo di ogni soggetto: questo approccio dovrebbe favorire lo sviluppo di un profilo funzionale personale mirato a migliorare le prestazioni dell'individuo sia attuali sia in prospettiva delle esigenze future.
La terza parte dell'ICIDH (individuata dalla lettera H = *Handicap*) classificazione riguarda l'handicap inteso "nel contesto delle conoscenze e delle opere sanitarie, come una condizione di svantaggio vissuta da una persona in conseguenza di una menomazione o di una disabilità che limita o impedisce la possibilità di ricoprire il ruolo normalmente proprio a quella persona in base all'età, al sesso, ai fattori culturali e sociali" (World Health Organization, 1980).
L'handicap riflette lo svantaggio che un individuo ha per la sua incapacità di uniformarsi ai modelli delle società in cui vive; l'handicap appare come l'espressione di una ripercussione

psicologica, familiare e sociale della disabilità e dei problemi ad essa connessi: si crea uno svantaggio per l'individuo nei confronti dei suoi simili.
Questo svantaggio dovrà necessariamente essere calcolato sia in termini personali riferiti unicamente al soggetto affetto, sia in termini sociali ed economici più ampi, riferiti all'intera comunità: un bambino costretto in sedia a rotelle da un grave deficit motorio, vivrà il suo handicap con minor svantaggio in un ambiente privo di barriere architettoniche, piuttosto che in un ambiente in cui queste esistono: eppure la sua disabilità è sempre la stessa.
La classificazione individua sei assi entro i quali codificare la presenza o meno di handicap:

1. Handicap nell'orientamento
2. Handicap nell'indipendenza fisica
3. Handicap nella mobilità
4. Handicap occupazionali
5. Handicap nell'integrazione sociale
6. Handicap nell'autosufficienza economica.

Le sei dimensioni chiave dell'esistenza sono indicate come "funzioni della sopravvivenza" e racchiudono in sé le attività che ci si aspetta normalmente da un individuo.
Per ogni soggetto tutte le dimensioni devono esser identificate ed analizzate, in considerazione del fatto che la presenza di una menomazione o di una disabilità può interferire su più versanti della sopravvivenza determinando a volte una tale complessità di svantaggi non prevedibili fin dall'inizio.

5.4.2 ICF (*International Classification of Functioning, Disabilities and Health*)

La classificazione ICF si compone di due parti: la prima esplora gli ambiti della salute ed indaga la persona in quanto individuo, la seconda esplora gli ambiti legati alla salute indagando la persona in quanto essere sociale sottoposto all'influenza di fattori ambientali che agiscono da ostacoli o da facilitatori.
La prima parte ha due componenti:

1. l'organo, che a sua volta comprende:
   - le funzioni d'organo (vedasi tabella 5.2);
   - le strutture d'organo (vedasi tabella 5.3);

2. le attività (ossia l'esecuzione di un compito) e la partecipazione (il fatto di prendere parte ad una situazione della vita reale). Sono state individuate nove aree, ciascuna delle quali può essere definita in termini di performance (ciò che un individuo riesce a fare nel quadro della sua vita reale) e/o in termini di capacità (descrive la possibilità di un individuo di effettuare un compito in un ambiente standard): apprendimento ed applicazione delle conoscenze, esecuzione di compiti ed esigenze generiche, comunicazione, mobilità, cura di sé, attività domestiche, relazioni interpersonali, grandi ambiti della vita, vita della comunità sociale e civica.

Tabella 5.2 Funzioni Corporee nell'ICF: ***funzioni fisiologiche dei sistemi corporei (incluse le funzioni psicologiche)***

1. Funzioni mentali
2. Funzioni sensoriali e del dolore
3. Funzioni della voce e del linguaggio
4. Funzioni dei sistemi cardiovascolare, ematologico, immunologico e dell'apparato respiratorio
5. Funzioni dell'apparato digerente e dei sistemi metabolico ed endocrino
6. Funzioni genitourinarie e riproduttive
7. Funzioni neuro-muscoloscheletriche e correlate al movimento
8. Funzioni della cute e delle strutture correlate

Tabella 5.3 Strutture Corporee nell'ICF: ***parti anatomiche del corpo, come gli organi, gli arti e le loro componenti***

1. Strutture del sistema nervoso
2. Occhio, orecchio e strutture correlate
3. Strutture coinvolte nella voce e nel linguaggio
4. Strutture dei sistemi cardiovascolare, ematologico, immunologico e dell'apparato respiratorio
5. Strutture correlate all'apparato digerente e ai sistemi metabolico ed endocrino
6. Strutture correlate ai sistemi genitourinario e riproduttivo
7. Strutture correlate al movimento
8. Cute e strutture correlate

Ogni componente di questa prima parte può assumere un'accezione positiva ed, in tal modo, tutti i componenti sono raggruppati sotto il termine ombrello di funzionamento, oppure negativa, ed i componenti sono cosi raggruppati sotto il termine generico di handicap.

La seconda parte dell'ICF prende in considerazione i fattori contestuali, ossia il quadro nel quale si svolge la vita di una persona. Questa sezione comprende:

1. i fattori personali: sono le caratteristiche di una persona che non fanno parte di un problema di salute o di uno stato funzionale, ad esempio l'età, il sesso, l'istruzione, il carattere, la personalità, ecc.
2. i fattori ambientali: sono le caratteristiche esterne alla persona. Costituiscono l'ambiente fisico, sociale e attitudinale nel quale la persona vive e possono rappresentare un ostacolo o essere dei facilitatori: tecnologia e prodotti, ambiente naturale e cambiamenti portati dall'uomo all'ambiente, legami e relazioni sociali, attitudini, servizi, politica e sistemi.

I fattori contestuali interagiscono con la "funzione d'organo", la "struttura d'organo" e con le "attività e partecipazioni": il funzionamento e l'handicap di una persona, quindi, sono determinati dalle complesse relazioni che si instaurano tra i componenti degli "ambiti legati alla salute".

Si è cercato, in questo modo, di liberare i termini disabilità e handicap dalla loro valenza negativa e inserire una terminologia più neutrale, cosicché ci si riferisca all'attività e non alla disabilità del paziente, alla sua partecipazione e non più all'handicap.

Lo scopo generale della classificazione ICF è quello di fornire un linguaggio standard e unificato che serva da modello di riferimento per la descrizione della salute e degli stati correlati ad essa (Giacobini, 2002).

5.5 Strumenti di valutazione e quantificazione delle capacità della persona disabile

L'ambito della valutazione e quantificazione delle capacità residue della persona disabile costituisce anch'esso, così come quello delle classificazioni, un insieme estremamente vasto ed eterogeneo di scale di misura, applicate ai contesti più diversi e modificate per le situazioni più particolari.

Si possono delineare tre grandi aree di applicazione delle scale:

a. <u>per macro aree di funzioni</u>.
   a1. Funzione cognitiva. Tale funzione è descritta dalle cosiddette scale di sviluppo fino ai 4 anni (scala di Brunet-Lezine, 1965; scala di Bayley, 1991; scala di Griffith, 1960), e dai Quozienti d'Intelligenza (QI) dopo i 4 anni.
   a2. Funzione motoria. Si distingue in motricità grossolana e motricità fine. La prima valuta la performance motoria relativamente a locomozione, postura e cammino.

Lo strumento più utilizzato è la Gross Motor Scale. Per la motricità fine esistono invece molteplici scale, che dipendono dallo specifico tipo di motricità a cui si fa riferimento.

a3. Comunicazione linguistica.

a4. Funzione visiva. Le scale di valutazione si dividono sostanzialmente in scale di acuità e scale di campo visivo. Sono definite delle griglie di ipovisione con dei valori soglia che consentono di determinare la percentuale di ipovisione.

a5. Funzione acustica. Tale funzione non necessita di scale, poiché esistono strumenti in grado di misurare con precisione il livello di diminuzione dell'udito.

a6. Funzione comportamentale. Vi sono scale (principalmente la Child Behaviour Check List e la Behaviour Check List) che misurano l'adeguatezza dei comportamenti in relazione agli ambienti di vita.

b. <u>Autonomia e funzione complessiva</u>.

b1. ADL (Activities of Daily Living). Tale scala misura la capacità di espletare le funzioni fondamentali della vita quotidiana, quali lavarsi, vestirsi, andare alla toilette, spostarsi, alimentarsi, continenza.

In questa accezione la disabilità è dunque un concetto complesso al cui interno si possono rintracciare tre dimensioni significative:

1. la dimensione delle "funzioni della vita quotidiana", che comprende le attività di cura della persona;
2. la dimensione "fisica", sostanzialmente relativa alle funzioni della mobilità e della locomozione (camminare, salire le scale, raccogliere oggetti da terra), che nella situazione limite si configura come "confinamento", cioè costrizione permanente dell'individuo in un letto, in una sedia o in una abitazione;
3. la dimensione "comunicativa", che comprende le funzioni della parola, della vista e dell'udito;

La scala ADL è ampiamente utilizzata sia a livello nazionale che internazionale. Viene utilizzata soprattutto in ambiente geriatrico, ma presenta una scarsa sensibilità rispetto ad altri indici successivamente elaborati, non include la mobilità esterna, l'occupazione, lo stato cognitivo e le relazioni sociali. È stata adottata dall'ISTAT fin dalle sue prime indagini sulle condizioni di salute.

b2. IADL (Instrumental Activities of Daily Life). La scala IADL va oltre la potenzialità di fornire una misura della disabilità già offerta dalle ADL. Questa

proposta consiste infatti in una lista di domande relative alla capacità di svolgere un insieme selezionato di funzioni e attività che richiedono un più alto livello di coordinazione motoria di quello necessario per lo svolgimento dell'insieme delle attività coperte dalla scala ADL. Si tratta di otto quesiti in cui si chiede agli adulti di 15 anni e più cosa sono capaci di fare normalmente (capacità di usare il telefono, di fare la spesa, di preparare il cibo, di avere cura della casa, di fare il bucato, di utilizzare i mezzi di trasporto, di prendere le proprie medicine, di gestire le finanze). Tutte le risposte ai quesiti si basano su una valutazione soggettiva da parte dell'intervistato (o di un suo diretto familiare) del grado di autonomia posseduto e pertanto risentono dell'influenza di fattori cognitivi, culturali o emozionali dell'intervistato stesso. Ogni quesito prevede più livelli di risposta: autonomia piena nello svolgimento dell'attività, autonomia con difficoltà (grave o lieve) e totale dipendenza da altri.

La scala IADL misura più propriamente l'handicap, poiché è in grado di fornire un indicatore che misura la capacità di svolgere normalmente il proprio ruolo sociale.

Anche la scala IADL, come l'ADL, vede un utilizzo a livello sia nazionale che internazionale, anche se non così vasto come l'ADL. A partire dall'Indagine sulla salute 1999-2000, l'ISTAT ha adottato anche questa scala.

b3. BINA (Breve Indice di Non Autosufficienza). È una scala di disabilità, particolarmente adatta alla popolazione anziana, che analizza 10 items, ognuno dei quali dotato di 4 modalità ordinate e dotate di un punteggio (min 10, max 100), che indica la gravità della disabilità. Gli items sono: medicazioni, necessità di prestazioni sanitarie, controllo sfinterico, disturbi comportamentali, comunicazione, deficit sensoriali, mobilità, attività della vita quotidiana, stato della rete sociale, fattori abitativi e ambientali. Il valore dell'indice corrisponde al punteggio totale riportato nella valutazione, e tale valore va confrontato col valore soglia, pari a 230. Sono considerati non autosufficienti i soggetti con punteggio superiore a 230.

b4. FIM (Functional Independence Measure). È stata messa a punto negli Stati Uniti da Granger e collaboratori nel 1986, nell'ambito di un lavoro promosso dall'American Congress of Rehabilitation Medicine (ACRM) e l'American Academy of Physical Medicine and Rehabilitation (AAPM&R) (1983), al fine di consentire la disponibilità di una misura funzionale in medicina riabilitativa,

applicabile in studi clinici e gestionali. La FIM, ideata per assegnare punteggi su quattro livelli gerarchici di autosufficienza, è stata sostituita nel 1997 dalla versione attuale che ne prevede sette. Tale scala è "patologia indipendente" e per questa sua particolare caratteristica può essere utilizzata da qualsiasi operatore clinico, indipendentemente dal suo ambito specialistico.
La scala FIM è composta complessivamente da 18 voci a cui viene attribuito un punteggio da 1 a 7 ed è comprensiva di valutazione sulla cura personale (nutrirsi, rassettarsi, lavarsi, vestirsi dalla vita in su, e vestirsi dalla vita in giù, igiene personale), controllo degli sfinteri (vescica, alvo), mobilità (trasferimenti letto-sedia-carrozzina, wc, vasca o doccia), locomozione (cammino, carrozzina, scale), cognitiva (soluzione di problemi, memoria). I livelli funzionali ed i punteggi si distinguono in: autosufficienza (7.completa o 6.con adattamenti), non autosufficienza parziale (5.supervisione o predisposizione/adattamenti, 4.assistenza con minimo contatto fisico, 3.assistenza moderata), non autosufficienza completa (2.assistenza intensa o 1.globale).

c. Qualità della vita e stress. Si tratta di scale e indicatori finalizzati a misurare la qualità della vita delle persone. Particolari indici sono definiti per le persone disabili, o per le persone affette da patologie che necessitano di terapie.

   c1. QALY (Quality Adjusted Life Years). misura la combinazione delle caratteristiche di prolungamento della vita dovuto ad una terapia (o la durata del beneficio di una terapia) con l'effetto della terapia stessa nella qualità della vita. Ad esempio, se un trattamento prolunga la vita di una persona per un anno, e consente una vita normale, allora il QALY varrà 1 (=1 anno * 1). Ma se l'anno aggiuntivo sarà invece trascorso in una sedia a rotelle, e se una qualche scala che misura la qualità della vita indica che la vita in sedia a rotelle ha una qualità dimezzata, allora il QALY varrà 0,5 (=1 anno * 0,5). Il QALY è quindi un indice che misura l'effetto complessivo di una terapia, poiché considera sia la sopravvivenza, che la qualità della vita.

   c2. DALY (Disability Adjusted Life Years). Ha una logica analoga al QALY; è la risultante dell'effetto congiunto della sopravvivenza e del livello di disabilità (misurato secondo una scala), e misura la distanza rispetto alla qualità di vita considerata normale.

5.6 Altre scale di valutazione dell'autonomia della persona disabile

5.6.1 L'indice di Barthel

L'indice di Barthel è una delle scale ADL maggiormente utilizzata. La prima versione dell'indice consisteva in una scala di 100 punti; successivamente tale scala è stata modificata da molti autori.

La scala originaria è composta da 15 fattori, sui quali deve essere espresso un punteggio a tre livelli. Gli *items* sono rivolti a misurare la disabilità nel: bere da una tazza, alimentarsi, vestirsi, relativamente alla parte superiore del corpo e relativamente alla parte inferiore del corpo, rassettarsi, pulizia personale, intestino, controllo degli sfinteri (alvo, vescica), sedersi ed alzarsi dalla sedia, sedersi ed alzarsi dal wc, sedersi ed alzarsi dalla vasca o doccia, camminare su terreno pianeggiante per 45 metri, salire e scendere le scale, muoversi con la sedia a rotelle.

Il punteggio più alto che può raggiungere l'indice di Barthel originario è pari a 100 ed indica il livello di indipendenza; il livello più alto di gravità è invece rappresentato dal valore 0. Secondo uno studio di Granger *et al.* (1979), i valori uguali o inferiori a 60 distinguono i soggetti con una minore disabilità da quelli con una disabilità più marcata; i valori uguali o inferiori a 40 rappresentano disabilità molto gravi; i valori uguali o inferiori a 20 sono caratterizzati da una dipendenza totale nella cura di sé e nella mobilità.

5.6.2 RAP (Rehabilitation Activities Profile)

I numerosi metodi standardizzati di valutazione della disabilità non sempre si focalizzano sulle informazioni fondamentali per il processo riabilitativo e spesso risultano di difficile compilazione. Inoltre, generalmente, non considerano la percezione della disabilità da parte del paziente. Secondo gli ideatori della scala RAP, l'opinione del paziente e le difficoltà che incontra nella vita quotidiana hanno invece una forte influenza sulla programmazione e sui risultati del processo di riabilitazione.

Il sistema RAP è un metodo di valutazione sviluppato recentemente (1991) quale strumento per l'équipe riabilitativa per la programmazione, l'erogazione e la valutazione delle prestazioni riabilitative. Tale strumento si basa sull'ICIDH ed è strutturato in due livelli: il primo livello è utile per identificare la disabilità del soggetto; il secondo per descrivere in dettaglio le disabilità riscontrate nel soggetto. Il metodo RAP può essere applicato a tutti i gruppi diagnostici sia in regime di degenza sia in regime ambulatoriale.

Il RAP distingue cinque aspetti funzionali: comunicazione, mobilità, cura di sé, attività, relazioni.

Il primo livello include 21 *item*. Il secondo 71. Si osserva che l'*item* della relazione è un aspetto riferibile all'handicap e non alle disabilità. Per calcolare la gravità vengono utilizzati tre diversi punteggi, basati sulla scala di Likert (1932):

1. gravità della disabilità;
2. gravità dell'handicap nel rapporto con gli altri;
3. gravità nella percezione dei problemi.

Il RAP è stato sviluppato per essere utilizzato da qualsiasi operatore del team riabilitativo sia per pazienti degenti sia per pazienti ambulatoriali, indipendentemente dalla loro diagnosi o menomazione.

5.7 Patologia e disabilità trattate al Children's Hospital dell'Università di Oulu

Il presente studio si occupa prettamente di spiegare i fattori genetici e clinici legati alla Sindrome da Distress Respiratorio neonatale (*Respiratory Distress Syndrome* RDS). Tale patologia viene classificata nell'ICD-9-CM all'interno del 15° macrogruppo "Alcune condizioni morbose di origine perinatale" che comprende i codici dal 760 al 779. Più precisamente il codice corrispondente a questa malattia è 769 nell'ICD-9-CM, mentre nell'ICD-10 il codice P22 indica appunto "Sofferenza (distress) respiratoria(o) del neonato" nel Settore XVI che ha cambiato numero (dal 15° al 16°) rispetto alla precedente classificazione, ma non ha cambiato dicitura: "Alcune condizioni morbose che hanno origine nel periodo perinatale" (P00-P96).

Se fosse appurato e confermato da numerosi ricercatori che la Sindrome da Distress Respiratorio è legata a geni specifici, è interessante ipotizzare che tale patologia venga classificata sotto altra dicitura o sotto altro gruppo: non più condizioni morbose che hanno origine nel periodo perinatale, ma per esempio condizioni morbose genetiche originate dalla prematurità.

Le possibili disabilità conseguenti alla Sindrome da Distress Respiratorio neontale non sono oggetto del presente studio, ma si accenna ad uno studio coorte attualmente in corso al Children's Hospital dell'Università di Oulu che mira ad individuare quali possano essere le menomazioni, disabilità o gli handicap dei pazienti affetti da RDS durante la loro età pediatrica.

5.8 Conclusioni

Il problema di classificazione dell'outcome non è ancora stato completamente risolto. Le versioni e revisioni di classificazioni, scale ed indici che vogliono dare un ordine nei diversi settori clinici, sono sempre più numerosi e tendono a creare una certa confusione nell'utente che deve prima studiarli per poi applicarli. La funzione dell'OMS è quella di diffondere le classificazioni adottate a livello internazionale, ma questo compito non è facile: in alcune realtà locali classificazioni ormai radicate da decenni quali l'ICD-IX-CM o l'ICIDH, sono difficili da espiantare al posto delle nuove ICD-X e ICF. È per tale motivo che l'OMS, attraverso dei progetti mirati nei differenti Paesi, sta cercando di formare il personale medico e paramedico dei centri pubblici e privati, nell'utilizzo delle nuove classificazioni. Un fattore di rallentamento di queste iniziative riguarda però alcune difficoltà tecniche nell'utilizzo delle nuove classificazioni: esse puntano, infatti, più all'esaustività degli argomenti trattati e all'utilità concettuale che ad una facilitazione nell'uso dei codici. Le vecchie classificazioni, già testate ed ormai utilizzate in molti flussi informativi correnti, prevedevano un codice identificativo per ogni patologia, sintomo, causa di mortalità o conseguenza di malattia. In alcuni casi, come per esempio nell'ICF, le nuove catalogazione, oltre a comportare un cambiamento di codifica con problematiche di tipo informatico, spesso richiedono l'uso di più di un codice per l'individuazione della specifica menomazione, disabilità, ecc. creando difficoltà anche dal punto di vista statistico. Serve quindi del tempo per riprogrammare e ritestare i numerosi registri e database informatizzati, i flussi informativi e gli archivi clinici, con le nuove (ormai non più così nuove) misure di classificazione.

Sezione D

Alcuni metodi di analisi

CAPITOLO 6 *L'analisi dei dati*

6.1 Premessa

Le decisioni più cruciali nel campo medico, scientifico, sociologico, politico, economico ed organizzativo sono prese in base all'analisi dei dati. I dati rappresentano informazioni che prese singolarmente sono di aiuto limitato, sebbene siano sintetizzati e riassunti in quantità specifiche.

Un sunto dei dati può difficilmente essere ottenuto soltanto guardando le singole righe di una tabella. Un minuzioso esame e un'accurata analisi dei dati può spesso fornire un'enorme quantità di informazioni preziose. Certamente, più complessa è la struttura dei dati, più sofisticata sarà l'analisi degli stessi (Khattree & Naik, 2000). La complessità di un dataset può dipendere da una grande varietà di ragioni. Per esempio, come si è visto nel capitolo 2, il dataset può contenere diversi valori mancanti, oppure può includere troppi valori anomali (outliers), la cui presenza nei dati non può essere giustificata da alcuna semplice spiegazione, quindi è necessario andare ad analizzare in profondità la struttura del database (Belsley, 1990). Situazioni in cui una semplice analisi basata sulle sole medie può non essere sufficiente, si verificano quando i dati su alcune variabili sono correlati o quando è presente un trend nei dati (i dati cioè vengono raccolti ripetutamente nel tempo).

L'analisi di dati si può dividere in due parti essenziali: l'analisi descrittiva, detta anche esplorativa, e l'analisi statistica multidimensionale. Gli obiettivi che si possono raggiungere con l'analisi esplorativa (non trattata nel presente testo) riguardano sia le stime e le verifiche d'ipotesi tradizionali, sia le rappresentazioni tabellari e grafiche dei dati.

Lo scopo del presente capitolo è quello di fornire una sintesi dei principali metodi dell'analisi multivariata che si ritengono utili negli studi clinici. L'analisi multivariata consiste in un insieme di metodi e tecniche appropriate per situazioni in cui si deve studiare contemporaneamente la variabilità casuale di più variabili. Data la vastità dell'argomento si segnalano alcuni testi utili per integrare le tematiche trattate in questo capitolo: Kendall, 1980; Chatfield & Collins, 1980; Krzanowski, 1988; Mardia et al., 1979.

Dopo aver fornito le principali proprietà dei modelli per l'analisi multivariata, vale a dire la simmetria tra variabili (paragrafo 6.2.1), il fatto che sia metrico o meno (paragrafo 6.2.2) e la linearità e monotonicità delle relazioni tra variabili (paragrafo 6.2.3), vengono presentate in questo capitolo l'analisi fattoriale (paragrafo 6.4), l'analisi delle componenti principali (paragrafo 6.5), l'analisi delle corrispondenze (paragrafo 6.6), la *cluster analysis* (paragrafo 6.7), l'analisi di regressione stepwise (paragrafo 6.8) e l'analisi discriminatoria (paragrafo

6.10). l'analisi di regressione logistica sarà argomento centrale del prossimo capitolo e qui solamente presentata (paragrafo 6.9). Il paragrafo 6.11 riporta infine le considerazioni effettuate nella scelta del metodo d'analisi al Children's Hospital dell'Università di Oulu.

6.2 Modelli per l'analisi statistica multidimensionale

Un modello di analisi dei dati è un costrutto teorico scelto per rappresentare le proprietà e le relazioni assunte tra gli elementi (più frequentemente tra le variabili, talvolta tra le unità statistiche) costitutivi dei dati in esame. Le proprietà tecniche dei modelli di analisi dei dati sono:

- la simmetria del legame tra le variabili;
- l'esistenza di condizioni sufficienti per svolgere analisi metriche sui dati;
- la linearità o la monotonicità della relazione funzionale che lega coppie o insiemi di più entità.

6.2.1 Simmetria della relazione tra le variabili

I metodi d'analisi statistica multidimensionale si classificano come simmetrici o asimmetrici in accordo con le relazioni supposte tra le variabili osservate.

Se l'obiettivo dell'analisi è quello di rendere palese il tipo o il rilievo di una supposta subordinazione – che si denomina dipendenza – tra due o più insiemi di variabili, il modello si dice asimmetrico. In un modello di dipendenza asimmetrica, le variabili osservate vengono ripartite in due insiemi, uno di variabili dipendenti, e uno di variabili esplicative o predittive o indipendenti. La relazione tra i due insiemi di variabili si esprime come $Y \leftarrow X$, dove Y denota l'insieme (vettore) delle variabili dipendenti, X quello delle esplicative e la freccia indica la direzione della relazione ipotizzata, e cioè che X determina Y (Fabbris, 1997). I modelli asimmetrici sono importanti per analisi che si propongono di individuare relazioni di "causa ed effetto" tra le variabili selezionate. Può succedere che esistano ulteriori relazioni tra le variabili esplicative: l'insieme X può, per esempio, essere composto da due sottoinsiemi X_1 e X_2, dei quali il secondo condiziona il primo e ambedue condizionano Y.

Se si considerano tutte le variabili sullo stesso piano causale, il modello si dice simmetrico. Le relazioni si presentano nella forma bidirezionale: $x_i \leftrightarrow x_j$ per ogni i, j. L'ipotesi di relazione simmetrica tra le variabili è più debole di quella asimmetrica, nel senso che il

modello simmetrico non include nelle finalità dell'analisi la ricerca di associazioni causali tra variabili.

6.2.2 Analisi metrica e non metrica

Metrica è l'analisi realizzata con dati quantitativi, ossia rappresentabili geometricamente, assumendo l'esistenza di una relazione diretta tra la natura quantitativa dei dati ed il metodo d'analisi applicato. L'analisi non metrica si basa su procedimenti di calcolo che, non assumendo il vincolo metrico per i dati osservati, sono applicabili qualunque sia la scala di misura delle variabili. La maggior parte delle tecniche d'analisi non metrica si conclude, comunque, con un risultato rappresentabile geometricamente.

Se si elaborano variabili qualitative con variabili quantitative, gli indici di relazione tra variabili devono considerarsi di natura essenzialmente non metrica.

6.2.3 Linearità e monotonicità delle relazioni tra variabili

Per analisi metriche, si assume, in genere, che l relazione che lega le variabili sia lineare. Per relazione lineare tra le variabili x e y s'intende, che y è esprimibile in funzione di x e di un termine residuale ε: $y = \alpha + \beta x + \varepsilon$, con α e β costanti opportune. Se il numero di variabili esplicative osservate ($x_j, j = 1,2,..., p$) è pari a p ($p \geq 2$), la relazione tra la variabile y e le p variabili esplicative è lineare se: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$, dove i coefficienti β_i (i=1,2,…,p) sono costanti opportune.

Tutte le relazioni che si presentano con una forma diversa da quella appena descritta (come l'esponenziale, la logaritmica, la sinusoidale e quelle polinomiali di grado superiore al primo) non sono lineari. Come si è visto nel capitolo 4, con opportune trasformazioni, una funzione non lineare può essere linearizzata. La linearità come la non linearità della relazione tra variabili si può assumere per analisi con modelli sia simmetrici che asimmetrici.

Una relazione tra due variabili si dice monotòna se, all'aumentare dei valori (per variabili quantitative) o dei ranghi (per variabili qualitative) di una delle variabili, i valori (o i ranghi) dell'altra tendono a crescere nella stessa direzione (funzione monotòna crescente) o a decrescere (funzione monotòna decrescente).

6.3 Metodi e tecniche d'analisi

Il *metodo d'analisi* consiste nella successione logica di passaggi per raggiungere un dato obiettivo di analisi dei dati. La *tecnica di analisi* è, invece, una soluzione operativa quasi sempre rappresentata da un algoritmo informatico, ideata per perseguire un obiettivo analitico. Per un preciso obiettivo d'analisi dei dati, un metodo è, dunque, generale e con valenza di schema, e una tecnica cui il metodo si riferisce è particolare e pratica.

Le possibili scelte di metodo non dipendono solo dagli obiettivi della ricerca, tipo, scala e distribuzione delle variabili, ma anche dalle proprietà tecniche dei modelli di analisi dei dati viste in precedenza. Si cerca ora di riassumere la gamma delle possibili scelte.

- Con un modello d'analisi metrico e simmetrico, e
  - con l'obiettivo della ricerca di investigare strutture latenti dei dati, si può scegliere tra il metodo d'*analisi fattoriale*, il metodo delle *componenti principali* e il metodo d'*analisi delle corrispondenze*;
  - con l'obiettivo di individuare tipologie o gruppi simili di dati si propone l'analisi di raggruppamento (*cluster analysis*).
- Con un modello d'analisi metrico e asimmetrico, e
  - con una sola variabile dipendente quantitativa, si propende per il metodo d'analisi di regressione multipla *stepwise*;
  - con una sola variabile dipendente dicotomica si sceglie il metodo d'*analisi di regressione logistica*;
  - con una sola variabile dipendente qualitativa (su scala ordinale) si opta per l'*analisi di regressione logistica multipla* (o polinomiale);
  - con più di una variabile dipendente si può scegliere tra l'*analisi discriminatoria* e l'*analisi di regressione logit*. Quest'ultima non trattata in questo capitolo.
- Con un modello d'analisi non metrico e simmetrico, esistono delle versioni non metriche di analisi fattoriale, analisi delle preferenze, analisi delle prossimità e della cluster analysis che non saranno affrontate in questo capitolo.

6.4 Analisi fattoriale

Nell'analisi fattoriale, le variabili osservate X_i, (con i=1,2,…,p) sono rappresentate come la combinazione lineare di un numero minore di variabili aleatorie f_j, j=1,2,…,q chiamati fattori. I fattori sono delle variabili latenti che generano le variabili osservate X_i.

Come le variabili osservate X_i, anche i fattori assumono valori diversi per ciascuna unità di osservazione. Diversamente dalle variabili X_i però, i fattori non possono essere direttamente misurati e osservati. L'ipotesi centrale dell'analisi fattoriale è che la correlazione tra le variabili è determinata da dimensioni non osservabili (i fattori) che generano le variabili osservate. L'analisi fattoriale esamina la varianza che le variabili hanno in comune, cioè la varianza comune.

Il modello di base dell'analisi stabilisce che il punteggio ottenuto da un soggetto in una variabile può essere espresso come la somma ponderata del punteggio ottenuto dallo stesso soggetto:

a) nei fattori comuni che riflettono ciò che le variabili hanno in comune;
b) in una componente unica, che riflette ciò che le variabili non condividono.

Questi fattori vengono individuati esattamente dall'analisi fattoriale, il cui modello matematico di base è il seguente:

$$X_i = a_{i1}f_1 + a_{i2}f_2 + ... + a_{iq}f_q + u_i c_i = \sum_{j=1}^{q} a_{ij}f_j + u_i c_i \quad (i=1,2,\ldots,p) \qquad [6.1]$$

dove

- X_i è la variabile osservata. La varianza delle X_i spiegata dal modello fattoriale è definita comunanza;
- a_{ij} sono i coefficienti, detti pesi del fattore, che misurano la relazione tra la variabile osservata X_i e le variabili latenti f_j;
- f_j sono i fattori comuni, che rappresentano la variabilità che il fattore j-esimo condivide con le altre variabili in analisi. Ogni fattore comune può influenzare più di una variabile osservata;
- c_i è il fattore unico della variabile X_i, non correlato con nessun altro fattore;
- u_i è il coefficiente del fattore unico c_i.

I pesi del fattore, le comunanze e il numero di fattori richiesti, vanno stimati partendo dai dati, solitamente dalla matrice di correlazione campionaria. Naturalmente non si possono utilizzare i metodi di regressione multipla dato che non sono noti i valori delle f_j.

I fattori comuni contribuiscono al punteggio di almeno due variabili, cioè saturano almeno due variabili, mentre i fattori unici contribuiscono al punteggio di una sola variabile. Questi

ultimi non sono correlati tra loro né con i fattori comuni (nei modelli ortogonali anche i fattori comuni non sono correlati).

Il modello [6.1] esprime le variabili in funzione dei fattori, stabilendo così le variabili come combinazione lineari dei fattori. Ogni fattore comune è una combinazione lineare di tutte le variabili osservate:

$$f_j = \sum_{i=1}^{p} w_{ij} x_i \quad (\text{j} = 1,2,\ldots,q) \qquad [6.2]$$

dove w_{ij} è il coefficiente fattoriale (*factor score coefficient*) che lega la variabile i-esima e il fattore j-esimo.

Adottando il modello d'analisi fattoriale si assumono dunque relazioni lineari ed additive tra le variabili osservate.

Questo metodo è utilizzato soprattutto nel campo della psicologia, ed è strettamente collegato all'analisi delle componenti principali.

6.5 Componenti principali

Si supponga di avere delle osservazioni su p variabili X_i, (con i=1,2,…,p) eseguite su ciascuno di n individui. È possibile combinare le X_i in un numero ristretto di altre variabili capaci di fornire quasi tutte le informazioni su come un individuo differisce da un altro, attraverso la definizione delle seguenti nuove variabili:

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

$$Y_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \qquad [6.3]$$

e così via, in modo che Y_1 abbia la più elevata varianza possibile, e rappresenti così, meglio di ogni altra combinazione lineare delle X_i, le differenze generali tra individui.

Si può dunque scegliere Y_2 in modo che non sia scorrelata con Y_1 e abbia la seconda varianza più elevata, e così via. Se p=3, le osservazioni individuali si possono visualizzare in un diagramma di dispersione tridimensionale con n punti, magari raggruppati in una nuvola a forma di dirigibile (Armitage & Berry 1996).

La prima componente principale, Y_1, rappresenta le distanze pungola lunghezza del dirigibile. La seconda componente, Y_2, rappresenta le maggiori distanze (da un lato all'altro) in direzione perpendicolare alla lunghezza; la terza e ultima componente, Y_3, rappresenta le distanze dalla cima al fondo del dirigibile. In generale, vi sono p componenti principali, tutte, eccetto alcune, poco variabili. Se, nel caso di p = 3, il dirigibile fosse molto piatto, quasi come un disco, Y_3

avrebbe una variabilità molto bassa e la posizione di un individuo sull'intero diagramma sarebbe determinata quasi esclusivamente da Y_1 e Y_2.

Il metodo di analisi richiede il calcolo degli autovalori delle covarianze o della matrice di correlazione. La maggior parte dei programmi contiene procedure ad hoc per il calcolo degli autovalori. È importante osservare che, cambiando la scala di misura di una qualsiasi variabile, anche solo per un fattore moltiplicativo, cambiano tutti i risultati. Questo problema ha indotto diversi ricercatori a standardizzare inizialmente ogni variabile, dividendole per la loro deviazione standard: ciò equivale a lavorare con una matrice di correlazione.

L'interpretazione di una componente qualsiasi, per esempio la j-esima, definita dalla [6.3], implica considerazioni sui valori relativi assunti dai coefficienti a_{ij} in corrispondenza di quella componente. Con tutti i coefficienti pressoché identici, si può interpretare la componente come media di tutte le variabili. Se alcuni coefficienti sono piccoli, si possono ignorare le variabili corrispondenti e interpretare la componente in termini di un sottoinsieme di variabili originali. Nel caso in cui le componenti abbiano un'interpretazione che sembri rappresentare qualche caratteristica pertinente al fenomeno di studio, l'applicazione del metodo consente solo un'utile riduzione nella dimensionalità dei dati. L'interpretazione risulta quindi soggettiva e richiede la conoscenza dei campi di applicazione.

Dato che l'interpretazione di ogni componente viene calcolata in termini di variabili originali, conviene valutare ogni componente nei termini delle sue correlazioni con le variabili originali, piuttosto che nei termini dei coefficienti a_{ij}. Per componenti principali, calcolate a partire dalla matrice di correlazione, tali correlazioni si ottengono moltiplicando ogni componente per la radice quadrata dell'autovalore corrispondente. Queste correlazioni sono dette pesi delle componenti.

Un altro aspetto dell'analisi concerne la scelta del numero di componenti da includere. Anche se sono stati proposti diversi metodi, non ne esiste uno universalmente accettato e, di solito, si decide secondo l'interpretazione delle componenti che il ricercatore ha in mente. Raramente vale la pena di includere componenti extra, non potendo fornire un'interpretazione sensata. Una volta deciso il numero di componenti da includere, è stato effettivamente ridotto il numero delle dimensioni dei dati; per esempio, accettando due componenti, si considerano le osservazioni come appartenenti ad uno spazio bidimensionale, tralasciando le rimanenti p-2 dimensioni delle variabili originali. Nello spazio bidimensionale un punto è rappresentato da due coordinate rispetto a due assi: usando la rappresentazione fornita dalla [6.3], il punto sarebbe (y_1, y_2). Tuttavia esistono molti altri metodi per costruire gli assi. Limitandosi agli assi ortogonali, è possibile far ruotare la coppia di assi attorno all'origine. Le coordinate di un

punto cambiano in (y'_1, y'_2), ma la configurazione geometrica dei punti resta inalterata. Riportandosi alla formula [6.3], si ottengono le variabili:

$$Y'_1 = a'_{11}X_1 + a'_{12}X_2 + ... + a'_{1p}X_p$$

$$Y'_2 = a'_{21}X_1 + a'_{22}X_2 + ... + a'_{2p}X_p \qquad [6.4]$$

e così via. Trovata una rotazione tale che le componenti definite nella [6.4] siano interpretabili più velocemente di quelle date dalla [6.3], l'analisi migliora. Sono stati proposti parecchi criteri per individuare una rotazione adatta, ma il più usato è il metodo *varimax*. Scopo del metodo è suddividere, per ogni componente, i coefficienti in due gruppi, in modo che un gruppo presenti i valori più alti possibili e l'altro i valori più vicini possibile a zero. In altre parole ogni componente viene espressa in termini di un sottoinsieme delle variabili originali con la sovrapposizione minima delle variabili non appartenenti al sottoinsieme.

Le equazioni [6.3] e [6.4] possono essere utilizzate per assegnare valori ad ogni componente per ogni individuo. Tali valori si chiamano punteggi delle componenti e possono essere utilizzati per esempio come insieme ridotto di variabili indipendenti per un'analisi di regressione di alcune altre variabili non considerate nell'analisi delle componenti principali.

6.6 Analisi delle corrispondenze

E' un'analisi di tipo fattoriale che ha come scopo quello di individuare dimensioni soggiacenti alla struttura dei dati, dimensioni intese a riassumere l'intreccio di relazioni di "interdipendenza" tra le variabili originarie.

Si consideri una tabella di frequenze 6.1 derivante dall'osservazione di due variabili (una di riga, A, e una di colonna, B) su qualsiasi scala. Il metodo per determinare le coordinate geometriche delle modalità poste sulle righe e di quelle poste sulle colonne al fine di evidenziare il pattern della dipendenza tra i due insiemi di modalità è chiamata analisi delle corrispondenze semplice.

Tabella 6.1 Tavola di frequenze

	B_1	…	B_j	…	B_q	Totale marginale di colonna
A_1	n_{11}	…	n_{1j}	…	n_{1q}	$n_{1.}$
…	…	…	…	…	…	…
A_i	n_{i1}	…	n_{ij}	…	n_{iq}	$n_{i.}$
…	…	…	…	…	…	…
A_p	n_{p1}	…	n_{pj}	…	n_{pq}	$n_{p.}$
Totale marginale di riga	$n_{.1}$	…	$n_{.j}$	…	$n_{.q}$	n

Dalla tabella di contingenza vengono ricavate le matrici dei profili-riga e dei profili-colonna, come indicato nelle tabelle 6.2 e 6.3.

Tabella 6.2 Profili riga

	B_1	…	B_j	…	B_q
A_1	$r_{11} = n_{11}/n_{1.}$	…	$r_{1j} = n_{1j}/n_{1.}$	…	$r_{1q} = n_{1q}/n_{1.}$
…	…	…	…	…	…
A_i	$r_{i1} = n_{i1}/n_{i.}$	…	$r_{ij} = n_{ij}/n_{i.}$	…	$r_{iq} = n_{iq}/n_{i.}$
…	…	…	…	…	…
A_p	$r_{p1} = n_{p1}/n_{p.}$	…	$r_{pj} = n_{pj}/n_{p.}$	…	$r_{pq} = n_{pq}/n_{p.}$
Centroide	$r_1 = n_{.1}/n$	…	$r_j = n_{.j}/n$	…	$r_q = n_{.q}/n$

Tabella 6.3 Profili colonna

	B_1	…	B_j	…	B_q	Centroide
A_1	$c_{11} = n_{11}/n_{.1}$	…	$c_{1j} = n_{1j}/n_{.j}$	…	$c_{1q} = n_{1q}/n_{.q}$	$c_1 = n_{1.}/n$
…	…	…	…	…	…	…
A_i	$c_{i1} = n_{i1}/n_{.1}$	…	$c_{ij} = n_{ij}/n_{.j}$	…	$c_{iq} = n_{iq}/n_{.q}$	$c_i = n_{i.}/n$
…	…	…	…	…	…	…
A_p	$c_{p1} = n_{p1}/n_{.1}$	…	$c_{pj} = n_{pj}/n_{.j}$	…	$c_{pq} = n_{pq}/n_{.q}$	$c_p = n_{p.}/n$

Il passaggio successivo è valutare la dispersione dei profili, riga e colonna, rispetto al loro centroide utilizzando la metrica del χ^2, nel modo seguente:

$$\chi^2 = \sum_{i=1}^{p} n_i . d_i^2 = \sum_{i=1}^{p} n_i . \sum_{j=1}^{q} \frac{\left(r_{ij} - r_j\right)^2}{r_j}$$

$$\chi^2 = \sum_{j=1}^{q} n_{.j} d_j^2 = \sum_{j=1}^{q} n_{.j} \sum_{i=1}^{p} \frac{\left(c_{ij} - c_i\right)^2}{c_i} \qquad [6.5].$$

La variabilità totale presente nella matrice dei dati viene denominata *inerzia* ed è proporzionale al χ^2 calcolato sulla tabella di partenza: $INERZIA = \frac{\chi^2}{n}$. L'obiettivo dell'analisi delle corrispondenze è spiegare questa inerzia (varianza). Come nell'analisi fattoriale, vengono estratti degli assi fattoriali (f_i), ortogonali tra loro, che spiegano ciascuno, in ordine decrescente, il massimo della variabilità della matrice dei dati. Esiste un numero massimo di fattori estraibili (k) dato dal minimo tra il numero di modalità riga e colonna meno uno (k = min(p, q) - 1). La quota di inerzia spiegata da ciascun fattore è proporzionale all'autovalore (λ_i) associato al fattore stesso: $INERZIA\ \ SPIEGATA\ \ (f_i) = \frac{\lambda_i}{INERZIA} = \frac{\lambda_i}{\sum_{i=1}^{k} \lambda_i}$.

L'analisi multipla, o composita, è invece l'analisi delle relazioni presenti in un numero qualsiasi di variabili, nel caso tipico di una matrice n x p con n unità statistiche e p variabili osservate su scala qualsiasi. La soluzione si determina applicando l'analisi fattoriale su un'opportuna trasformazione dei dati di partenza. L'analisi delle corrispondenze è una metodica di analisi più duttile di quella fattoriale, considerato che dà la possibilità di scegliere le unità e le variabili su cui svolgere le elaborazioni statistiche, utilizzando le altre unità e/o le altre variabili osservate a fini di conferma o di approfondimento dell'esito delle analisi. Le variabili osservate si possono cioè ripartire in due categorie, quelle dette "*attive*", da impiegare per la ricerca delle componenti principali, e quelle dette "*illustrative*", utilizzabili per un supplemento di analisi per la ricerca di relazioni con le modalità del primo insieme. Nella suddivisione in due parti dell'insieme di unità in esame, una è usata per la ricerca della soluzione (ricerca delle componenti), l'altra per paragone (proiezione sugli stessi assi per analisi suppletive).

Dal punto di vista matematico, l'analisi delle corrispondenze consiste nella ricerca dei fattori presenti nei dati osservati attraverso la metodica della ricerca degli autovalori e autovettori di trasformate dei dati osservati. Per questo è appropriato chiamarla analisi fattoriale delle corrispondenze.

6.7 Cluster analysis

L'analisi di raggruppamento si distingue dall'analisi fattoriale perché la prima è pertinente per raggruppare entità, mentre la seconda è appropriata per lo studio delle relazioni tra variabili. Inoltre, l'analisi fattoriale assume che le relazioni tra le variabili inserite nel modello d'analisi, siano lineari, mentre la forma delle relazioni tra variabili è trascurabile nella cluster analysis.

Questo non esclude che si possa arrivare a identiche conclusioni adottando l'uno o l'altro metodo. Con opportune elaborazioni, si possono rendere evidenti, dopo aver eseguito una cluster analysis, le variabili più discriminanti tra le entità, e dopo un'analisi dei fattori, le unità che più sono simili o dissimili con riferimento ai fattori trovati.

Le tecniche per il raggruppamento di entità si possono suddividere in due grandi categorie, secondo che i gruppi, che rappresentano l'esito dell'analisi, siano o no gerarchizzabili.

In un'analisi gerarchica dei gruppi, ogni classe fa parte di una classe più ampia, la quale è contenuta a sua volta in una classe di ampiezza superiore, e così in progressione fino alla classe che contiene l'intero insieme di entità analizzate.

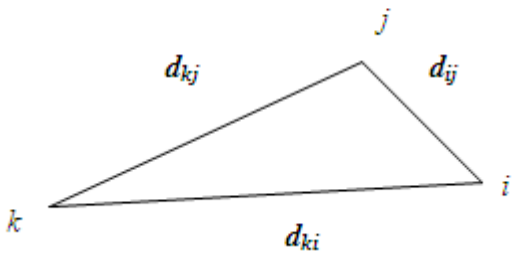
Non gerarchiche sono le tecniche che generano gruppi non gerarchizzabili. Per queste, si deve decidere a priori il numero di gruppi, oppure si deve eseguire l'analisi in modo da ottenere soluzioni per un numero di gruppi variabile.

Le tecniche di analisi gerarchica si possono ulteriormente distinguere in:

- agglomerative, se procedono ad una successione di fusioni delle n unità, a partire dalla soluzione di base nella quale ognuna costituisce un gruppo a sé stante e fino allo stadio n-1 nel quale si forma un gruppo che le comprende tutte.
- scissorie, quando l'insieme delle n unità, in n-1 passi, si ripartisce in gruppi che sono, ogni passo dell'analisi, sottoinsiemi di un gruppo formato allo stadio di analisi precedente, e che termina con la situazione in cui ogni gruppo è composto da un'unità.

La tecnica d'analisi più utilizzata è quella agglomerativa. Data una matrice simmetrica di prossimità (o di vicinanze) tra n entità, si trova la coppia di entità più prossime e con queste si forma un gruppo. Tra le entità che fanno parte del gruppo si assume distanza nulla, e la distanza tra questo nuovo gruppo e le rimanenti entità è unica. Il modo in cui si calcola la distanza tra il gruppo e le rimanenti entità dipende dalla strategia di aggregazione scelta. Per aggregare un'altra entità, si individua nella matrice di prossimità, diventata di ordine $n - 1$, l'entità più prossima. Si formerà così un altro gruppo e si calcoleranno di nuovo le distanze tra il gruppo formato le entità rimaste. L'individuazione dell'entità più prossima e il ricalcolo delle distanze si ripeterà $n - 1$ volte finché tutte le unità faranno parte di un unico gruppo. Si supponga, ad un certo stadio dell'analisi, di aggregare le entità i e j (si può trattare di unità singole o di gruppi composti da più unità) e di voler calcolare la distanza tra il nuovo gruppo (i, j) e una qualsiasi entità esterna k. Le tre entità di dimensione (numerosità) n_i, n_j, n_k, possono essere raffigurate su un triangolo avente per vertici i, j, k e per cateti le misure di distanza d_{ki}, d_{kj} e d_{ij} tra le entità (figura 6.1).

Figura 6.1 Rappresentazione grafica delle distanze tra le entità i, j, k



La distanza $d_{k(ij)}$ tra l'entità k e il gruppo (i, j) si calcola combinando le distanze d_{ki}, d_{kj} e d_{ij} con pesi che differiscono in ragione del criterio di aggregazione scelto. Alcuni di questi metodi si presentano qui di seguito.

– *Metodo della media di gruppo* per cui la distanza $d_{k(ij)}$ è data dalla media aritmetica delle distanze d_{ki} e d_{kj} ponderate con la numerosità delle unità appartenenti ai gruppi *i* e *j*:

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} \text{ con } i \neq j \neq k = 1,2,...,n \qquad [6.6]$$

dove $\alpha_i = n_i /(n_i + n_j)$, $\alpha_j = n_j /(n_i + n_j)$, mentre d_{ki} e d_{kj} sono due misure qualsiasi di dissomiglianza.

– *Metodo del centroide*, con cui la distanza tra due gruppi è la distanza euclidea tra i centroidi dei gruppi (Sokal & Michener, 1958):

$$d_{k(ij)} = \sqrt{\alpha_i d_{ik}^2 + \alpha_j d_{jk}^2 - \alpha_i \alpha_j d_{ij}^2} \text{ con } i \neq j \neq k = 1,2,...,n \qquad [6.7].$$

– *Metodo della mediana*. La distanza tra due entità *i* e *j* che si fondono nel gruppo (i, j) ed una terza entità *k* è data da (Grower, 1966):

$$d_{k(ij)} = d_{ki}/2 + d_{kj}/2 - d_{ij}/4 \text{ con } i \neq j \neq k = 1,2,...,n \qquad [6.8]$$

dove le distanze sono euclidee.

– *Metodo del legame singolo* (o *del vicino più prossimo*). Con questa strategia la distanza tra l'entità *k* e la nuova fusione (i, j) è la distanza minore tra *k* e le due entità aggregate (Johnson, 1967):

$$d_{k(ij)} = \min\{d_{ki}, d_{kj}\} \text{ con } i \neq j \neq k = 1,2,...,n \qquad [6.9],$$

la quale in riferimento alla figura 6.1 è d_{kj}.

– *Metodo del legame completo* (o *del vicino più lontano*). Si contrappone, come logica e risultati a quello del legame singolo. Tra l'entità esterna k e il gruppo di nuova formazione (i, j), la distanza è, infatti, rappresentata dal valore più elevato tra d_{ki} e d_{kj} (MacNaughton-Smith, 1965):

$$d_{k(ij)} = \max\{d_{ki}, d_{kj}\} \text{ con } i \neq j \neq k = 1,2,...,n \qquad [6.10],$$

che in riferimento alla figura 6.1 è rappresentata da d_{ki}.

Il concatenamento tra le classi generate da un qualsiasi criterio di aggregazione gerarchico si può rappresentare mediante un diagramma ad albero (*dendrogramma*) su un sistema di assi cartesiani, con le entità in ascissa ed i livelli delle prossimità tra entità, volta per volta aggregate, in ordinata.

6.8 Analisi di regressione *stepwise*

Si supponga di aver osservato una variabile di outcome Y e p variabili esplicative X_i (i = 1,2,…,p). L'analisi di regressione *stepwise* cerca di determinare la funzione di regressione lineare $y = f(x)$ che, pur contenendo il minor numero di variabili esplicative, è capace di interpretare nel migliore dei modi in senso statistico, la variabilità di Y. La procedura sviluppata da Garside (1965) mira a selezionare il sottoinsieme di variabili ottimale tra quelli possibili, immettendo o togliendo dall'equazione di regressione una variabile esplicativa alla volta. Il modello di analisi di regressione *stepwise* (o a gradini) assume che ogni osservazione y_h sia esprimibile come una combinazione lineare delle x con coefficienti β e di una variabile non osservata ε, interpretabile come errore residuale della regressione:

$$y_h = \beta_0 + \sum_{i=1}^{p} \beta_i x_{hi} + \varepsilon_h \qquad (h = 1,2,\ldots,n) \qquad [6.11].$$

Per realizzare un'analisi di regressione *stepwise* è necessario compiere alcune scelte prima di iniziare. Dopo aver deciso quali sono le variabili esplicative candidate alla selezione, si deve:

1. decidere quale criterio di selezione adottare. I principali criteri proposti in letteratura sono tre:
   – la selezione progressiva (in inglese *forward selection*) che consiste nell'inserire una variabile esplicativa per volta nell'equazione di regressione che inizialmente è semplicemente $y_h = \beta_0 + \varepsilon_h$ dove β_0 è il termine noto. La selezione si basa sul contributo della variabile inserita alla spiegazione della

variabilità di *y*, individuato come la massima riduzione di devianza. Una volta entrata la prima variabile il modello si presenta come $y_h = \beta_0 + \beta_1 x_{h1} + \varepsilon_h$. Il processo di selezione continuerà fino a quando non viene soddisfatto un criterio di arresto della procedura, che può essere un limite massimo di predittori inseriti, il raggiungimento di una data frazione di devianza complessivamente spiegata (o spiegata dall'ultimo predittore), la valutazione della significatività statistica del contributo dei predittori all'interpretazione della variabilità di *y* con il test F di Snedecor.

- L'eliminazione a ritroso (in inglese *backward elimination*) che consiste nel rimuovere una variabile alla volta dall'equazione di regressione con *p* variabili, in ragione della minore perdita di capacità esplicativa della variabilità di *y* conseguente all'eliminazione della variabile. Il processo si arresta quando viene soddisfatto un criterio per troncare il processo di eliminazione.
- La regressione *stepwise* convenzionale (in inglese *stepwise regression analysis*), che è una combinazione delle due procedure precedenti. Una variabile candidata è inclusa nell'equazione se, in una fase del processo, dà il contributo più significativo all'interpretazione della variabilità di *y*, ma può venire rimossa, in un'altra fase d'analisi, qualora la sua capacità esplicativa risulti surrogata da altri predittori entrati nel frattempo. Quando si parla di regressione *stepwise* si fa riferimento a questo metodo.

2. Definire l'insieme di dati su cui svolgere l'analisi. Si tratta, in particolare, di stabilire quali sono le variabili *dummy* da creare con le variabili qualitative, se e come trasformare le variabili per cui fosse evidente la non-linearità della relazione con la variabile di outcome. Questi aspetti sono stati affrontati nel capitolo 4.

3. Stabilire quali parametri seguire nel valutare gli esiti del processo. In alcune ricerche può capitare che il modello di regressione multipla si dimostri insoddisfacente, sia per l'esiguità della varianza spiegata dall'equazione, sia per l'instabilità delle stime dei coefficienti di regressione, e ciò va contro l'intento di estendere l'equazione a modelli estrapolativi dei risultati dell'analisi. Quando il modello è poco significativo, nella maggior parte delle volte è a causa delle variabili esplicative osservate che risultano inadeguate a riprodurre il fenomeno oggetto di studio, oppure a causa della scarsa accuratezza dei dati di base. Nell'interpretare i risultati dell'analisi, per maggior

sicurezza, si dovrebbe ripeterla su un insieme simile di dati e le conclusioni dovrebbero essere tratte sulla base di risultati concordi nelle due analisi. Uno dei modi di procedere per ottenere insiemi confrontabili di unità è la suddivisione casuale dell'insieme di dati osservato in due metà, che dovranno poi essere analizzate con le medesime regole (Fabbris, 1997).

6.9 Analisi di regressione logistica

Questo metodo d'analisi verrà ampiamente trattato nel prossimo capitolo, quindi si fornirà ora solamente una spiegazione ristretta del suo utilizzo. L'analisi di regressione logistica è una delle applicazioni del più generale metodo di analisi della regressione. Si applica quando la variabile di outcome *y* è dicotomica, ossia rappresenta un attributo (1 = possesso/presenza, 0 = non possesso/assenza), e si vuole spiegare il cosiddetto *logit* della frequenza di *y* nella popolazione. Per *logit* di π si intende il logaritmo del rapporto $\pi/(1-\pi)$, dove π è la frequenza attesa dell'attributo di *y*. La funzione di regressione logistica si presenta come segue:

$$logit(\pi(x)) = \beta_0 + \sum_{i=1}^{q} \beta_i x_i = X\beta \qquad [6.12].$$

Si spendono due parole per presentare molto brevemente un tipo di analisi di regressione logistica nel caso in cui la variabile di outcome sia qualitativa e presenti più di due modalità. In questa situazione si è visto che si ricorre all'analisi di regressione logistica multinomiale. L'analisi di regressione logistica multinomiale procede col confronto dell'effetto delle variabili esplicative sulla possibilità di spiegare ogni $n - 1$ categoria rispetto ad una categoria di riferimento della variabile di outcome. Questo equivale ad analizzare $n - 1$ modelli di regressione logistica con variabile dipendente dicotomica.

6.10 Analisi discriminatoria

L'analisi discriminatoria, o analisi discriminante, tratta uno specifico problema di analisi multivariata: assegnare un'osservazione ad un gruppo con un baso tasso d'errore. Si supponga che vi siano *k* gruppi con n_i unità nel gruppo i-esimo e che su ogni unità siano rilevate *p* variabili $X_1, X_2, \ldots, X_p$. Si richiede una regola per discriminare tra i gruppi, cioè una regola che si possa applicare ad ogni nuova unità, della quale si sappia che proviene da uno dei gruppi, ma non precisamente da quale, e che permetta di assegnarlo al gruppo corretto. Nell'analisi discriminante si deve partire con l'idea che in qualche modo si è in grado di definire i gruppi,

individuando le variabili che permettano di stabilirli. A volte i gruppi possono essere determinati con procedure molto sofisticate, ma difficili da operare; l'analisi discriminatoria vuole verificare l'attendibilità degli stessi risultati utilizzando le informazioni delle *X* più economiche e meno impegnative per il paziente. Per esempio, in uno studio sul cancro, una biopsia potrebbe essere usata per definire i gruppi "presenza di cancro" e "assenza di cancro", ma dato che tale procedura è molto costosa e scomoda, risulta preferibile effettuarla solo su quei pazienti che hanno una più alta probabilità di manifestare la patologia: questa probabilità è individuabile attraverso l'analisi discriminante.

Prima di iniziare l'analisi vera e propria è importante verificare la validità dell'assunto di normalità multivariata dei gruppi per specifiche variabili. Se si usa l'analisi discriminante in situazioni per cui non è verificato l'assunto di normalità multivariata, si può rischiare di ottenere risultati errati e non veritieri (Lachenbruch, 1975).

Si prende in considerazione in questo paragrafo il caso con $k = 2$, denotando i due gruppi con Π_1 e Π_2. Si osserva un gruppo di variabili per ciascuna unità: questo insieme viene identificato dal vettore *x* di dimensioni *m* x 1 e si vuole assegnare l'unità, le cui misure sono date da *x*, a Π_1 o Π_2. Serve una regola per assegnare *x* a Π_1 o Π_2. Se i parametri delle distribuzioni di *x* in Π_1 e Π_2 sono noti, è possibile utilizzare questa informazione nella costruzione della regola d'assegnazione, altrimenti è possibile usare due campioni di dimensioni n_1 e n_2 presi da Π_1 e Π_2 per stimarli. Serve poi un criterio di bontà della classificazione. Fisher (1936) ha suggerito l'uso di una combinazione lineare delle osservazioni e la scelta dei coefficienti in modo che il rapporto della differenza delle medie della combinazione lineare nei due gruppi diviso la loro varianza sia massimizzato. Denotando la combinazione lineare con $Y = \lambda' x$, la media di *Y* è data da $\lambda'\mu_1$ in Π_1 e $\lambda'\mu_2$ in Π_2 mentre la varianza è data da $\lambda'\Sigma\lambda$ in ogni gruppo se si assume che le matrici di covarianze $\Sigma_1 = \Sigma_2 = \Sigma$. Si vuole scegliere λ in modo da massimizzare la seguente equazione:

$$\phi = \frac{(\lambda'\mu_1 - \lambda'\mu_2)^2}{\lambda'\Sigma\lambda} \qquad [6.13].$$

Differenziando ϕ rispetto a λ si ottiene:

$$\frac{\partial\phi}{\partial\lambda} = \frac{2(\mu_1 - \mu_2)\lambda'\Sigma\lambda - 2\Sigma\lambda(\lambda'\mu_1 - \lambda'\mu_2)}{(\lambda'\Sigma\lambda)^2} = 0 \qquad [6.14]$$

che dà:

$$\mu_1 - \mu_2 = \Sigma\lambda\left(\frac{\lambda'\mu_1 - \lambda'\mu_2}{\lambda'\Sigma\lambda}\right) \qquad [6.15].$$

Dal momento che si usa λ solo per separare i gruppi, è possibile moltiplicare λ per una qualsiasi costante. Di conseguenza λ è proporzionale a $\Sigma^{-1}(\mu_1 - \mu_2)$. Se i parametri non sono noti si possono stimare con $\bar{x}_1, \bar{x}_2$ ed S. La procedura di assegnazione concerne l'assegnare un'unità al gruppo Π_1 se $\overline{Y_1} = (\bar{x}_1 - \bar{x}_2)' S^{-1} \bar{x}_1$ ed assegnare l'unità al gruppo Π_2 altrimenti.
Per controllare se sono presenti differenze significative tra i due gruppi si può usare la distribuzione di $D^2 = Var(Y) = \lambda' S \lambda$, per cui la variabile

$$F = \frac{n_1 n_2 (n_1 + n_2 - k - 1)}{(n_1 + n_2)(n_1 + n_2 - 2)k} D^2 \qquad [6.16]$$

dove n_1 e n_2 sono le numerosità campionarie dei gruppi Π_1 e Π_2, rispettivamente, e k è il numero di variabili, ha una distribuzione F con k e $n_1 + n_2 - k - 1$ gradi di libertà.
Il metodo appena descritto risulta ottimale nel caso in cui le osservazioni hanno una distribuzione multivariata normale.
Altri metodi per la bontà d'adattamento sono stati presentati: si citano Welch (1939), che suggerisce di minimizzare la probabilità totale di errata classificazione (in inglese *misclassification*), e Von Mises (1945) il quale consiglia invece di minimizzare la massima probabilità di *misclassification* nei due gruppi. Una chiara discussione su queste possibili scelte è riportata in Anderson (1958).

6.11 Applicazione. Scelta del metodo d'analisi dei dati nel lavoro svolto al Children's Hospital dell'Università di Oulu – Finlandia

Lo studio progettato come caso-controllo evidenzia una netta distinzione tra variabile dipendente e variabile/i indipendente/i, indicando quindi una asimmetria del modello d'analisi. In particolare si vuole spiegare il manifestarsi della patologia RDS attraverso le covariate cliniche e genetiche raccolte come precedentemente descritto. Si punta quindi ad un'analisi di regressione, escludendo i metodi d'analisi che pongono la totalità delle variabili sullo stesso piano mirando alla ricerca di una o più relazioni "nascoste". Il fatto che la variabile dipendente sia dicotomica suggerisce la scelta dell'analisi di regressione logistica multipla. Nel caso in cui la variabile di outcome fosse stata quantitativa si sarebbe optato per la regressione lineare multipla, mentre nel caso si fosse trattato di una variabile dipendente categoriale non dicotomica si sarebbe scelta l'analisi di regressione logistica con metodo politomico (o multinomiale). Esistono dei metodi specifici di regressione logistica per il

Matched Case-Control Study che è appunto quello che è stato affrontato nel lavoro al Children's Hospital dell'Università di Oulu.
Nel prossimo capitolo si presenterà l'analisi di regressione logistica in un primo momento, e in un secondo momento si approfondirà l'aspetto del *Matched Case-Control Study*.

6.12 Conclusioni

Si sono presentati i principali metodi d'analisi multivariata utilizzati negli studi clinici. Altre volte, però, l'obiettivo della ricerca può richiedere una più semplice tecnica per rispondere ai quesiti centrali dello studio, come per esempio una verifica d'ipotesi applicando dei test statistici di significatività. È quindi fondamentale preparare un disegno di studio che specifichi nel migliore dei modi gli obiettivi e le ipotesi di partenza che si vogliono testare, in modo che la raccolta delle informazioni sia legata a tali scopi, e, di conseguenza, anche l'analisi venga condotta sulla base dei dati ottenuti e delle ipotesi da vagliare.
Tutti i metodi d'analisi presentati in questo capitolo sono molto complicati se affrontati manualmente con grandi quantità di dati. Esistono diversi package statistici che svolgono il lavoro più arduo, lasciando al ricercatore il compito della programmazione e dell'interpretazione dei risultati sotto forma di output. I sistemi di programmazione più noti e diffusi nel campo clinico epidemiologico sono il Sas System, l'SPSS, STATA, il package STATISTICA, S-PLUS ed il software R.
Per le elaborazioni svolte nell'applicazione al Children's Hospital dell'Università di Oulu sono stati utilizzati i programmi Sas System ed SPSS.

Sezione E

Esempio di metodo di analisi: la regressione logistica

CAPITOLO 7 *Analisi di Regressione Logistica*

7.1 Premessa

L'utilizzo dei modelli di regressione logistica è esploso durante le ultime decadi. Dalla sua originale accettazione nella ricerca epidemiologica, il metodo è oggi comunemente usato in diversi campi, quali la ricerca biomedica, l'ecologia, le politiche sociali, l'ingegneria, la finanza, la criminologia e la linguistica (Hosmer & Lemeshow, 2000). Nello stesso tempo c'è stato un egual incremento di ricerche su tutti gli aspetti statistici relativi al modello di regressione logistica. Se i metodi di regressione sono la componente di ogni analisi concernente la relazione tra una variabile risposta ed una o più variabili esplicative, l'analisi di regressione logistica si adatta ai casi in cui la variabile dipendente assume due valori, si adatta cioè ad una variabile risposta dicotomica o binaria. In realtà il modello di regressione logistica può essere facilmente modificato per trattare il caso in cui la variabile risposta sia nominale con più di due livelli. Questo modello di regressione logistica è detto *multinomiale* e non verrà trattato in questo lavoro. L'obiettivo dell'analisi di regressione logistica è trovare il miglior adattamento e il modello più conveniente e parsimonioso per descrivere la relazione tra una variabile risposta e una variabile esplicativa (regressione logistica univariata) o un set di variabili esplicative (regressione logistica multipla o multivariata).

Nel presente capitolo si presenterà il modello d'analisi di regressione multipla (paragrafo 7.2), il suo adattamento (paragrafo 7.3) attraverso la stima degli r parametri β_0, β_1,...,β_r con il metodo della *log verosimiglianza* e l'analisi della sua significatività (paragrafo 7.5) attraverso il *test del rapporto di verosimiglianza*. Attraverso la stima degli errori standard per i coefficienti β stimati (paragrafo 7.4) è possibile calcolarne anche gli intervalli di confidenza (paragrafo 7.6). La seconda parte del capitolo mira a descrivere il modello d'analisi di regressione logistica per il disegno *Matched Case-Control Study* (paragrafo 7.7) adottato nel lavoro al Children's Hospital dell'Università di Oulu e necessario data la stratificazione del campione di pazienti secondo le variabili sesso, età gestazionale e terapia glucocorticoide. Tali variabili di disturbo non vengono inserite nel modello d'analisi che viene denominato "di regressione logistica condizionata" perché utilizza la verosimiglianza condizionata nel calcolo dei coefficienti.

L'ultimo passaggio del presente capitolo concerne l'applicazione del modello di regressione logistica condizionata ai dati dello studio al Children's Hospital dell'Università di Oulu attraverso l'utilizzo del programma statistico Sas System (paragrafo 7.8).

7.2 Definizione del modello d'analisi di Regressione Logistica Multipla

Data l'esigenza di disporre di un modello matematico che esprima il legame tra X (vettore di variabili esplicative) e Y (variabile dipendente) statisticamente si ricorre all'analisi di regressione (Magagnoli U., 1993).

Innanzitutto occorre definire la spezzata di regressione [μ_y (x_i)], la quale indica come varia in media Y al mutare di X. Proprio perché si è alla ricerca di un modello matematico, l'obiettivo dell'analisi di regressione sarà quello di trovare una funzione analitica, y = y(x), in grado di sostituire la spezzata di regressione osservata. Più in particolare si giungerà alla formulazione seguente:

$$y(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_r x_r \qquad [7.1],$$

avente come secondo membro dell'equazione una combinazione lineare di r variabili con r + 1 parametri.

La [7.1] si ottiene attraverso il procedimento di interpolazione per punti della spezzata di regressione, ovviamente senza passare però per tutti i punti in questione, in quanto diverrebbe troppo oneroso e ciò comporterebbe difficoltà soprattutto dal punto di vista interpretativo: si procede di conseguenza definendo un "criterio di accostamento" dei dati al modello, propriamente detto criterio dei minimi quadrati, il quale si specifica nel minimizzare la seguente funzione di perdita quadratica:

$$\sum \left[y(x_i) - \mu_y(x_i) \right]^2 \cdot n_i \qquad [7.2],$$

dove n_i sta ad indicare la numerosità delle entità osservate per ogni modalità della variabile X.

Stabilite queste semplici nozioni introduttive va però sottolineata una considerazione rilevante: data la [7.1] sorge un problema quando la variabile dipendente è una probabilità. Infatti, proprio perché si sta trattando una probabilità, la grandezza che rappresenta il fenomeno di cui si vuole studiare la dipendenza da altri fattori può assumere valori compresi unicamente tra 0 ed 1. Questa condizione, però, non può certo essere rispettata dal secondo membro della [7.1], che invece può raggiungere valori esterni all'intervallo 0-1.

Con lo scopo di superare questo ostacolo occorre allora attuare una trasformazione della variabile Y, in modo da poter trattare quella particolare forma di regressione che in statistica prende il nome di regressione logistica.

Quello di regressione logistica è dunque un caso speciale dell'analisi di regressione, che trova applicazione quando la variabile dipendente è dicotomica, mentre l'analisi di regressione

lineare si applica se la variabile dipendente è quantitativa. L'attributo studiato può essere dicotomico in natura o dicotomizzato a fini d'analisi (Fabbris, 1997).
Oltre che per la scala di misura della variabile dipendente, l'analisi della regressione logistica si differenzia da quella lineare perché per quest'ultima si ipotizza una distribuzione normale di Y, mentre se Y è dicotomica la sua distribuzione è ovviamente binomiale.
Dato il vettore **x'** di q variabili predittive, la stima di Y nell'analisi della regressione logistica assume, come già si è detto, il significato di probabilità che Y sia uguale a 1: $P(Y = 1|\mathbf{x'}) = \pi(\mathbf{x'})$.
La funzione di regressione logistica si presenta come segue:

$$logit(\pi(\mathbf{x'})) = \beta_0 + \sum_{i=1}^{q} \beta_i x_i = \mathbf{x'\beta'} \qquad [7.3],$$

dove $logit(\pi(\mathbf{x}))$ denota il logaritmo naturale del rapporto fra probabilità di "successo" e probabilità di "insuccesso" dato il vettore **x** di q variabili predittive:

$$logit(\pi(\mathbf{x'})) = \ln\left[\frac{\pi(x')}{1-\pi(x')}\right] \qquad [7.4]$$

e $\pi(\mathbf{x'})$ denota la probabilità che Y valga 1 in funzione del vettore di variabili esplicative **x'**.
La scelta del *logit* per descrivere la funzione che lega la probabilità di Y alla combinazione delle variabili predittive è determinata dalla constatazione che la probabilità si avvicina ai limiti 0 e 1 gradualmente e descrive una figura a S (detta sigmoide) che assomiglia alla cumulata della distribuzione casuale degli errori detta "funzione logistica".
La probabilità si può infatti scrivere come una funzione logistica:

$$\pi(\mathbf{x'}) = \frac{e^{x'\beta'}}{1+e^{x'\beta'}} \qquad [7.5].$$

Pur non essendo il *logit* l'unica funzione che consente di modellare la probabilità di un fenomeno, essa è privilegiata, dato che è una trasformata del rapporto tra due probabilità complementari, ovvero tra il numero di successi per ogni insuccesso del fenomeno in esame.
Si apre ora una parentesi relativamente al tipo di variabili indipendenti che si prendono in considerazione. Se alcune covariate sono espresse su scala nominale, come per esempio il sesso, la professione o il gruppo di trattamento, non è corretto includerle nel modello di

regressione logistica come se fossero variabili su scala ordinale. I numeri utilizzati per rappresentare le diverse categorie di queste variabili nominali sono meri identificatori e non hanno alcun significato numerico. In questa situazione il metodo più opportuno riguarda l'utilizzo di una serie di variabili di disegno, meglio conosciute come *variabili dummy*. Le *variabili dummy* sono variabili dicotomiche che presentano valore 1 per la categoria di appartenenza e 0 per tutte le altre, mentre per la categoria di riferimento si avranno tutti valori pari a 0. Se si hanno k categorie in cui si presenta una variabile, basterà creare (k-1) *variabili dummy*, escludendo la categoria presa come riferimento.

7.3 Adattamento del modello di Regressione Logistica Multipla

Si assuma di avere un campione di n osservazioni indipendenti ($\mathbf{x}_i$, y_i), i=1,2,…,n. L'adattamento del modello richiede che si ottengano le stime del vettore $\boldsymbol{\beta}' = (\beta_0, \beta_1, ..., \beta_r)$ con il metodo della massima verosimiglianza. Un modo conveniente di esprimere il contributo alla funzione di verosimiglianza della coppia ($\mathbf{x}_i$, y_i) è dato dalla seguente espressione:

$$\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \qquad [7.6].$$

Dal momento che le osservazioni sono supposte essere indipendenti, la funzione di verosimiglianza è ottenuta come il prodotto dei termini dati nell'espressione [7.6] come segue:

$$l(\boldsymbol{\beta}') = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \qquad [7.7].$$

Il principio di massima verosimiglianza sta nell'usare come stima di β il valore che massimizza l'espressione [7.7]. Tuttavia è matematicamente più semplice lavorare con il logaritmo naturale dell'equazione [7.7]. Questa espressione, detta *log verosimiglianza*, è definita come:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\} \qquad [7.8].$$

Ci saranno $r + 1$ equazioni di verosimiglianza e per trovare il valore di $\hat{\beta}$ che massimizzi $L(\beta)$ si differenzia $L(\beta)$ rispetto agli $r + 1$ coefficienti. Le *equazioni di verosimiglianza* che risultano possono essere espresse come segue:

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0$$

e

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0 \text{ per } j=1,2,\ldots,r.$$

Sia $\hat{\beta}$ la soluzione di queste equazioni. Di conseguenza, i valori adattati al modello di regressione logistica sono $\hat{\pi}(x_i)$, vale a dire il valore dell'espressione [7.5] ottenuto utilizzando $\hat{\beta}$ e x_i.

7.4 Stima degli Errori Standard per i coefficienti β stimati

Il metodo di stima di varianze e covarianze dei coefficienti stimati segue la teoria della stima della massima verosimiglianza (Rao, 1973). Questa teoria attesta che gli stimatori sono ottenuti dalla matrice delle derivate seconde parziali della funzione di *log verosimiglianza*. Queste derivate parziali hanno la seguente forma generale:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j{}^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i) \qquad [7.9]$$

e

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1 - \pi_i) \qquad [7.10]$$

per $j,\ l = 0,1,2,\ldots,r$ dove π_i denota $\pi(x_i)$. Si denoti come $\mathbf{I}(\boldsymbol{\beta})$ la matrice $(r + 1) \cdot (r + 1)$ contenente i termini negativi delle equazioni [7.9] e [7.10]. Questa matrice è chiamata *matrice delle informazioni osservate*. Le varianze e le covarianze dei coefficienti stimati sono ottenute dalla matrice inversa di $\mathbf{I}(\boldsymbol{\beta})$ che si denota con $\text{Var}(\boldsymbol{\beta}) = \mathbf{I}^{-1}(\boldsymbol{\beta})$. Si utilizza la notazione $\text{Var}(\beta_j)$ per esprimere il j^{mo} elemento diagonale di questa matrice, che rappresenta la varianza di $\hat{\beta}_j$, e

Cov(β_j,β_l) per denotare un elemento arbitrario fuori dalla diagonale della matrice: quest'ultimo elemento rappresenta la covarianza di $\hat{\beta}_j$ e $\hat{\beta}_l$. Gli stimatori di varianze e covarianze, che verranno denotati rispettivamente da $\hat{Var}(\hat{\beta}_j)$ e $\hat{\text{Cov}}(\hat{\beta}_j, \hat{\beta}_l)$ $j, l = 0, 1, 2,\ldots,$ r, sono ottenuti valutando Var(β_j) con $\hat{\beta}_j$ e Cov(β_j, β_l) con $\hat{\beta}_j$ e $\hat{\beta}_l$. Si utilizzeranno principalmente gli errori standard stimati dei coefficienti stimati, che si denotano con:

$$\hat{SE}(\hat{\beta}_j) = [\hat{Var}(\hat{\beta}_j)]^{1/2} \text{ per } j = 0, 1, 2, \ldots, \text{r.}$$

Una formulazione della matrice delle informazioni osservate è $I(\hat{\beta}) = X'VX$ dove X è una matrice $n\cdot(r + 1)$ contenente i dati per ogni soggetto, e V è una matrice diagonale $n\cdot n$ con gli elementi generali $\hat{\pi}_i (1 - \hat{\pi}_i)$. La matrice X è la seguente:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots x_{1r} \\ 1 & x_{21} & x_{22} & \ldots x_{2r} \\ : & : & : & \ldots : \\ 1 & x_{n1} & x_{n2} & \ldots x_{nr} \end{bmatrix}$$

e la matrice V è:

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \ldots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \ldots & 0 \\ : & 0 & \ldots & : \\ 0 & \ldots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

7.5 Analisi della significatività del modello (significatività dei coefficienti)

Una volta adattato un particolare modello multivariato di regressione logistica, si inizia il processo di valutazione del modello stesso. Il primo passo in questo processo è solitamente l'analisi della significatività delle variabili nel modello. Solitamente questo implica la formulazione e la verifica di un'ipotesi statistica per determinare se le variabili indipendenti nel modello siano "significativamente" relazionate con la variabile risposta. Il metodo per eseguire questo test di verifica d'ipotesi è abbastanza generale e differisce da un modello a un altro solo per dettagli specifici.

L'approccio su cui si basa il test sulla significatività del coefficiente di ciascuna variabile in ogni modello è legato alla seguente domanda: *il modello che include la variabile in questione fornisce maggiori informazioni circa la variabile risposta rispetto al modello che non include quella stessa variabile?* (Hosmer & Lemeshow, 2000). La risposta a questa domanda si trova confrontando i valori osservati sulla variabile dipendente con quelli attesi per ciascuno dei due modelli, il primo con la variabile indipendente in questione, il secondo senza. La funzione matematica utilizzata per comparare i valori osservati con quelli attesi dipende dal problema particolare. Se i valori attesi nel modello con la variabile inclusa sono migliori, o più accurati in un certo senso, di quelli ottenuti quando la variabile non è nel modello, allora la variabile in questione è "significativa". È importante sottolineare che non si sta valutando se i valori attesi siano una rappresentazione accurata dei valori osservati in un senso assoluto (verrebbe chiamata *bontà d'adattamento*), ma la domanda è posta in un senso relativo.

Il metodo generale per valutare la significatività delle variabili è facilmente spiegato nel modello di regressione lineare. Un confronto tra questo approccio e quello utilizzato nell'analisi di regressione logistica evidenzierà le differenze tra modelli con variabile di risposta continua e modelli con variabile risposta dicotomica.

Nella regressione lineare, la valutazione della significatività relativa ai coefficienti dell'inclinazione della curva viene svolta attraverso la *tabella d'analisi della varianza*. Questa tabella divide in due parti le somme totali delle deviazioni al quadrato riferite alle medie delle osservazioni:

- la somma delle deviazioni al quadrato relative alle osservazioni, e
- la somma del quadrato dei valori attesi basati sul modello di regressione.

Questo è solo un metodo appropriato per esprimere il confronto tra i valori osservati e i valori attesi sotto due modelli. Nella regressione lineare il confronto tra valori osservati e valori attesi è basato sul quadrato della distanza tra i due. Se y_i denota il valore osservato e $\hat{y}_i$

denota il valore atteso per l'individuo *i* sotto il modello, allora la statistica usata per stimare questo confronto è:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

Nel modello in cui non è contenuta la variabile indipendente l'unico parametro è β_0, e $\beta_0 = \bar{y}$, cioè la media della variabile risposta. In questo caso $\hat{y}_i = \bar{y}$ e *SSE* è uguale alla varianza totale. Quando invece viene inclusa una variabile indipendente nel modello ogni decremento di *SSE* sarà dovuto al fatto che il coefficiente della curva riferita alla variabile indipendente non è zero. La variazione del valore di *SSE* è dovuta alla fonte di variabilità della regressione, denotata con *SSR*. Quest'ultima, detta *somma dei quadrati residuale*, è espressa come:

$$SSR = \left[\sum_{i=1}^{n} (y_i - \bar{y}_i) \right] - \left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right].$$

Nella regressione lineare si focalizza l'interesse sulla dimensione di *SSR*. Un valore elevato suggerisce che la variabile indipendente è importante, mentre un valore basso suggerisce che la variabile indipendente non è utile nel predire la variabile risposta.

Il principio guida nella regressione logistica è lo stesso: *comparare i valori osservati della variabile risposta con i valori attesi ottenuti dai modelli con o senza la variabile esplicativa in questione*. Nella regressione logistica il confronto tra valori osservati e valori attesi è basata sulla funzione di *log verosimiglianza* definita nella formula [7.8]. Per meglio capire questo confronto, è concettualmente utile pensare a un valore osservato della variabile risposta come se fosse un valore atteso risultante da un *modello saturato*. Un modello saturato è un modello che contenga tanti parametri quante sono le osservazioni.
Il confronto tra valori osservati e attesi usando la funzione di verosimiglianza è basato sulla seguente espressione:

$$D = -2 \ln \left[\frac{(verosimiglianaza\ del\ MODELLO\ ADATTATO)}{(verosimiglianaza\ del\ MODELLO\ SATURATO)} \right] \qquad [7.11].$$

La quantità all'interno delle parentesi quadrate nell'espressione [7.11] è chiamata *rapporto di verosimiglianza*. La componente '-2ln' è necessaria per ottenere una quantità la cui distribuzione sia nota e che possa quindi essere utilizzata per scopi di verifica d'ipotesi. Un test di questo tipo è detto *test del rapporto di verosimiglianza*. Utilizzando l'equazione [7.8], la formula [7.11] diventa:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right], \qquad [7.12]$$

dove $\hat{\pi}_i = \hat{\pi}(x_i)$.

La statistica D nell'equazione [7.12] è chiamata *devianza* (McCullagh & Nelder, 1989) e gioca un ruolo centrale in alcuni approcci nella valutazione della bontà d'adattamento. La devianza per la regressione logistica gioca lo stesso ruolo della somma dei quadrati residuale nella regressione lineare.

Inoltre, in un database per cui i valori della variabile risposta sono 0 o 1, la verosimiglianza del modello saturato è 1. Specificatamente, dalla definizione di modello saturato segue che $\hat{\pi}_i = y_i$ e che la verosimiglianza è:

$$l(MODELLO\ SATURATO) = \prod_{i=1}^{n} y_i^{\,y_i} \cdot (1-y_i)^{(1-y_i)} = 1.$$

Segue dall'equazione [7.11] che la devianza è:

$$D = -2\ln(verosimiglianza\ del\ MODELLO\ ADATTATO). \qquad [7.13]$$

Un Software di programmazione statistica come il SAS System riporta il valore della devianza espresso come nella [7.13] piuttosto che la *log verosimiglianza* per il modello adattato.

Per gli scopi di stima della significatività di una variabile indipendente si confrontano i valori di *D* con e senza la variabile indipendente nell'equazione. La variazione del valore di *D* dovuta all'inclusione della variabile indipendente nel modello è ottenuta come:

$$G = D(\text{modello senza la variabile}) - D(\text{modello con la variabile}).$$

Questa statistica nella regressione logistica gioca lo stesso ruolo del numeratore del test *F* parziale nella regressione lineare. Siccome la verosimiglianza del modello saturato è comune ad entrambi i valori di *D, G* può essere espressa come:

$$G = -2\ln\left[\frac{\left(verosimiglianza\ \ SENZA\ la\ VARIABILE\right)}{\left(verosimiglianza\ \ CON\ \ la\ VARIABILE\right)}\right]. \qquad [7.14]$$

Nel caso di una singola variabile, è facile dimostrare che quando la variabile non è nel modello la stima di massima verosimiglianza per β_0 è $\ln\left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1-y_i)}\right)$ e il valore atteso è costante, $\frac{\sum_{i=1}^{n} y_1}{n}$. In questo caso il valore di *G* è:

$$G = -2\ln\left[\frac{\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^{\sum_{i=1}^{n} y_i}\left(\frac{\sum_{i=1}^{n}(1-y_i)}{n}\right)^{\sum_{i=1}^{n}(1-y_i)}}{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{(1-y_i)}}\right] \qquad [7.15]$$

oppure:

$$G = 2\left\{\sum_{i=1}^{n}\left[y_i\ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)\right] - \left[\sum_{i=1}^{n} y_i \cdot \ln\left(\sum_{i=1}^{n} y_i\right) + \sum_{i=1}^{n}(1-y_i)\cdot\ln\left(\sum_{i=1}^{n}(1-y_i)\right) - n\ln(n)\right]\right\} \qquad [7.16].$$

Il *test del rapporto di verosimiglianza* per la significatività globale degli *r* coefficienti per le variabili indipendenti nel modello è rappresentato esattamente nello stesso modo appena visto per il caso univariato. Il test è basato sulla statistica *G* data nell'equazione [7.16], con l'unica

differenza che i valori adattati, $\hat{\pi}$, sotto il modello, sono basati sul vettore contenente $r + 1$ parametri, $\hat{\beta}$. Sotto l'ipotesi nulla che gli r coefficienti delle covariate nel modello siano uguali a zero, G si distribuirà come un Chi-Quadrato con r gradi di libertà.

Il calcolo della *log verosimiglianza* e del test del rapporto di verosimiglianza sono aspetti standard di tutti i software sulla regressione logistica. Questo permette di controllare facilmente la significatività dell'aggiunta di una nuova covariata nel modello. Oltre al test del rapporto di verosimiglianza, altre due statistiche equivalenti vengono spesso fornite: il *test di Wald* e il *test Score*. Le assunzioni necessarie per questi test sono identiche a quelle viste per il test del rapporto di verosimiglianza. Il test di Wald (nel caso univariato) è ottenuto comparando la stima di massima verosimiglianza del parametro, $\hat{\beta}_j$, con una stima del suo errore standard.

$$W = \frac{\hat{\beta}_j}{\hat{SE}\ (\hat{\beta}_j)}$$

Il rapporto risultante, sotto l'ipotesi che $\beta_j = 0$, seguirà una distribuzione normale standard.

Il caso multivariato del test di Wald è ottenuto dalla seguente espressione vettore-matriciale:

$$W = \hat{\beta}' \left[\hat{Var}(\hat{\beta}) \right]^{-1} \hat{\beta} = \hat{\beta}' (XVX) \hat{\beta},$$

che si distribuisce come un Chi-Quadrato con $r + 1$ gradi di libertà sotto l'ipotesi nulla che ognuno degli $r + 1$ coefficienti è uguale a zero.

Hauck and Donner (1977) hanno studiato il comportamento del test di Wald e hanno trovato che si comporta in un modo aberrante, spesso fallendo nel rigetto dell'ipotesi nulla quando il coefficiente era significativo. Si raccomandano quindi che venga usato il test del rapporto di verosimiglianza.

Sia il test del rapporto di verosimiglianza che il test di Wald richiedono il calcolo della stima di massima verosimiglianza per β_j. Un test per la significatività di una variabile che non richiede questa computazione è il test Score. Il test Score è basato sulla teoria distributiva delle derivate della *log verosimiglianza* e la statistica test per il test Score è la seguente:

$$Score = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sqrt{\bar{y} \cdot (1-\bar{y}) \sum_{i=1}^{n} (x_i - \bar{x})^2}} \ .$$

Riassumendo le informazioni fin qui fornite, il metodo per testare la significatività dei coefficienti di una variabile nella regressione logistica è simile all'approccio usato nella regressione lineare; tuttavia questo metodo usa la funzione di verosimiglianza per una variabile risposta dicotomica.

7.6 Stima degli intervalli di confidenza

Un'aggiunta importante al test di significatività del modello è il calcolo e l'interpretazione degli intervalli di confidenza per i parametri d'interesse. Come nel caso della regressione lineare si possono ottenere gli intervalli di confidenza per la curva, l'intercetta e il *logit*. In alcuni casi è interessante fornire anche stime degli intervalli per i valori adattati, cioè le probabilità attese.

La base per la costruzione degli stimatori d'intervallo è la stessa teoria statistica utilizzata per formulare i test per la significatività del modello. In particolare, gli stimatori per gli intervalli di confidenza della curva e dell'intercetta sono basati sul loro rispettivo test di Wald. Gli estremi di un intervallo di confidenza al 100(1-α)% per un coefficiente $\hat{\beta}_j$ sono:

$$\hat{\beta}_j \pm z_{1-\alpha/2} \hat{SE}(\hat{\beta}_j) \qquad [7.17],$$

dove $z_{1-\alpha/2}$ è il più elevato valore al 100(1-α)% della distribuzione normale standard e $\hat{SE}(\hat{\beta}_j)$ denota uno stimatore dell'errore standard (basato sul modello) del rispettivo stimatore parametrico.

Il *logit* è la parte lineare del modello di regressione logistica e come tale è più simile alla retta stimata del modello di regressione lineare. Un'espressione generale per la stima del *logit* in un modello con *r* covariate è:

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_r x_r \ . \qquad [7.18]$$

Un modo alternativo per esprimere lo stimatore del *logit* nella [7.18] è attraverso l'uso della notazione vettoriale: $\hat{g}(x) = \mathbf{x'}\ \hat{\beta}\mathbf{'}$, dove il vettore $\hat{\beta}\mathbf{'} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_r)$ denota lo stimatore degli r + 1 coefficienti e il vettore $\mathbf{x'} = (x_0, x_1, x_2, ...x_r)$ rappresenta la costante e un set di valori delle r covariate nel modello, dove $x_0 = 1$.

Lo stimatore della varianza dello stimatore del *logit* è invece dato dalla formula seguente:

$$\hat{Var}[\hat{g}(x)] = \sum_{j=0}^{r} x_j^2\, \hat{Var}(\hat{\beta}_j) + \sum_{j=0}^{r} \sum_{k=j+1}^{r} 2x_j x_k\, \hat{Cov}(\hat{\beta}_j, \hat{\beta}_k)\,. \qquad [7.19]$$

È possibile esprimere questo risultato con una formula più concisa, utilizzando l'espressione matriciale per lo stimatore della varianza dello stimatore dei coefficienti. Dall'espressione per la matrice delle informazioni osservate, si ha che

$$\hat{Var}(\hat{\beta}) = (\mathbf{X'VX})^{-1}. \qquad [7.20]$$

Segue dalla [7.20] che un'espressione equivalente per lo stimatore della [7.19] è:

$$\hat{Var}[\hat{g}(x)] = \mathbf{x'}\hat{Var}(\hat{\beta})\ \mathbf{x} = \mathbf{x'(X'VX)}^{-1}\mathbf{x}. \qquad [7.21]$$

Fortunatamente tutti i pacchetti software di buona qualità forniscono l'opzione per l'utente di creare una nuova variabile contenente i valori stimati dell'espressione [7.21] o l'errore standard per tutti i soggetti del dataset.

7.7 Regressione Logistica per il *Matched Case-Control Study*

7.7.1 Introduzione

Un rapporto dettagliato sul Matched Case-Control Study può essere trovato in testi di epidemiologia quali Breslow & Day (1980), Schlesselman (1982), Kelsey et al. (1986) e Rothman & Greenland (1998).

Nel disegno *Matched Case-Control Study* i soggetti vengono stratificati sulla base di variabili che si presumono associate con l'outcome. All'interno di ogni strato viene scelto un campione di casi (y = 1) e uno di controlli (y = 0). Il numero di casi e controlli non deve necessariamente essere costante tra gli strati, ma in questo capitolo si affronterà il disegno per l'accoppiamento 1-1 che è quello utilizzato nel lavoro sulla RDS.

È possibile trattare le variabili di stratificazione includendole nel modello, ma questo approccio funziona adeguatamente solo se il numero di soggetti in ogni strato è ampio. Invece, in un tipico studio caso-controllo per appaiamento 1-1 con n coppie di casi-controlli, si riscontrano solo due osservazioni per ciascuno strato. Di conseguenza, in un'analisi completamente stratificata con p covariate, sarà richiesto di stimare $n + p$ parametri identificati dalla costante, dai p coefficienti per le covariate e dagli $n - 1$ coefficienti specifici per le variabili di stratificazione utilizzando un campione di dimensione $2n$. Le proprietà ottimali del metodo di massima verosimiglianza, derivate dal rendere ampia la numerosità del campione, valgono solo quando il numero di parametri rimane fisso. Questo non è chiaramente il caso di uno studio per appaiamento 1-M. Con un'analisi completamente stratificata, il numero di parametri cresce all'aumentare della dimensione del campione. Può essere dimostrato che in un modello contenente una covariata dicotomica, l'errore nella stima del suo coefficiente è pari al 100% quando si tratta un disegno d'appaiamento 1-1 attraverso una verosimiglianza completamente stratificata (Breslow & Day, 1980). Se si considerano le variabili di stratificazione come parametri di disturbo e si può rinunciare al loro inserimento nel modello, allora è possibile usare metodi d'inferenza condizionata nella creazione della funzione di verosimiglianza che producano stimatori di massima verosimiglianza per i coefficienti nel modello di regressione logistica che siano consistenti ed asintoticamente distribuiti normalmente.

Si supponga di avere K strati con n_{1k} casi e n_{0k} controlli nello strato k-esimo (k =1,2,…,K). Il modello di regressione logistica specifico per lo strato k-esimo è dato da:

$$\pi_k(x) = \frac{e^{\alpha_k + \beta' x}}{1 + e^{\alpha_k + \beta' x}}, \qquad [7.22]$$

dove α_k denota il contributo al *logit* di tutti i termini costanti all'interno del k-esimo strato (cioè le variabili di stratificazione o d'appaiamento). Il vettore dei coefficienti, $\boldsymbol{\beta}$, contiene solo i p coefficienti, $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$. Ne consegue che ogni coefficiente indica la variazione nel *log-odds* per un incremento di un'unità nella covariata tenendo costanti tutte le alte variabili esplicative in ogni strato.

La verosimiglianza condizionata per il *k*-esimo strato è ottenuta come probabilità dei dati osservati condizionati allo strato per il numero totale di casi osservati. In questo contesto si tratta della probabilità dei dati osservati relativa alla probabilità dei dati per tutte le possibili assegnazioni degli n_{1k} casi e degli n_{0k} controlli rispetto agli $n_k = n_{1k} + n_{0k}$ soggetti. Il numero di possibili assegnazioni dello stato di caso agli n_{1k} soggetti tra gli n_k soggetti, denotata con c_k, è dato dall'espressione matematica seguente:

$$c_k = \binom{n_k}{n_{1k}} = \frac{n_k!}{n_{1k}!(n_k - n_{1k})!} .$$

Si indichi ora con l'indice *j* scritto in basso una qualsiasi di queste assegnazioni c_k. Per ogni assegnazione, inoltre, i soggetti da 1 a n_{1k} corrispondano ai casi ed i soggetti da $n_{1k} + 1$ a n_k corrispondano ai controlli. Questo viene indicizzato con *i* per i dati osservati e con i_j per la *j*-esima possibile assegnazione. La verosimiglianza condizionata è:

$$l_k(\beta) = \frac{\prod_{i=1}^{n_{1k}} P(x_i | y_i = 1) \prod_{i=n_{1k}+1}^{n_k} P(x_i | y_i = 0)}{\sum_{j=1}^{c_k} \left\{ \prod_{i_j=1}^{n_{1k}} P(x_{ji_j} | y_{i_j} = 1) \prod_{i_j=n_{1k}+1}^{n_k} P(x_{ji_j} | y_{i_j} = 0) \right\}} . \qquad [7.23]$$

La completa verosimiglianza condizionata è il prodotto delle $l_k(\beta)$ sui *K* strati, vale a dire:

$$l(\beta) = \prod_{k=1}^{K} l_k(\beta) . \qquad [7.24]$$

Se si assume che il modello di regressione logistica nella [7.22] è corretto allora l'applicazione del teorema di Bayes ad ogni termine di P(x|*y*) della [7.23] porta a:

$$l_k(\beta) = \frac{\prod_{i=1}^{n_{1k}} e^{\beta' x_i}}{\sum_{j=1}^{c_k} \prod_{i_j}^{n_{1k}} e^{\beta' x_{ji_j}}} \qquad [7.25].$$

I software per effettuare i calcoli necessari per ottenere lo stimatore di massima verosimiglianza condizionata sono disponibili in molti pacchetti statistici. Nel SAS System, per esempio, viene utilizzata la procedura PHREG con alcuni accorgimenti.

7.7.2 Analisi di Regressione Logistica per 1-1 *Matched Study*

Il disegno d'appaiamento più frequentemente usato è quello in cui ogni caso è abbinato ad un singolo controllo, di conseguenza ci sono due soggetti in ogni strato. Per semplificare la notazione, $\mathbf{x}_{1k}$ denoti il vettore di dati per i casi e $\mathbf{x}_{0k}$ il vettore per i controlli riferito al k-esimo strato (o coppia). Usando questa notazione, la verosimiglianza condizionata per il k-esimo strato deriva dalla [7.25] ed è:

$$l_k(\beta) = \frac{e^{\beta' x_{1k}}}{e^{\beta' x_{1k}} + e^{\beta' x_{0k}}} \qquad [7.26].$$

Attribuiti valori specifici a β, $\mathbf{x}_{1k}$ e $\mathbf{x}_{0k}$, l'equazione [7.26] rappresenta la probabilità che un soggetto identificato come caso sia effettivamente un caso sotto le assunzioni che:

1. si abbiano due soggetti uno dei quali è un caso, e
2. il modello di regressione logistica nell'equazione [7.22] sia il modello corretto.

Per esempio si supponga di avere un modello con una singola covariata dicotomica e che β sia pari a 0.8. Se i valori osservati sono $\mathbf{x}_{1k} = 1$ e $\mathbf{x}_{0k} = 0$, allora il valore dell'equazione [7.26] è:

$$l_k(0.8) = \frac{e^{0.8\times 1}}{e^{0.8\times 0} + e^{0.8\times 1}} = 0.690 .$$

Di conseguenza, la probabilità che un soggetto con $x = 1$ sia un caso è 0.69 rispetto ad un soggetto con $x = 0$. D'altro canto se $\mathbf{x}_{1k} = 0$ e $\mathbf{x}_{0k} = 1$ allora

$$l_k(0.8) = \frac{e^{0.8\times 0}}{e^{0.8\times 1} + e^{0.8\times 0}} = 0.310$$

rappresenta la probabilità che un soggetto con $x = 0$ sia un caso rispetto ad un soggetto con $x = 1$. Dalla [7.26] segue anche che se i dati per il caso e per il controllo sono identici, $\mathbf{x}_{1k} = \mathbf{x}_{0k}$, allora $l_k(\beta) = 0.5$ per qualsiasi valore di β (i dati per il caso e per il controllo sono equamente probabili sotto il modello). Le coppie di casi-controlli con lo stesso valore in una

delle covariate sono, quindi, identicamente informative per la stima del coefficiente specifico di quella covariata.

Dividendo il numeratore e denominatore dell'espressione [7.26] per $e^{\beta' x_{0k}}$ si ottiene:

$$l_k(\beta) = \frac{e^{\beta'(x_{1k} - x_{0k})}}{1 + e^{\beta'(x_{1k} - x_{0k})}}. \qquad [7.27]$$

Ne consegue che la verosimiglianza condizionata totale può essere espressa come il prodotto delle verosimiglianze per gli strati individuali:

$$l(\beta) = \prod_{k=1}^{K} \frac{e^{\beta'(x_{1k} - x_{0k})}}{1 + e^{\beta'(x_{1k} - x_{0k})}}. \qquad [7.28]$$

Si noti che la verosimiglianza condizionata per dati di coppie appaiate è uguale alla verosimiglianza non condizionata per un modello di regressione logistica in cui la variabile risposta è sempre pari ad 1, i valori delle covariate sono uguali alle differenze tra i valori per casi e controlli, e non c'è intercetta. Questo significa che è possibile utilizzare un programma per la regressione logistica standard configurando appropriatamente i dati ed eliminando l'intercetta. È necessario seguire le seguenti procedure (Stokes et al., 1995):

- creare l'unità di campionamento per ogni coppia appaiata che rappresenterà un singolo record, e individuare le variabili esplicative come differenze tra i valori dei casi ed i valori dei controlli.
- Porre la variabile risposta pari ad 1 (o qualunque altro valore).
- Porre l'intercetta del modello uguale a zero.

Il processo concettuale per modellare dati appaiati è identico a quello per i dati non appaiati. Se si sviluppano le strategie viste in precedenza nel modello d'abbinamento 1-1 come se si avesse un disegno di non appaiamento e se poi si usa la verosimiglianza condizionata, si è sicuri di procedere sempre correttamente.

7.8 Applicazione ai dati dello studio al Children's Hospital dell'Università di Oulu usando SAS System

I 137 casi (nati prematuri affetti da RDS) sono stati abbinati ai 137 controlli (nati prematuri sani) con stessa età gestazionale, stesso sesso e che hanno avuto, o meno, un trattamento glucocorticoide. Le informazioni sono state raccolte sulle variabili cliniche viste in precedenza.

L'obiettivo dell'analisi è determinare se particolari caratteristiche genetiche dei pazienti sono relazionate con l'RDS. Di base si sono effettate due analisi di regressione logistica standard con il programma statistico SAS System. La prima analisi ha mirato all'individuazione delle eventuali associazioni genetiche con l'RDS, la seconda ha preso come covariate le variabili cliniche peso e apgar. Ulteriori analisi di regressione logistica sono state portate a termine prendendo in esame sotto-popolazioni diversificate per durata gestazionale, in particolare si sono analizzate le relazioni tra covariate genetiche ed RDS in sottocampioni di età gestazionale < 32 e ≥ 32 settimane.

Per effettuare l'analisi relativa alle associazioni genetiche con l'RDS sono state dicotomizzate le dieci variabili relative ai polimorfismi dei singoli nucleotidi dei geni SP-A1, SP-A2 e SP-B ponendo uguale ad 1 la modalità ipotizzata essere quella di rischio (coppia di nucleotidi eterozigotica) nel manifestarsi della RDS, e pari a 0 le due modalità considerate non a rischio (coppie di nucleotidi omozigotiche). Per esempio nel caso del gene SP-A1, polimorfismo del singolo nucleotide per il codone 19, la modalità TC è stata presa come fattore di rischio, mentre le alternative geniche CC e TT sono state poste come non a rischio.

Riconducendosi alla forma di database descritta nel precedente paragrafo, ogni coppia di casi e controlli è stata trasformata in una singola osservazione, dove il valore delle variabili esplicative per ogni record è rappresentato dalla differenza tra i corrispondenti valori per il caso ed il controllo. La variabile risposta RDS è stata imposta pari ad un valore *k* identico per tutte le coppie di osservazioni, mentre le variabili per cui era stato effettuato l'appaiamento (durata gestazionale, sesso e trattamento glucocorticoide) non sono state considerate nel modello d'analisi essendo uguali per il caso e il controllo di ciascuna coppia. Infine nel programma d'analisi di regressione logistica realizzato con la "*proc logistic*" del SAS System è stato specificato di non includere l'intercetta attraverso l'istruzione "*noint*". Nella procedura per la logistica si è scelto un modello di selezione *stepwise*, con un livello di significatività d'entrata della variabile pari a 0.10 ed un livello di significatività per la rimozione pari a 0.05.

Si riporta ora il pezzo di programma effettuato con SAS System per l'analisi di regressione logistica:

proc logistic descending;
model RDS = variabili esplicative rappresentate dalla differenza dei valori di casi e controlli / noint selection=stepwise SLE=0.10 SLS=0.05;
run;

Un altro modo per effettuare l'analisi di regressione logistica condizionata col SAS (che ha portato agli stessi risultati della procedura precedente) è stato attraverso la "*proc phreg*". Questa procedura è designata ad un modello per l'analisi di sopravvivenza dei dati con un modello di Cox per i rischi proporzionali, ma certe equivalenze computazionali ne permettono l'uso anche per effettuare analisi di regressione logistica condizionate. È possibile utilizzare la "*proc phreg*" per analizzare dati appaiati 1-1 senza dover prima creare un database con variabili relative alle differenze delle covariate di casi e controlli. In questa situazione la variabile risposta (RDS) è stata definita come dicotomica, con valore 1 per i casi e 0 per i controlli, in modo che la probabilità di essere un caso potesse essere modellata. Inoltre è stato necessario specificare la variabile identificante ciascun record (*id*) come variabile identificativa attraverso l'istruzione "*strata*". Di seguito si riporta il frammento principale del programma:

proc phreg;
strata id;
model RDS = variabili esplicative / selection = forward details;
run;

7.9 Risultati dello studio

Prima di passare ad esaminare i risultati dell'analisi di regressione logistica condizionata, si presentano alcuni esiti ottenuti con l'analisi descrittiva dei dati. Confronti di frequenze alleliche sono stati eseguiti attraverso l'uso della stima dell'OR grezzo per cui vengono riportati anche gli intervalli di confidenza e le rispettive significatività in termini di *p-value*. Le distribuzioni degli alleli nei neonati con RDS e nei controlli sono state confrontate attraverso l'uso di tabelle 2 x k, mentre le frequenze dei singoli alleli sono state confrontate con tabelle 2 x 2.

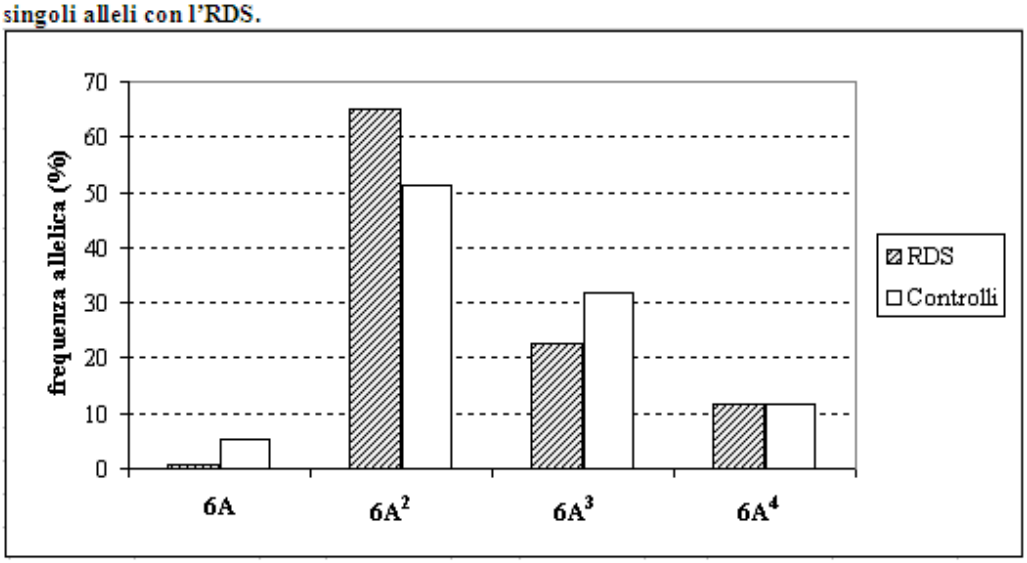
7.9.1 Associazione allelica tra i geni SP-A e RDS

Sono stati analizzati 137 campioni di DNA dai neonati prematuri con RDS e dai rispettivi controlli appaiati per grado di prematurità, sesso e terapia glucocorticoide prenatale. Dall'analisi complessiva del Chi-quadrato per un confronto delle distribuzioni alleliche di casi e controlli (senza differenziarli per età gestazionale) è risultata una differenza significativa per SP-A1 ($p = 0.017$), ma non per SP-A2 ($p = 0.23$). La figura 7.1.A mostra come l'allele $6A^2$ di SP-A1 è risultato sovrarappresentato nei neonati con RDS rispetto ai controlli sani (le frequenze sono rispettivamente 0.65 e 0.51; $p = 0.017$). Inoltre l'allele $6A^3$ di SP-A1 è sottorappresentato nei neonati con RDS rispetto ai controlli (frequenza: 0.23 nei casi vs. 0.32 nei controlli; $p = 0.034$) (figura 7.1.A). Dalla figura 7.1.B si riscontra che i neonati con RDS presentano un trend verso la sovrarappresentazione dell'allele $1A^0$ rispetto ai controlli (le frequenze sono rispettivamente 0.63 e 0.54; $p = 0.11$).

La frequenza dell'aplotipo $6A^2/1A^0$ di SP-A1/SP-A2 è risultato significativamente più elevata nei neonati con RDS rispetto ai controlli (0.62 vs. 0.50 rispettivamente) (figura 7.1.C).

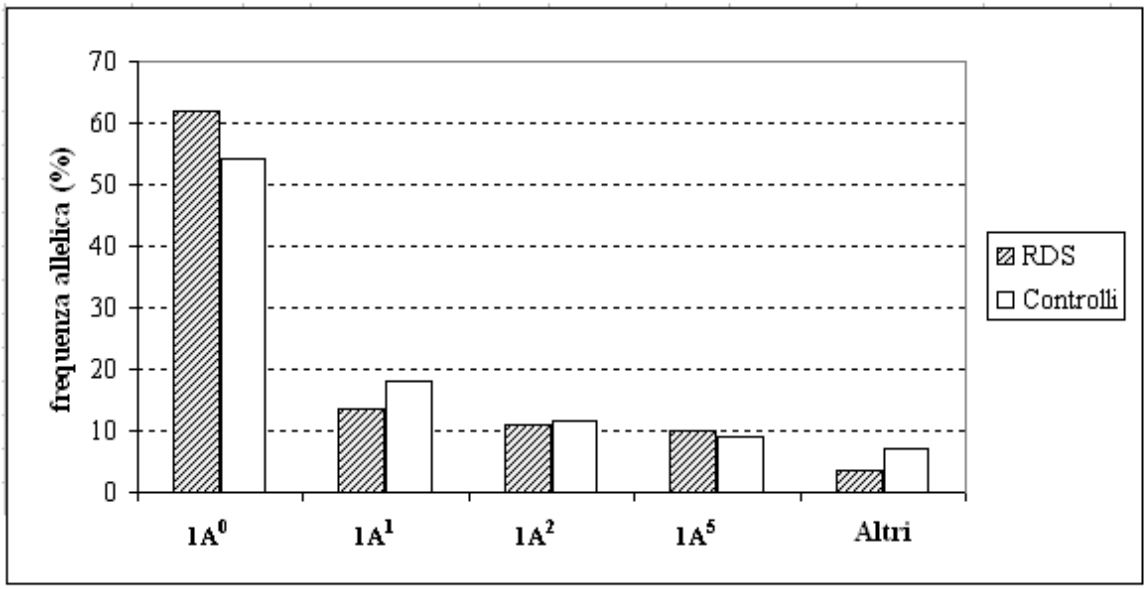
La frequenza del genotipo omozigotico $6A^2/6A^2$ di SP-A1 tende ad essere differente tra i neonati con RDS ed i controlli (0.43 vs. 0.31, rispettivamente; $p = 0.86$), mentre la frequenza del genotipo eterozigotico $6A^2/*$ è simile nei due gruppi (0.43 vs. 0.43; $p = 1.0$). Similmente la frequenza del genotipo omozigotico $6A^3/6A^3$ differisce tra i neonati con RDS ed i controlli (0.34 vs. 0.15 rispettivamente; $p = 0.009$, OR=0.20, IC 95% 0.06-0.74), mentre la frequenza del genotipo eterozigotico $6A^3/*$ non appare differire nei due gruppi (rispettivamente 0.41 vs. 0.39; $p = 0.76$). La frequenza del genotipo omozigotico $1A^0/1A^0$ di SP-A2 tende ad essere differente nei pazienti con RDS rispetto ai controlli (rispettivamente 0.40 vs. 0.28; $p = 0.11$), mentre la frequenza del genotipo eterozigotico $1A^0/*$ risulta simile (rispettivamente 0.45 vs. 0.51; $p = 0.45$).

Figura 7.1.A. Distribuzione degli alleli del gene SP-A1 per i 137 neonati con RDS e per i rispettivi 137 controlli accoppiati per grado di prematurità, terapia glucocorticoide prenatale e sesso. Sotto il grafico sono riportati i *p-value*, gli OR con i rispettivi intervalli di confidenza, che illustrano l'associazione dei singoli alleli con l'RDS.



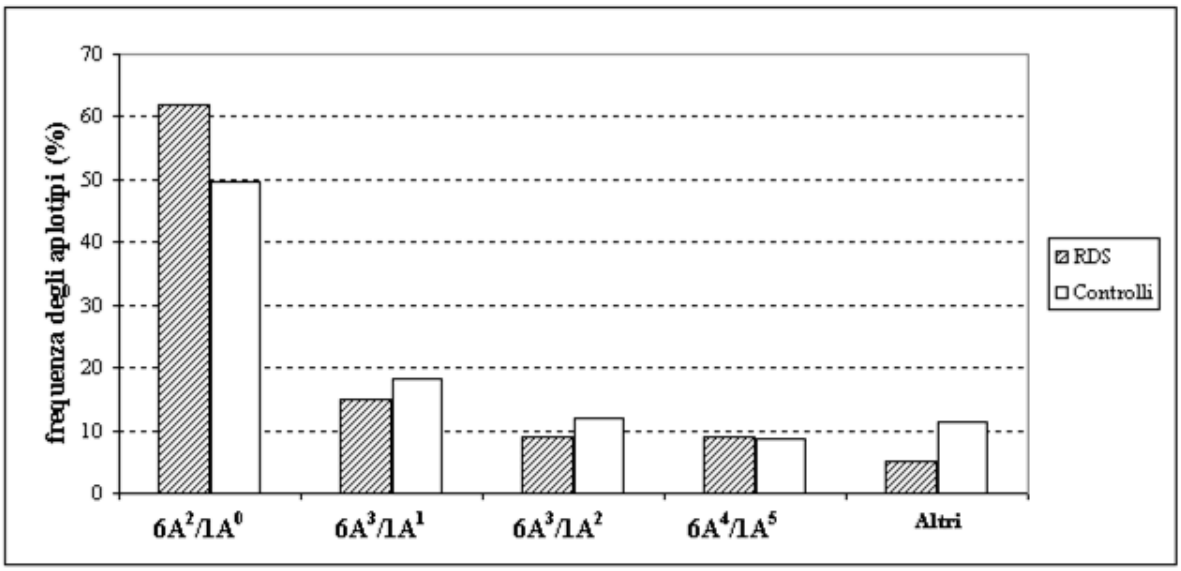
p-value	0,032	0,017	0,034	0,58
OR	0,21	1,68	0,61	1,22
CI	0,05 - 1,00	1,09 - 2,58	0,38 - 0,97	0,60 - 2,51

Figura 7.1.B. Distribuzione degli alleli del gene SP-A2 per i 137 neonati con RDS e per i rispettivi 137 controlli accoppiati per grado di prematurità, terapia glucocorticoide prenatale e sesso. Sotto il grafico sono riportati i *p-value*, gli OR con i rispettivi intervalli di confidenza, che illustrano l'associazione dei singoli alleli con l'RDS.



p-value	0,11	0,2	0,87	0,55	0,066
OR	1,42	0,69	0,95	1,27	0,41
CI	0,93 - 2,17	0,39 - 1,22	0,49 - 1,82	0,58 - 2,80	0,15 - 1,09

Figura 7.1.C. Distribuzione dei principali aplotipi SP-A1/SP-A2 per i 137 neonati con RDS e per i rispettivi 137 controlli accoppiati per grado di prematurità, terapia glucocorticoide prenatale e sesso. Sotto il grafico sono riportati i *p-value*, gli OR con i rispettivi intervalli di confidenza, che illustrano l'associazione dei singoli alleli con l'RDS.



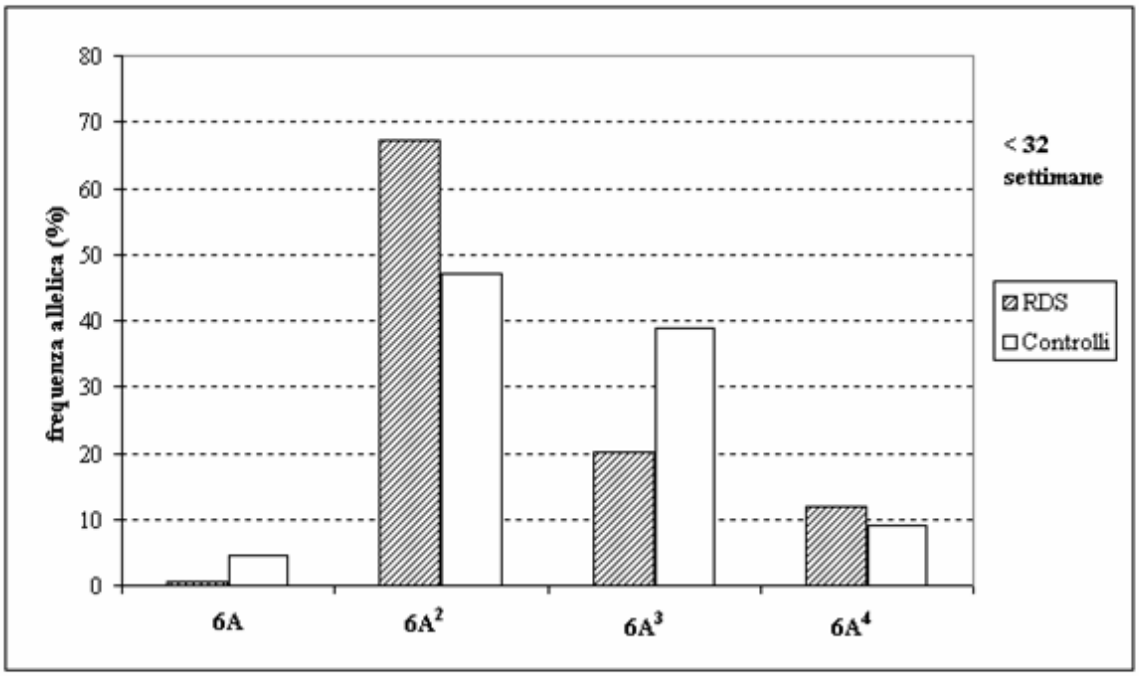
p-value	0,018	0,32	0,20	0,84	0,24
OR	1,66	0,75	0,62	1,09	0,49
CI	1,09 - 2,54	0,43 - 1,32	0,30 - 1,29	0,48 - 2,46	0,14 - 1,65

7.9.2 Influenza dei fattori di rischio nell'associazione tra RDS e alleli SP-A e tra RDS e gli aplotipi

La prematurità è il fattore di rischio più serio per l'RDS. Le frequenze alleliche di entrambi i geni SP-A sono state determinate nei neonati con RDS e nei controlli, separatamente per età gestazionale <32 settimane e ≥32 settimane. Si è scelta tale settimana gestazionale per separare le due analisi in quanto consiste nella mediana della popolazione oggetto di studio. L'associazione tra RDS ed il locus del gene SP-A sembra essere dipendente dal grado di prematurità (figure 7.2.A e 7.2.B).

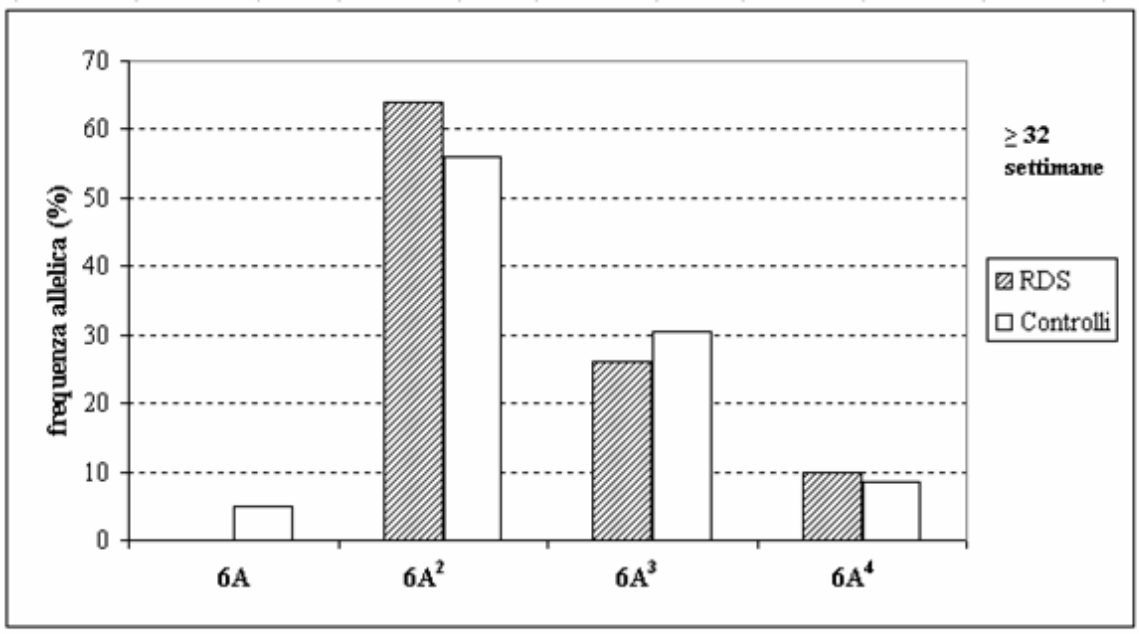
Neonati con RDS, con età gestazionale ≥32 settimane, non presentano differenze significative nella distribuzione allelica dei geni SP-A1 ($p = 0.56$) ed SP-A2 ($p = 0.53$) rispetto ai controlli. Al contrario, tra i neonati con durata gestazionale <32 settimane, è presente una differenza significativa tra casi e controlli nella distribuzione allelica di SP-A1 ($p = 0.036$), ma non in quella di SP-A2 ($p = 0.26$). Considerando l'età gestazionale <32 settimane, la frequenza dell'allele $6A^2$ di SP-A1 è risultata pari a 0.66 nei neonati con RDS, mentre nei controlli è stata calcolata pari a 0.49 ($p = 0.018$; figura 7.2.A). In aggiunta la frequenza dell'allele $6A^3$ di SP-A1 è risultata essere significativamente più bassa nei casi rispetto ai controlli (0.21 vs. 0.37, rispettivamente; $p = 0.015$).

Figura 7.2.A. Distribuzione degli alleli del gene SP-A1 per i 67 neonati con RDS e per i rispettivi 67 controlli nati con età gestazionale inferiore alle 32 settimane. Sotto il grafico sono riportati i *p-value*, gli OR con i rispettivi intervalli di confidenza, che illustrano l'associazione dei singoli alleli con l'RDS.



p-value	0,43	0,018	0,015	0,73
OR	0,46	2,06	0,44	1,19
CI	0,08 - 2,55	1,13 - 3,77	0,23 - 0,86	0,45 - 3,17

Figura 7.2.B. Distribuzione degli alleli del gene SP-A1 per i 70 neonati con RDS e per i rispettivi 70 controlli nati con età gestazionale superiore o uguale alle 32 settimane. Sotto il grafico sono riportati i *p-value*, gli OR con i rispettivi intervalli di confidenza, che illustrano l'associazione dei singoli alleli con l'RDS.



p-value	0,06	0,31	0,59	0,68
OR		1,37	0,84	1,25
CI		0,75 - 2,51	0,43 - 1,61	0,43 - 3,61

Nei nati estremamente pretermine (< 32 settimane), la frequenza del genotipo omozigote $6A^2/6A^2$ di SP-A1 tende a differenziarsi tra neonati con RDS e controlli (la frequenza per i casi è 0.48 rispetto alla frequenza per i controlli che risulta pari a 0.30; $p = 0.09$). La frequenza del genotipo eterozigote $6A^2/*$ sembra invece non presentare differenze tra casi e controlli (0.37 vs. 0.37; $p = 0.98$). La frequenza del genotipo omozigote $6A^3/6A^3$ differisce significativamente nel confronto tra neonati con RDS e neonati sani (0.022 per i primi e 0.21 per i secondi; $p = 0.006$), mentre la frequenza di $6A^3/*$ non differisce tra i due gruppi (0.37 per i casi e 0.30 per i controlli; $p = 0.50$). Un orientamento simile si è ottenuto pure per la frequenza del genotipo omozigote $1A^0/1A^0$ di SP-A2 (0.44 vs. 0.28, rispettivamente nei neonati con RDS e nei controlli; $p = 0.13$) e per la frequenza del genotipo eterozigotico $1A^0/*$ (0.43 vs. 0.47, rispettivamente per casi e controlli; $p = 0.77$)

L'incidenza di RDS è più elevata nei maschi che nelle femmine (Farrell & Wood, 1976). Tenendo presente questa considerazione, si sono determinate, separatamente per maschi (80 coppie) e femmine (57 coppie), le differenze di frequenze per gli alleli di SP-A nei neonati con RDS e nei controlli. Nei pazienti con RDS, la distribuzione di frequenza di SP-A1 è risultata simile tra maschi e femmine ($p = 0.40$). Sia per i maschi che per le femmine, la frequenza di $6A^3$ tende ad essere più bassa nei neonati con RDS rispetto ai controlli, mentre la frequenza di $6A^2$ tende ad essere più alta nei casi che nei controlli.

La terapia glucocorticoide in parti con minaccia di prematurità riduce il rischio di RDS. Tra i neonati con età gestazionale inferiore alle 32 settimane si è confrontata la frequenza dell'allele $6A^2$ del gene SP-A1 per il gruppo di pazienti che avevano ricevuto la terapia e che poi hanno sviluppato l'RDS (0.70) con la frequenza dello stesso allele nei pazienti a cui non è stata somministrata la terapia e che non hanno sviluppato la malattia (0.22) ($p = 0.0002$). La frequenza dell'allele $6A^3$ del gene SP-A1 ha presentato una marcata differenza tra questi due gruppi (0.72 per i pazienti con RDS vs. 0.17 per i controlli; $p < 0.0001$). Anche la distribuzione degli alleli di SP-A2 sono risultati diversi tra casi e controlli. Nei controlli, nati cioè prima delle 32 settimane, che non hanno ricevuto la terapia glucocorticoide e per cui non si è sviluppata l'RDS, la frequenza di $1A^0$ è risultata inferiore ($p = 0.005$) e le frequenze di $1A^1$ ($p = 0.010$) e $1A^2$ ($p = 0.048$) si sono presentate superiori rispetto alle corrispondenti frequenze nei neonati con RDS.

Il grado di prematurità è risultato essere il maggior determinante dell'associazione allelica osservata. Le frequenze degli alleli di SP-A1 e SP-A2, come le frequenze degli aplotipi di SP-A1 e SP-A2 sono risultate significativamente differenti tra il gruppo di prematuri con RDS ed il gruppo di controllo nati prima delle 32 settimane gestazionali, mentre non si sono

riscontrate differenze significative tra il gruppo con RDS ed il gruppo di controlli per età gestazionali più elevate (≥ 32 settimane). L'allele più frequente di SP-A1, il $6A^2$, è sovra rappresentato nei nati con RDS rispetto ai controlli con età gestazionale < 32 settimane (p=0.018, OR: 2.06, 95% CI 1.13 – 3.77), mentre il secondo allele più frequente di SP-A1, il $6A^3$, è sotto rappresentato (p=0.015, OR: 0.44, 95% CI 0.23 – 0.86).

7.10 Risultati dell'analisi di regressione logistica condizionata

La durata gestazionale nella popolazione dello studio varia dalle 23 alle 37 settimane, con una media approssimativa di 32 settimane gestazionali. 85 delle 137 coppie studiate sono state trattate con profilassi glucocorticoide. Come si è più volte specificato i ruoli del grado di prematurità, sesso e trattamento glucocorticoide in caso di prematurità sono stati tenuti costanti tramite l'appaiamento, attraverso il modello di regressione logistica condizionata per evitare qualsiasi effetto di confondimento nella loro eventuale associazione con gli alleli di SP-A, SP-B e con l'RDS.

Sono state effettuate due analisi di regressione logistica condizionate, la prima utilizzando le variabili cliniche come covariate, la seconda utilizzando i predittori genetici, vale a dire i polimorfismi dei singoli nucleotidi dei geni SP-A ed SP-B.

La tabella 7.1 presenta i risultati della prima analisi in cui i predittori candidati all'entrata nel modello sono peso, lunghezza, apgar ed ospedale di nascita. Solamente l'apgar entra nel modello (*p-value* del test sul rapporto di verosimiglianza = 0.0022) e sembra che al diminuire del punteggio di apgar alla nascita aumenti la probabilità di osservare un neonato affetto da RDS rispetto ad un neonato sano. Nello specifico, applicando la formula $OR(-1\,punto) = \exp\left[(-1\,punto)\times(-0{,}486)\right]$, risulta che per ogni decremento pari a 1 nel punteggio apgar, la probabilità di osservare un neonato affetto da RDS rispetto ad uno sano aumenti di 1,63 volte. Questo è abbastanza comprensibile dato che il punteggio apgar indica la salute del bambino alla nascita e se un nato è affetto da RDS avrà un punteggio inferiore a parità di età gestazionale e sesso.

Tabella 7.1 Risulati dell'analisi di regressione logistica condizionata tra RDS e le covariate cliniche

	Stima di massima verosimiglianza	Standard Error	p-value	OR	IC 95%
Apgar	-0.486	0.189	0.01	0.615	0.425 - 0.891

IC 95% = intervalli di confidenza al 95%

È importante sottolineare la non entrata nel modello della variabile ospedale. Ciò sta ad indicare che il fatto di essere nato in un ospedale piuttosto che in un altro non aumenta il rischio di RDS. Nemmeno le variabili peso e lunghezza sono risultate significative, probabilmente perchè sono associate all'età gestazionale. Considerato che casi e controlli hanno stessa età gestazionale, verosimilmente avranno anche un peso ed una lunghezza affini.
L'analisi di regressione logistica condizionata relativa alle caratteristiche genetiche delle 137 coppie di nati prematuri (casi e controlli) ha evidenziato un'associazione tra un solo SNP del gene SP-A1 e l'RDS, cioè il polimorfismo per il singolo nucleotide del codone 133 nel gene SP-A1. Presentare una coppia di nucleotidi eterozigotica, AG, nello SNP del codone 133 nel gene SP-A1 aumenta il rischio di RDS di 2,6 volte rispetto a pazienti con una coppia monozigotica (tabella 7.2). Il *p-value* del test sul rapporto di verosimiglianza è pari a 0.038.

Tabella 7.2 Risulati dell'analisi di regressione logistica condizionata tra RDS e le covariate genetiche (SNPs)

	Stima di massima verosimiglianza	Standard Error	p-value	OR	IC 95%
SPA1aa19	0.955	0.326	0.042	2.60	1.372 - 1.894

IC 95% = intervalli di confidenza al 95%

Si è anche provato a realizzare un'analisi di regressione logistica condizionata per le sole coppie con età gestazionale inferiore alle 32 settimane, ma non si sono ottenuti risultati significativi, nel senso che nessuna covariata soddisfa il criterio limite per entrare nel modello.

7.9 Conclusioni

Se l'analisi di regressione logistica multipla è il metodo più appropriato nella maggior parte degli studi biomedici e genetici che investigano i rischi associati ad una particolare patologia il cui esito è misurato con assenza o presenza nel paziente, l'analisi di regressione logistica multipla condizionata risulta il modello più efficace nel caso in cui lo studio consideri coppie di pazienti (un controllo sano per ogni caso affetto da malattia) abbinate secondo una o più variabili considerate di confondimento per l'associazione tra fattori di rischio e patologia oggetto di ricerca. Si è visto che i coefficienti che massimizzano l'espressione di *log verosimiglianza* rappresentano il miglior adattamento del modello. La funzione della *log verosimiglianza* viene utilizzata anche nel calcolo della significatività dei coefficienti nel confronto tra valori osservati e valori attesi. Nel programma Sas System utilizzato nel presente lavoro, però, viene utilizzato il valore della devianza D (equazione 7.13) come base per calcolare la significatività dei coefficienti: la variazione del valore di D dovuta all'inclusione di una specifica variabile esplicativa nel modello rappresenta l'indice di significatività adottato dal Sas System. Attraverso opportune istruzioni, viste in questo capitolo, nella *proc logistic* o nella *proc phreg*, e/o attraverso apposite trasformazioni dei dati, il programma Sas System permette l'applicazione dell'analisi di regressione logistica condizionata di cui si vedranno nel prossimo capitolo i risultati.

Sezione F

Bibliografia, riassunto e abstract

BIBLIOGRAFIA

Acock A.C., *Working with missing values*, Journal of Marriage and Family 67: 1012-1028; 2005.

Adams T.D., Gress R.E., Smith S.C., Halverson R.C., Simper S.C., Rosamond W.D., Lamonte M.J., Stroup A.M., Hunt S.C., *Long-term mortality after gastric bypass surgery*, The New England Journal of Medicine 357(8): 753-761; 2006.

Ake C.F., *Rounding after multiple imputation with non-binary categorical covariates*, SUGI 30 Proceedings 112 (30): 1-11; 2005.

Agresti A. & Finlay B., *Statistical Methods for the Social Sciences*, Prentice Hall, Upper Saddle River, New Jersey, Third Edition; 1999.

Allison P.D., *Logistic Regression using the SAS System: Theory and Application*, SAS Institute Inc., Cary, North Carolina, USA; 1999.

Allison P.D., *Missing Data*, Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-136, Thousand Oaks, 2001.

Anderson T.W., *An introduction to multivariate statistical analysis*, New York, Jon Wiley & Sons Inc., 1958.

Armitage P. & Berry G., *Statistica Medica – metodi statistici per la ricerca in medicina*, Terza Edizione; McGraw-Hill Libri Italia; 1996.

Armstrong J.B. & Mayda J.E., *Estimation of record linkage models using dependent data*, Proceedings of the Section on Survey Research Methodology. American Statistical Association 1992 pp. 853–858; 1992.

Avery M.E. & Mead J., *Surface properties in relation to atelectasis and hyaline membrane disease*, American Journal of Diseases of Children 97: 517-523; 1959.

Barzi F. & Woodward M., *Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies*, American Journal of Epidemiology 160 (1): 34; 2004.

Bayley N., *Bayley Scales of Infant Development*, 2nd edition San Antonio, TX: The Psychological Corporation, 1993.

Belsley D.A., *Conditioning Diagnostics: Collinearity and Weak data in Regression*, John Wiley & Sons, Inc, 1991.

Bickenbach J.E., Chatterji S, Badley E.M., Üstün T.B. *Models of disablement, universalism and the international classification of impairments, disabilities and handicaps*. Social Science & Medicine 48: 1173-1187; 1999.

Bohmer R.M., Newell J., Torchiana D.F., *The effect of decreasing length of stay on discharge destination and readmission after coronary bypass operation*, Surgery 132 (1): 10-15; 2002.

Bosio C.A., *Grazie no! : il fenomeno dei non rispondenti in Politica e Sondaggio*, a cura di P. Ceri, Rosenberg & Sellier, Torino; 1997.

Breslow N.E. & Day N.E., *Statistical Methods in Cancer Research.* Vol. 1: The Analysis of case-control studies. International Agency on Cancer, Lyon, France; 1980.

Breslow N.E., *Statistics in epidemiology: the case-control study*, Journal of the American Statistical Association 91: 14-28; 1996.

Bruyère S.M., Van Looy S.A., Peterson D.B. *The International Classification of Functioning, Disability and Health: Contemporary Literature Overview*. Rehabilitation Psychology 2005; 50 (2): 113-121; 2005.

Brunet O. & Lezine I., *Le Dévelopment Psychologique de la Première Enfance*, Paris: Presses Universitaries de France, 1965.

Chamberlayne R., Green B., Barer M.L., Hertzman C., Lawrence W.J., Sheps S.B., *Creating a population-based linked health database: a new resource for health services research*, Canadian Journal of Public Health 89 (4): 270-273; 1998.

Chatfield C. & Collins A.J., *Introduction to Multivariate Analysis*, London: Chapman and Hall, 1980.

Choi B.C & Pak A.W., *Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness*, Clinical and investigative medicine, 29(6): 351-364 Review; 2006.

Clements J.A., *Functions of the alveolar lining*, American Review of Respiratory Disease 115: 67-71; 1977.

Cochrane A.L., *Effectiveness and Efficiency: Random Reflections on Health Services*, London: Nuffield Provincial Hospitals Trust, 1972.

College M.J., Johnson J.H., Pare R., Sande I.J., *Large scale imputation of survey data*, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 431-436; 1978

Crowley P. A., *Antenatal corticosteroid therapy: a meta-analysis of the randomized trials, 1972 to 1994*, American Journal of Obstetrics and Gynecology 173: 322-335; 1995.

Dempster P., Laird N. M., Rubin D.B., *Maximum Likelihood from Incomplete Data via The EM Algorithm*, Journal of Royal Statistical Society, 39: 1-38; 1977.

DiAngelo S., Lin Z., Wang G., Phillips S., Rämet M., Luo J., Floros J., *Novel, non-radioactive, simple and multiplex PCR-cRFLP method for genotyping human SP-A and SP-D marker alleles*, Disease Markers 15: 269-281; 1999.

Duckworth D. *The need for a standard terminology and classification of disablement*. In: C. V. Granger & G. E. Gresham (Eds.). Functional assessment in rehabilitation medicine. Baltimore: Williams & Wilkins, pp 1-13; 1984.

Durrant G.B., *Imputation Methods for Handling Item-Nonresponse in the Social Sciences : A Methodological Review*, ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, 2005.

Emrie P.A., Jones C., Hofmann T., Fisher J.H., *The coding sequence for the human 18,000-dalton hydrophobic pulmonary surfactant protein is located on chromosome 2 and identifies a restriction fragment length polymorphism*, Somatic Cell and Molecular Genetics 14: 105-110; 1988.

Fabbris L., *Statistica Multivariata, analisi esplorativa dei dati*, McGraw-Hill, Milano, 1997;

Fabbris L., *L'indagine campionaria. Metodi, disegni e tecniche di campionamento*, La Nuova Italia Scientifica, Roma; 1996.

Facchin P., Boccuzzo G., Visonà Dalla Pozza L., Salmaso L., *Il complesso percorso che dalla menomazione porta all'handicap: analisi delle correlazioni e dei nessi causali*, Modelli e metodi per l'analisi di rischi sociali e sanitari pp. 185-206, CLEUP Editrice, Padova; 2002.

Farrell P.M. & Wood R.E., *Epidemiology of hyaline membrane disease in the United States: analysis of national mortality statistics*, Pediatrics 58: 167-176; 1976.

Fay R.E., *Alternative Paradigms for the Analysis of Imputed Survey Data*, Journal of the American Statistical Association, 91, 434, 490-498; 1996.

Fellegi I.P. & Sunter A.B., *A Theory for record linkage*, Journal of the American Statistical Association 1969, 64: 1183-1210; 1969.

Fisher R.A., *The use of multiple measurements in taxonomic problems*, Annals of eugenics 7: 179-188; 1936.

Floros J., Veletza S.V., Kotikalapudi P., Krizkova L., Karinch A.M., Friedman C., Buchter S., Marks K., *Dinucleotide repeats in the human surfactant protein-B gene and respiratory distress syndrome*, Biochemical Journal 305: 583-590; 1995.

Floros J., DiAngelo S., Koptides M., Karinch A.M., Rogan P.K., Nielsen H., Spragg R.G., *Human SP-A locus: allele frequencies and linkage disequilibrium between the two surfactant protein A genes*, America Journal of Respiratory Cell and Molecular Biology 15: 489-498; 1996.

Friedman G.D., *Epidemiologia per Discipline Bio-mediche*, Edizione italiana a. cura di L. Sebastiano Annichiarico Petruzzelli, McGraw-Hill Libri Italia, Milano; 1995.

Garside M.J., *The best sub-set in multiple regression analysis*, Applied Statistics 14: 196-200; 1965.

Giacobini C., *L'Handicap cambia nome*, Mobilità 19:1- 4; 2002.

Glantz S.A., *Statistica per discipline bio-mediche*, McGraw-HillLibri Italia, Milano, 1997.

Glantz L.H., *Researchers' ethical duties are not to be outsourced*, Nature 449 (7159):139; 2007.

Granger C.V., Dewis L.S., Peters N.C., Sherwood C.C., Barrett, J.E., *Stroke rehabilitation: analysis of repeated Barthel Index measures*, Archives of Physical Medicine and Rehabilitation 60(1): 14-17; 1979.

Greenland S., *Response and follow-up bias in cohort studies*, American Journal of Epidemiology 106: 184-187; 1977.

Goldenberg R.L. & Rouse D.J., *Prevention of premature birth*, New England Journal of Medicine 339: 313-320; 1998.

Gomatam S., Carter R., Ariet M., Mitchell G., *An empirical comparison of record linkage procedures*, Statistics in Medicine, 21:1485–1496; 2002.

Gomez R., Ghezzi F., Romero R., Munoz H., Tolosa J.E., Rojas I., *Premature labor and intra-amniotic infection. Clinical aspects and role of the cytokines in diagnosis and pathophysiology*, Clinics in Perinatology 22: 281-342; 1995.

Graham J.W. & Donaldson S.I., *Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and the use of follow-up data*, Journal of Applied Psychology, 78, pp. 119-128; 1993.

Grande E. & Luzi O., *Metodologie per l'imputazione delle mancate risposte parziali: analisi critica e soluzioni disponibili in ISTAT*, ISTAT, Servizio delle Metodologie di Base per la produzione statistica; 2002.

Grannis S.J., Overhage J.M., McDonald C.J., *Analysis of identifier performance using a deterministic linkage algorithm*, Proc AMIA Symp. (2002) 305-309; 2002.

Graven S.N. & Misenheimer H.R., *Respiratory Distress Syndrome and the high risk mother*, American Journal of Diseases of Children 109: 489-494; 1965.

Greci L.S., Gilson G.J., Nevils B., Izquierdo L.A., Qualls C.R., Curet L.B., *Is amniotic fluid analysis the key to preterm labor? A model using interleukin-6 for predicting rapid delivery*, American Journal of Obstetrics and Gynecology 179: 172-178; 1998.

Grower J.C., *Some distance properties of latent root and vector methods used in multivariate analysis*, Biometrika 53: 325-338, 1966.

Gurel O., Ikegami M., Chroneos Z.C., Jobe A.H., *Macrophage and type II cell catabolism of SP-A and saturated phosphatidylcholine in mouse lungs*, American Journal of Physiology. Lung Cellular and Molecular Physiology 280: L1266-L1272; 2001.

Guseo R., *Istruzioni di Statistica. Lezioni*, Cedam – Padova, 1997.

Haagsman H.P & Golde L.M., *Synthesis and assembly of lung surfactant*, Annual Review of Physiology 53: 441-464; 1991.

Hafez M., el-Sallab S., Khashaba M., Risk M.S., el-Morsy Z., Bassiony M.R., el-Kenawy F., Zaghloul W., *Evidence of HLA-linked susceptibility gene(s) in respiratory distress syndrome*, Disease Markers 7: 201-208; 1989.

Hallman M., Arjomaa P., Mizumoto M., Akino T., *Surfactant proteins in the diagnosis of fetal lung maturity. I. Predictive accuracy of the 35 kD protein, the lecithin/sphingomyelin ratio, and phosphatidylglycerol*, American Journal of Obstetrics and Gynecology 158: 531-535; 1988.

Hallman M., Merritt T.A., Akino T., Bry K., *Surfactant protein A, phosphatidylcholine, and surfactant inhibitors in epithelial lining fluid. Correlation with surface activity, severity of respiratory distress syndrome, and outcome in small premature infants*, The American Review of Respiratory Disease 144: 1376-1384; 1991.

Hallman M., *Cytokines, pulmonary surfactant and consequences of intrauterine infection*, Biology of the Neonate 76 Suppl 1: 2-9; 1999.

Hallman M., Marttila R., Pertile R., Ojaniemi M., Haataja R., *Genes and environment in common neonatal lung disease*, Neonatology 91: 298-302; 2007.

Hatzis D., Deiter G., deMello D.E., Floros J., *Human surfactant protein-C: genetic homogeneity and expression in RDS; comparison with other species*, Experimental Lung Research 20: 57-72; 1994.

Hauck W.W. & Donner A., *Wald's test as applied to hypotheses in logit analysis*, Journal of the American Statistical Association, 72, 851-853; 1977.

Hawgood S. & Clements J.A., *Pulmonary surfactant and its apoproteins*, The Journal of Clinical Investigation 86: 66-70; 1990.

Heitjan D.F. & Rubin D.B., *Inference from Coarse Data via Multiple Imputation with Application to Age Heaping*, Journal of the American Statistical Association, 85, 410, 304-314; 1990.

Hennekens C.H. & Buring J.E., *Epidemiology in Medicine*, Little, Brown & co. Boston/Toronto, 1987.

Hill A. Bradford & Hill I.D., *Bradford Hill's Principles of Medical Statistics*, 12th edn, London: Arnold, 1991.

Horton N.J. & Lipsitz S.R., *Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables*, The American Statistician, 55, 3, 244-254; 2001.

Hosmer D.W. & Lemeshow S., *Applied Logistic Regression*, Second Edition. Wiley Series in Probability and Statistics, Wiley, Inc., New York, 2000.

Howe G.R., *Use of computerized record linkage in cohort studies.* Epidemiologic Reviews 1998; 20:112-21; 1998.

Hulsey T.C., Alexander G.R., Robillard P.Y., Annibale D.J., Keenan A., *Hyaline membrane disease: the role of ethnicity and maternal risk characteristics*, American Journal of Obstetrics and Gynecology 168: 572-576; 1993.

Johansson J. & Curstedt T., *Molecular structures and interactions of pulmonary surfactant components*, European Journal of Biochemistry 244: 675-693; 1997.

Johnson S.C., *Hierarchical clustering schemes*, Psychometrika 32: 241-254, 1967.

Johnston M & Pollard B. *Consequences of disease: testing the WHO International Classification of Impairments, Disabilities and Handicaps (ICIDH) model*. Social Science & Medicine 2001; 53: 1261-1273; 2001.

Juster F.T. & Smith J.P., *Improving the quality of economic data: Lessons from the HRS and AHEAD*, Journal of the American Statistical Association 92, 27; 1998.

Kala P., Koptides M., DiAngelo S., Hoover R.R., Lin Z., Veletza V., Kouretas D., Floros J., *Characterization of markers flanking the human SP-B locus*, Disease Markers 13: 153-167; 1997.

Kala P., Ten Have T., Nielsen H., Dunn M., Floros J., *Association of pulmonary surfactant protein A (SP-A) gene and respiratory distress syndrome: interaction with SP-B*, Pediatric Research 43: 169-177; 1998.

Kalton G., *Compensating for Missing Survey Data*. Michigan,1983.

Kalton G. & Kasprzyk D., *Imputing for Missing Survey Responses.* Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 22-31; 1982

Kalton G. & Kasprzyk D., *The Treatment of Missing Survey Data*. Survey Methodology, Vol. 12, n. 1, pp. 1-16; 1986

Kazanjian A., *Understanding women's health through data development and data linkage: implications for research and policy*, CMAJ 159 (4) (1998) 342-345; 1998.

Kelsey J.L., Thompson W.D., Evans A.S., *Methods in Observational Epidemiology*, Oxford University Press, New York; 1986.

Kendall M.G., *Multivariate Analysis*, 2nd edition. London: Griffin, 1980.

Khattree R. & Naik D.N., *Multivariate Data Reduction and Discrimination with SAS Software*, SAS Institute Inc., 2000.

Khoury M.J., Marks J.S., McCarthy B.J., Zaro S.M., *Factors affecting the sex differential in neonatal mortality: the role of respiratory distress syndrome*, American Journal of Obstetrics and Gynecology 151: 777-782; 1985.

Kim H., Golub G.H., Park H, *Missing value estimation for DNA microarray gene expression data: local least squares imputation*, Bioinformatics, 22(11):1410-1411; 2006.

King R.J., *Pulmonary surfactant*, Journal of Applied Physiology 53: 1-8; 1982.

Knuth D., *The art of computer programming: sorting and searching*, Reading Massachusetts: Addison-Wesley; 1973.

Koskinen R., Meriläinen J., Gissler M., Virtanen M., *Finnish Perinatal Statistics 1997-1998*. Statistical report 41/1999, 1-128. Helsinki, Finland, National Research and Development Centre for Welfare and Health; 1999.

Krzanowski W.J., *Principles of Multivariate Analysis: A User's Perspective*, Oxford: Clarendon Press, 1988.

Lachenbruch P.A., *Discriminant analysis*, Hafner Press, A Division of Macmillan Publishing Co., Inc, 1975.

Larsen M.D. & Rubin D.B., *Iterative automated record linkage using mixture models*, Journal of the American Statistical Association 2001; 96:32–41; 2001.

Lee K., Khoshnood B., Wall S.N., Chang Y., Hsieh H.L., Singh J.K., *Trend in mortality from respiratory distress syndrome in the United States, 1970-1995*, Journal of Pediatric 134: 434-440; 1999.

Leonardi M. *ICF: la Classificazione Internazionale del Funzionamento, della Disabilità e della Salute dell'Organizzazione Mondiale della Sanità. Proposte di lavoro e di discussione per l'Italia*. MR, Giornale Italiano di Medicina Riabilitativa 17 (1): 53-59; 2003.

Lessler J.T. & Kalsbeek W.D., *Nonsampling Error in Surveys*, New York, Chichester, 1992.

Li B., Quan H., Fong A., Lu M., *Assessing record linkage between health care and Vital Statistics databases using deterministic methods*. BMC Health Services Research 6: 48; 2006.

Likert R., *A Technique for the Measurement of Attitudes*, Archives of Psychology 140: 1-55; 1932.

Lipsitz S.R., Zhao L.P., Molenberghs G., *A Semiparametric Method of Multiple Imputation*, Journal of the Royal Statistical Society, Series B, Statistical Methodology, 60, 1, 127-144; 1998.

Little R.J.A., *Survey nonresponse adjustments.* Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 1-10; 1984

Little R.J.A. & Rubin D.B., *Statistical analysis with missing data*, Wiley & Sons, New York, 1987.

Little R.J.A. & Rubin D.B., *Statistical Analysis with Missing Data*, Second Edition, New York, 2002.

Liu S. & Wen S.W., *Development of record linkage of hospital discharge data for the study of neonatal readmission*, Chronic Diseases in Canada 20 (2): 77-81; 1999.

MacMahon B. & Pugh T.F., *Epidemiology: Principles and Methods*, Boston, Mass.: Little Brown, 1970.

MacNaughton-Smith P., *Some statistical and other numerical techniques for classifying individuals*, Home Office Research Unit Report No 6, HMSO London, 1965.

Magagnoli U., *Elementi di statistica descrittiva*, CLUEB, Bologna; 1993.

Mardia K.V., Kent J.T., Bibby J.M., *Multivariate Analysis*, London: Academic Press, 1979.

Mason R.J., Greene K., Voelker D.R., *Surfactant protein A and surfactant protein D in health and disease*, American Journal of Physiology 275: 1-13; 1998.

McCormick S.M., Boggaram V., Mendelson C.R., *Characterization of mRNA transcripts and organization of human SP-A1 and SP-A2 genes*. American Journal of Physiology, 266: 354-366; 1994.

McCullagh P. & Nelder J.A., *Generalized Linear Models*, Second Edition, Chapman & Hall, London, 1989.

Mazzetti A., *Reti Neurali Artificiali. Introduzione ai principali modelli e simulazione su Personal Computer*, Rai/Apogeo, 1991.

Mc Cusker J., *Epidemiology in community health*, Nairobi; AMREF, 1978.

Miettinen O.S., *Estimability and estimation in case-referent studies*, American Journal of Epidemiology 103: 226-235, 1976.

Microsoft Office, *Manuale dell'utente*, *Sistema di gestione di database relazionali per Windows*. Microsoft Corporation, Ireland, 1992.

Newcombe H.B., Kennedy J.M., Axford S.J., James A.P., *Automatic linkage of vital and health records*. Science (1959), 130: 954-959; 1959.

Newcombe H.B., *Handbook of record linkage: methods for health and statistical studies, administration and business.* Oxford, England: Oxford University Press, 1988.

Nogee L.M., *Genetics of the hydrophobic surfactant proteins*, Biochimica et Biophysica Acta 1408: 323-333; 1998.

Oosterlaken-Dijksterhuis M.A., Haagsman H.P., van Golde L.M., Demel R.A., *Characterization of lipid insertion into monomolecular layers mediated by lung surfactant proteins SP-B and SP-C*, Biochemistry 30: 10965-10971; 1991.

Pates R.D., Scully K.W., Einbinder J.S., Merkel R.L., Stukenborg G.J., Spraggins T.A., Reynolds C., Hyman R., Dembling B.P., *Adding value to clinical data by linkage to a public death registry*, Medinfo. 10 (Pt 2) (2001) 1384-1388; 2001.

Pocock S.J., *Clinical Trials. A Practical Approach*, John Wiley & Sons Ltd, 1989.

Raghunathan T.E., *What do we do with missing data? Some options for analysis of incomplete data*, Annual Review of Public Health, 25, 99-117; 2004.

Rämet M., Haataja R., Marttila R., Floros J., Hallman M., *Association between the Surfactant Protein A (SP-A) Gene Locus and Respiratory-Distress Syndrome in the Finnish Population*, American Journal of Human Genetics 66: 1569-1579; 2000.

Rao C.R., *Linear Statistical Inference and Its Application*, Second Edition. Wiley, Inc., New York, 1973.

Rooney S.A., Young S.L., Mendelson C.R., *Molecular and cellular processing of lung surfactant*, FASEB Journal 8: 957-967; 1994.

Rothman K.J., *Modern Epidemiology*, Little, Brown and company edition, Boston/Toronto; 1986.

Rothman K.J. **& Greenland S.**, *Modern Epidemiology*, Third Edition, Lippincott-Raven, Philadelphia; 1998.

Rubin D.B., *Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys*. Journal of the American Statistical Association. 72, 538-543; 1977.

Rubin D.B., *Multiple Imputation for Nonresponse in Surveys*, New York, Chichester, 1987.

Rubin D.B., *Multiple Imputation after 18+ Years*, Journal of the American Statistical Association, 91, 434, 473-489; 1996.

Rubin D.B. & Schenker N., *Multiple Imputation for Interval Estimation from Simple Random Samples With Ignorable Nonresponse*, Journal of the American Statistical Association, 81, 394, 366-374; 1986.

SAS Institute Inc, SAS Procedures Guide, Version 8, Cary, NC: SAS Institute Inc, 1999.

Scavalli E., *Edit and Imputation Using MLP Neural Networks in SARs Data*. EUREDIT internal report, 2002

Schafer J. L., *Analysis of incomplete multivariate data*. London: Chapman & Hall, 1997.

Schafer J. L. & Olsen M. K., *Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective*, Multivariate Behavioral Research, 33, 545-571; 1998.

Schafer J.L. & Graham J.W., *Missing Data: Our View of the State of the Art*, Psychological Methods, 7, 2, 147-177; 2002.

Schlesselman J.J., *Case-Control Studies. Design, Conduct, Analysis*. Oxford University Press, New York; 1982.

Shannon C.E. & Weaver W., *The mathematical theory of communication*, Urbana, Uiv. Of Illinois Press, 1949.

Shill W., Jöckel K. H., Drescher K., Timm J., *Logistic analysis in case-control studies under validation sampling*, Biometrika 1993 80(2):339-352; 1993.

Schulte Nordholt E., *Imputation: Methods, Simulation, Experiments and Practical Examples*. International Statistical Review, Vol. 66, n. 2, pp. 159-180; 1998.

Sheehe P.R., *Dynamic risk analysis in retrospective matched pair studies of disease*, Biometrics 18: 323-341, 1962.

Simeonsson R. J., Lollar D., Hollowell J., Adams M., *Revision of the International Classification of Impairments, Disabilities and Handicaps. Developmental Issues.* Journal of Clinical Epidemiology; 53: 113-124; 2000.

Sinharay S., Stern H.S., Russell D. *The Use of Multiple Imputation for the Analysis of Missing Data*, Psychological Methods, 6, 317-329; 2001.

Sokal R.R. & Michener C.D., *A statistical method for evaluating systematic relationships*, Kansas University Science Bulletin, 38: 1409-1438, 1958.

Soliani A., *Statistica applicata alla ricerca e alle professioni scientifiche*, Dispense 2005 dal sito http://www.dsa.unipr.it/soliani.

Stokes M.E., Davis C.S., Koch G.G., *Categorical Data Analysis using the SAS System*, SAS Institute Inc., SAS Campus Drive, Cary, North Carolina, USA, 1995.

Sutherland H.J., Meslin E.M., Till J.E., *What's missing from current clinical trial guidelines? A framework for integrating science, ethics, and the community context*, The Journal of clinical ethics 5(4): 297-303; 1994.

Thuriaux M.C. *Les conséquences de la maladie et leur mesure: introduction*. World Health Statistics Quaterly; 42 (3): 110-114; 1989.

Tanner M.A. & Wong W.H., *The Calculation of Posterior Distributions by Data Augmentation*, Journal of the American Statistical Association, **82**, 528-550; 1987.

Timm N. H. & Mieczkowski T. A., *Univariate & Multivariate General Linear Model: theory and applications using SAS software*. SAS Institute Inc., Cary, NC, USA, 1997.

Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., Altman R.B., *Missing value estimation methods for DNA microarrays*, Bioinformatics, 17(6):520-525; 2001.

Vajani L., *Statistica Descrittiva*, EtasLibri Editore, 1997.

Vamvakopoulos N.C., Modi W.S., Floros J., *Mapping the human pulmonary surfactant-associated protein B gene (SFTP3) to chromosome 2p12-->p11.2*, Cytogenetics and Cell Genetics 68: 8-10; 1995.

Van Den Brandt P.A., Schouten L.J., Goldbohm R.A., Dorant E., Hunen P.M.H., *Development of a record linkage protocol for use in the Dutch Cancer Registry for epidemiological research.* International Journal of Epidemiology 1990; 19: 553-8; 1990.

Veletza S.V., Rogan P.K., Ten Have T., Olowe S.A., Floros J., *Racial differences in allelic distribution at the human pulmonary surfactant protein B gene locus (SP-B)*, Experimental Lung Research 22: 489-494; 1996.

Von Mises R., *On the classification of observation data into distinct groups*, Annual of Mathematic and Statistics 16: 68-73; 1945.

Waien S.A., *Linking large administrative databases: a method for conducting emergency medical services cohort studies using existing data.* Academic Emergency Medicine (official journal of the Society for Academic Emergency Medicine) 4: 1087-95; 1997.

Warr R.G., Hawgood S., Buckley D.I., Crisp T.M., Schilling J., Benson B.J., Ballard P.L., Clements J.A., White R.T., *Low molecular weight human pulmonary surfactant protein (SP5): isolation, characterization, and cDNA and amino acid sequences*, Proceedings of the National Academy of Sciences U S A 84: 7915-7919; 1987.

Weiner M., Stump T.E., Callahan C.M., Lewis J.N., McDonald C.J., *A practical method of linking data from Medicare claims and a comprehensive electronic medical records system*, International Journal of Medical Informatics 71 (1): 57-69; 2003.

Welch B.L., *Note on discriminant functions*, Biometrika 31: 218-220; 1939.

Williams M.C., *Conversion of lamellar body membranes into tubular myelin in alveoli of fetal rat lungs*, The Journal of Cell Biology 72: 260-277; 1977.

Winkler W.E., *Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage*, Proceedings of the Section on Survey Research Methodology. American Statistical Association: 1988; 667–671; 1988.

World Health Organization, *International Classification of Diseases 9th Revision – Clinical Modification (ICD9-CM)*, Geneva, Switzerland: WHO, 1975.

World Health Organization. *The International Classification of Impairments, Disability and Handicap: a manual of classification relating to consequences of diseases*. Geneva, Switzerland: WHO, 1980.

World Health Organization. *ICF Classificazione Internazionale del Funzionamento, della Disabilità e della Salute*. Ginevra-Milano: WHO, 2002.

World Health Organization. *The International Classification of Functioning, Disability and Health*. Geneva, Switzerland: WHO, 2001.

Wright J.R. & Clements J.A., *Metabolism and turnover of lung surfactant*, The American Review of Respiratory Disease 136: 426-444; 1987.

Xian W., Ao L., Zhaohui J., Huanqing F., *Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme*, Bioinformatics, 22;7:32; 2006.

Yaruss J. S. & Quesal R. W. *Stuttering and the International Classification of Functioning, Disability and Health (ICF): An update*. Journal of Communication Disorders 2004; 37(1): 35-52; 2004.

RIASSUNTO

Negli studi clinici le unità statistiche su cui viene condotta la ricerca sono chiamate casistiche cliniche, vale a dire gruppi di pazienti o volontari sani che prendono parte allo studio con lo scopo di migliorare le conoscenze, le terapie ed i trattamenti in campo biomedico. Lo studio clinico serve a chiarire una serie programmata di domande e viene condotto secondo uno schema ben definito che consiste nel disegno di studio. Esistono due principali tipologie di disegni di studio statistici, il disegno di studio sperimentale o *clinical trial* e l'indagine o *survey*. Il *clinical trial* richiede un'interferenza pianificata secondo il corso naturale degli eventi, tale da poterne osservare gli effetti. Nella maggior parte dei casi un *clinical trial* viene adottato in studi atti a testare un nuovo farmaco o una nuova terapia su un particolare gruppo di soggetti, vale a dire pazienti affetti da una specifica patologia. Nell'indagine, o *survey*, il ricercatore è un osservatore meno attivo, che interferisce il meno possibile con i fenomeni da registrare. Il disegno e la conduzione di qualsiasi tipo di studio clinico richiedono sicuramente di esaminare ed indagare importanti questioni bio-mediche, ma richiedono anche di essere basati su una rigorosa metodologia che possa fornire una risposta corretta al quesito di ricerca. Per questo motivo il presente lavoro è stato suddiviso in cinque sezioni principali seguendo uno schema che possa guidare il lettore nell'analisi degli aspetti critici che si possono affrontare, in primo luogo nella gestione dei dati, in secondo luogo nella progettazione, nello sviluppo e nelle fasi finali di uno studio clinico. La sezione *A, Problemi di gestione di banche dati*, propone un capitolo sul *Record Linkage* ed un secondo sui *Missing Values*. Il *Record Linkage* consiste in una serie di tecniche, prevalentemente informatico-statistiche, per l'integrazione di dati e per la fusione di più dataset sulla base di variabili identificative dei pazienti. Il problema dei *missing values* (valori mancanti) è inevitabilmente presente in uno studio: quando un database è affetto da valori mancanti si può ricorrere a specifici sistemi che, attraverso programmi statistici, riescono a calcolare un opportuno valore per l'unità priva d'informazione.

Nella sezione *B, Disegni di studio*, vengono analizzati in un primo tempo i *clinical trial*, in un secondo tempo i *survey*, scorporando quest'ultimi in studi coorte e studi caso-controllo. In uno studio coorte (altrimenti detto studio *follow-up* o studio prospettico) uno o più gruppi di individui vengono definiti secondo l'esposizione o meno a uno o più fattori considerati di rischio per una patologia o una gamma di patologie. Gli individui oggetto di studio vengono seguiti prospetticamente nel tempo, in modo da poterne osservare l'incidenza di malattia/e e correlarla ai fattori eziologici. Uno studio caso-controllo (o studio retrospettivo) fornisce una strategia di ricerca per investigare fattori che possono prevenire o causare una particolare

malattia. Il metodo comporta un confronto tra pazienti affetti dalla patologia (casi) con un gruppo di controllo (persone sane). Il confronto ha l'obiettivo di scoprire i fattori che possono differire nei due gruppi spiegando il manifestarsi della patologia nei pazienti.

La sezione *C* concerne il *trattamento delle variabili* e le *possibili classificazioni della variabile di outcome*. Nel primo caso si considerano le tipologie di variabili che si possono incontrare in uno studio clinico, le loro scale di misura, la sintetizzazione attraverso indicatori e le possibili trasformazioni che a volte risultano necessarie. Nel secondo caso si presentano le classificazioni della variabile di outcome (o variabile risposta), che riguardano una specifica patologia, i sintomi riportati dal paziente oppure una conseguenza di malattia, vale a dire una menomazione, una disabilità o un handicap.

Le sezioni *D* ed *E* affrontano i problemi nella scelta del metodo d'analisi multivariata. La sezione *D* mira a dare una sintesi dei principali metodi d'analisi, suddividendoli in base alla simmetria o meno delle variabili. Nel caso vi sia una distinzione tra variabile risposta e variabili esplicative si propone un metodo d'analisi di regressione multipla (*stepwise* o *logistica*), o l'analisi discriminatoria, sulla base del tipo (e della quantità) di variabile/i risposta; se invece le variabili in analisi sono tutte sullo stesso piano si propongono l'*analisi fattoriale*, il metodo delle *componenti principali*, l'*analisi delle corrispondenze* e la *cluster analysis* sulla base degli obiettivi specifici dello studio. Nella sezione *E* viene sviluppata l'analisi di regressione logistica, cogliendone gli aspetti centrali ed esaminando il caso dell'analisi di regressione logistica condizionata per il disegno *Matched Case-Control Study*.

Ogni capitolo è strutturato in modo tale che nella prima parte sia riportata una sintesi dei metodi teorici e pratici proposti dalla letteratura per risolvere il problema in questione, mentre alla fine di ogni capitolo è stata inserita una sezione che spiega come il problema trattato è stato affrontato ed applicato in uno specifico studio clinico-genetico al Children's Hospital dell'Università di Oulu – Finlandia. In tale Centro è stato svolto un lavoro sostanziale sulla creazione di un database inerente a dati clinici e genetici di nati prematuri nei tre principali ospedali della Finlandia centro-settentrionale, Oulu, Tampere e Seinäjoki. Attraverso l'utilizzo del database totale costituito dai pazienti dei tre nosocomi, è stato condotto uno studio caso-controllo (*match* 1-1) clinico-genetico sull'associazione tra Sindrome da Distress Respiratorio (in inglese *Respiratory Distress Syndrome*, RDS) ed i polimorfismi dei singoli nucleotidi dei geni SP-A e SP-B. L'intero progetto è stato diretto dal Prof. M. Hallman.

ABSTRACT

In clinical studies the statistical units on which the research is carried out are called clinical cases, i.e. groups of patients or healthy volunteers participating in the study with the goal to improve knowledge, therapies and treatments in biomedicine. A clinical study is useful to clarify some planned questions and it follows a well defined scheme (the study design). There are two main types of statistical study designs, the *clinical trial* and the *survey*. The clinical trial requires a planned interference by the researcher during the natural course of events, so that it's possible to study the effects. In most cases a clinical trial is adopted to test a new drug or therapy on a particular group of patients with a specific disease. In a survey the researcher is a less active observer and he meddles as less as he can in the phenomena to record. The design of each clinical study surely requires to investigate important biomedical problems, but it also needs to be based on a strong methodology giving a correct answer to the research questions. For this reason this PhD thesis has been divided in five main sections following a scheme guiding the reader in the analysis of critical aspects it's possible to find, first in data management, then in the study design and during the development and the final phases of a clinical study.

Section *A*, *Problems in database management*, proposes a chapter about *Record Linkage* and a second chapter about *Missing Values*. *Record Linkage* consists in a series of statistical and data processing techniques for data integration and for the union of more datasets on the basis of patients' identifying variables. The problem of *missing values* is frequently present in a study: when a database contains missing values it's possible to turn to specific systems useful to calculate a suitable value for the unit without information, through statistical programs.

In section *B*, *Study design*, first *clinical trials* are analysed, then *surveys* are described and divided in *cohort studies* and *case-control studies*. In a cohort study (also called follow-up or prospective study) one or more groups of subjects are defined in accordance with the exposure to risk factors for one or more diseases, or not. Subjects are prospectively followed to study the disease(s) incidence and to observe if the disease(s) is(are) correlated with the etiological factors. A case-control study (or retrospective study) provides a research strategy to investigate possible factors preventing or causing a particular disease. The method implies a comparison between patients with the disease (cases) and a control group (healthy subjects). The comparison ends to find out factors which can be different in the two groups, explaining the presence of the disease in patients.

Section *C* concerns *Variables treatment* and *Outcome variable classifications*. In the first part, types of variables, their measurement scales, their reduction in indicators and their

transformations are analysed. In the second part the classifications of the outcome variable is described. These classifications may regard a specific disease (ICD IX and X), the patient's symptoms (ICD) or a disease effect, i.e. an impairment, a disability or a handicap (ICIDH and ICF).

Sections *D* and *E* deal with choice of the *multivariate analysis method*. Section *D* aims to give a synthesis of the main analysis methods, dividing them in accordance with the variables symmetry. If there is a distinction between outcome variable and covariates, a multiple regression (stepwise or logistic) analysis method is suggested; discriminant analysis is useful if there are more outcome variables. If variables are all on the same level, factors analysis, principal components analysis, correspondences analysis and cluster analysis are proposed in conformity with the specific analysis aims. In section *E logistic regression analysis* is developed, pointing out the main aspects and considering the conditional logistic regression analysis for a *Matched Case-Control Study* design.

Every chapter is organized in two parts: in the first one a synthesis of theory and practical methods are given to explain how literature deals with the specific problem in a clinical study; in the second one a specific application is presented to show how the author solved the methodology problems in a particular clinical-genetic study carried out at the Children's Hospital of Oulu University – Finland. In this Institute a substantial work on a preterm infants database creation has been performed. The database collected all the clinical and genetic information about preterm infants born in one of the three main hospitals in Northern-central Finland, Oulu, Tampere and Seinäjoki. Through this final database containing all patients from the three hospitals, a clinical-genetic case-control study (*match* 1-1) on the association between *Respiratory Distress Syndrome* (RDS) and single nucleotide polymorphisms (SNPs) has been carried out using a conditional logistic regression analysis. The project was directed by Prof. M. Hallman.

Sezione G

Allegati

ALLEGATO I

PEER-REVIEWED ARTICLE: PUBLISHED IN *NEONATOLOGY* 2007, vol. 91, n. 4: 298-302

Genes and Environment in Common Neonatal Lung Disease

Review

Neonatology
formerly Biology of the Neonate

Neonatology 2007;91:298–302
DOI: 10.1159/000101345

Published online: June 7, 2007

Genes and Environment in Common Neonatal Lung Disease

Mikko Hallman[a] Riitta Marttila[b] Riccardo Pertile[c] Marja Ojaniemi[a]
Ritva Haataja[a]

[a]Department of Pediatrics, Biocenter Oulu, University of Oulu, Oulu, [b]Central Hospital of Southern Osthrobotnia, Seinäjoki, Finland; [c]University of Padua, Padua, Italy

Key Words
Respiratory distress syndrome · Bronchopulmonary dysplasia · Preterm birth · Prematurity · Genetics · Genome-wide screening · Biobank

Abstract
Respiratory distress syndrome (RDS) and bronchopulmonary dysplasia (BPD) are common, serious lung diseases in preterm infants. Polymorphism of the genes involved in basic lung function and alveolar stability, lung differentiation and pulmonary host defense may influence the risk. Natural selection has refined the genes responsible for cardiopulmonary adaptation and resistance against pneumonia in term and near-term infants. Before the era of antibiotics, however, virtually all very preterm infants died of asphyxia, respiratory failure or infections. Today, the degree of prematurity plays a dominant role in susceptibility to serious lung disease. In addition, genetic polymorphism and constitution modulate the risk of RDS and BPD that have different, partly overlapping predisposition. According to twin studies, the genetic impact on the risk of RDS and BPD among preterm and very preterm infants is 35–65%. Individual disease genes generally have low penetrance. Large-scale genetic studies are required as part of neonatal and perinatal research in order to learn about the risk factors and to investigate pharmacogenetics. The aim in the future is to individualize therapies.

Introduction

Perinatal respiratory adaptation is an effective sequence of complex events. However, respiratory failure is the most common cause of death in early infancy. Successful respiratory adaptation in near-term infants has been an evolutional advantage. Previously, however, the population of very prematurely born infants had additional lethal diseases in the event of them not dying of asphyxia or respiratory distress syndrome (RDS) shortly after birth. As a result of new prenatal and neonatal treatments, the prognosis of these infants has improved dramatically. Because the limit of viability has been lowered by 8–10 weeks within 50 years, prematurity continues to be the prominent basic cause of perinatal and infant death and the major cause of chronic disability, particularly due to bronchopulmonary dysplasia (BPD) and neurosensory diseases.

RDS presents as a transient deficiency of alveolar surfactant that is influenced by polymorphisms of the genes playing a critical role in homeostasis of type 2 alveolar cells. Apart from surfactant deficiency, a number of interactive constitutional, environmental and genetic factors disturb neonatal respiratory adaptation and delay the recovery from RDS, and may thus influence the genetic predisposition. The susceptibility to BPD is likely to be affected by polymorphisms of genes influencing a number of critical pulmonary functions ranging from

KARGER
Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1661–7800/07/0914–0298$23.50/0

Accessible online at:
www.karger.com/neo

Mikko Hallman, MD
Department of Pediatrics, Biocenter Oulu Laboratory, University of Oulu
PO Box 5000
FIN–90014 Oulu (Finland)
Tel. +358 8 315 5100, Fax +358 8 315 5559, E-Mail mhallman@cc.oulu.fi

the maintenance of gas exchange to host defense, lung growth and differentiation.

The susceptibility of very preterm infants to RDS or BPD has not been directly under the selection pressure of evolution, as all these infants used to die early. It is therefore not very surprising to observe that the genotypes increasing the risk have generally a high frequency. In the present brief review, we discuss the genetic susceptibility to common lung diseases after premature birth.

Diagnosis of Lung Disease in Small Preterm Infants

The definition of lung diseases in very preterm (VLGA; <32 weeks of gestation) infants is not always accurate. However, it is important to systematically define these diseases and to try to improve diagnostic methods.

BPD is prevalent among VLGA infants, especially among those born extremely preterm (ELGA; <28 weeks of gestation). BPD in VLGA infants is defined as a generalized lung disease requiring supplemental oxygen at 28 days of age (CLD or 'mild' BPD) and at 36 weeks of postmenstrual age (moderate to severe BPD) [1, 2]. A longer follow-up may give a more accurate definition, as some infants with BPD have low morbidity later in infancy. Others many suffer from recurrent hospitalizations due to lung disease, which contributes to delay in growth and neurodevelopment.

The diagnosis of RDS has become problematic due to the various treatments available at or after birth. Surfactant therapy in the delivery room is considered an exclusion criterion for genetic studies of RDS, since this treatment may conceal the characteristic symptoms and diagnostic changes in chest x-rays. On the other hand, antenatal glucocorticoid treatment, which decreases the risk of RDS, is a confounder rather than a contraindication for genetic studies. The major haplotype of *SP-A1* 6A2 is a risk factor among VLGA infants. This haplotype of SP-A is very prominent among those VLGA infants who develop RDS despite antenatal glucocorticoid treatment, and is underrepresented among those VLGA infants with neither antenatal corticosteroid treatment nor RDS [3].

Other Lung Diseases

The susceptibility to other neonatal pulmonary diseases, including infections or meconium aspiration syndrome, is also likely to have a genetic component. Genetic susceptibility to severe respiratory syncytial virus infection during epidemics has been studied [4, 5]. In mice, the susceptibility to group B β-hemolytic streptococcal (GBS) pneumonia appears to have a genetic background [6] and genetic predisposition to other neonatal pulmonary infections is likely [7]. However, very few studies have been published so far.

Differentiation between Genetic and Environmental Factors

The close association between genetic and environmental factors complicates the evaluation of any genetic impact on a risk of disease. Another factor affecting the evaluation is the extreme phenotype. For instance, the generalized multiorgan immaturity in ELGA infants born before 23–25 weeks' gestation is due to their extreme vulnerability and lack of adequate host defense. RDS is virtually always diagnosed in these cases. In addition, BPD and diseases involving the central nervous system, the cardiovascular system and the gastrointestinal tract are prevalent. Inclusion of these cases may decrease the penetrance of the disease genes or result in an evaluation of the genetic risk of extreme prematurity rather than the associated diseases. Genetic association with RDS or BPD may become detectable first when the population is limited on the basis of the degree of prematurity or of other environmental stress factors.

RDS and BPD are separate, partly overlapping clinical entities. However, the genetic and environmental risk factors and their interactions are not likely to be homogenous. The phenotype of RDS in the population of near-term infants is distinguishable from that among VLGA infants. The alleles, genotypes and haplotypes associating with RDS in near-term infants are mostly different from the alleles associating with RDS in VLGA infants. The disease genotypes in near-term infants tend to have a low frequency and rather high penetrance [8], whereas the disease genotypes in the RDS of VLGA infants tend to have a high frequency and low penetrance [3, 9, 10]. In contrast, certain rare mutations with very high penetrance are associated with fatal respiratory failure in mostly term-born infants [11].

The well-known environmental (e.g. gestational age) and constitutional (e.g. gender) factors influencing the risk of the lung disease are usually analyzed as independent variables. An association between the genotype, the risk factor (or protective factor) and the phenotype may, at best, give a clue to the function of the disease genotype.

Family Studies. Studies on sibs or other relatives help to link the common hereditary factors with the risk of disease, providing a numerical estimate of hereditability. Genome-wide linkage analysis of large families with a high risk of the disease could optimally define the linkage between the disease and a DNA locus. However, it would be difficult to recruit a sufficient study population, since these diseases are predominant only in preterm infants. Although the recurrence rate of preterm birth increases 3-fold after a single premature birth and 6-fold after two premature births, the variation in the degree of prematurity complicates the assessment of the risk. Inclusion of the families with recurrent prematurity will limit the population to cases with a genetic predisposition to preterm birth. In twin pregnancies, on the other hand, the environmental and constitutional differences appear to be small, and the difference in consanguinity between dizygotic (DZ) and monozygotic (MZ) twin pairs (50% among DZ twin pairs, 100% among MZ) consolidates the evaluation of hereditability [12]. Thus, the concordance difference between MZ and DZ twins provides a direct estimate of the hereditability of the disease. Twin studies have yielded evidence indicating that the genetic risks of BPD [13] and RDS [14] are considerable, with a wide range in estimates.

There are caveats in twin studies, however. A difference in the mean gestational age between MZ and DZ twin pairs and the weight or gender difference within twin pairs perturbs the concordance difference estimates. Moreover, the calculated hereditablity figure may not be specific, since unidentified environmental factors disturb the concordance difference. For instance, leukemia cells may be transmitted non-genetically to a MZ twin pair via placental anastomoses [15]. Likewise, the multiple pregnancy setting perturbs the concordance of RDS, as the presenting twin has a lower risk of RDS than the non-presenting twin. The acceleration of lung maturation of the presenting twin is under genetic control that involves the *SP-B* Ile131Thr polymorphism [16]. The low RDS risk is confined to the presenting fetuses, who were carriers of the common *SP-B* Ile131 allele, whereas the RDS risk of the non-presenting fetus is not influenced by the *SP-B* Ile131Thr genotypes.

Studies Identifying the Genetic Locus

Studies on genetic susceptibility generally require a large population sample because of the abundance of dependent variables and the generally low penetrance of genetic variants. The power estimate is mandatory, and the limits of power need to be realized. Prospectively designed case-control studies are preferred whenever the cases and controls show nearly equal population representation, and the major dependent variables (at least the length of gestation) can be equated. On the other hand, population cohorts with minimal selection bias make it possible to evaluate the overall impact of genetic factors. In post-hoc analysis, the independent variables need to be carefully considered. Penetrance, zygosity and genotype frequency influence the risk assessment.

The genetic background of the population requires consideration. The difference in frequencies of individual genotypes as compared between the different populations may influence the genetic risk in a more complex way than anticipated on the basis of the different frequencies (effect of genetic constitution). Therefore, ethnic/racial variation needs to be considered independently. Genetically homogenous populations, particularly when there is evidence on a recent confounder effect [17], are suitable for the identification of a genetic risk. Generally, a smaller sample of a homogenous than a heterogenous population is required for the study.

To ascertain the heredity of a trait, inclusion of the family triad (mother, father and child) as an addition to the case-control setting is preferred. This allows examination of the linkage between the gene and the disease using the transmission disequilibrium test (TDT). In the TDT the aim is to evaluate whether the transmission of the putative disease genotype to individuals with the disease is significantly favored. It is also possible to study the transmission of the protective genotype to healthy 'supernormal' individuals (i.e. healthy despite environmental stress) [18].

Demonstration of a genetic association with the disease requires both case-control studies and those involving studies of genetic linkage (such as TDT). However, positive association needs to be confirmed in other independent studies. Discrepancies in the results are not uncommon, and meta-analyses are required to judge the overall significance of the genetic association.

Candidate Genes

Studies on the pathogenesis of a disease help to define the candidate genes. The evidence on lethal neonatal respiratory failure caused by the lack-of-function mutation in a gene serves as an indication for studies on the gene polymorphism that may increase the risk of multifactorial lung disease (RDS or BPD). Such evidence has been

Table 1. Candidate genes that have a proposed association with susceptibility to RDS and BPD

Disease/gene	Risk (protective) allele or haplotype	References
RDS		
SP-A	SP-A1 6A2 (6A3) haplotype	3, 9, 10, 20
SP-B	131Thr	9, 10, 16, 21, 22
SP-C	186Asn	23, 24
GPRA	H4/5 (H1) haplotype	8
ABCA3	rs13332514 in exon 10	25
IFN-γ	874T	26
BPD		
SP-B	i4del, or various alleles	24, 27
ABCA3	rs13332514 in exon 10	25
TNF-α	(–308A), or no association	28, 29
IFN-γ	(874T)	26
ACE	Deletion allele	30

obtained for at least *SP-B* and *ABCA3*, and several other genes critical for lung function and preferentially expressed in the lung may become interesting targets for research on multifactorial lung diseases [19]. Haplotype analysis yields information about DNA stretches with an exceptionally predictable set of sequences that are in close linkage disequilibrium. Common alleles, haplotypes and genotypes of representative genes are often studied. On the other hand, rare alleles and genotypes may occasionally be significant risk factors even of common multifactorial lung diseases, provided that the detrimental effect of the allele has a high penetrance.

Table 1 lists many of the genes and alleles proposed as candidate genes for RDS and BPD. All studies are preliminary or controversial, and all associations need to be confirmed. In most studies the association is limited to a phenotype that is restricted by the length of gestation at birth.

Genomic Studies

The microsatellites that contain variable repeat sequences of 2–4 base pairs have been commonly used in linkage analysis. Other techniques are increasingly applied. The microchip techniques currently allow the analysis of single nucleotide polymorphisms (SNP) in the range of 1 million. In molecular biology these genomic and proteomic techniques are already commonly applied. In common diseases of adults (diabetes, cardiovascular diseases) genomic approaches are increasingly used and will likely be applied in neonatal medicine as well. The current price of a single SNP (about 10 cents) is perhaps misleading, since besides the association study, a remarkable amount of additional research is required to define a single disease genotype and its function. Selective genomic analysis involving genes with similar functional category (e.g. innate immunity) are considered as well.

Comment

In view of the aim to increase quality-adjusted life years of sick newborn infants, research on genetics of common neonatal diseases is justified. However, such research projects require prioritization, dedicated networks, multidisciplinary teams and sufficient resources. A biobank linked with a database would, in theory, provide accurate genomic, constitutional and hospital data. However, the ethical issues, legislation and acceptance within the community need to be overcome. Computational research and large projects involving translational and clinical research would be required before approaching potential applications. The aim to understand and define the disease genes influencing susceptibility to neonatal lung diseases should not supersede the continuous research and development in other aspects of neonatal medicine. Despite possible delays and problems, genetic studies are likely to expand and diversify. We predict that eventually they will involve trials with pharmacogenetic incentives to develop a new generation of individualized therapies.

References

1 Jobe AH, Bancalari E: Bronchopulmonary dysplasia. Am J Respir Crit Care Med 2001; 163:1723–1729.

2 Walsh MC, Yao Q, Gettner P, Hale E, Collins M, Hensman A, Everette R, Peters N, Miller N, Muran G, Auten K, Newman N, Rowan G, Grisby C, Arnell K, Miller L, Ball B, McDavid G, National Institute of Child Health and Human Development Neonatal Research Network: Impact of a physiologic definition on bronchopulmonary dysplasia rates. Pediatrics 2004;114:1305–1311.

3 Ramet M, Haataja R, Marttila R, Floros J, Hallman M: Association between the surfactant protein A gene locus and respiratory-distress syndrome in the Finnish population. Am J Hum Genet 2000;66:1569–1579.

4 Lofgren J, Ramet M, Renko M, Marttila R, Hallman M: Association between surfactant protein A gene locus and severe respiratory syncytial virus infection in infants. J Infect Dis 2002;185:283–289.
5 Lahti M, Lofgren J, Marttila R, Renko M, Klaavuniemi T, Haataja R, Ramet M, Hallman M: Surfactant protein D gene polymorphism associated with severe respiratory syncytial virus infection. Pediatr Res 2002; 51:696–699.
6 Mancuso G, Midiri A, Beninati C, Biondo C, Galbo R, Akira S, Henneke P, Golenbock D, Teti G: Dual role of TLR2 and myeloid differentiation factor 88 in a mouse model of invasive group B streptococcal disease. J Immunol 2004;172:6324–6329.
7 Hartel C, Konig I, Koster S, Kattner E, Kuhls E, Kuster H, Moller J, Muller D, Kribs A, Segerer H, Wieg C, Herting E, Gopel W: Genetic polymorphisms of hemostasis genes and primary outcome of very low birth weight infants. Pediatrics 2006;118:683–689.
8 Pulkkinen V, Haataja R, Hannelius U, Helve O, Pitkanen OM, Karikoski R, Rehn M, Marttila R, Lindgren CM, Hastbacka J, Andersson S, Kere J, Hallman M, Laitinen T: G protein-coupled receptor for asthma susceptibility associates with respiratory distress syndrome. Ann Med 2006;38:357–366.
9 Haataja R, Ramet M, Marttila R, Hallman M: Surfactant proteins A and B as interactive genetic determinants of neonatal respiratory distress syndrome. Hum Mol Genet 2000;9: 2751–2760.
10 Marttila R, Haataja R, Guttentag S, Hallman M: Surfactant protein A and B genetic variants in respiratory distress syndrome in singletons and twins. Am J Respir Crit Care Med 2003;168:1216–1222.
11 Nogee LM: Genetic mechanisms of surfactant deficiency. Biol Neonate 2004;85:314–318.
12 Teikari JM, Kaprio J, Koskenvuo MK, Vannas A: Heritability estimate for refractive errors – a population-based sample of adult twins. Genet Epidemiol 1988;5:171–181.
13 Bhandari V, Bizzarro MJ, Shetty A, Zhong X, Page GP, Zhang H, Ment LR, Gruen JR, Neonatal Genetics Study Group: Familial and genetic susceptibility to major neonatal morbidities in preterm twins. Pediatrics 2006; 117:1901–1906.
14 Hallman M, Haataja R: Surfactant protein polymorphisms and neonatal lung disease. Semin Perinatol 2006;30:350–361.
15 Greaves MF, Maia AT, Wiemels JL, Ford AM: Leukemia in twins: lessons in natural history. Blood 2003;102:2321–2333.
16 Marttila R, Haataja R, Ramet M, Lofgren J, Hallman M: Surfactant protein B polymorphism and respiratory distress syndrome in premature twins. Hum Genet 2003;112:18–23.
17 Varilo T, Paunio T, Parker A, Perola M, Meyer J, Terwilliger JD, Peltonen L: The interval of linkage disequilibrium detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. Hum Mol Genet 2003;12:51–59.
18 Haataja R, Marttila R, Uimari P, Lofgren J, Ramet M, Hallman M: Respiratory distress syndrome: evaluation of genetic susceptibility and protection by transmission disequilibrium test. Hum Genet 2001;109:351–355.
19 Maeda Y, Dave V, Whitsett JA: Transcriptional control of lung morphogenesis. Physiol Rev 2007;87:219–244.
20 Kala P, Ten Have T, Nielsen H, Dunn M, Floros J: Association of pulmonary surfactant protein A (SP-A) gene and respiratory distress syndrome: interaction with SP-B. Pediatr Res 1998;43:169–177.
21 Floros J, Veletza SV, Kotikalapudi P, Krizkova L, Karinch AM, Friedman C, Buchter S, Marks K: Dinucleotide repeats in the human surfactant protein B gene and respiratory-distress syndrome. Biochem J 1995;305:583–590.
22 Floros J, Thomas NJ, Liu W, Papagaroufalis C, Xanthou M, Pereira S, Fan R, Guo X, Diangelo S, Pavlovic J: Family-based association tests suggest linkage between surfactant protein B (SP-B) (and flanking region) and respiratory distress syndrome (RDS): SP-B haplotypes and alleles from SP-B-linked loci are risk factors for RDS. Pediatr Res 2006;59: 616–621.
23 Lahti M, Marttila R, Hallman M: Surfactant protein C gene variation in the Finnish population – association with perinatal respiratory disease. Eur J Hum Genet 2004;12:312–320.
24 Rova M, Haataja R, Marttila R, Ollikainen V, Tammela O, Hallman M: Data mining and multiparameter analysis of lung surfactant protein genes in bronchopulmonary dysplasia. Hum Mol Genet 2004;13:1095–1104.
25 Karjalainen M, Haataja R, Hallman M: Tagging SNP selection and haplotype analysis of ABCA3 gene polymorphisms in relation to respiratory distress syndrome in preterm infants. Proc Am Thoracic Soc 2007, in press.
26 Bokodi G, Derzbach L, Banyasz I, Tulassay T, Vasarhelyi B: Association of interferon-γ T+874A and interleukin-12 p40 promoter CTCTAA/GC polymorphism with the need for respiratory support and perinatal complications in low birthweight neonates. Arch Dis Child Fetal Neonatal Ed 2007;92:F25–F29.
27 Makri V, Hospes B, Stoll-Becker S, Borkhardt A, Gortner L: Polymorphisms of surfactant protein B encoding gene: modifiers of the course of neonatal respiratory distress syndrome? Eur J Pediatr 2002;161:604–608.
28 Kazzi SN, Kim UO, Quasney MW, Buhimschi I: Polymorphism of tumor necrosis factor-α and risk and severity of bronchopulmonary dysplasia among very low birth weight infants. Pediatrics 2004;114:e243–e248.
29 Adcock K, Hedberg C, Loggins J, Kruger TE, Baier RJ: The TNF-α -308, MCP-1 -2518 and TGF-β_1 +915 polymorphisms are not associated with the development of chronic lung disease in very low birth weight infants. Genes Immun 2003;4:420–426.
30 Kazzi SN, Quasney MW: Deletion allele of angiotensin-converting enzyme is associated with increased risk and severity of bronchopulmonary dysplasia. J Pediatr 2005;147: 818–822.

ALLEGATO II

ARTICLE: submitted to *Hormone Research*

A case study of young patients affected by Turner syndrome: psycho emotional and relational characteristics, psychopathological aspects and evaluation of treatments

A case study of young patients affected by Turner syndrome: psycho emotional and relational characteristics, psychopathological aspects and evaluation of treatments

*Gatta M., *Pertile R., **Ramaglioni E., **Bertossi E., **Nigri B., *Battistella P.A., *Condini A

*Department of Paediatrics - University of Padua (Italy),
**Neuropsychiatric Unit - Azienda ULSS 16 Padua (Italy)

Key words: Turner syndrome, psychopathology, multidisciplinary interventions

ABSTARCT

AIM In several previous studies women affected by Turner syndrome were shown to have peculiar phenotypic, psychopathological and psychosocial characteristics. The aim of this study was to evaluate, with a case-control study, psychological, affective-relational, psychosocial characteristics in a group of Turner syndrome affected pre adolescents and adolescents.
SAMPLE A group of 47 Turner syndrome affected girls were enrolled in this study. The age of the patients was between 10 and 19years. They attend the Pediatric Department of Padua University.
METHODS In order to evaluate psychological, affective-relational and psychosocial aspects in Turner women, a specific questionnaire, with 42 questions, was submitted to evaluate personal data, family, psycho-social, extra scholastic habits and therapeutic interventions (pediatric and psychological- neuropsychiatric). Both this questionnaire and the Youth Self Report by T. Achenbach were submitted to 47 girls affected by Turner syndrome and to a case control group of age-matched healthy females. The descriptive analysis included the observed frequencies calculation with the respective percentages for each variable collected in the questionnaire, separately for girls with Turner Syndrome and for the control group. Moreover the crude odds ratios (ORs) estimate with confidence intervals and p-values was carried out. Categorical data are given as numbers with percentages and continuous data as medians with ranges. Then a bivariate analysis was performed to study the possible relations between couples of variables, for both cases and controls. Since data are independent, the Chi-Square Test was used to compare proportions in m*n tables (without the Yates correction) and the Fisher exact Test was selected for 2*2 tables. A p-value<0.05 was considered significant. Once found a significant relation between two variables a standardized residuals analysis was carried out to locate the significant differences between the attributes of the considered variables.
Multivariate analysis was performed using a stepwise logistic regression analysis (significance level for entering = 0.15 and significance level for removing = 0.10) to identify those psychological, affective-relational, psychosocial factors that had a statistically significant correlation with Turner Syndrome.
RESULTS This study highlighted the following aspects regarding Turner girls:
1. 52 % of the girls were unsatisfied by their look.
91 % of the girls were unsatisfied because of their short height and obesity (44.7% was within 51-75 percentile). 83% of patients wished a different physical appearance.
2. The relationship with parents influences both the acceptance of the illness and several other psycho-social aspects.

3. As regards the relations with the opposite sex only 31.9% of girls had a relationship, whereas 66% wanted a family and children. Occupational and educational choices were influenced by this desire.
4. 61.7% of the girls underwent a psychiatric and psychological consultation and 58.6% considered it useful.
5. Turner patients following pharmacological therapy show a better integration in the scholastic group and have positive feelings towards their illness.
By comparing the group of patients with the control group, the most outstanding difference stands in the relationship with parents and in the psychological support. According to this study in the group of healthy girls both the relationship with parents and the psychological consultations did not influence self acceptance and other psychosocial aspects, while in Turner syndrome patients both factors did.
The most common way of interpreting a Logistic Regression Analysis is to convert each maximum likelihood estimate to an Odds Ratio (OR) using the *exp( )* function. The OR represents an estimate of the Relative Risk and the value 1 represents a full statistical independence between the dependent variable and the covariate.
All the variables collected from the questionnaire for both cases and controls were considered as possible predictors in the model explaining the Turner syndrome. Categorical variables were dichotomized, taking an attribute of each variable as reference category.
RESULTS OF THE LOGISTIC REGRESSION ANALYSIS: maximum likelihood estimates (b_i) with the corresponding standard errors (SE) and statistical significances (p-values), ORs with the corresponding confidence intervals (CI) and the p-value of the likelihood ratio test for the goodness of the model are reported. The seven variables entered in the model are height, weight, favourite scholastic subject, integration with the classmates, brothers and/or sisters, social withdrawal and aggressive behaviour. Analysing the maximum likelihood estimates for the two continuous variables, patients affected by Turner syndrome have a lower *height* (b_1 = -0,517, p = 0,0004) and a higher *weight* (b_2 = 0,304, p =0,0028) than controls. In other words, looking at the ORs, for every decrease of 1 centimeter in stature, the probability to observe a patient affected by Turner syndrome increases 1.7 times (the formula is: $OR(-1cm) = \exp[(-1cm) \times (-0{,}517)]$); for every increase of 1 kilogram in weight, the probability to observe a girl with Turner syndrome increases 1,356 times. The very small confidence interval resulted for these two variables indicates the great significance in explaining the correlation with Turner Syndrome.
As far as the variable '*favourite subject*' (b_3 = 2,677, p = 0,0425), if we know a patient answered that *Italian* is her favourite subject, we have a probability 14,5 times higher to observe a girl affected by Turner syndrome than a normal girl. The fourth variable entered in the model is '*integration with classmates*' (b_4 = 4,243, p-value = 0,035): girls affected by Turner syndrome have a higher probability to present a *low integration* compared to controls. In details, an answer '*low integration*' may belong to a girl affected by Turner syndrome with a probability 69,6 times higher than a control.
A very important variable in explaining the Turner syndrome is '*brothers and/or sisters*'(b_5 = 3,434, p-value = 0,0072): a patient affected by Turner syndrome has a probability 31 times higher than a control to be only daughter.
CONCLUSION An early diagnosis is fundamental for programming an effective medical therapy and a psychological support to both Turner affected girls and their parents, who have a great influence on the psychological and social aspects of these patients.

ALLEGATO III

ARTICLE: submitted to *Journal of Clinical and Experimental Neuropsychology*

Use of CPT II test, Conner's Continuous Performance Test II, in ADHD diagnosis during the period of development, in an Italian sample

Use of CPT II test, Conner's Continuous Performance Test II, in ADHD diagnosis during the period of development, in an Italian sample

M. Ronchese**, G. de Renoche**, L. Bianchin**, R. Pertile *A. Condini*

*Department of Paediatrics - University of Padua (Italy),
**Neuropsychiatric Unit - Azienda ULSS 16 Padua (Italy)

Key words: ADHD, CPT, diagnosis.

ABSTRACT

Conners' Continuous Performance Test (CPT II) Version 5.1 for Windows is an effective instrument for the ADHD diagnosis. In the study, the CPT II test has been given out to a sample of 142 subjects arrived at the Neuropsychiatric Unit for attention - behavioural problems.

Multivariate analysis was performed using a stepwise logistic regression analysis (significance level for entering = 0.15 and significance level for removing = 0.10) to identify those indexes from the CPT II test and those factors related to IQ that had a statistically significant correlation with Attention Deficit Hyperactivity Disorder (ADHD). For patients with ADHD an univariate linear regression analysis, between age and each index obtained from the CPT II test, was performed.

Results of the Logistic Regression Analysis: maximum likelihood estimates (b_i) with the corresponding standard errors (SE) and statistical significances (p-values), ORs with the corresponding confidence intervals (CI) and the p-value of the likelihood ratio test for the goodness of the model are reported. The three variables entered in the model are Confidence Index, Detectability and Reaction Time ISI Block Change. The p-value of likelihood ratio test < 0,0001 indicates a high goodness of the final model. The percentages of sensitivity and the specificity are respectively 83,7 and 51,0.