UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova
Dipartimento della Salute della Donna e del Bambino

Corsi di Dottorato di Ricerca in Medicina dello Sviluppo e Scienze della Programmazione Sanitaria
Curriculum: Scienze della Programmazione Sanitaria
CICLO XXIX

PROGETTAZIONE DI UNA CARTELLA GENERALIZZATA UTILE A COSTRUIRE REGISTRI DI PATOLOGIE RARE

Coordinatore: Ch.mo Prof. Carlo Giaquinto
Supervisore: Ch.ma Prof.ssa Paola Facchin
Co-Supervisore: Dott.ssa Monica Mazzucato

Dottorando: Elena Rizzardi

A Giulia e Sofia…
il colore del mondo e la luce di ogni giorno…

"Quando curi una persona puoi vincere o perdere… Quando ti prendi cura di una persona, vinci sempre…"

Hunter Doherty "Patch" Adams

INDICE

Abstract

Presupposti e scopi: Nel campo delle malattie rare, e delle malattie genetiche in particolare, l'introduzione della tecnologia di *Next-Generation Sequencing* (NGS) ha rappresentato una rivoluzione senza precedenti, le cui implicazioni per la ricerca e la medicina nel suo complesso sono state dirompenti. L'impulso all'acquisizione di nuove conoscenze è stato notevole, basti pensare che il numero di descrizioni fenotipiche riportate in OMIM, per le quali è noto il meccanismo molecolare sottostante, è più che raddoppiato dal 2007 al 2014, così come è cresciuto esponenzialmente il numero di geni associati a malattie rare.

Al crescere della disponibilità di dati ricavabili con le nuove tecniche di sequenziamento, è nel contempo cresciuta la necessità di disporre di descrizioni chiare dei tratti fenotipici degli stessi soggetti. L'importanza dell'osservazione del fenotipo, infatti, gioca un ruolo fondamentale non solo nel processo diagnostico riferito al singolo individuo, ma anche in ambito di ricerca. Per tutti questi motivi, è quindi enormemente cresciuto l'interesse per la descrizione del fenotipo umano e per il raggiungimento di una maggiore standardizzazione. Infatti, le descrizioni fenotipiche presenti in Letteratura, relative alle stesse entità di malattia, spesso presentano delle differenze, anche molto significative.

I tentativi di sistematizzazione mirati a produrre terminologie e/o ontologie del fenotipo sono relativamente recenti. Il loro sviluppo ha richiesto competenze non solo mediche, ma anche statistiche e bio-informatiche.

Nelle malattie rare esiste spesso una oggettiva maggiore difficoltà a porre diagnosi. Sono poche le malattie rare che si presentano con segni clinici patognomonici. Talvolta i segni ed i sintomi di presentazione sono relativamente comuni, ma quello che è caratteristico è la loro associazione. Ne deriva il rischio di mancata o ritardata diagnosi di malattia. Questo ritardo si traduce in un danno difficilmente quantificabile non solo a livello individuale, ma anche a livello di sistemi sanitari, in

termini di ricorso ad indagini inappropriate, somministrazione di trattamenti inefficaci o anche dannosi, e ricoveri ripetuti. Il problema dei pazienti non diagnosticati è divenuto sempre più emergente.

L'idea del nostro progetto è nata col proposito di creare un registro generalizzato, valido cioè per tutte le patologie rare, adattabile alle singole realtà locali, sostenibile in termini di costi e di risorse umane, utile anche per lo studio della popolazione di pazienti rari senza diagnosi. Il progetto si è articolato in più fasi: quella di creazione della Cartella Clinica generalizzata e dell'arruolamento pazienti; quella di analisi della popolazione dei pazienti rari; quella di messa a punto di sistemi matematici esperti che, imparando dai dati di Letteratura e dalla popolazione di pazienti rari inclusa nello studio, possano, in futuro, essere utilizzati per studiare la popolazione di pazienti senza diagnosi.

Materiali e metodi: La realizzazione dello strumento analitico si è avvalsa del modello di registro già in uso nella nostra regione, il Registro per le Malattie Rare del Veneto. La cartella clinica informatizzata è stata creata con il supporto del database ORACLE, situato all'interno del Centro di coordinamento della Regione Veneto e collocato su differenti network proprietari SPC o GARR. Il prodotto finale è un'interfaccia utente di tipo JAVA. Per la creazione della cartella sono stati utilizzati sistemi di classificazione basati su ICD o ORPHANET RD Classification. La cartella informatizzata ha una logica gerarchica, basata sulla possibilità di inserire informazioni principali (anamnesi, segni, sintomi, sequele, esami diagnositici e genetici) e secondarie (gli attributi: data, localizzazioni, severità, ecc.) in grado di relazionarsi le une con le altre in numero potenzialmente infinito.

I dati dei pazienti con diagnosi e senza diagnosi, sono stati inseriti nel Registro da parte di specialisti afferenti a Centri Accreditati partecipanti allo studio, all'interno di un programma ministeriale dedicato ai pazienti senza diagnosi: "*A multicenter collaborative research network for the identification and study of rare undiagnosed patients: the impact on the rare disease National Health Service network (UnRareNet)*". La numerosità del campione dei pazienti rari è stata arricchita utilizzando

pazienti già arruolati nel Registro delle Malattie Rare del Veneto. Analisi più approfondite sono state effettuate mediante package statistico SAS System 9.4 connesso ad Oracle. Per la popolazione di pazienti affetti da malattia rara e arruolati nel progetto è stata effettuata un'analisi descrittiva.

Tale modello statistico-matematico è stato utilizzato per la generazione della popolazione di pazienti simulati, previa creazione di profili di malattia, riportanti la prevalenza delle caratteristiche cliniche per ciascuna patologia studiata, tratti dalla Letteratura scientifica più aggiornata.

Per la creazione dei sistemi esperti si è utilizzata una procedura (NEURAL) disponibile nel SAS System. Il sistema è stato applicato dapprima sul gruppo di tre patologie: Sindrome di DiGeorge, Associazione CHARGE e Sindrome di Smith-Lemli-Opitz; successivamente su un gruppo più ampio costituito da otto patologie: Sindrome di DiGeorge, Associazione CHARGE, Sindrome Leopard, Sindrome di Noonan, Sindromi Noonan-like, Sindrome di Kabuki, Sindrome di Sotos, Sindrome di Williams. Il sistema ha subito una doppia validazione in entrambe le popolazioni.

Una validazione semplice è stata ottenuta suddividendo il dataset dei pazienti simulati in due gruppi: un primo insieme (pari al 75% dei casi) è stato utilizzato per la fase di *training* della rete; il restante campione (25%) per la validazione del sistema. Una seconda validazione è stata poi ottenuta istruendo il sistema (training) con tutta la popolazione di pazienti simulati (100%) e testandolo successivamente sui casi con diagnosi nota già inseriti nel database dai Centri Accreditati (pazienti reali).

Risultati: Il database è stato creato nel 2014 e i partecipanti al progetto hanno iniziato l'inserimento dati nel Settembre dello stesso anno, dopo corsi di formazione specifici e dopo adeguata validazione del programma. Il sistema della cartella clinica informatizzata è stato inoltre implementato, in base ai suggerimenti ricevuti. E' risultato flessibile e adattabile alle singole realtà coinvolte, accessibile 24 ore/365 giorni, capace di garantire la sicurezza del dato e la sua tracciabilità.

La popolazione di pazienti coinvolti nel progetto è attualmente pari a 1427 soggetti (518 afferenti al Progetto UnRareNet, i restanti appartenenti al Registro delle Malattie Rare del Veneto); l'89% di questi pazienti è affetto da patologia rara nota (le più rappresentate: Distrofia di Duchenne 22,7%, Distrofia di Becker 10,7%, sindrome di DiGeorge 6,8%), il restante da patologia non nota. L'età media dei pazienti per le due popolazioni (pazienti con e senza diagnosi) è pari a 27,1 e 38,9 anni rispettivamente. La maggior parte dei soggetti è risultato essere di sesso maschile (F:M=872:401).

Attraverso l'utilizzo di software statistici e dati clinici di prevalenza estratti dalla Letteratura, è stata contemporaneamente generata la popolazione di pazienti simulati rispettivamente per il gruppo di tre e otto patologie. La popolazione di pazienti simulati è stata utilizzata per istruire il sistema esperto che, in un secondo momento, è stato validato in due occasioni: su un gruppo ristretto di pazienti simulati (25% del totale) e sulla popolazione di pazienti con malattia rara diagnosticata arruolati nel progetto. I dati preliminari della validazione nel gruppo di tre patologie sono stati molto incoraggianti: il sistema ha infatti dimostrato di avere un alto grado di apprendimento (100%) e una bassa percentuale di errore nel riconoscimento della patolgia (1.13%). Solo in un caso (1/88), infatti, il sistema ha errato la diagnosi.

In un secondo momento, si è proceduto ad implementare il sistema con un numero più consistente di patologie (gruppo a otto patologie). Anche in questa occasione, è stata generata una popolazione di pazienti simulati (10.000 per ciascuna entità) con caratteristiche cliniche presenti in percentuale sovrapponibile alle prevalenze della Letteratura. Il sistema esperto è stato quindi istruito utilizzando questa popolazione (pazienti simulati) e validato inizialmente su una parte di essa (25%) e successivamente sulla totalità della popolazione di casi reali, come già avvenuto in precedenza. Anche in questa occasione il tasso di apprendimento è stato ottimale (98-99%), il tasso di riconoscimento è risultato essere soddisfacente per 5/8 patologie. L'errore tuttavia è risultato elevato (32.9%). L'analisi delle errate diagnosi ha permesso di capire come il 43.3% (23/53) dei pazienti erano descritti da informazioni

cliniche troppo scarse o generiche. Si è quindi proceduto ad eliminare tale sottogruppo. Nei restanti (30), il sistema ha attribuito una maggiore importanza a quei segni clinici presenti nella popolazione reale in percentuale diversa rispetto ai dati di Letteratura (discrepanza tra la nostra popolazione e i valori di prevalenza della Letteratura scientifica).
L'ultima analisi (quella effettuata senza il sottogruppo di "pazienti poco descritti"), ha mostrato ottima capacità di apprendimento del sistema (99.8%) e tasso di errore più contenuto (23%).

Conclusioni: La cartella clinica informatizzata è risultata essere uno strumento semplice, nonostenate la grande mole di *records* conentuti, flessibile e adattabile alle singole realtà coinvolte, capace di garantire la sicurezza del dato e la sua tracciabilità. L'implementazione della stessa, in base alle considerazioni degli esperti partecipanti al progetto, è risultata utile per semplificare il percorso descrittivo e sostenibile. La cartella clinica informatizzata rappresenta un esempio unico nel suo genere e si è rivelata essere un valido strumento conoscitivo.
L'analisi dei dati ha permesso una valutazione preliminare di pazienti con e senza diagnosi. Tale casistica consentirà, in un prossimo futuro, studi epidemiologici approfonditi.
Nonostante la complessità della materia trattata, le prime validazioni del sistema esperto hanno dato risultati promettenti individuando una diagnosi corretta nel 77% dei pazienti con malattia rara presenti nel sistema. L'analisi dell'errore riscontrato dal sistema, ha delineato l'importanza di un'accurata descrizione delle informazioni cliniche inserite e di criteri di selezione dei dati tratti dalla Letteratura più verosimili ai contesti di realtà.
La complessità del problema dei pazienti senza diagnosi può beneficiare della messa a punto di sistemi innovativi che indirizzino il clinico verso la diagnosi di malattia attraverso la conoscenza e l'uso di algoritmi complessi come quelli inerenti le reti neurali e la logica fuzzy, capaci di autoapprendimento e funzioni filtro, come nel nostro studio. La messa a punto di questi sistemi esperti, istruiti con informazioni estratte dalla Letteratura scientifica e testati su una popolazione di pazienti reali, richiede tuttavia una descrizione analitica, sistematica ed accurata, delle

informazioni immesse nel database al fine di predisporre il buon funzionamento di questi algoritmi. Da qui, come già sottolineato da più parti, l'importanza di sviluppare termini e definizioni condivisi ed integrabili tra loro, avvalendosi anche delle nuove ontologie sviluppate nell'ambito di malattie rare.

1. INTRODUZIONE

1.1 La rivoluzione genomica e le malattie rare

Mai come prima si è assistito ad una vera e propria crescente disponibilità di dati potenzialmente utili alla migliore comprensione dei meccanismi patogenetici alla base di molte malattie, e quindi, auspicabilmente legati allo sviluppo di nuovi approcci terapeutici.

Nel contempo, si sta anche assistendo ad una vera e propria proliferazione, anch'essa senza precedenti, di sistemi per la raccolta di tali dati. Espressione di questo fenomeno sono gli andamenti temporali dell'utilizzo in letteratura dei termini genotipo e fenotipo, i quali sono esponenzialmente cresciuti negli ultimi anni.

A fronte di questo scenario, in rapida evoluzione, occorre rilevare che la difficoltà per i ricercatori è oggi quella di integrare tutte queste informazioni, al fine di accelerare la produzione non solo di dati, ma di nuova conoscenza che da essi può derivare. Nel campo delle malattie rare, e delle malattie genetiche in particolare, l'introduzione della tecnologia di *Next-Generation Sequencing* (NGS) ha rappresentato una rivoluzione senza precedenti, le cui implicazioni per la ricerca e la medicina nel suo complesso sono state dirompenti [1].

Basti pensare che mentre il progetto di sequenziamento del DNA umano, completato nel 2003, ha richiesto una collaborazione internazionale della durata di 10 anni e ha avuto un costo di circa 2.7 miliardi di dollari [2], oggi è possibile sequenziare in un tempo molto limitato, passato rapidamente da settimane a giorni, ad un costo di circa 2.000 dollari, i 34 Mb di esoma, la parte codificante del DNA, corrispondente a circa l'1.2% del genoma umano.

Alcuni Autori hanno individuato l'accesso e il costo come i due principali aspetti da considerare per comprendere l'evoluzione della ricerca sulle

malattie rare [3]. L'introduzione delle nuove tecniche di sequenziamento ha avuto un impatto su entrambi tali aspetti, agendo sia sull'accesso di un maggior numero di pazienti, sia portando con sé una diminuzione dei costi legati all'utilizzo di queste metodiche. Gli stessi Autori hanno introdotto un parallelismo interessante per descrivere le conseguenze dell'impatto delle nuove tecniche di sequenziamento nell'ambito della ricerca sulle malattie rare. Il modello proposto è quello noto in economia come regola di Pareto. Il modello descrive come spesso il 20% dei prodotti di un'industria sia collegato all'80% dei suoi guadagni. Tradotto, la ricerca tradizionale si è focalizzata sul 20% delle cause genetiche collegate all'80% delle malattie note, mentre ora l'attenzione si sta spostando sul restante e non trascurabile 20%, investendo nel quale, i risultati in termini di avanzamenti della conoscenza saranno considerevoli.

I primi esempi di utilizzo di tecniche di sequenziamento di nuova generazione per lo studio di malattie a trasmissione mendeliana hanno dimostrato che era possibile identificare geni causativi di malattia sequenziando l'esoma di un numero ristretto di individui. In alcuni casi anche l'analisi dell'esoma di un singolo individuo si è rivelata sufficiente per definire l'alterazione genetica responsabile di un determinato quadro clinico. L'impulso all'acquisizione di nuove conoscenze è stato notevole, basti pensare che il numero di descrizioni fenotipiche riportate in OMIM, per le quali è noto il meccanismo molecolare sottostante, è più che raddoppiato dal 2007 al 2014, così come è cresciuto esponenzialmente il numero di geni associati a malattie rare.

Accanto alle prospettive straordinarie offerte dallo sviluppo e dall'affinamento delle nuove tecnologie indaganti il genoma umano, sono state evidenziate da più parti anche potenziali criticità di utilizzo, soprattutto quando da alcuni è stata proposta la loro potenziale applicazione anche a soggetti senza una storia personale o familiare di malattia genetica. Sono emersi aspetti di complessità, prima solo ipotizzati.

Anche nell'utilizzo di metodi di indagine più tradizionali, come l'amplificazione PCR ed il Sequenziamento Sanger, uno degli aspetti più critici è l'interpretazione delle varianti di significato incerto "*variants of*

unknown significance" (VUS). Il tentativo di classificare tali varianti in neutrali o correlate ad una determinata patologia rappresenta uno sforzo notevole, ancorché necessario, richiedente la confrontabilità e la condivisione di dati su scala internazionale.

Un altro aspetto di complessità, rivelato proprio con l'avvento della tecnologia NGS, riguarda il grado di variabilità del DNA umano. Un'intuizione in tal senso era stata espressa in maniera lungimirante già da Victor McCusick nel 1981 quando affermava che: "anche laddove si conoscesse l'anatomia del genoma umano fino all'ultimo nucleotide, non sarà possibile conoscere la funzione specifica di ogni singola parte del DNA" [4].

Attualmente si sa che ciascun esoma, corrispondente all'1,2 % dell'intero genoma, può contenere fino a 20.000 variazioni di singoli nucleotidi, 500 delle quali con una frequenza allelica molto bassa, private o specifiche a livello di singolo individuo. Un aspetto di tale variabilità è costituito dai cosiddetti "incidental findings", divenuti un tema dai risvolti non solo scientifici, ma anche di ordine etico. Questi possono essere definiti in genetica come variazioni del genoma identificate da tecniche di sequenziamento, ma non correlate al quesito diagnostico per il quale il soggetto è stato sottoposto all'indagine [5]. Il dibattitto sulla loro gestione, soprattutto rispetto alla possibilità di comunicarli o meno, è ancora molto aperto e con soluzioni proposte diverse a seconda dei contesti, e a volte anche dei Paesi.

Il terzo aspetto di complessità riguarda il determinismo genetico di malattie complesse. L'ipotesi inizialmente percorsa era che la maggior parte delle malattie più frequenti fosse determinata a livello genetico dalla combinazione di varianti genetiche, ciascuna associata ad un rischio di modesta entità.

L'ipotesi alternativa che si va affermando con l'avvento della tecnologia NGS è che l'ereditarietà di molte malattie complesse sia dovuta a varianti genetiche rare, tuttavia conferenti un rischio rilevante.

La conclusione che deriva da tutti e tre questi macro-aspetti è che, per dipanare questa matassa di estrema e crescente complessità, occorra investire nello sviluppo di strumenti nuovi e/o nell'adattamento di quelli

esistenti a nuove esigenze. Questa considerazione ha investito, di riflesso, parallelamente al genoma, l'ambito del fenotipo.

1.2 Il ritrovato interesse per il fenotipo

Al crescere della disponibilità di dati ricavabili con le nuove tecniche di sequenziamento, riferiti ai genomi di un numero crescente di individui, è nel contempo cresciuta la necessità di disporre di descrizioni chiare dei tratti fenotipici degli stessi soggetti. A questo proposito è necessario constatare come i termini che vengono utilizzati per descrivere il fenotipo hanno avuto uno sviluppo che si potrebbe definire non articolato.

La definizione stessa di fenotipo si è modificata nel tempo, potendosi rilevare anche delle incongruenze rispetto ai significati attribuiti a questo termine, a seconda degli Autori e del contesto degli studi. Una revisione di Nachtomy et Al. [6] ha analizzato e confrontato 5 diverse definizioni di fenotipo utilizzate in Letteratura. I tre usi più frequenti del termine potevano essere riferiti ai seguenti tre ambiti:

- l'insieme delle caratteristiche di un organismo o di uno dei suoi sottosistemi;
- un organismo caratterizzato da uno specifico, e generalmente parziale, fenotipo;
- una classe di organismi aventi in comune lo stesso "generalmente parziale fenotipo".

La definizione di fenotipo elaborata dagli Autori di questo lavoro, risalente al 1997, può essere tradotta con "l'insieme delle caratteristiche osservabili di un organismo, sia morfologiche che relative alla fisiologia, a livello di cellula, organo, corpo, e di comportamento, incluse caratteristiche quali i profili di espressione genica in risposta a stimoli ambientali".

In ambito clinico, il termine fenotipo è utilizzato principalmente per definire uno scostamento dalla normalità, sia che essa si riferisca alla morfologia, alla fisiologia o al comportamento di un individuo.

La descrizione appropriata del fenotipo è un processo fondamentale in medicina. Da tale descrizione, basata su dati rilevabili dall'esame obiettivo, derivati dalla diagnostica per immagini, da test di laboratorio o

da test specifici per la valutazione di alcune funzioni, si formulano poi le ipotesi diagnostiche di partenza.

L'importanza dell'osservazione del fenotipo gioca un ruolo fondamentale non solo nel processo diagnostico riferito al singolo individuo, ma anche in ambito di ricerca. Ad esempio, certe malattie presentano un fenotipo simile, come la sindrome di Marfan (codice ORPHA558) e l'aracnodattilia contratturale congenita (codice ORPHA115). I geni mutati in tali sindromi sono infatti rispettivamente, FBN2 e FBN1, i quali appartengono alla stessa famiglia e presentano similitudini di funzione. D'altra parte, l'osservazione che alcune malattie genetiche presentano quadri fenotipici sovrapponibili ha portato a sviluppare il concetto non solo di famiglie di geni, ma anche di "famiglie di malattie", dove la similitudine può essere dovuta all'alterazione, ad esempio, di una via di regolazione. Pertanto, una corretta ed approfondita descrizione fenotipica può essere determinante in molti casi per la comprensione di meccanismi patogenetici o del funzionamento di sistemi a livello cellulare, identificando gruppi di geni coinvolti in essi, il cui funzionamento alterato determina conseguenze simili sul fenotipo.

Lo studio del fenotipo può anche riferirsi alla descrizione dettagliata ed aggiornata dello spettro delle alterazioni fenotipiche associate ad entità di malattia, sia di nuova definizione che già note. Un processo fondamentale del ragionamento clinico consiste, infatti, nel rilevare un'alterazione fenotipica e stabilire se questa sia espressione di una patologia o rappresenti un'alterazione isolata in un individuo. La descrizione dettagliata delle alterazioni fenotipiche rilevate in un soggetto è di fondamentale importanza non solo per formulare ipotesi diagnostiche, ma assume anche in molti casi un significato prognostico, laddove la presenza di un'alterazione sia indicativa di una determinata rispondenza ad un trattamento o si associ ad una maggiore probabilità di sviluppare una complicanza. Gli sforzi dedicati a produrre descrizioni migliori e più dettagliate del fenotipo possono essere anche ricondotti allo sviluppo relativamente recente del concetto di "medicina di precisione". Lo scopo della medicina di precisione è, infatti, quello di favorire l'accesso a trattamenti sempre più efficaci per il singolo individuo e/o mirati a ridurre al

massimo gli effetti secondari del trattamento, sulla base delle diverse caratteristiche che distinguono un paziente da altri con la stessa presentazione clinica [7].

Per tutti questi motivi, è quindi enormemente cresciuto l'interesse per la descrizione del fenotipo umano e per il raggiungimento di una maggiore standardizzazione. Infatti, le descrizioni fenotipiche presenti in Letteratura, relative alle stesse entità di malattia, spesso presentano delle differenze, anche significative. Tali differenze sono attribuibili ai diversi scopi per i quali tali descrizioni sono state prodotte, ai diversi profili degli utilizzatori e ai contesti in cui le stesse possono essere utilizzate.

Ad esempio, diverso è il modo in cui l'informazione relativa ad un paziente viene raccolta nelle cartelle cliniche rispetto al modo in cui questa è presentata in articoli scientifici pubblicati per descrivere una determinata malattia, siano essi revisioni di ampie casistiche o *case-report*. Ad essere diverso non è solo il livello di dettaglio, ma il contesto stesso che determina la scelta di quali termini utilizzare per descrivere la stessa alterazione. Si è quindi affermata una progressiva necessità di standardizzazione.

I tentativi di sistematizzazione mirati a produrre terminologie e/o ontologie del fenotipo sono relativamente recenti. Il loro sviluppo ha richiesto competenze non solo mediche, ma anche statistiche e bio-informatiche. Verranno di seguito descritti brevemente alcuni strumenti che hanno rappresentato, in vario modo, il tentativo di arrivare ad elaborare un linguaggio comune per la descrizione del fenotipo umano.

1.3 Strumenti per la descrizione del fenotipo

1.3.1 London Dysmorphology Database

Il *London Dysmorphology Database* (LDDB) [8] ha rappresentato per molti anni un punto di riferimento fondamentale ed una risorsa per la diagnosi delle malattie genetiche. Il database è basato su un vocabolario di circa 1.000 termini descriventi alterazioni suddivise per parte anatomica interessata, organi e apparati interessati. I termini sono rappresentati in maniera gerarchica nell'interfaccia del database. Tuttavia, ad essi non è assegnato un identificativo visibile all'utilizzatore. L'utilità di questo

strumento risiede nell'associazione tra alterazioni e malattie in cui sono state descritte, con la possibilità di utilizzare uno strumento di ricerca per segni e sintomi con funzioni avanzate, quali la possibilità di determinare il carattere obbligatorio o facoltativo dell'alterazione, e di escludere un segno con la funzione "NOT" nella fase di ricerca. Per ciascuna alterazione descritta è disponibile anche una breve definizione. Parte integrante del database sono anche immagini, a corredo delle descrizioni di malattie presenti nel database. L'ultima versione disponibile del database risale al 2002.

1.3.2 OMIM (Online Mendelian Inheritance in Man)

Da 50 anni, OMIM (*Online Mendelian Inheritance in Man*) è uno strumento di classificazione delle malattie genetiche basato sullo studio delle relazioni tra geni e loro varianti molecolari e fenotipi associati [9]. L'assegnazione di un nome e la classificazione di un quadro fenotipico è il processo fondamentale alla base della costruzione di una nosologia. In tal senso, per le malattie genetiche OMIM ha svolto un ruolo fondamentale, sin dalla sua prima edizione cartacea prodotta nel 1966. La classificazione delle malattie è un processo basato anzitutto sulla definizione dello spettro di caratteristiche rilevabili, identificando quelle che distinguono una condizione da un'altra. Le entità attorno alle quali il database OMIM è costruito sono quadri fenotipici che possono descrivere malattie monogeniche, oppure quadri più complessi nei quali l'alterazione di un singolo gene determina in modo significativo il fenotipo. Le mutazioni note per essere causative di un determinato fenotipo sono classificate nella sezione delle varianti alleliche del gene causativo.

Per ciascuna entità è stata prodotta nel tempo una sinossi dei segni e dei sintomi descritti in Letteratura. Alcuni dei limiti principali di tale strumento sono: il fatto che non esista un thesaurus consultabile di tutti i segni e sintomi registrati, il fatto che essi non siano compresi in una struttura gerarchica di relazioni ed il fatto che possono essere presenti termini diversi usati per descrivere la stessa alterazione.

1.3.3 Elements of Morphology

Abbiamo già accennato al fatto che, sia in ambito clinico che di ricerca, si sia diffuso l'utilizzo di termini specifici per descrivere alterazioni di parti del corpo o di organi e apparati, senza che si fosse stabilita una loro sistematizzazione. Per superare questo limite, un gruppo di dismorfologi diede vita ad uno sforzo collaborativo internazionale per cercare di standardizzare la terminologia utilizzata in questo ambito per descrivere alterazioni fenotipiche. Le attività si focalizzarono sulla descrizione di alterazioni a livello di cranio, faccia, piedi e mani e sono poi proseguite fino ad arrivare ad una più ampia terminologia per le malformazioni congenite nel loro complesso [10]. Per ciascun termine è stata elaborata una definizione in lingua inglese, e una descrizione utile a rilevare correttamente e in maniera il più possibile standardizzata un'alterazione, laddove appropriato. Oltre ai termini e alla loro definizione, il gruppo di lavoro ha anche condiviso delle immagini per rappresentare una parte considerevole delle alterazioni descritte. Questo sforzo è esitato in una serie di pubblicazioni apparse sull'*American Journal of Medical Genetics* [11-12-13]. Lo scopo di questo lavoro non è mai stato quello di fornire uno strumento di diagnosi differenziale, associando le alterazioni alle malattie in cui queste possono riscontrarsi, quanto piuttosto quello di produrre un inventario e definire una terminologia comune da utilizzare per descrivere le alterazioni fenotipiche riscontrabili nell'ambito delle malformazioni congenite.

1.3.4 Human Phenotype Ontology

Human Phenotype Ontology (HPO) è un'ontologia di alterazioni fenotipiche sviluppata da un gruppo di ricercatori facenti capo all'Istituto di Genetica Medica di Berlino. Lo scopo di questo sforzo collaborativo divenuto internazionale era all'inizio quello di registrare tutte le alterazioni fenotipiche attribuibili a diverse malattie, a partire da quelle rare [14]. Il punto di partenza è stato proprio il thesaurus di segni e sintomi presenti nelle sinossi delle entità che popolano il database OMIM. Un lavoro preliminare è consistito nel ricondurre le diverse descrizioni della stessa alterazione ad uno stesso termine nella classificazione HPO. Il secondo

passo è consistito nel costruire le relazioni tra termini, creando una struttura gerarchica. Il terzo passo è stato quello di associare le alterazioni alle entità di malattia. A ciascun termine HPO è assegnato un codice identificativo univoco. Nell'ultima versione scaricabile aggiornata al gennaio 2015 [15], HPO conteneva più di 10.500 classi (concetti) e circa 16.000 termini per descrivere alterazioni fenotipiche, compresi circa 6.200 sinonimi, consultabili oltre al termine indicato come principale. Uno sviluppo più recente di questa risorsa *open-source* è l'ampliamento dell'interesse, dalle malattie rare a malattie più comuni. Attualmente sono codificati più di 132.000 annotazioni fenotipiche relative a 3.145 malattie non rare [16].

1.3.5 SNOMED-CT

Se HPO può essere considerato al momento una risorsa utilizzata prevalentemente in ambito di ricerca, SNOMED-CT rappresenta un esempio di terminologia clinica utilizzata in ambito assistenziale [17]. La terminologia fornisce termini, sinonimi e relazioni riferite ai seguenti ambiti: malattie, segni e sintomi e procedure. Dal 2002 sono stati prodotti, con cadenza semestrale, 22 aggiornamenti. Nel 2009 è stato costituito un Consorzio internazionale (*International Health Terminology Standards Development Organisation* - IHTSDO) per coordinare le attività ed aggiornare la terminologia. Al momento, 19 Paesi hanno identificato SNOMED-CT come terminologia da utilizzare preferibilmente nella compilazione di cartelle cliniche informatizzate. La versione disponibile al 2015 conteneva circa 300.000 concetti.

1.3.6 UMLS

Lo *Unified Medical Language System* (UMLS) è un sistema creato ed aggiornato dalla *National Library of Medicine* statunitense. A differenza delle altre terminologie, esso rappresenta una meta-terminologia, nel senso che mira a integrare terminologie già esistenti, in modo da rappresentare con lo stesso termine i medesimi concetti, identificati con descrittori diversi in diverse fonti. Proprio per questa sua caratteristica si

presta ad essere un possibile strumento da utilizzare per valutare e testare l'interoperabilità tra le diverse terminologie ed ontologie sviluppate.

1.3.7 Orphanet e l'ontologia Onto-Orpha

Orphanet (www.orpha.net) è il portale di riferimento per le informazioni sulle malattie rare. Orphanet nasce dall'intuizione di Segolene Aymè, medico genetista, e nel 1997 diventa una collaborazione tra INSERM, l'Istituto francese di salute e ricerca in medicina, ed il Ministero della Salute francese. Dal 2000 è finanziato a livello europeo con progetti e Joint-actions sulle malattie rare, oltre che da istituzioni pubbliche e private di vari Paesi, dove esistono collaborazioni nazionali. Il database è multilingue e fornisce informazioni su circa 7.000 malattie rare. Per ciascuna malattia, le informazioni si riferiscono, ai Centri di riferimento individuati nei singoli Paesi, ai laboratori dove effettuare i test genetici, ai *trials* clinici in corso, ai registri istituiti, alle bio-banche, alle associazioni di pazienti. Per ciascuna entità è anche presente un thesaurus di segni e sintomi, derivati dalla Letteratura e dal contributo di esperti. Tutta l'informazione è organizzata attorno ad un database relazionale, dove le entità del database sono le malattie.

Dal 1997 al 2007, la funzione principale del database è stata quella di essere un inventario di malattie rare, secondo la definizione europea, e di informazioni ad esse riferite. Successivamente, è emersa l'esigenza di adattare i contenuti a profili diversi di utilizzatori. Tale maggiore complessità, relativa all'organizzazione dell'informazione secondo livelli diversi di granularità, ha determinato l'evoluzione verso una rappresentazione gerarchica delle entità di malattia e di loro macro-gruppi di appartenenza. Il primo passo è stata la produzione di classificazioni di malattie rare per branca, aventi la caratteristica di essere multi-assiali. Le malattie rare sono infatti spesso malattie multi-sistemiche. La stessa entità di malattia, definita da un codice alfanumerico univoco chiamato orpha-code, può quindi ritrovarsi in più alberi classificatori, a seconda delle sue manifestazioni fenotipiche e/o dei meccanismi patogenetici. Questo sforzo classificatorio è stato alla base della creazione da parte dell'Organizzazione Mondiale della Sanità di un *Topic Advisory Group*

(TAG) specifico per le malattie rare, coordinato da Orphanet, al fine di migliorare la rappresentazione complessiva di queste malattie nella versione undicesima della classificazione ICD, tutt'oggi in divenire.

Il *feedback* ricevuto dagli utilizzatori dopo il lancio della versione di Orphanet basata su tale nuovo approccio classificatorio multi-gerarchico e multi-assiale, ha determinato, a partire dal 2008, un'ulteriore evoluzione, derivante dalla necessità di rendere la rappresentazione dell'informazione contenuta in Orphanet più interoperabile con altri sistemi di classificazione e terminologie utilizzati per la descrizione delle entità di malattie o dei loro segni e sintomi. È stata quindi realizzata una vera e propria ontologia, denominata Onto-Orpha [18]. Il lavoro di mappatura tra contenuti di Orphanet e altre terminologie e classificazioni è accessibile gratuitamente in una sezione dedicata del portale (www.orphadata.org) ed è tuttora in corso. Esso riguarda le seguenti terminologie e classificazioni: ICD-10, UMLS, SNOMED CT, MeSH, MedDRA e più recentemente anche HPO. Anche Orphanet quindi sta investendo molte risorse nell'interoperabilità, anche dei contenuti relativi alla descrizione fenotipica delle malattie rare.

1.4 Database per la raccolta di dati sul fenotipo

Da un'analisi della letteratura si evince come esista una certa confusione anche nella scelta dei termini utilizzati per descrivere gli strumenti sviluppati per rispondere all'esigenza di raccogliere dati in maniera tale che siano a posteriori integrabili e confrontabili tra loro. Per esempio, in più articoli gli strumenti sopra descritti sono indicati come *database*, quando sarebbe più corretto parlare di "terminologie mediche" o, in alcuni casi, di vere e proprie "ontologie" [19].

Parallelamente allo sviluppo di tali strumenti, si è registrato un proliferare di database strutturati e realizzati per raccogliere dati sul fenotipo di pazienti, sia con diagnosi di malattie genetiche note, sia senza diagnosi. Tali database hanno lo scopo principale di correlare i dati del fenotipo con quelli del genotipo, assumendo una valenza di ricerca, ma anche con ricadute assistenziali, laddove lo strumento consenta di facilitare l'identificazione di mutazioni causative, effettuando una prioritizzazione delle varianti rilevate con NGS [20].

Gli studi di associazione di tutto il genoma (*Genome-Wide Association Studies*, GWAS) hanno rivelato migliaia di associazioni tra varianti genetiche ed un ampio spettro di fenotipi, svelando i meccanismi eziopatogenetici alla base di molte malattie. Un altro aspetto promettente di questo tipo di approccio, che migliora la comprensione dei meccanismi patogenetici, è la possibile riproposizione a fini terapeutici di molecole già note, agenti su tali meccanismi.

I database disegnati per indagare la correlazione genotipo-fenotipo possono essere centrati sulla raccolta di dati su una malattia o su un gruppo di malattie correlate, così come avere un orientamento non malattia-specifico.

Esempi di quest'ultimo tipo di database sono PhenoDB e PhenoTIPS [21-22]. Il limite di tali esperienze può essere individuato nella loro sostenibilità. Non solo economica, aspetto comunque rilevante, considerato che la maggior parte di tali database sono risorse non a pagamento. La sostenibilità riguarda anche la necessità di disporre di risorse professionali esperte, sia sul versante bio-informatico che sul versante medico, per la definizione e la manutenzione dei contenuti, in continua espansione. Un altro aspetto critico è la fruibilità di tali strumenti e la loro capacità di evoluzione, da un ambito strettamente di ricerca ad un ambito clinico-assistenziale. Esiste infatti una necessità di integrazione con altri sistemi di raccolta di dati, utilizzati correntemente in ambito medico. L'interoperabilità dei sistemi riguarda, infatti, non solo gli aspetti tecnici, ma anche quelli semantici, laddove per "interoperabilità semantica" si intende il processo che garantisce che il significato di un'informazione sia comprensibile da ogni sistema o applicazione che non sia stato inizialmente sviluppato per raccoglierla" [23].

Diversi sono gli sviluppi recenti che hanno preso impulso da queste considerazioni. Da una parte, sono stati condotti molti studi mirati a valutare le caratteristiche delle terminologie mediche e delle classificazioni di malattie utilizzate in diversi contesti, al fine di procedere con una loro linearizzazione. Orphanet, come già detto, ha investito molto in questa complessa attività.

Un altro sviluppo molto interessante e promettente è l'iniziativa Monarch (https://monarchinitiative.org). Tale collaborazione internazionale mira a superare l'eterogeneità che caratterizza sia i sistemi di raccolta di dati sul genotipo che sul fenotipo [24]. La novità è che questo tentativo di integrazione, riguarda non solo i sistemi di raccolta dati riferiti alla specie umana, ma anche alle altre specie, partendo dalla considerazione che difficilmente un unico modello animale da solo può riprodurre tutte le caratteristiche fenotipiche determinate nell'uomo dalla malattia oggetto di studio.

1.5 Le malattie rare e i malati senza diagnosi

Le malattie rare rappresentano un gruppo eterogeneo di malattie che potenzialmente possono interessare qualsiasi organo o apparato. Il carattere spesso multi-sistemico si accompagna alla loro gravità dal punto di vista clinico. Si tratta di malattie trasversali che possono interessare tutte le classi di età, con eziologie differenti anche se esiste una prevalente componente genetica (in circa l'80%).

Le malattie rare hanno un decorso quasi sempre severo, e sono responsabili di disabilità sia fisiche che psichiche, determinanti una riduzione dell'aspettativa di vita degli individui colpiti. In Europa una malattia è definita rara secondo un criterio di prevalenza, quando colpisce meno di 5 abitanti su 10.000. Nonostante venga spesso citato che il numero di persone con malattia rara rappresenti dal 6 all'8% della popolazione europea, in realtà è appurato che tale dato costituisca una stima non basata su uno studio di popolazione. Uno studio condotto nella Regione Veneto e riferito al periodo 2002-2014 di attività di un registro *web-based* di popolazione sulle malattie rare ha stabilito che il numero di malati rari possa essere quantificato tra l'1,3% e il 2% della popolazione europea, a seconda che si considerino o meno i tumori rari [25]. Nonostante questi dati ridimensionino le stime generalmente riportate, lo stesso studio ha rilevato che l'impatto di tali malattie è rilevante. Le malattie rare determinano infatti complessivamente una percentuale di anni di vita perduti circa doppia rispetto a quella attribuibile al diabete. Un recente studio australiano ha confermato gli stessi dati relativi alla

diffusione in popolazione delle malattie rare e ha quantificato anche il loro impatto in termini di ospedalizzazione di questi pazienti [26]. L'impatto è rilevante, non solo in termini economici, ma anche per gli elevati costi umani e sociali collegati. Questi dati supportano la considerazione che le malattie rare rappresentino un tema rilevante di salute pubblica, per la combinazione di numero comunque consistente di persone interessate e danno subìto.

Nelle malattie rare esiste spesso una oggettiva maggiore difficoltà a porre diagnosi. Sono poche le malattie rare che si presentano con segni clinici patognomonici. Talvolta i segni ed i sintomi di presentazione sono relativamente comuni, ma quello che è caratteristico è la loro associazione. In altri casi quello che è peculiare non è l'associazione in sé dei segni o dei sintomi, quanto il loro manifestarsi nel tempo. La difficoltà nel giungere ad una diagnosi amplifica il peso, già considerevole, che queste condizioni pongono sia sull'individuo colpito, che sulla sua famiglia. Una mancata diagnosi aggiunge una componente di danno, per esempio determinando il mancato accesso a terapie in grado di modificare la storia naturale. A questo può associarsi anche il danno dovuto alla mancata diagnosi in altri componenti del nucleo familiare.

Il ritardo diagnostico si traduce in un danno difficilmente quantificabile non solo a livello individuale, ma anche a livello di sistemi sanitari, in termini di ricorso ad indagini inappropriate, somministrazione di trattamenti inefficaci o anche dannosi, e ricoveri ripetuti. Le difficoltà diagnostiche possono avere due conseguenze: una diagnosi errata o l'assenza di diagnosi. Graber et al. [27] hanno indagato in una serie di lavori le cause di errori diagnostici ed hanno osservato che spesso la causa degli errori diagnostici non risiede nella mancanza di conoscenza, quanto piuttosto nell'incapacità di sintesi dell'informazione disponibile.

Quando si parla di malattie rare esiste, comunque, anche un problema di mancanza di informazioni e conoscenze che può determinare errori e ritardi diagnostici. Secondo uno studio condotto da Eurordis su circa 6.000 pazienti, nel 25% dei casi erano intercorsi da 5 fino ad un massimo di 30 anni tra l'insorgenza dei primi sintomi e il raggiungimento della diagnosi definitiva ed il 40% dei pazienti aveva ricevuto una diagnosi errata prima

che la condizione venisse correttamente diagnosticata. I tempi di diagnosi possono variare molto da malattia a malattia o anche all'interno di una stessa condizione. Alcuni Studi [28] hanno evidenziato come, per alcune patologie rare selezionate, il 25% dei pazienti abbia atteso da 5 a 29 anni prima di avere conferma della diagnosi e il 40% abbia inizialmente ricevuto una diagnosi errata. Un 25% dei soggetti considerati nello studio, riferiva inoltre di aver dovuto cambiare regione per ottenere la diagnosi e, in un 2%, i pazienti si erano recati all'estero per lo stesso scopo. Dallo studio emerge come, a causa della mancata/errata diagnosi, un paziente su sei è stato sottoposto a un intervento chirurgico non necessario e un paziente su dieci è stato sottoposto a terapia psichiatrica. Allo stesso tempo, gli Autori vanno a considerare il carico per le famiglie dei pazienti e il mancato sostegno durante l'iter diagnostico; segnalano inoltre l'assenza di *counselling* genetico con nascita di fratelli anch'essi malati. I pazienti riportano un alto grado di sfiducia nei vari sistemi sanitari. Tutto questo genera ansia, incomprensione, depressione e isolamento nel paziente e nella sua famiglia. Studi autoptici, inoltre, hanno dimostrato come fino a un 5% di persone decedute, sarebbero sopravvissute se fosse stata formulata una diagnosi corretta [29]. Alcuni Autori hanno analizzato i tempi delle ritardate diagnosi e gli errori che hanno condotto alla mancata formulazione delle stesse [30]. Ad esempio, per alcune distrofie muscolari [31], la diagnosi è stata posta mediamente dopo 1,5 anni dalla prima visita medica del paziente e 2,5 anni dopo la comparsa dei primi sintomi. In un altro Lavoro [32], viene riportato come, a causa di manifestazioni sfumate o della comparsa dei sintomi in età adulta, la diagnosi nei pazienti affetti da sindrome di Williams venga posta mediamente con 2,7 anni di ritardo dalla comparsa della sintomatologia. Le conseguenze sullo stato di salute del paziente possono essere molto gravi: ad esempio, in una popolazione di pazienti affetti da sclerosi tuberosa, condizione caratterizzata da quadri fenotipici estremamente variabili, il ritardo nella formulazione di una diagnosi di malattia può arrivare anche a 15 anni con conseguente maggior rischio di comorbidità quali insufficienza respiratoria, insufficienza renale, cardiopatie, etc. [33].

La mancata diagnosi può essere legata a varie condizioni: in alcuni casi i pazienti sono affetti da malattie rare note che non vengono riconosciute, in altri, si tratta di quadri di presentazione rari di patologie più comuni o di fenotipi comuni a più patologie. Raramente, infine, il paziente è affetto da una malattia che non è ancora stata descritta in Letteratura [34].

I malati rari senza diagnosi rappresentano quindi un gruppo eterogeneo di pazienti. Recentemente, sono stati oggetto di una crescente attenzione, che ha combinato l'interesse scientifico alla necessità di una risposta a questo fenomeno da parte dei sistemi sanitari. La raccolta di dati su questi pazienti rappresenta anche una preziosa opportunità per i ricercatori, laddove l'indagine sui meccanismi alla base di entità cliniche non comuni si è dimostrata potenzialmente utile per svelare meccanismi patogenetici alla base anche di malattie più comuni. La rivoluzione genomica ha positivamente investito questo gruppo di pazienti. Sono stati condotti alcuni studi che possono indicare quale sia l'impatto potenziale dell'utilizzo delle nuove tecnologie per lo studio di pazienti con malattie rare senza diagnosi.

L'esperienza descritta da Yang et al. [35] riguarda 250 probandi con sospetto di malattia genetica riferiti ad un laboratorio certificato per effettuare il sequenziamento dell'esoma (WES). Circa l'80% dei probandi presentava un fenotipo con segni neurologici. La definizione del difetto genetico causativo grazie all'utilizzo di WES è stata possibile per circa il 25% di questa popolazione selezionata di pazienti.

Un altro esempio è lo studio inglese DDD (*Deciphering Development Disorder project*), che ha coinvolto, a partire dal 2010, circa 180 clinici e 24 servizi regionali di genetica in Gran Bretagna ed Irlanda, allo scopo di effettuare il sequenziamento dell'esoma in bambini con difetti dello sviluppo non diagnosticati e nei loro familiari [36]. Lo studio ha arruolato un numero complessivamente molto consistente di soggetti, circa 8000 famiglie in 3 anni. I criteri di inclusione erano la presenza di: difetti dello sviluppo di grado severo e/o malformazioni congenite, alterazioni dei parametri di crescita, dismorfismi e aspetti comportamenti atipici. L'informazione clinica è stata raccolta utilizzando i termini della *Human Phenotype Ontology* (HPO), inseriti in un database (Decipher). Questo

database è uno strumento accessibile on-line sviluppato per raccogliere dati su varianti genetiche, allo scopo di facilitare l'identificazione e l'interpretazione di tali varianti in soggetti con malattie rare. Più di 200 centri clinici di genetica medica contribuiscono al suo aggiornamento. Attraverso la sua consultazione possono essere identificati altri pazienti con caratteristiche simili di rapporto genotipo-fenotipo. Per la descrizione fenotipica dei soggetti sono stati utilizzati nel complesso 1.435 termini dei 10.000 disponibili nella *Human Phenotype Ontology*. Considerando la coorte composta da 1.133 trii (probando, madre e padre), gli Autori dello studio hanno riferito un miglioramento della diagnosi nel 27% dei casi con l'applicazione del WES.

Accanto a questi studi, sono stati creati dei veri e propri programmi dedicati a questi pazienti [37-38]. Negli Stati Uniti, in particolare, a partire dal 2008, è stato messo a punto un programma specifico per i pazienti senza diagnosi (*Undignosed Diseases Program*, UDP). Lo scopo di questo progetto, realizzato in collaborazione con il *National Institute of Health Clinical Center*, era quello di comparare i dati clinici di pazienti senza diagnosi con l'esito di tecniche genetiche di nuova generazione (*Whole exome sequencing*, WES; *Whole genome sequencing*, WGS) al fine di giungere ad una diagnosi di malattia [39]. Nei primi due anni di operato, il programma ha analizzato 1191 cartelle cliniche, escludendone il 59% e includendo nello studio 326 pazienti. Di questi 160 è stato sottoposto a valutazioni e indagini eseguite in regime di ricovero settimanale. Un team di 60 esperti ha inoltre analizzato periodicamente i singoli casi. Il 47% del campione era rappresentato da bambini, il 55% del campione aveva sesso femminile e il 53% dei pazienti presentava, al momento della valutazione, disturbi di tipo neurologico. Il programma ha portato a diagnosi in 39/160 pazienti considerati (24%). Nella maggior parte dei casi la diagnosi è stata effettuata in età adulta (32/94 adulti vs 7/66 bambini). Di questi, 9 sono risultati essere affetti da patologie comuni e note, 7 da due nuove entità nosologiche, i restanti da patologia rara nota non precedentemente riconosciuta. Nello studio sono stati identificati vari gruppi di pazienti con caratteristiche simili. Per quanto riguarda l'analisi genetica, tre patologie sono state identificate utilizzando l'analisi SNP,

altre tre utilizzando il *Whole exome sequencing*. Ne è emersa l'importanza dell'utilizzo di tecniche avanzate di sequenziamento genomico (NGS) nella ricerca della corretta diagnosi. Il programma è risultato tuttavia molto dispendioso in termini di tempo e di risorse umane e sproporzionato rispetto al numero di pazienti diagnosticati. Esso ha tuttavia posto l'attenzione sulla grande quantità di richieste formulate per l'accesso al progetto. Sulla base di ciò, il *National Institute of Health Common Fund* si è attivato, mettendo a disposizione risorse per la creazione di un *network* di pazienti rari senza diagnosi (*Undiagnosed Diseases Network*, UDN). Nel suo primo anno di vita, questo network [40] ha mostrato risultati promettenti fornendo strumenti di condivisione delle informazioni all'interno della comunità scientifica. Il suo approccio pionieristico, ha permesso inoltre di identificare quadri di presentazione estremamente rari e si è proposto di aumentare rapidamente la numerosità del campione di pazienti arruolati al fine di approfondire i meccanismi patogenetici di queste patologie.

Ad oggi, non sono disponibili dati che quantifichino il fenomeno dei pazienti senza diagnosi, né rispetto alla loro frequenza in popolazione né relativamente alla loro proporzione rispetto al totale dei pazienti con malattie rare. Si può ragionevolmente affermare che il loro numero dipenda anche dalla presenza o meno di reti di assistenza dedicate per le persone con malattie rare e dalla loro capacità di indirizzare, nei percorsi diagnostici più appropriati, i malati con sospetto di malattia rara.

Le scenario statunitense è in questo senso radicalmente diverso da quello europeo. Gli Stati Uniti hanno orientato da subito le loro politiche a favore delle malattie rare, supportando in particolare la ricerca su queste condizioni più che lo sviluppo di reti formali di assistenza.

Le politiche europee hanno assunto progressivamente una crescente valenza di salute pubblica. L'impulso, come negli Stati Uniti, è giunto dallo sviluppo di politiche a favore dei farmaci orfani. Successivamente è cresciuta la consapevolezza che, laddove la scarsità di conoscenze e il numero limitatissimo di persone affette rappresentano elementi peculiari, il contesto sovra-nazionale è il più adeguato per intraprendere interventi e stabilire principi di programmazione sanitaria. In questo senso le malattie

rare rappresentano un esempio paradigmatico di ambito di salute pubblica in cui le azioni, se supportate a livello di Unione Europea, possono assumere un valore aggiunto considerevole.

Sulla scia di questo orientamento, sono stati intrapresi ad oggi diversi interventi a livello europeo, ferma restando comunque l'autonomia decisionale in materia sanitaria dei singoli Stati Membri.

Questo scenario, in progressivo mutamento, vede ora affacciarsi all'orizzonte la concreta applicazione della Direttiva sull'assistenza sanitaria transfrontaliera n.24/2011. A livello europeo si stanno attualmente concludendo le fasi di selezione per il bando per la creazione delle reti europee di eccellenza (ERN) per le malattie rare. Tali reti sono strutturate per macro-gruppi di malattie, complessivamente 21. Compito preliminare degli Stati Membri era l'identificazione dei prestatori di assistenza sanitaria da candidare per la partecipazione a tali reti. All'inizio del processo si era ipotizzata la creazione di una rete europea di riferimento per i pazienti senza diagnosi. Tuttavia, vista la trasversalità di questo tema rispetto ai vari gruppi di patologie attorno ai quali ruotano le ERN, tale rete non è al momento tra quelle che potenzialmente verranno istituite. Tuttavia, il tema dei pazienti senza diagnosi sarà di estremo interesse per alcune di queste reti, in primis quella dedicata alle anomalie dello sviluppo e disabilità intellettive. Molte appaiono le potenzialità e allo stesso tempo criticità legate allo sviluppo di tali reti. Ovviamente, l'effetto di queste collaborazioni e le effettive ricadute assistenziali sui pazienti potranno essere valutate solo nel lungo periodo.

1.6 Clustering analisi, reti neurali e logica fuzzy

La Letteratura scientifica ha di recente mostrato interesse in campo biomedico nei confronti di sistemi matematici elaborati, già utilizzati in altri ambiti scientifici, per il calcolo delle probabilità e per il confronto tra popolazioni/campioni [41-44].

1.6.1 Clustering analisi

Il clustering o analisi dei gruppi è un insieme di tecniche di analisi multivariata dei dati volte a raggruppare elementi in classi omogenee in

modo tale che gli elementi della stessa classe siano il più simili possibile e gli elementi di classi differenti siano il più possibile diversi [45] .

Il clustering può essere pensato come una forma di compressione dei dati, dove un ampio numero di campioni sono convertiti in un piccolo numero di gruppi rappresentativi. A seconda dei dati e dell'applicazione possono essere usate diverse misure per identificare le classi, ad esempio, la distanza, la connettività e l'intensità [46].

Possiamo identificare due principali tipi di clustering:

a) *Hard Clustering o crisp clustering* dove l'assegnazione di ogni elemento ad un gruppo è esclusiva, pertanto i *clusters* risultanti non possono avere elementi in comune;

b) *Soft Clustering o Fuzzy Clustering* [47] in cui un elemento può appartenere a più cluster con gradi di appartenenza diversi. Questo approccio elimina i confini netti dell'*hard clustering*: ogni elemento appartiene a più cluster con un grado di appartenenza diverso. I gradi di appartenenza vengono calcolati attraverso algoritmi iterattivi. Uno degli algoritmi fuzzy di clustering ampiamente usato è il Fuzzy C-Means (FCM) [48].

Questo tipo di caratteristiche permette di applicare efficacemente la logica fuzzy negli ambiti che sono intrinsecamente mal definiti e dove i parametri considerati hanno limiti realmente sfumati, si pensi, ad esempio, in campo medico al processo diagnostico di malattie o sottogruppi di malattie mal definite.

Nello stesso modo la logica fuzzy viene applicata in sistemi esperti e a strategie decisionali come pure nella valutazione dei rischi naturali e nella modellistica dove entrano in gioco variabili e parametri che per vari motivi hanno una certa distribuzione e grado di incertezza.

1.6.2 La logica fuzzy

Nella logica fuzzy, introdotta da Zadeh, proprio come in logica classica, esistono gli insiemi di elementi detti "*fuzzy set*". La principale differenza tra la logica fuzzy e quella classica sta nel fatto che nella prima gli insiemi non sono più separati tra loro da confini netti e gli elementi non sono più caratterizzati dall'appartenenza o dalla non appartenenza esclusiva ad un

insieme. Al contrario, ad ogni elemento viene attribuito un grado di appartenenza ossia un valore numerico compreso nell'intervallo [0, 1], che dà un'indicazione di quanto un elemento appartenga ad un determinato insieme.

Questa possibilità di definire campi sfumati per i limiti di un insieme permette una grande flessibilità nella manipolazione e gestione delle informazioni, flessibilità che è impedita dalla logica classica che è per sua natura dicotomica (0 o 1, Vero o Falso, appartiene oppure NON appartiene all'insieme). Nella logica fuzzy, ad esempio, un elemento potrebbe contemporaneamente avere un grado di appartenenza del 30% ad un insieme dato mentre ad un altro del 70%.

Per effettuare un ragionamento in logica fuzzy è necessario stabilire delle regole (*fuzzy rules*) che permettano di ragionare sui dati di un problema [49-56]. L'associazione di un dato effettivo descrivente lo stato di un sistema (esempio segno, sintomo di una malattia) con un valore che indica il grado di appartenenza di un predicato viene effettuata da funzioni chiamate funzioni di appartenenza.

La funzione di appartenenza è una legge la cui distribuzione può avere la forma più varia (ad esempio: triangolare, trapezoidale, gaussiana etc.) che ha il compito di associare un dato effettivo, per esempio una temperatura di 70 gradi, con un valore che indica il grado di verità di un predicato, per esempio 0.85.

Le regole fuzzy sono tratte dall'esperienza. In altre parole, una persona esperta nell'affrontare un certo problema, descrive a parole come risolverebbe quel certo problema. Uno svantaggio, dunque, della logica fuzzy è legata al fatto che la *performance* dipende dalla regola base e dai dati raccolti.

1.6.3 Le reti neurali

Le reti neurali sono un'applicazione dell'intelligenza artificiale che permette di risolvere problemi la cui natura li rende impossibili da gestire con metodi computazionali classici.

Le reti neurali, come si può evincere dal nome, sono strutturate in modo tale da simulare il funzionamento del cervello umano. Esse permettono di

riconoscere specifici pattern attraverso la taratura dei parametri immessi nel sistema; ciò consente di minimizzare la possibilità di errore attraverso l'apprendimento da fonti specifiche [57-59]. Sono costituite da numerosi dispositivi (l'equivalente dei neuroni) connessi tra loro da collegamenti (l'equivalente di assoni e sinapsi). Questi dispositivi (*Processing Elements* o PE) chiamati neuroni di Hopfield, costituenti una rete neurale, sono dispositivi elettronici a più ingressi e ad una sola uscita.

Ogni Processing Elements riceve diversi segnali e può inviare il suo segnale d'uscita. La funzione dell'interconnesione tra vari Processing Elements, consiste nel dare un "peso" al segnale del neurone moltiplicandolo per un peso di interconnessione che varia tra 0 e 1.

Le reti neurali possono essere suddivise in reti supervisionate e non supervisionate.

Le prime sono sottoposte ad addestramento da parte di un supervisore che conosce in anticipo gli output che ci si aspetta in risposta ad un certo input. L'addestramento della rete consiste nell'inviare alla rete un input e osservare l'output ottenuto. Poi, conoscendo in anticipo l'output che si vuole ottenere, si modificano i pesi dei collegamenti in modo da ottenere i risultati voluti. Per istruire le reti neurali vengono utilizzate diverse metodiche tra cui la *backpropagation* (abbreviazione di "*backward propagation of errors*"), capace di addestrare la rete in modo tale da ottenere il minor gradiente di funzione di perdita (funzione di costo). Il fine ultimo di qualsiasi algoritmo di apprendimento è infatti quello di identificare la migliore funzione matematica in grado di associare una serie di imputs a degli output corretti, attribuendo un peso a ciascuna relazione. La discrepanza tra gli output determina inoltre l'errore della funzione individuata. Il sistema, utilizzando pesi diversi calcola un gradiente di funzione di costo e, attraverso lo stesso, riesce a discriminare tra una funzione e l'altra, apprendendo quella che meglio definisce la corrispondenza tra imputs e outputs corretti.

Le reti neurali non supervisionate (si conosce solo l'input) vengono utilizzate in genere come meccanismo di classificazione degli input in quanto sono in grado di riconoscere ed organizzare gli input simili tra loro.

Le regole di apprendimento sono le regole dove i pesi di interconnessione variano in base all'attività della rete [60-63].

1.6.4 Reti neuro-fuzzy

Di recente è nata un' "unione" tra le reti neuronali e le reti fuzzy, dove, i punti di forza di una, colmano i punti di debolezza dell'altro. La rete neuronale si comporta agli occhi dell'utente come una scatola nera, di cui si conoscono solo ingressi e uscite. In un sistema fuzzy verifichiamo in modo diretto il suo funzionamento. Quindi un sistema ottimale deve possedere l'elasticità tipica della logica fuzzy e la capacità di apprendimento che contraddistingue le reti neuronali o gli algoritmi genetici.

I sistemi che integrano le reti neurali con la logica fuzzy, detti ANFIS (*adaptive network based fuzzy inference system)*, acquisiscono conoscenza dai dati mediante gli algoritmi di funzionamento tipici delle reti neurali e la rappresentano mediante regole di tipo fuzzy.

L'utilizzo di questi sistemi in campo medico è relativamente recente e viene considerato sia in ambito clinico che in ambito diagnostico-terapeutico [64-67]. L'applicazione di reti neurali ha riguardato per esempio la corretta classificazione di un gruppo di tumori con caratteristiche istologiche in parte sovrapponibili (neuroblastoma, rabdomiosarcoma, linfoma non-Hodkin e tumore di Ewing) [57]. La difficoltà nel porre una diagnosi corretta tra tumori appartenenti a categorie diagnostiche differenti, ha indotto gli Autori a progettare reti neurali in grado di discriminare tra una patologia e l'altra attraverso l'analisi di campioni istologici e l'analisi genetica. E' così stato possibile stimare con precisione la probabilità di avere o non avere una determinata forma tumorale utilizzando la rete neurale per decifrare le alterazioni genetiche delle cellule tumorali (SRBCT, *round blue-cell tumors*). Il sistema è stato quindi utilizzato per creare una classificazione diagnostica. Altri campi in cui sono state applicate le reti neurali, sono, per esempio, la diagnosi dell'infarto acuto del miocardio e delle aritmie in base all'elettrocardiogramma o l'interpretazione delle immagini di radiografie e

risonanza magnetica nucleare. Esistono poi svariati altri esempi di reti neuro-fuzzy.

2. SCOPO DELLO STUDIO

Lo scopo del nostro lavoro è stato:

1. Creazione di uno strumento di descrizione analitica dei diversi profili fenotipici e genotipici di malattia rara

L'eterogeneità dei quadri di presentazione fenotipica delle malattie rare e le innumerevoli varianti geniche ad esse associate, rappresentano uno degli ostacoli maggiori alla corretta identificazione delle patologie stesse e ai meccanismi conoscitivi di questa complessa materia. A tutto ciò va aggiunta la difficoltà di utilizzare termini e definizioni di malattia condivisi e uniformi per le diverse realtà.

Da queste premesse è nata l'idea di creare uno strumento di raccolta delle informazioni di tipo universale, valido cioè per per tutte le patologie e adattabile ai diversi contesti clinici. Tale strumento prevede l'utilizzo dell'informazione secondo una logica gerarchica basata su entità prinicipali e secondarie (gli attributi) in modo tale da dare vita a un numero potenzialmente infinito di relazioni tra le entità stesse. Le entità, a loro volta, prevedono un'organizzazione in moduli utili ad organizzare il lavoro di progettazione e importanti nel definire al maggior dettaglio desiderato le informazioni.

Partendo quindi da un modello già esistente, il Registro delle Malattie Rare della Regione Veneto, già validato e in uso, si è pensato di implementare il sistema con nuovi moduli descrittivi dei profili fenotipici e genotipici.

Tale modello, partendo da una logica clinica, deve essere fruibile anche in ambito di ricerca. Esso deve, allo stesso tempo, consentire la definizione di una singola patologia o di patologie anche molto diverse tra loro al più

alto grado di completezza e accuratezza, garantendo flessibilità. In quest'ottica esso si deve caratterizzare per l'uso di informazioni non predefinite in modo tale da consentire ampio margine descrittivo anche di eventuali nuove entità nosologiche. Un'accurata progettazione dello stesso, inoltre, deve predisporre all'uso di informazioni precodificate, valide cioè per tutti i contesti in modo da uniformare la terminaologia descrittiva adottata.

La Cartella Clinica Generalizzata viene quindi ad assumere una logica "aperta" nei confronti delle diverse malattie e delle possibilità di descrizione dei singoli elementi clinici. Questa sua "disponibilità" nei confronti di ciascuna realtà clinica, favorisce una raccolta accurata delle informazioni e analisi più approfondite mediante *package* statistico.

Allo stesso tempo, lo strumento deve garantire: ampia fruibilità del sistema da parte degli operatori, accessibilità in tempo reale, privacy dei pazienti e protezione delle informazioni in esso contenute, tracciabilità dei dati.

2. Analisi della popolazione di pazienti con malattia rara arruolati nel sistema

Dopo progettazione e validazione del sistema, il progetto prevede l'arruolamento di pazienti affetti da malattia rara afferenti a Centri Accreditati coinvolti nel progetto UnRareNet ("*A multicenter collaborative research network for the identification and study of rare undiagnosed patients: the impact on the rare disease National Health Service network (UnRareNet)*". La raccolta delle informazioni relative a questa popolazione prevede la collaborazione con specialisti che necessitano di opportuno addestramento mediante incontri di confronto con il team di coordinamento e lezioni teorico-pratiche di tipo frontale. Tale premessa sembra infatti facilitare non solo la collaborazione degli operatori coinvolti, ma anche una loro maggiore preparazione nel descrivere con più accuretezza le informazioni cliniche, evitando possibili fonti di errore.

Al fine di ottenere un gran numero di dati utili a vari tipi di analisi, il campione può essere implementato con la popolazione già presente nel database del Registro delle Malattie Rare del Veneto.

I dati di popolazione, per le singole patologie o per l'intero campione, potranno quindi essere descritti o eleaborati a seconda delle necessità. In particolare sarà possibile andare ad analizzare le caratteristiche cliniche del gruppo di pazienti rari con diagnosi formulata e del gruppo di pazienti senza diagnosi.

3. Sviluppo del Sistema Esperto

Come già sottolineato, lo sviluppo di sistemi matematici intelligenti, capaci cioè di apprendere da dati informativi e selezionare il tipo di informazione utile a seconda del contesto e del tipo di analisi, rappresenta il futuro e la speranza di molti pazienti nell'ambito delle malattie rare. A questo proposito, il progetto prevede, come step finale, quello di utilizzare sistemi algoritmici complessi per l'analisi di varaibili. Tale progettazione richiede tuttavia una conoscenza approfondita della materia e la possibilità di istruire il sistema con dati qualitativamente elevati. Si è pertanto pensato di generare una popolazione di pazienti simulati, riportante caratteristiche di malattia in percentuale sovrapponibile a quanto rportato in Letteratura. Si procede quindi a validazione del sistema sulla popolazione di pazienti reali arruolati.

3. MATERIALI E METODI

3.1 La cartella clinica informatizzata (CCI)

La cartella clinica informatizzata è stata costruita partendo da un modello base già in utilizzo nel Registro per le Malattie Rare della Regione Veneto. Lo stesso è stato implementato con i contenuti relativi alle caratteristiche cliniche (fenotipiche e genotipiche) utili a definire una patologia. Il disegno sperimentale ha previsto la collaborazione con alcuni Centri Accreditati inseriti in un programma ministeriale dedicato ai pazienti senza diagnosi: "*A multicenter collaborative research network for the identification and study of rare undiagnosed patients: the impact on the rare disease National Health Service network (UnRareNet)*" (Ospedale Pediatrico Bambino Gesù, Dipartimento di Pediatria Azienda Ospedaliera Federico II Napoli, Istituto Neurologico "C. Besta", Istituto Dermopatico dell'Immacolata-IRCCS Roma, Università di Padova).

Ciascuno specialista referente è stato dotato di *login* e *password* per accesso al sistema informatico. E' stato richiesto il consenso scritto dei pazienti.

Il progetto di creazione della cartella generalizzata si è articolato in più fasi:

1. identificazione e condivisione dei contenuti e delle definizioni di malattia attraverso incontri e riunioni tra i rappresentanti dei centri accreditati coinvolti;
2. realizzazione e condivisione in rete del database;
3. arruolamento dei pazienti da parte dei Centri Accreditati collaboranti, dopo specifico corso di *training* con gli utenti del sistema;

4. implementazione del sistema e dei moduli della cartella clinica informatizzata in base alle richieste dei singoli Centri Accreditati e agli errori riscontrati nel dataset da parte del centro coordinatore;

3.1.1 Sistemi classificatori

Le variabili pazienti, intese come soggetti affetti/probabilmente affetti o non affetti da patologia rara e le informazioni cliniche ad essi correlate, sono state catalogate utilizzando una logica classificatoria di tipo gerarchico a struttura nodulare comprendente svariati moduli specifici per i diversi contenuti. Ciascun modulo è strutturato in modo da contenere entità principali ed entità secondarie capaci di comunicare tra loro attraverso innumerevoli relazioni.

3.1.2 Entità e relazioni

Il sistema creato è strutturato su:

- entità primarie: anagrafica, informazione nosologica, nota o sospetta; segni, sintomi, comorbidità, menomazione/danno funzionale, indagini diagnostiche (esami ematochimici, biochimici, istologici, radiologici, ecc.), genotipo;
- entità secondarie, i cosiddetti attributi, quali: data di rilevazione, severità, localizzazione, dimensioni, numero, colore, ecc.
- relazioni tra entità primarie e secondarie, capaci di creare legami complessi uno-molti e molti-molti.

3.1.3 Fonti

I sistemi di classificazione utilizzati sono:

- ICD, gerarchica con caratteristiche di unidimensionalità e con livello di granularità limitato al profilo di aggregazione dei gruppi di malattie correlate;
- OMIM, gerarchica con caratteristiche di unidimensionalità e con livello di granularità non limitato al profilo di aggregazione dei gruppi di malattie correlate;

- ORPHANET, con caratteristiche di multidimensionalità, capace di garantire l'associazione di un codice univoco (Orphacode) a ciascuna entità.

3.1.4 Contenuti

La realizzazione dello strumento analitico si è strutturata su un modello già esistente: il Registro per le Malattie Rare della Regione Veneto. Attraverso la collaborazione di esperti del Centro di Coordinamento Regionale per le Malattie Rare del Veneto è stato possibile creare e mantenere l'infrastruttura informatica del database, organizzando i contenuti secondo la logica prescelta, coordinando i vari Centri Accreditati partecipanti allo studio, supervisionando i contenuti e garantendo la loro sicurezza.
La cartella clinica informatizzata è stata creata con il supporto del database ORACLE, situato all'interno del Centro di coordinamento della Regione Veneto e collocato su differenti network proprietari SPC o GARR. Il sistema garantisce backups automatici che assicurano l'archiviazione ottica in caso di *disaster recovery*. Il prodotto finale è un'interfaccia utente di tipo JAVA [68].

3.1.5 Prodotti

Le informazioni raccolte nel data base vengono elaborate in dataset capaci di riportare:

- i pazienti nelle righe;
- le entità principali e secondarie nelle colonne, identificando le relazioni tra le stesse.

  Questi prodotti sono stati studiati al fine di facilitare l'analisi della popolazione studiata e delle sue peculiarità.

3.1.6 Implementazione

L'implementazione dei contenuti, sia riguardo alle caratteristiche genotipiche che a quelle fenotipiche, è stata realizzata attraverso le

competenze degli specialisti operanti nei Centri Accreditati coinvolti nello studio.

Le criticità del mezzo di supporto informatico, sono state risolte con l'aiuto del personale esperto operante nel Centro di Coordinamento delle Malattie Rare del Veneto.

3.2 Software di analisi

Le informazioni inserite nel database da parte degli operatori dei Centri Accreditati, sono state estratte usando il software SQL Developer di Oracle, in modo da renderli disponibili dapprima in formato Excel, utile per una prima visione ed elaborazione dei dati. In un secondo momento, si sono eseguite delle analisi più approfondite mediante package statistico SAS System 9.4 connesso direttamente ad Oracle. Ad una prima analisi descrittiva sono seguite analisi esplorative dei dati mediante la creazione di modelli ad hoc.

3.3 Profili di malattia dalla Letteratura

Per la creazione di profili di malattia che tengano conto di tutto lo spettro fenotipico presentato da ciascuna patologia, si è proceduto a una revisione sistematica della Letteratura scientifica consultando i dati contenuti in Orphanet, OMIM, MeSh. I criteri di selezione hanno privilegiato i lavori più recenti in ordine cronologico, con maggiore numerosità di campione, maggiore dettaglio nella descrizione della popolazione analizzata e migliore bontà del dato descritto. A ciascuna caratteristica clinica è stata quindi attribuita una prevalenza, in base al dato grezzo riportato nei lavori selezionati (dato quantitativo) o in base a una valutazione del gruppo revisore (dati qualitativi).

Per ciascuna caratteristica riscontrata si è proceduto ad uniformare la terminologia sia tra quanto riscontrato in Letteratura tra le diverse patologie sia tra quanto trovato in Letteratura e quanto disponibile nella Cartella Clinica Informatizzata. La presenza/assenza di ciascuna caratteristica clinica è stata individuata nella Cartella andando ad esaminare tra tutti i segni, sintomi, comorbidità e/o menomazioni. A

ciascuna caratteristica, quindi, sono stati associati un nome e gli identificativi già utilizzati in Cartella. In alcuni casi, caratteristiche in parte sovrapponibili o comuni ad uno stesso sintomo sono state raggruppate in una singola voce (gruppo).

In una prima fase del progetto, al fine di semplificare il percorso di programmazione, è stato deciso di concentrare le attenzioni su un gruppo ristretto di patologie che presentano caratteristiche cliniche in parte analoghe, ossia che entrano tra loro in diagnosi differenziale. Le patologie selezionate per questa prima analisi sono: la Sindrome di DiGeorge, che nella popolazione del database ha una buona numerosità di campione, l'Associazione CHARGE e la Sindrome di Smith-Lemli-Opitz.

In un secondo momento, si è proceduto a integrare il sistema con un numero più ampio di patologie (Sindrome di DiGeorge, Associazione CHARGE, Sindrome Leopard, Sindrome di Noonan, Sindromi Noonan-like, Sindrome di Kabuki, Sindrome di Sotos, Sindrome di Williams).

3.4 Identificazione dei casi e modelli

3.4.1 Simulazione dei casi

Per ciascuna patologia presa in esame, sono stati generati ad hoc dei pazienti con caratteristiche cliniche sovrapponibili ai dati di prevalenza della Letteratura.

Tale simulazione è avvenuta mediante il programma SAS System, che permette la generazione di un numero potenzialmente infinito di pazienti.

La numerosità del nostro campione, scelta a priori dall'equipe, è di 10.000 casi simulati per ciascuna patologia.

Sono stati quindi ottenuti dataset contenenti la popolazione dei pazienti sviluppata in righe e la descrizione della sintomatologia riportata (con ID identificativo per ciascuna voce) in colonne.

3.4.2 I casi reali

È stata successivamente eseguita un'analisi descrittiva dei dati inseriti dai Centri Accreditati nella Cartella Clinica Informatizzata, relativa alle patologie considerate.

Si sono ricavati dei dataset analoghi ai precedenti, con riportata la presenza/assenza delle diverse caratteristiche cliniche ricavate dai segni, sintomi, menomazioni e comorbidità imputate.

3.4.3 Sistemi esperti

Si è proceduto con l'implementazione dei sistemi esperti che sfruttano le proprietà della logica fuzzy e quella delle reti neurali.

Si è utilizzata una procedura (NEURAL) disponibile nel SAS System.

Il sistema è stato applicato dapprima sul gruppo di tre patologie considerate e descritte sopra (Sindrome di DiGeorge, Associazione CHARGE e Sindrome di Smith-Lemli-Opitz), e successivamente sulle otto patologie (Sindrome di DiGeorge, Associazione CHARGE, Sindrome Leopard, Sindrome di Noonan, Sindromi Noonan-like, Sindrome di Kabuki, Sindrome di Sotos, Sindrome di Williams).

Il sistema è stato messo a punto eseguendo svariate prove, differenti tra loro per le diverse combinazioni di parametri di input scelti, tra cui i principali sono: il numero di HIDDEN (nodi) inseriti, il numero di seme (RANDOM), la funzione di attivazione e il tipo di tecnica (TECHNIQUE) selezionata. Tra tutti i modelli testati è stato selezionato quello maggiormente capace di garantire la miglior convergenza, il maggiore apprendimento possibile e il minor errore nel riconoscimento della diagnosi di malattia.

Ciascun gruppo di pazienti, quello a tre malattie e quello a otto, è stato validato due volte.

Una validazione semplice è stata ottenuta suddividendo il dataset dei pazienti simulati in due gruppi: un primo insieme (pari al 75% dei casi) è stato utilizzato per la fase di *training* della rete; il restante campione (25%) per la validazione del sistema.

Una seconda validazione è stata poi ottenuta istruendo il sistema (training) con tutta la popolazione di pazienti simulati (100%) e testandolo successivamente sui casi con diagnosi nota già inseriti nel database dai Centri Accreditati (pazienti reali).

4. RISULTATI

4.1 Lo strumento: la cartella clinica informatizzata (CCI)

La cartella clinica informatizzata è stata pensata in modo tale da essere:

- unica: valida per tutte le patologie e per tutti i centri collaboranti;
- flessibile: per dare ascolto alle esigenze dei diversi specialisti collaboranti al progetto;
- analitica;
- sostenibile nel tempo e nei costi;
- accessibile 24 ore al giorno e 365 giorni l'anno;
- sicura nel garantire la tracciabilità del dato e la privacy dei pazienti.

Al fine di garantire una facile fruibilità del sistema si è pensato a un modello di cartella clinica "aperto", capace cioè di descrivere nel dettaglio le variabili cliniche dei pazienti, attraverso l'utilizzo di "entità primarie", come i segni, i sintomi e gli esami diagnostici, e "entità secondarie", i cosiddetti attributi, capaci di definire nel dettaglio le entità primarie.

La cartella è stata articolata in moduli:

- anagrafica (nome, cognome, data di nascita, residenza, Ulss di appartenenza, diagnosi certa/sospetta);
- entità nosologiche;
- trattamenti pregressi e piano terapeutico assistenziale;
- valutazioni e controlli clinici;
- diagnostica.

Uno sforzo aggiuntivo è stato fatto per il modulo delle localizzazioni, trovando soluzioni e logiche di classificazione capaci di collocare un segno/sintomo in qualsiasi parte dell'organismo, da quelle macroscopiche a quelle più piccole.

Tutto ciò ha consentito una maggiore libertà d'azione nella compilazione dei contenuti da parte dei vari specialisti coinvolti nell'inserimento dati, aumentando la qualità dell'informazione presente nel database. Va segnalato tuttavia come, nonostante la grande mole di voci presenti nel database, la selezione è risultata veloce e facilmente comprensibile.

Si riportano di seguito alcune immagini tratte dalla cartella clinica informatizzata (Immagine 1-5); risulta evidente come la "selezione a tenda" favorisca una più veloce e immediata selezione delle voci cercate.

Immagine 1. Modulo contenente i dati anagrafici della cartella clinica informatizzata.

Immagine 2. Selezione della diagnosi nel modulo dati anagrafici della cartella clinica informatizzata.

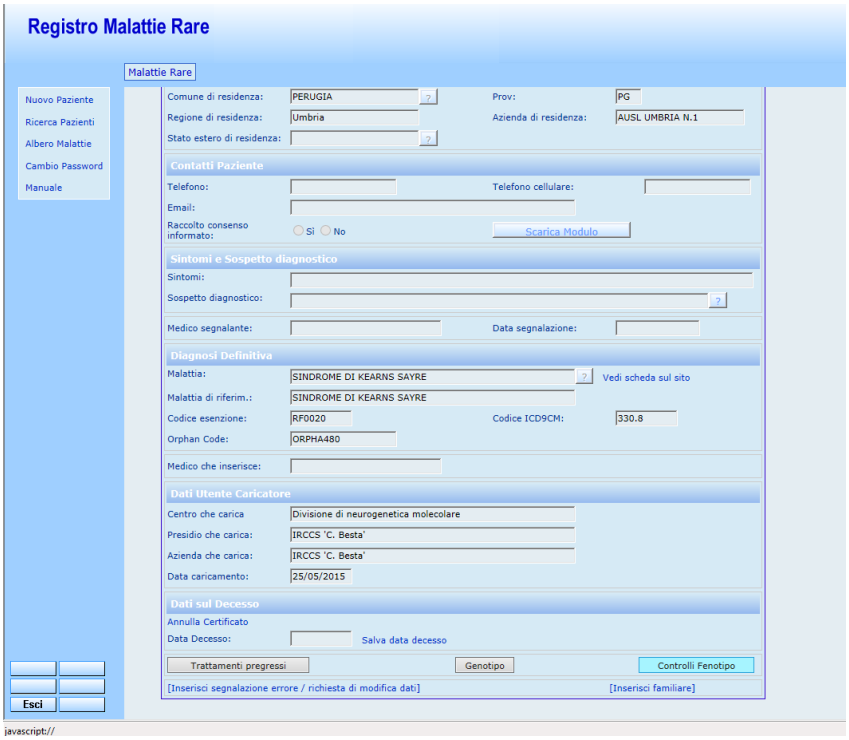


Immagine 3. Accesso all'inserimento dati clinici attraverso la selezione del campo "Controlli Fenotipo".

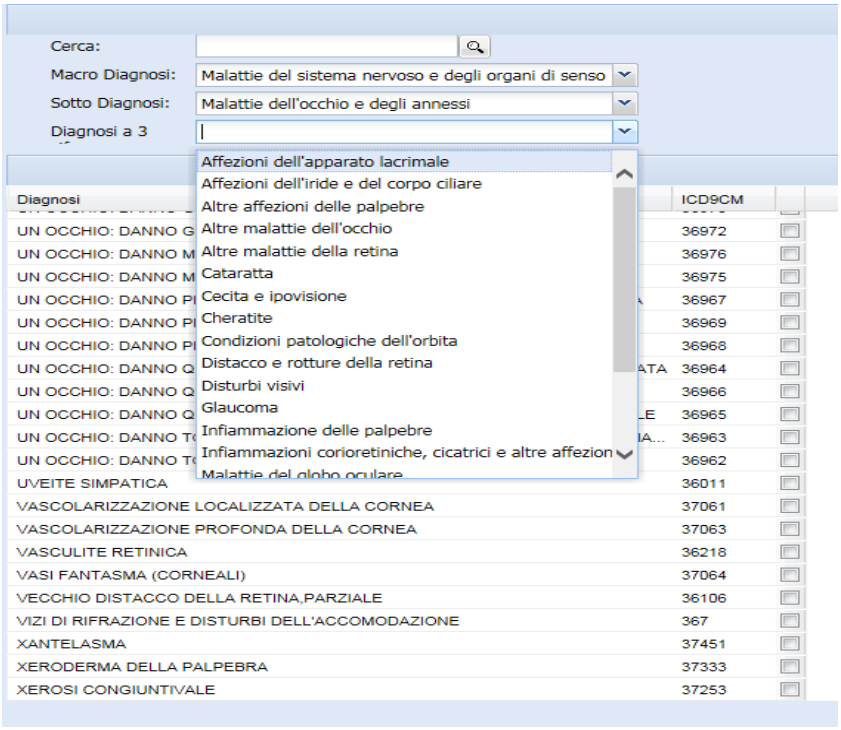


Immagine 4. Esempio di Modulo Segni all'interno del campo "Controlli Fenotipo".

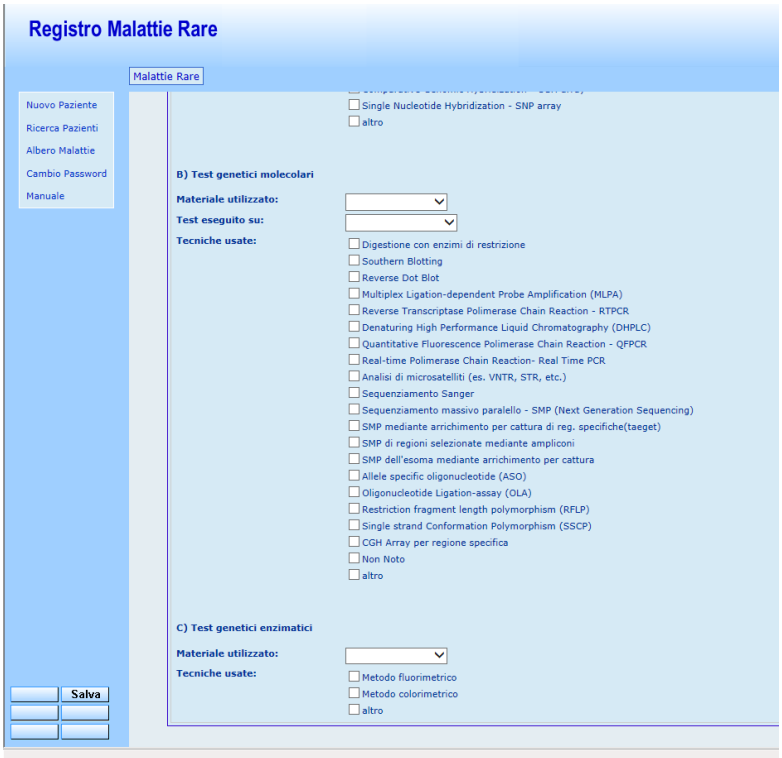


Immagine 5. Esempio del Modulo Genetica con le indagini diagnostiche correlate.

I contenuti del database sono stati estratti e riportati in tabelle Excel capaci di facilitare il lavoro di confronto e revisione dei dati da parte del personale che coordina il progetto e utili per eventuali elaborazioni. Le tabelle riguardano:

- “Elenco Segni”: entità principali dei segni distribuite per righe e i dati relativi a ciascuna entità distribuiti per colonna secondo le variabili ID progressivo, nome di sistema, alterazione e segno;
- “Elenco Sintomi”: entità principali dei sintomi distribuite per righe e i dati relativi distribuiti per colonna secondo le variabili ID progressivo e nome sistema, e sintomo;
- “Elenco Comorbidità”: entità principali di comorbidità distribuite per righe e i dati relativi distribuiti per colonna secondo le variabili ID progressivo e nome di macrogruppo e microgruppo di appartenenza nel sistema classificativo ICD;
- “Elenco Menomazioni”: entità principali di menomazioni distribuite per righe e i dati relativi distribuiti per colonna secondo le variabili ID progressivo e nome di macrogruppo e microgruppo di appartenenza nel sistema classificativo ICD, descrizione e codice della menomazione;
- “Elenco Indagini”: entità principali di indagini diagnostiche distribuite per righe e i dati relativi distribuiti per colonna secondo le variabili ID progressivo e nome di macrogruppo, microgruppo e indagine diagnostica;
- “Elenco Attributi”: entità secondarie di attributi distribuite per righe e i dati relativi distribuiti per colonna secondo le variabili ID progressivo, nome, tipo e descrizione;
- “Elenco Localizzazioni”: entità secondarie di attributi di localizzazione distribuite per righe e i dati relativi distribuiti per colonna secondo le variabili ID progressivo, descrizione della localizzazione, livello numerico corrispondente nella struttura albero delle localizzazioni, presenza o assenza della localizzazione;
- “Elenco Segni Localizzazioni”: indica le relazioni tra segni e attributi di localizzazione, riportando le entità segni in righe e i dati relativi a ciascuna distribuiti in colonna indicando ID e nome di sistema, alterazione, segno, presenza, descrizione o assenza della localizzazione, livello numerico

corrispondente della struttura albero delle localizzazioni, presenza o assenza di lateralizzazione;

- “Elenco Segni Sistemi Alterazioni Attributi”: indica le relazioni tra le entità segni, sistemi, alterazioni e le entità attributi, riportando le entità principali segni in righe e i dati di ciascuna entità in colonna secondo le variabili ID e nome di sistema, alterazione, segno, ID e nome dell’attributo;
- “Elenco Sintomi Localizzazioni”: indica le relazioni tra le entità sintomi e localizzazione, riportando le entità principali in righe e i dati di ciascuna entità in colonna secondo le variabili ID e nome di sistema, sintomo, descrizione o assenza di localizzazione;
- “Elenco Sintomi Sistemi Attributi”: indica la relazione tra le entità sintomi, alterazioni e le entità attributi, riportando le entità principali sintomi in righe e i dati di ciascuna entità in colonna secondo le variabili ID e nome di sistema e sintomo, ID e nome dell’attributo;
- “Elenco Indagini Segni”: indica le relazioni tra le entità principali indagini diagnostiche e le entità principali segni, riportando le entità principali indagini diagnostiche in righe e i dati di ciascuna entità in colonna secondo le variabili ID e nome del macrogruppo, microgruppo e indagine diagnostica, ID e nome del segno.

La numerosità delle voci per tabella viene qui riassunta (Tabella 1).

Segni	5014
Sintomi	299
Comorbidità	13861
Menomazioni	278
Indagini diagnostiche	1272
Attributi	910
Localizzazioni	18202
Relazioni Segni/Attributi	25834
Relazioni Sintomi/Attributi	2081

Tabella 1. Numerosità delle voci presenti nella Cartella Clinica Informatizzata e riportate nelle tabelle Excel estratte dall’applicativo.

L'errore più frequentemente riscontrato nell'inserimento dati dei pazienti arruolati è risultato essere la plurima digitazione di uno stesso campo per selezione di tutti i nodi utili ad arrivare ad un oggetto finale anziché del solo oggetto finale. Questo errore, facilmente riconoscibile dal centro di coordinamento del progetto, è stato corretto e ulteriori errori sono stati ovviati attraverso un confronto personale con il personale addetto all'inserimento dati.

4.2 Pazienti

L'inserimento dati da parte dei Centri accreditati coinvolti (specialisti autorizzati, dotati di login e password di accesso al registro) (Grafico 1) è iniziato nel Settembre 2014, dopo corsi di formazione specifica e dopo validazione del programma da parte degli esperti del Centro di Coordinamento Regionale per le Malattie Rare del Veneto. Il sistema è stato inoltre implementato con l'inserimento di una popolazione di pazienti affetti da malattia rara già inclusi nel Registro delle Malattie Rare della Regione Veneto.

Nei primi due anni di arruolamento pazienti (Settembre 2014-Settembre 2016), il numero complessivo di soggetti inseriti nel database è risultato pari a 1427 (518 afferenti al Progetto UnRareNet, i restanti appartenenti al Registro delle Malattie Rare del Veneto).

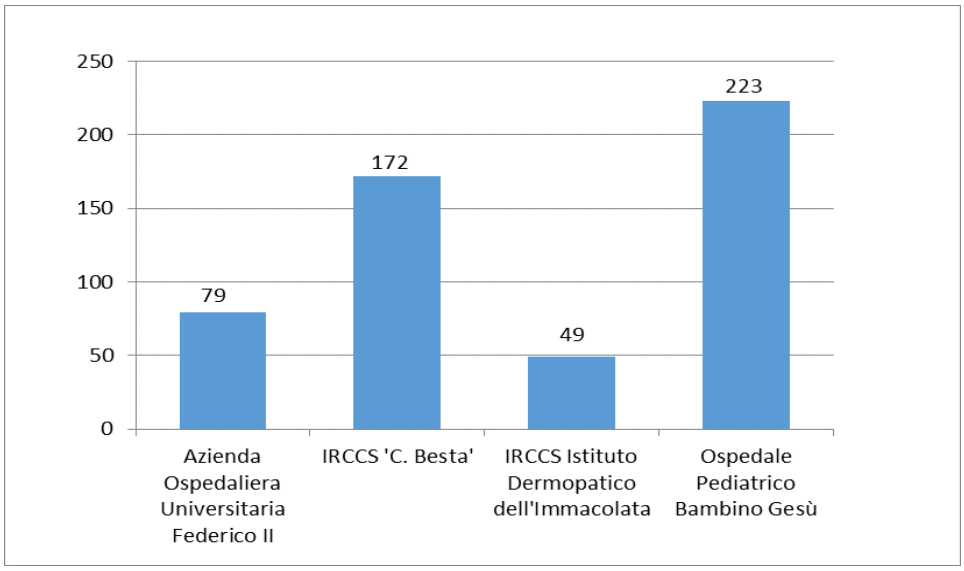


Grafico 1. Pazienti arruolati nel progetto UnRareNet, suddivisi per ciascun Centro Accreditato partecipante.

La media delle informazioni cliniche inserite per paziente è riportata nel seguente grafico (Grafico 2):

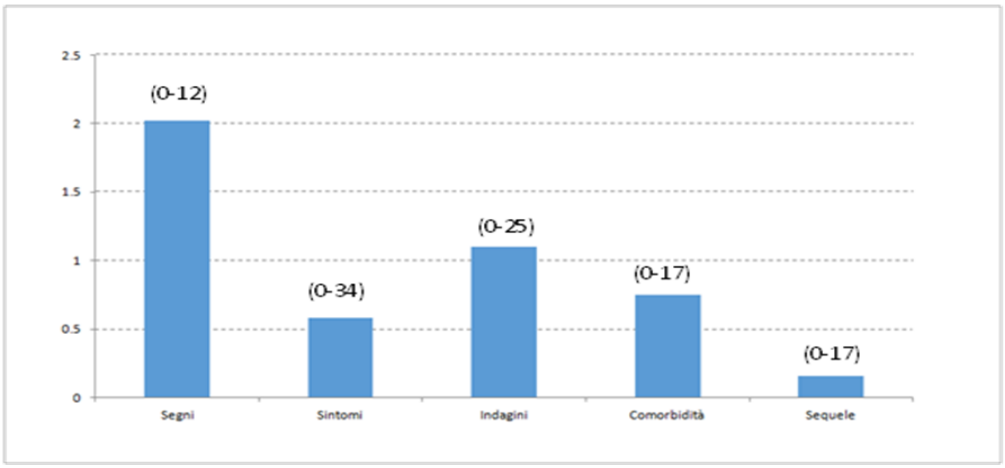


Grafico 2. Media (range) di segni, sintomi, indagini, comorbidità e sequele per paziente arruolato nel database.

Il consenso informato è stato raccolto nel 79,9% dei casi (1114 pazienti), rimandato a un successivo momento nel 12,2% dei pazienti (175), non noto nel 9,6% dei casi (138).

La popolazione di pazienti arruolati è stata analizzata suddividendola in due gruppi: quella dei pazienti con diagnosi di malattia rara nota e quella dei pazienti senza una diagnosi formulata.

4.2.1 Pazienti con diagnosi nota

La popolazione di pazienti con diagnosi di malattia rara formulata è pari al 89% della popolazione arruolata (1273/1427, di cui: 755 arruolati dal Registro delle Malattie Rare della Regione Veneto, 518 dal programma UnRareNet). Le diagnosi di questi pazienti vengono riportate nel seguente Grafico a torta (Grafico 3).

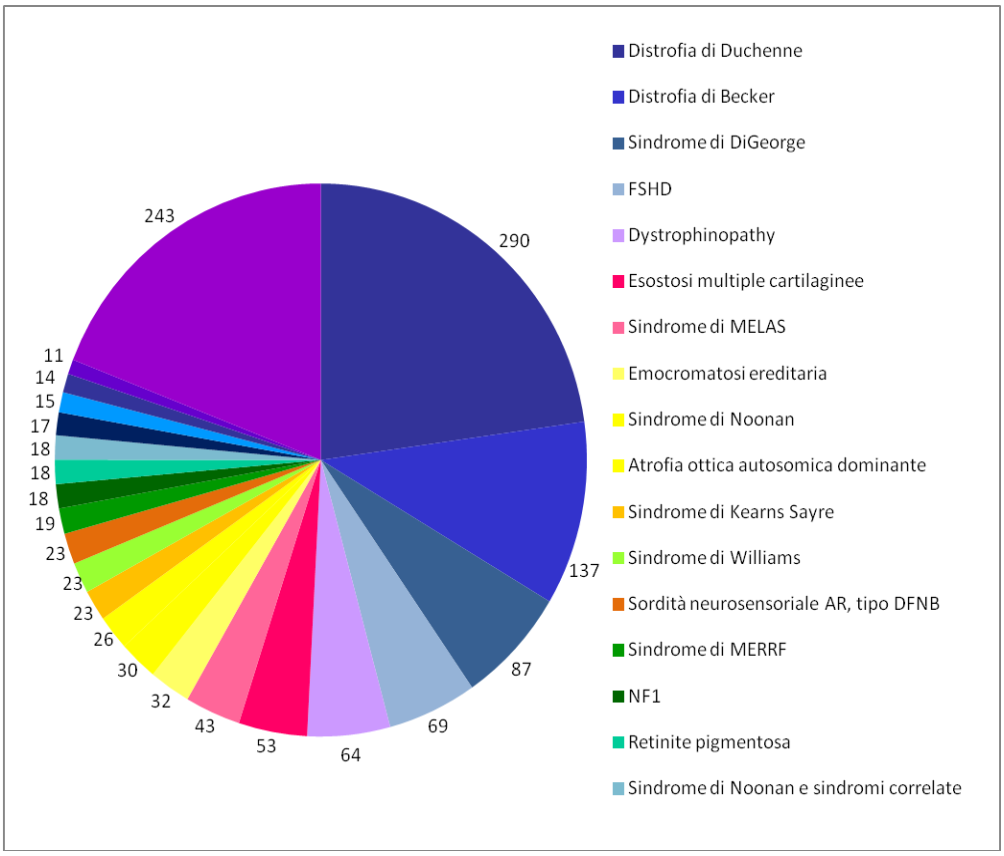


Grafico 3. Distribuzione per patologia dei pazienti arruolati.

Di seguito la rappresentazione della distribuzione dei pazienti per patologia suddivisi tra quelli afferenti al Registro delle Malattie Rare della Regione Veneto e quelli del Progetto UnRareNet (Grafico 4-5).

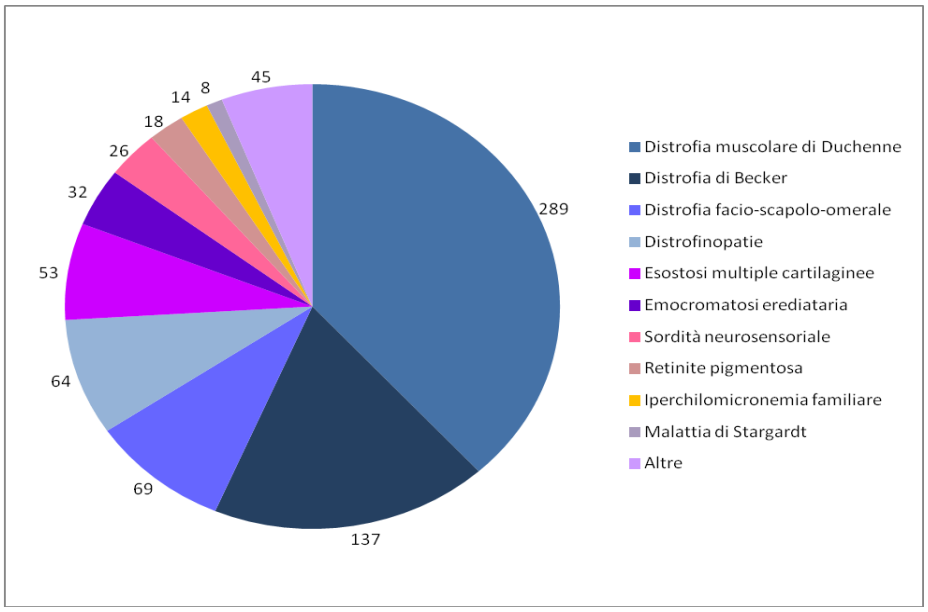


Grafico 4. Distribuzione per patologia dei 755 pazienti con diagnosi nota del Registro delle Malattie Rare del Veneto.

Le patologie più rappresentate tra i pazienti del Registro delle Malattie Rare della Regione Veneto sono quelle del gruppo delle distrofie (di Duchenne, di Becker, facio-scapolo-omerale), interessando circa due terzi del campione.

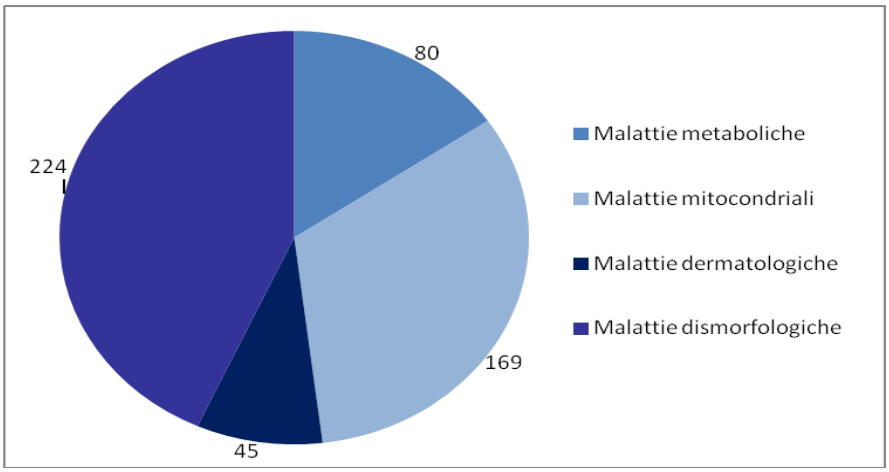


Grafico 5. Distribuzione per patologia dei 518 pazienti con patologia nota afferenti al progetto UnRareNet.

Per quanto riguarda il progetto UnRareNet le malattie più rappresentate sono risultate: la Sindrome di DiGeorge (16,7%), la Sindrome MELAS (8,3%), la Sindrome di Noonan (5,7%) e l'Atrofia ottica autosomica dominante (4,8%).

La maggior parte dei soggetti della popolazione arruolata è di sesso maschile (F:M=401:872). L'età media della popolazione è pari a 27,1 anni (range: 1,3-86,4; mediana: 19,05). La distribuzione per età è illustrata nel Grafico seguente (Grafico 6).

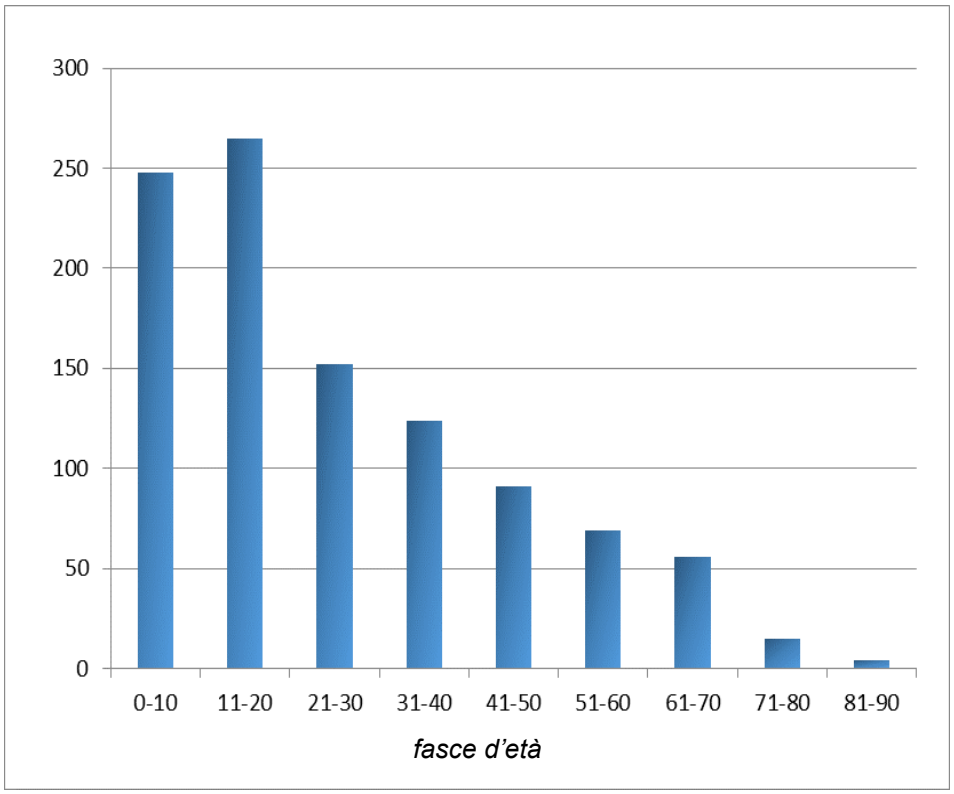


Grafico 6. Distribuzione per età dei pazienti con diagnosi formulata, arruolati nello studio.

Nel gruppo di pazienti con diagnosi sono state eseguite 1484 indagini genetiche (1,16 indagini/paziente).

Le tecniche utilizzate sono di seguito riportate in ordine di frequenza (Tabella 2).

Tecnica	N	%
Sequenziamento Sanger	556	37,4
Multiplex Ligation-dependent Probe Amplification (MLPA)	440	29,6
Southern Blotting	90	6,1
Denaturing High Performance Liquid Chromatography (DHPLC)	84	5,7
Real-time Polimerase Chain Reaction- Real Time PCR	64	4,3
Digestione con enzimi di restrizione	61	4,1
SMP di regioni selezionate mediante ampliconi	36	2,4
Cariotipo metafasico	34	2,3
Fluorescent In-Situ Hybridization (FISH)	27	1,8
Restriction fragment length polymorphism (RFLP)	26	1,75
Sequenziamento massivo paralello - SMP (Next Generation Sequencing)	24	1,6
SMP mediante arrichimento per cattura di reg. specifiche(taeget)	16	1,1
Reverse Transcriptase Polimerase Chain Reaction - RTPCR	10	0,7
Comparative Genomic Hybridization - CGH array	7	0,5
CGH Array per regione specifica	4	0,2
Metodo colorimetrico	1	0,06
Metodo fluorimetrico	1	0,06

Quantitative Fluorescence Polimerase Chain Reaction - QFPCR	1	0,06
Reverse Dot Blot	1	0,06
SMP dell'esoma mediante arrichimento per cattura	1	0,06
TOT	1484	100

Tabella 2. Tecniche genetiche utilizzate nella popolazione dei pazienti con diagnosi formulata e numerosità delle indagini effettuate per ciascuna tecnica.

4.2.2 Pazienti senza diagnosi

L'11% della popolazione arruolata, pari a 154 soggetti su 1427, è rappresentato da pazienti che non hanno ricevuto una diagnosi formulata (pazienti senza diagnosi) o per i quali è disponibile solo un sospetto diagnostico.

Le patologie sospettate dai clinici per questo gruppo di pazienti sono di seguito riportate (Tabella 3).

Malattia sospetta	N	%
Neurofibromatosi tipo 1	48	31,3
Emocromatosi ereditaria	41	26,8
Sordità neurosensoriale non sindromica autosomica recessiva, tipo DFNB	20	13,1
Malattia di Stargardt	16	10,4
Distrofia vitelliforme di Best	10	6,5
Retinite pigmentosa	6	3,9
Malattia mitocondriale	3	1,9
Epidermolisi bollosa	2	1,3
Central areolar choroidal dystrophy	1	0,6
Distrofia pigmentosa retinica	1	0,6
Esostosi multiple cartilaginee	1	0,6
Familial drusen	1	0,6
Ittiosi congenita	1	0,6
Sindrome di Kindler	1	0,6
Sindrome di Waardenburg, tipo 1	1	0,6
Sindrome Melas	1	0,6
TOT	154	100

Tabella 3. Sospetto diagnostico per i 154 soggetti senza diagnosi arruolati nel sistema.

L'età media di questa popolazione è risultata essere pari a 38,9 anni (range: 1,8-83; mediana: 22,2); anche in questo caso, la maggior parte dei soggetti è di sesso maschile (F: M=63:90).

La distribuzione per età di questa seconda popolazione è riportanta nel grafico seguente (Grafico 7).

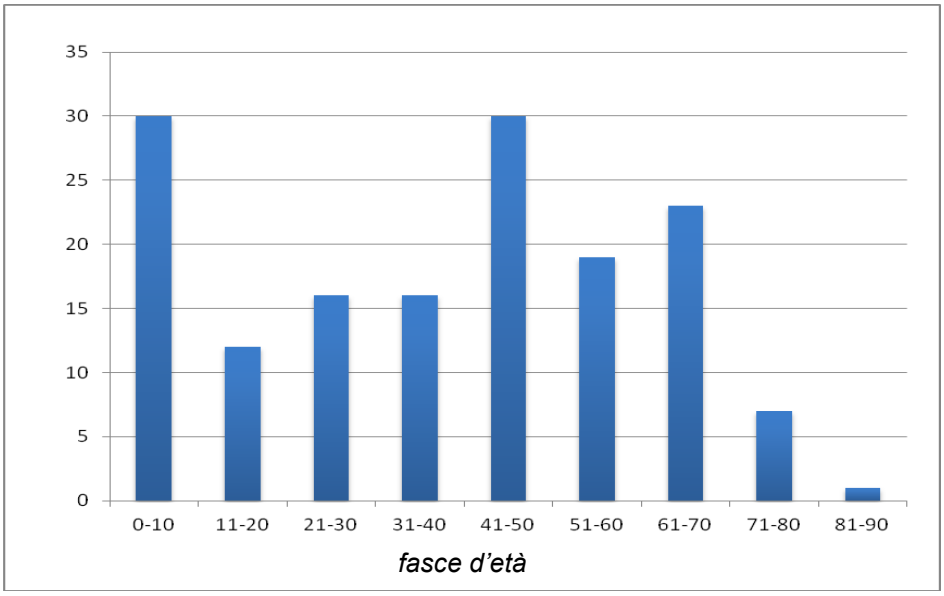


Grafico 7. Distribuzione per età dei pazienti senza diagnosi formulata, arruolati nello studio.

In questo gruppo di pazienti sono state eseguite 161 indagini genetiche. Le tecniche utilizzate sono riportate nella seguente Tabella (Tabella 4).

Tecnica	N	%
Sequenziamento Sanger	100	62,1
Denaturing High Performance Liquid Chromatography (DHPLC)	38	23,6
SMP mediante arrichimento per cattura di reg. specifiche(taeget)	16	9,9
Real-time Polimerase Chain Reaction- Real Time PCR	3	1,8
Digestione con enzimi di restrizione	1	0,6
Multiplex Ligation-dependent Probe Amplification (MLPA)	1	0,6
Real-time Polimerase Chain Reaction- Real Time PCR	1	0,6
SMP dell'esoma mediante arrichimento per cattura	1	0,6
TOT	161	100

Tabella 4. Tecniche genetiche utilizzate nella popolazione dei pazienti senza diagnosi formulata e numerosità delle indagini effettuate per ciascuna tecnica.

Circa il 60% di questo gruppo (70 pazienti) è stato sottoposto a un'unica indagine genetica, il 35,8% (pari a 42 pazienti) a due indagini/paziente, il 3,4% (4 pazienti) a 3 indagini, il 0,85% (1 paziente) a 4 indagini. Ciò nonostante, non si è riusciti a giungere alla formulazione di diagnosi di malattia.

4.2.3 Pazienti simulati

Contemporaneamente all'arruolamento di pazienti reali (con e senza una diagnosi), è stata eseguita una revisione della Letteratura scientifica più aggiornata riguardante un gruppo ristretto di patologie tra loro simili per alcune caratteristiche fenotipiche e genotipiche, cioè con diagnosi differenziale comune. Le patologie selezionate, come già anticipato, sono: la Sindrome di DiGeorge, che nella popolazione del database ha una buona numerosità di campione (87 pazienti, di cui due non sono stati inclusi nell'analisi per mancanza di dati fenotipici descrittivi inseriti nel database), l'Associazione CHARGE (1 paziente) e la Sindrome di Smith-Lemli-Opitz (2 pazienti). Di queste patologie, sono state estratte dalla Letteratura misure teoriche di frequenza delle caratteristiche cliniche e degli esami diagnostici.

Nel Grafico seguente (Tabella 5) si riportano le caratteristiche cliniche comuni a tutte le tre patologie (rosso), comuni a solo alcune delle tre patologie (giallo, verde, arancione), peculiari di ciascuna patologia (bianco).

DIGEORGE	
PARAMETRO	PREV.
anomalie_naso	49
bassa_statura	15
basso_peso_inf	13
difetto_setto_ventr	17
dismorf_facciali	30
dist_alim_deglutiz	**74**
idronefrosi	17.5
ipoacusia	30
labiopalatoschisi	29
micrognazia	35
palatoschisi	29
ritardo_psicomot	46.5
strabismo	18
anomalie_palpebre	34
emivertebre_torac	4.7
ipocalcemia	49.5
ipotiroidismo	20.5
refl_vesc_ureter	12
scoliosi	47
tetralogia_fallot	30
trasp_grandi_arter	1.75
agen_displ_renale	46
colelitiasi	19
ipertelorismo	32
ipospadia	10
microcefalia	6
ptosi	4
acne_severa	23
ambliopia	4
anemia_emol_autoim	0.5
ano_imperforato	5
anomalia_b_p	58.5
anomalie_coste	2.3
anomalie_tronco	2
aplasia_ipopl_timo	35
arco_aortico_destro	43
artrite_idiop_giov	2
artrite_reumatoide	0.5
depressione	5
diff_apprend	56.5
dita_affusolate	63
embriotoxon_poster	49
ernia_ombelicale	23
inclinaz_nervo_ret	1
insuff_velofaringea	66
inter_arco_aortico	18.5
iperattivita	5
ipertiroidismo	5
ipocalcemia_neonat	36
ipoparatiroidismo	19
ipopl_smalto_dent	5
malrotaz_intest	2.3
obesita_eta_ad	35
or_anom_art_succl	20
orecchie_picc_acc	17
porpora_tromb_idiop	4
psoriasi	0.5
ritardo_linguaggio	60
schizofrenia	22.6
seborrea	35
spina_bifida	4.7
suscettib_infez	56
tetania_convulsioni	46
tortuos_vasi_ret	39
tronco_arterioso	11
vena_cava_sup_sn	11
vitiligine	0.5

CHARGE	
PARAMETRO	PREV.
anomalie_naso	**19**
bassa_statura	70
basso_peso_inf	70
difetto_setto_ventr	16
dismorf_facciali	28
dist_alim_deglutiz	90
idronefrosi	22
ipoacusia	62
labiopalatoschisi	26
micrognazia	**38**
palatoschisi	26
ritardo_psicomot	72
strabismo	78
anomalie_palpebre	5
emivertebre_torac	22
ipocalcemia	25
ipotiroidismo	10
refl_vesc_ureter	22
scoliosi	10
tetralogia_fallot	24
trasp_grandi_arter	32
ambiguita_genit	5
criptorchidismo	40
difetto_setto_atr	9.5
malformazioni_SNC	40
micropene	50
orecchie_imp_basso	85
PDA	30
sindatt_clinodatt	18
rit_cresc_intraut	22
agenesia_dent	5
alteraz_olfatto	100
anisometropia	89
anom_par_nervi_cranici	40
anomalie_laringe	24
asimm_linea_mediana	70
asimmetria_fac	**47**
astigmatismo	78
atresia_coane	50
atrofia_cer	94
collo_corto	**19**
coloboma	80
dispnea	74
faccia_squadrata	**24**
lesioni_isch_emorr	47

SLO	
nome_SAS	PREV.
anomalie_naso	15
bassa_statura	75
basso_peso_inf	75
difetto_setto_ventr	25
dismorf_facciali	70
dist_alim_deglutiz	15
idronefrosi	5
ipoacusia	5
labiopalatoschisi	45
micrognazia	35
palatoschisi	45
ritardo_psicomot	60
strabismo	10
agen_displ_renale	15
colelitiasi	5
ipertelorismo	15
ipospadia	50
microcefalia	80
ptosi	10
ambiguita_genit	20
criptorchidismo	50
difetto_setto_atr	15
malformazioni_SNC	5
micropene	15
orecchie_imp_basso	20
PDA	5
sindatt_clinodatt	60
rit_cresc_intraut	80
affollam_dent	10
andat_anomala	15
cataratta	20
cisti_renali	5
coartaz_aortica	10
comp_autistico	50
comp_autol_agr	25
convulsioni	5
costipazione	20
dist_sonno	15
epicanto	15
fotosens_sev	60
incompl_lob_polm	20
ipertonia	40
ipotonia	25
lingua_ipopl	5
lussaz_anca	20
metatarso_add	5
monorene	5
mov_fet_ridotti	15
orecchie_ruot_post	15
polidattilia	36
pollici_corti	10
pollici_prossimali	20
polmoni_ipoplas	15
stenosi_piloro	25
suscettib_inf	10
vomito	5

Tabella 5. Caratteristiche cliniche comuni a tutte le tre patologie considerate (evidenziate in rosso), comuni a solo alcune delle tre patologie (giallo, verde, arancione), peculiari di ciascuna patologia (bianco).

Attraverso l'utilizzo di queste misure di frequenza e di software statistici (SAS System v.9.4), è stata quindi creata la popolazione di pazienti simulati: 30.000 pazienti (10.000 casi per ciascuna patologia) con caratteristiche cliniche sovrapponibili per frequenza a quelle già descritte in Letteratura (Immagine 6).

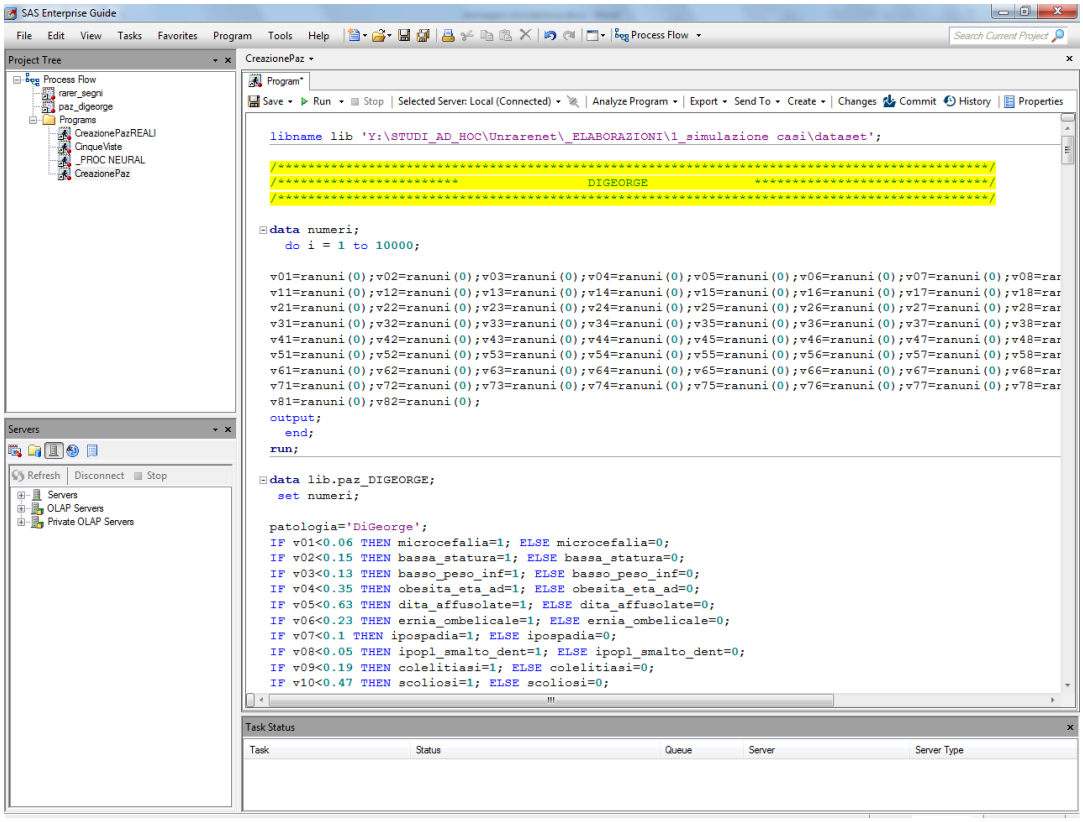


Immagine 6. Interfaccia SAS relativa alla simulazione di 10.000 pazienti con Sindrome di DiGeorge.

Il sistema ha generato dataset che riportano la popolazione di pazienti simulati nelle righe e la presenza/assenza (valori 0/1) delle varie caratteristiche cliniche nelle colonne. I valori di frequenza di queste caratteristiche, sono stati assegnati in maniera casuale dal sistema in modo da restituire il profilo di probabilità atteso e coerente con i dati della Letteratura (Immagine 7).

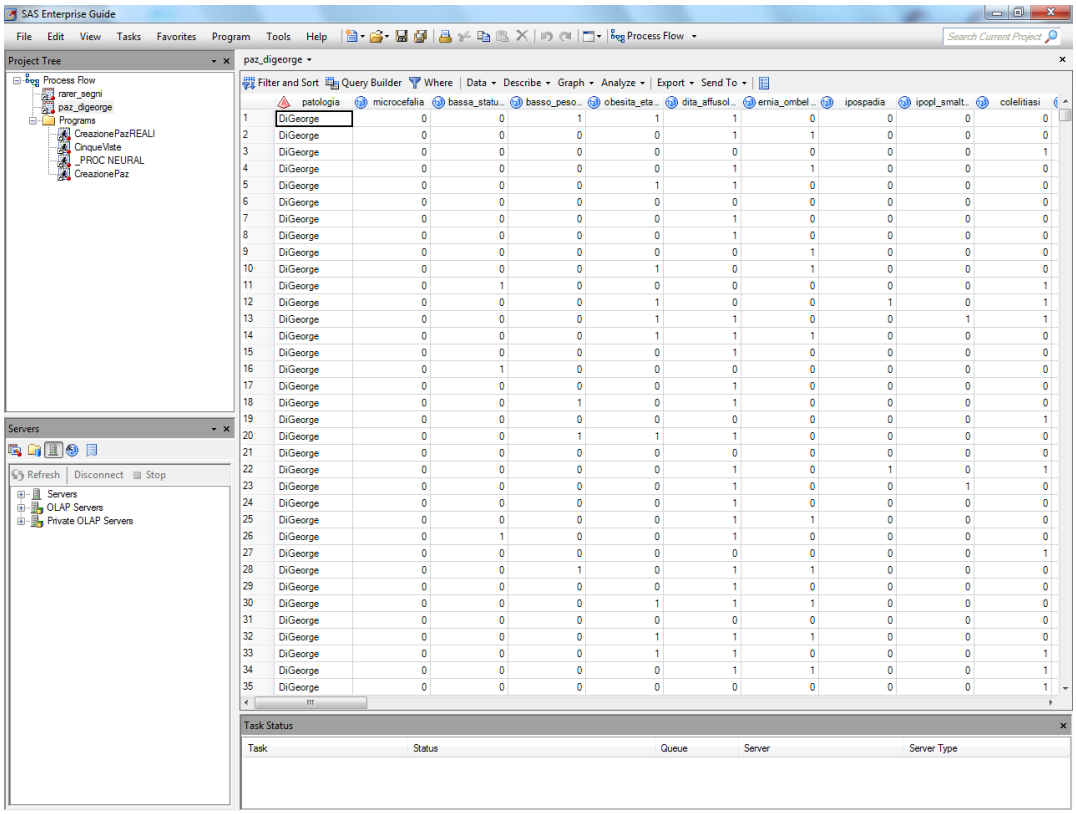


Immagine 7. Interfaccia SAS relativo alla popolazione di 10.000 pazienti simulati affetti da Sindrome di DiGeorge (righe) con le rispettive caratteristiche cliniche (colonne).

In un secondo momento, si è proceduto a creare dataset per un gruppo più ampio di patologie: Sindrome di DiGeorge, Associazione CHARGE, Sindrome Leopard, Sindrome di Noonan, Sindromi Noonan-like, Sindrome di Kabuki, Sindrome di Sotos, Sindrome di Williams. Per la raccolta dei dati di Letteratura, si è proceduto utilizzando analoghi criteri di scelta dei vari lavori. Il sistema (SAS System v.9.4) ha quindi generato una popolazione di pazienti simulati (10.000 per ciascuna patologia) aventi analoghe caratteristiche rispetto a quelle riportate nei vari lavori scientifici.

4.3 Sistema esperto

L'ultima fase del progetto ha previsto la creazione di un sistema matematico esperto, capace cioè, dopo un adeguato addestramento, di riconoscere all'interno di una popolazione, le patologie per cui è stato istruito attraverso un'analisi probabilistica delle caratteristiche cliniche dei pazienti della popolazione stessa. Il sistema esperto è stato creato utilizzando modelli matematici esistenti (SAS v.9.4) che sfruttano la logica fuzzy e le reti neurali.

La popolazione di pazienti simulati, con valori di frequenza che rispecchiano quelli della Letteratura scientifica, è stata utilizzata per istruire il sistema esperto (fase di *training*).

Si è quindi proceduto a due tipi di validazione.

Una prima validazione ha previsto l'addestramento del sistema esperto con il 75% della popolazione di pazienti simulati e la validazione sulla restante quota di pazienti. Dopo diverse prove, il sistema esperto ottimale è risultato quello con: 10 HIDDEN, RANDOM 1234 e funzione TECHNIQUE CONGRA.

Una seconda validazione ha previsto l'addestramento del sistema con l'intera popolazione di pazienti generati (100%) e la validazione sulla popolazione di pazienti reali.

4.3.1. Gruppo a tre patologie

Per quanto riguarda il primo gruppo di patologie testate, quello più ristretto (Sindrome di DiGeorge, Associazione CHARGE, Sindrome di Smith-Lemli-Opitz), il sistema ha evidenziato un'ottima capacità di apprendimento (pari al 100%) e ha riconosciuto correttamente tutte le diagnosi (stima dell'errore pari a 0) (Immagine 8).

Misclassification Table

Table of F_patologia by I_patologia					
		I_patologia(Into: patologia)			
		CHARGE	DIGEORGE	SLO	Total
F_patologia(From: patologia)					
CHARGE	Frequency	7500	0	0	7500
	Percent	33.33	0.00	0.00	33.33
	Row Pct	100.00	0.00	0.00	
	Col Pct	100.00	0.00	0.00	
DIGEORGE	Frequency	0	7500	0	7500
	Percent	0.00	33.33	0.00	33.33
	Row Pct	0.00	100.00	0.00	
	Col Pct	0.00	100.00	0.00	
SLO	Frequency	0	0	7500	7500
	Percent	0.00	0.00	33.33	33.33
	Row Pct	0.00	0.00	100.00	
	Col Pct	0.00	0.00	100.00	
Total	Frequency	7500	7500	7500	22500
	Percent	33.33	33.33	33.33	100.00

Misclassification Table - VALIDATION

Table of F_patologia by I_patologia

		I_patologia(Into: patologia)			
		CHARGE	DIGEORGE	SLO	Total
F_patologia(From: patologia)					
CHARGE	Frequency	2500	0	0	2500
	Percent	33.33	0.00	0.00	33.33
	Row Pct	100.00	0.00	0.00	
	Col Pct	100.00	0.00	0.00	
DIGEORGE	Frequency	0	2500	0	2500
	Percent	0.00	33.33	0.00	33.33
	Row Pct	0.00	100.00	0.00	
	Col Pct	0.00	100.00	0.00	
SLO	Frequency	0	0	2500	2500
	Percent	0.00	0.00	33.33	33.33
	Row Pct	0.00	0.00	100.00	
	Col Pct	0.00	0.00	100.00	
Total	Frequency	2500	2500	2500	7500
	Percent	33.33	33.33	33.33	100.00

Fits Statistics for the Training Data Set

Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Maximum Absolute Error	Train: Root Final Prediction Error	Train: Misclassification Rate	Train: Number of Wrong Classifications
2485.89	.000002930	0.084483	.001758877	0	0

Fits Statistics for the Training Data Set - VALIDATION

Valid: Average Squared Error	Valid: Maximum Absolute Error	Valid: Mean Squared Error	Valid: Misclassification Rate	Valid: Number of Wrong Classifications
.000002917	0.062400	.000002917	0	0

Immagine 8. Interfaccia del Sistema SAS con i risultati relativi alla prima validazione del modello esperto sul gruppo a tre patologie.

La seconda validazione, quella di confronto con la popolazione di casi reali (85 soggetti con Sindrome DiGeorge, 1 paziente con Associazione CHARGE e 2 pazienti con Sindrome di Smith-Lemli-Opitz) ha mostrato una capacità di apprendimento pari al 100% e ha riconosciuto correttamente 87/88 diagnosi (errore pari a 1,13%).

Anche in questo caso, dopo diverse prove, il sistema esperto ottimale è risultato quello con: 10 HIDDEN e RANDOM 123 e funzione TECHNIQUE

CONGRA. Questa seconda validazione ha dato risultati molto incoraggianti, individuando tutti i pazienti con Sindrome di Smith Lemli Opitz (1/1), tutti i pazienti con Sindrome di DiGeorge (85/85). Solo il paziente con Associazione CHARGE è stato riconosciuto come avente un profilo di probabilità di malattia più alto per Sindrome di DiGeorge, come mostrato dalle tabelle qui riportate (Immagine 9).

Misclassification Table

Table of F_patologia by I_patologia					
		I_patologia(Into: patologia)			
		CHARGE	DIGEORGE	SLO	Total
F_patologia(From: patologia)					
CHARGE	Frequency	10000	0	0	10000
	Percent	33.33	0.00	0.00	33.33
	Row Pct	100.00	0.00	0.00	
	Col Pct	100.00	0.00	0.00	
DIGEORGE	Frequency	0	10000	0	10000
	Percent	0.00	33.33	0.00	33.33
	Row Pct	0.00	100.00	0.00	
	Col Pct	0.00	100.00	0.00	
SLO	Frequency	0	0	10000	10000
	Percent	0.00	0.00	33.33	33.33
	Row Pct	0.00	0.00	100.00	
	Col Pct	0.00	0.00	100.00	
Total	Frequency	10000	10000	10000	30000
	Percent	33.33	33.33	33.33	100.00

Misclassification Table - VALIDATION

Table of F_patologia by I_patologia				
		I_patologia(Into: patologia)		
		DIGEORGE	SLO	Total
F_patologia(From: patologia)				
CHARGE	Frequency	1	0	1
	Percent	1.14	0.00	1.14
	Row Pct	100.00	0.00	
	Col Pct	1.16	0.00	
DIGEORGE	Frequency	85	0	85
	Percent	96.59	0.00	96.59
	Row Pct	100.00	0.00	
	Col Pct	98.84	0.00	
SLO	Frequency	0	2	2
	Percent	0.00	2.27	2.27
	Row Pct	0.00	100.00	
	Col Pct	0.00	100.00	
Total	Frequency	86	2	88
	Percent	97.73	2.27	100.00

Fits Statistics for the Training Data Set

Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Maximum Absolute Error	Train: Root Final Prediction Error	Train: Misclassification Rate	Train: Number of Wrong Classifications
2536.77	.000007358	0.22767	.002768446	0	0

Fits Statistics for the Training Data Set - VALIDATION

Valid: Average Squared Error	Valid: Maximum Absolute Error	Valid: Mean Squared Error	Valid: Misclassification Rate	Valid: Number of Wrong Classifications
.007838427	1.00000	.007838427	0.011364	1

Immagine 9. Interfaccia del Sistema SAS con i risultati relativi alla seconda validazione del modello esperto sul gruppo a tre patologie.

4.3.2 Gruppo a otto patologie

Un secondo tipo di analisi ha previsto l'implementazione del sistema messo a punto con un maggior numero di patologie rare, con caratteristiche cliniche in parte sovrapponibili tra loro e selezionate tra quelle della cartella clinica informatizzata. Le patologie selezionate, come già detto, sono otto:

- Sindrome di DiGeorge (85 pazienti);
- Associazione CHARGE (1 paziente);
- Sindrome Leopard (3 pazienti);
- Sindrome di Noonan (29 pazienti);
- Sindromi Noonan-like (18 pazienti);
- Sindrome di Kabuki (11 pazienti);
- Sindrome di Sotos (9 pazienti);

  Sindrome di Williams (23 pazienti).

La selezione delle stesse si è basata sulla maggiore rappresentazione delle popolazioni di malattia presenti nel database (numerosità di pazienti, riportata di fianco al nome di patologia) e sulla maggiore descrizione della casistica (informazioni inserite nel database).

Per ciascuna di esse, dopo aver creato profili di malattia riportanti la prevalenza di ciascuna caratteristica clinica (n=277), è stato utilizzato il sistema già in uso per creare una popolazione di pazienti generati pari a 10.000 individui per ciascuna entità nosologica (totale di 80.000 pazienti

generati). Nella Tabella 6 si riporta un estratto delle prevalenze riscontrate in Letteratura nelle 8 patologie relative alle prime 60 voci:

caratteristica clinica	DiGeorge	CHARGE	Leopard	NOONAN	NOONAN LIKE	KABUKI	SOTOS	WILLIAMS
disturbi_linguaggio	61	10	9	20	10	10	30	95
ritardo_psicomot	30	71	40	40	40	92	95	95
strabismo	18	78	5	48	15	23	25	54
scoliosi	47	10	5	12		17	41	51
dismorf_facciali	30	28		59	95	95	50	40
diff_apprend	56		30	11	15	14	45	66
iperlassita_ligam	10		5	5	5	76	20	5
ipotonia	9		51	5	25	71	69	5
criptorchidismo	15	40	50	40	20		5	
ipoacusia	30	63	19	10		22		8
bassa_statura	15	70	50		61	73		51
difetto_setto_atr	4	10		19	20	18	33	
difetto_setto_ventr	17	15		25	25	24	9	
rime_palp_obl_antim	20	5		45	25	92	90	
dist_alim_deglutiz	74	90		15	10	37		30
anom_bocca_labbr_fil	58		5	22	15	22		40
ipertelorismo	32		95	65	25		95	26
piede_piatto	10			10	20	20	46	5
piede_valgo	10			10	5	10	10	5
ritardo_cresctia_SP	14	50	44		49			30
epicanto	15	30		39	5			29
anom_padigl_auric	17		50	15	10	77		
ptosi_palpebrale	4		15	47	15	56		
malf_valv_polmonare	15		32	73	25			25
naso_infossato	3		10	10	10			10
RGE	40			10	10	59		25
coloboma		80	5	5	1	5		
orecchie_imp_basso		85	51	22	25	5		
ipospadia	10	10	30		3			
refl_vesc_ureter	12	22				40	20	
ipotiroidismo	21	10				16		17
microcefalia	6			15	5	43		
micrognazia	35			44	5	16		
malf_valvola_aort	8			8	1			61
narici_anteverse	3			10	25			15
collo_corto		19	31	87	20			
dimagr_sottopeso	13	30		40				
emivertebre_torac	4	22		27				
astigmatismo	3	78			2			
ipermetropia	2	2			2			
labiopalatoschisi	29	51				39		
tetralogia_fallot	29	24				6		
trasp_grandi_arter	2	32				9		
elice_piegato	10			9	10			
nevi	5			20	1			
dita_affusolate	62			5		10		
punta_naso_bulbosa	49			5				40
ang_bocca_basso	5				5	10		
labbro_sup_sottile	5				5	10		
naso_largo	5				20	20		
epilessia	5				2		5	
radice_naso_larga	3				6		5	
carie	5				8			31
ernia_ombelicale	23				5			5
iperattivita	5				31			21
radice_naso_schiacc	3				5			20
ernia_inguinale	5					5		5
stipsi	6					10		36
dist_sonno	5						10	5
PDA		30				17	18	

Tabella 6. Percentuali riscontrate in Letteratura per le prime 60 caratteristiche cliniche, suddivise per patologia.

Nella Tabella 7 si riporta l'analogo estratto delle prevalenze riscontrate nella popolazione dei casi reali:

caratteristica clinica	DiGeorge n=85	CHARGE n=1	LEOPARD n=3	NOONAN 29	NOONAN _LIKE n=18	KABUKI n=11	SOTOS n=9	WILLIAMS n=23
disturbi_linguaggio	42	100	33	17	17	64	56	35
ritardo_psicomot	68	100	33	17	44	64	100	87
strabismo	2	100	33	3	6	27	22	26
scoliosi	25			7			33	13
dismorf_facciali	27			38	11			39
diff_apprend	34		33	17	11	27	44	4
iperlassita_ligam	4			7	11	27	22	13
ipotonia	6			3		18	22	
criptorchidismo	7			34	39		11	
ipoacusia	8			10	6	9		4
bassa_statura	8		33	72	6	9		
difetto_setto_atr	18	100	33	17	22	36		9
difetto_setto_ventr	26			10	11	9		9
rime_palp_obl_antim	2	100		14	11		11	
dist_alim_deglutiz	1			3			11	4
anom_bocca_labbr_fil	33		67	28	22	45		48
iperterlorismo	5		67	14	17	9		
piede_piatto	5			3	6	18		9
piede_valgo	1				6	9	11	9
ritardo_cresctia_SP	2		33		56	45		17
epicanto	4	100		14	28			
anom_padigl_auric	20		67	17	28	45		
ptosi_palpebrale	1		33	34	39	18		
malf_valv_polmonare	11		33	41	67			26
naso_infossato	2		67	21	50			9
RGE	8			7	22	9		9
coloboma				3	6			
orecchie_imp_basso			33	3	44	9		
ipospadia		100			6			
refl_vesc_ureter								
ipotiroidismo	7					9		17
microcefalia	2			10	6	45		
micrognazia	9			3	11			
malf_valvola_aort	6			3	6	9		30
narici_anteverse	2			7	39			48
collo_corto			33		6			
dimagr_sottopeso	2		33	24				
emivertebre_torac								
astigmatismo	4			3	6			
ipermetropia	2				6	27		
labiopalatoschisi	9							
tetralogia_fallot	21							
trasp_grandi_arter								
elice_piegato	13			7	22			
nevi	2			7	6	9		
dita_affusolate	7			3		18		
punta_naso_bulbosa	24			3				
ang_bocca_basso	4				6	9		
labbro_sup_sottile	4				6	18		
naso_largo	1					9		
epilessia	5				6		11	
radice_naso_larga	2				6		11	
carie	15				11			
ernia_ombelicale	8				11		11	9
iperattivita	2		33	7	6			4
radice_naso_schiacc	1				6			4
ernia_inguinale	4					9		13
stipsi	2					9		4
dist_sonno	6						11	
PDA					11	9	22	

Tabella 7. Percentuali riscontrate nei casi reali per le prime 60 caratteristiche cliniche, suddivise per patologia.

Nella Tabella 8 si riporta la differenza tra la percentuale di prevalenza riscontrata nei casi reali rispetto ai dati di Letteratura:

caratteristica clinica	DiGeorge n=85	CHARGE n=1	LEOPARD n=3	NOONAN 29	NOONAN _LIKE n=18	KABUKI n=11	SOTOS n=9	WILLIAMS n=23
disturbi_linguaggio	-18	90	24	-3	7	53	25	-60
ritardo_psicomot	39	29	-7	-22	4	-28	5	-8
strabismo	-16	22	28	-45	-9	4	-3	-28
scoliosi	-22	-10	-5	-5	0	-17	-7	-38
dismorf_facciali	-3	-28	0	-21	-84	-95	-50	0
diff_apprend	-22	0	3	7	-4	13	0	-62
iperlassita_ligam	-7	0	-5	2	6	-49	2	8
ipotonia	-4	0	-51	-1	-25	-52	-47	-5
criptorchidismo	-8	-40	-50	-6	19	0	6	0
ipoacusia	-22	-63	-19	1	6	-13	0	-3
bassa_statura	-7	-70	-17	72	-55	-64	0	-51
difetto_setto_atr	14	90	33	-2	2	19	-33	9
difetto_setto_ventr	9	-15	0	-14	-13	-15	-9	9
rime_palp_obl_antim	-18	95	0	-31	-14	-92	-79	0
dist_alim_deglutiz	-73	-90	0	-12	-10	-37	11	-26
anom_bocca_labbr_fil	-25	0	62	6	7	24	0	8
ipertelorismo	-27	0	-28	-52	-8	9	-95	-26
piede_piatto	-6	0	0	-7	-15	-2	-46	3
piede_valgo	-9	0	0	-10	0	-1	1	4
ritardo_cresctia_SP	-12	-50	-11	0	6	45	0	-13
epicanto	-11	70	0	-25	23	0	0	-29
anom_padigl_auric	3	0	17	2	18	-31	0	0
ptosi_palpebrale	-3	0	18	-13	24	-38	0	0
malf_valv_polmonare	-4	0	2	-31	41	0	0	1
naso_infossato	0	0	56	11	40	0	0	-1
RGE	-32	0	0	-3	13	-50	0	-16
coloboma	0	-80	-5	-1	5	-5	0	0
orecchie_imp_basso	0	-85	-17	-18	19	4	0	0
ipospadia	-10	90	-30	0	3	0	0	0
refl_vesc_ureter	-12	-22	0	0	0	-40	-20	0
ipotiroidismo	-14	-10	0	0	0	-7	0	0
microcefalia	-4	0	0	-4	1	3	0	0
micrognazia	-26	0	0	-40	6	-16	0	0
malf_valvola_aort	-2	0	0	-5	4	9	0	-30
narici_anteverse	-1	0	0	-3	14	0	0	33
collo_corto	0	-19	3	-87	-15	0	0	0
dimagr_sottopeso	-10	-30	33	-16	0	0	0	0
emivertebre_torac	-4	-22	0	-27	0	0	0	0
astigmatismo	1	-78	0	3	4	0	0	0
ipermetropia	0	-2	0	0	4	27	0	0
labiopalatoschisi	-20	-51	0	0	0	-39	0	0
tetralogia_fallot	-8	-24	0	0	0	-6	0	0
trasp_grandi_arter	-2	-32	0	0	0	-9	0	0
elice_piegato	3	0	0	-2	12	0	0	0
nevi	-3	0	0	-13	5	9	0	0
dita_affusolate	-55	0	0	-2	0	8	0	0
punta_naso_bulbosa	-26	0	0	-2	0	0	0	-40
ang_bocca_basso	-1	0	0	0	1	-1	0	0
labbro_sup_sottile	-1	0	0	0	0	8	0	0
naso_largo	-4	0	0	0	-20	-11	0	0
epilessia	0	0	0	0	4	0	6	0
radice_naso_larga	-1	0	0	0	0	0	6	0
carie	10	0	0	0	3	0	0	-31
ernia_ombelicale	-15	0	0	0	6	0	11	4
iperattivita	-3	0	33	7	-25	0	0	-17
radice_naso_schiacc	-2	0	0	0	0	0	0	-15
ernia_inguinale	-2	0	0	0	0	4	0	8
stipsi	-4	0	0	0	0	-1	0	-31
dist_sonno	1	0	0	0	0	0	1	-5
PDA	0	-30	0	0	11	-8	4	0

Tabella 8. Differenze percentuali riscontrate nei casi reali rispetto ai dati di Letteratura per le prime 60 caratteristiche cliniche, suddivise per patologia.

Anche in questo gruppo, si è proceduto a una prima validazione istruendo il sistema con il 75% della popolazione generata, testandolo successivamente con la restante quota di pazienti.

Come già in precedenza, il sistema esperto ottimale è risultato quello con: 10 HIDDEN, RANDOM 1234 e funzione TECHNIQUE CONGRA.

Il sistema ha dimostrato un'ottima capacità di apprendimento (99%) e un buon riconoscimento delle patologie in questione rilevando solo lo 0.12% di tasso di errore (Immagine 10).

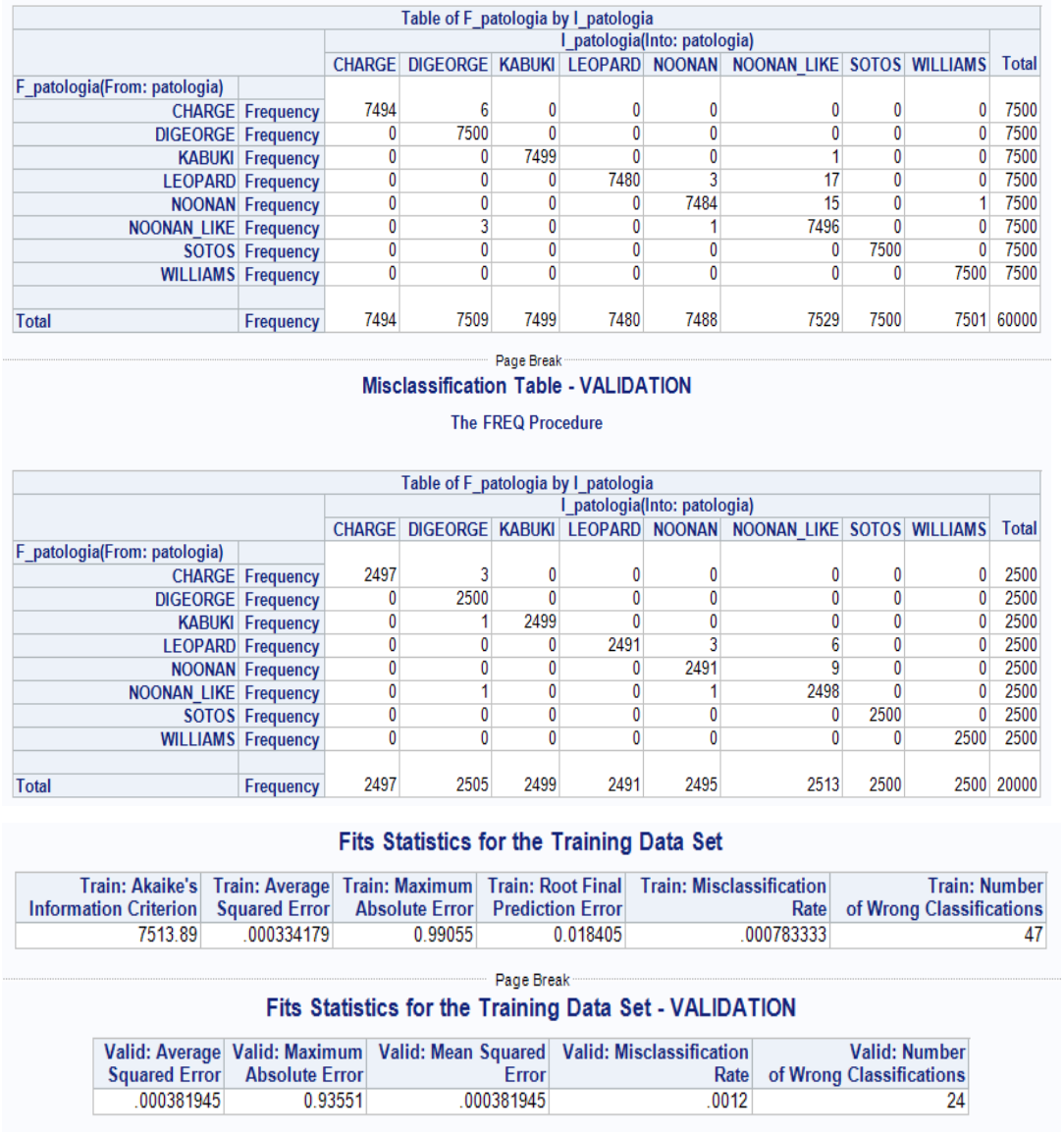


Table of F_patologia by I_patologia

		I_patologia(Into: patologia)								
		CHARGE	DIGEORGE	KABUKI	LEOPARD	NOONAN	NOONAN_LIKE	SOTOS	WILLIAMS	Total
F_patologia(From: patologia)										
CHARGE	Frequency	7494	6	0	0	0	0	0	0	7500
DIGEORGE	Frequency	0	7500	0	0	0	0	0	0	7500
KABUKI	Frequency	0	0	7499	0	0	1	0	0	7500
LEOPARD	Frequency	0	0	0	7480	3	17	0	0	7500
NOONAN	Frequency	0	0	0	0	7484	15	0	1	7500
NOONAN_LIKE	Frequency	0	3	0	0	1	7496	0	0	7500
SOTOS	Frequency	0	0	0	0	0	0	7500	0	7500
WILLIAMS	Frequency	0	0	0	0	0	0	0	7500	7500
Total	Frequency	7494	7509	7499	7480	7488	7529	7500	7501	60000

Page Break

Misclassification Table - VALIDATION

The FREQ Procedure

Table of F_patologia by I_patologia

		I_patologia(Into: patologia)								
		CHARGE	DIGEORGE	KABUKI	LEOPARD	NOONAN	NOONAN_LIKE	SOTOS	WILLIAMS	Total
F_patologia(From: patologia)										
CHARGE	Frequency	2497	3	0	0	0	0	0	0	2500
DIGEORGE	Frequency	0	2500	0	0	0	0	0	0	2500
KABUKI	Frequency	0	1	2499	0	0	0	0	0	2500
LEOPARD	Frequency	0	0	0	2491	3	6	0	0	2500
NOONAN	Frequency	0	0	0	0	2491	9	0	0	2500
NOONAN_LIKE	Frequency	0	1	0	0	1	2498	0	0	2500
SOTOS	Frequency	0	0	0	0	0	0	2500	0	2500
WILLIAMS	Frequency	0	0	0	0	0	0	0	2500	2500
Total	Frequency	2497	2505	2499	2491	2495	2513	2500	2500	20000

Fits Statistics for the Training Data Set

Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Maximum Absolute Error	Train: Root Final Prediction Error	Train: Misclassification Rate	Train: Number of Wrong Classifications
7513.89	.000334179	0.99055	0.018405	.000783333	47

Page Break

Fits Statistics for the Training Data Set - VALIDATION

Valid: Average Squared Error	Valid: Maximum Absolute Error	Valid: Mean Squared Error	Valid: Misclassification Rate	Valid: Number of Wrong Classifications
.000381945	0.93551	.000381945	.0012	24

Immagine 10. Interfaccia del Sistema SAS con i risultati relativi alla prima validazione del modello esperto sul gruppo a otto patologie.

Una seconda validazione di questo gruppo di patologie è stata ottenuta addestrando il sistema con la totalità dei pazienti simulati (80.000 soggetti), procedendo a successiva validazione sulla totalità della popolazione di pazienti reali (179 soggetti).

Anche in questa occasione, il modello di sistema esperto utilizzato è stato quello con: 10 HIDDEN, RANDOM 1234 e funzione TECHNIQUE CONGRA.

Anche in questa occasione il grado di apprendimento è risultato molto buono (98%). L'analisi dei risultati ha riscontrato un buon grado di riconoscimento per 5/8 patologie: Sindrome di DiGeorge nel 75% dei casi, Sindrome Leopard nel 66.6% dei casi, Sindrome di Noonan nel 72% dei casi, Sindrome di Sotos nel 77% dei casi, Sindrome di Williams nel 73% dei casi.

L'errore del sistema è risultato tuttavia elevato, pari al 32,9%, ossia 59 pazienti/179. Le patologie meno riconosciute dal sistema sono le rimanenti:

- Associazione CHARGE: 1 paziente identificato come Sindrome di Williams,
- Sindrome di Kabuki, riconosciuta solo nel 36% dei casi;
- Sindromi Noonan-like riconosciute solo nel 27% dei casi (Immagine 11).

Table of F_patologia by I_patologia

		I_patologia(Into: patologia)								
		CHARGE	DIGEORGE	KABUKI	LEOPARD	NOONAN	NOONAN_LIKE	SOTOS	WILLIAMS	Total
F_patologia(From: patologia)										
CHARGE	Frequency	9999	0	0	0	0	0	0	1	10000
DIGEORGE	Frequency	6	9979	1	5	9	0	0	0	10000
KABUKI	Frequency	0	0	10000	0	0	0	0	0	10000
LEOPARD	Frequency	0	0	0	9993	0	7	0	0	10000
NOONAN	Frequency	2	4	0	12	9956	21	0	5	10000
NOONAN_LIKE	Frequency	1	13	10	18	35	9923	0	0	10000
SOTOS	Frequency	0	0	0	0	0	1	9999	0	10000
WILLIAMS	Frequency	3	0	0	0	0	0	0	9997	10000
Total	Frequency	10011	9996	10011	10028	10000	9952	9999	10003	80000

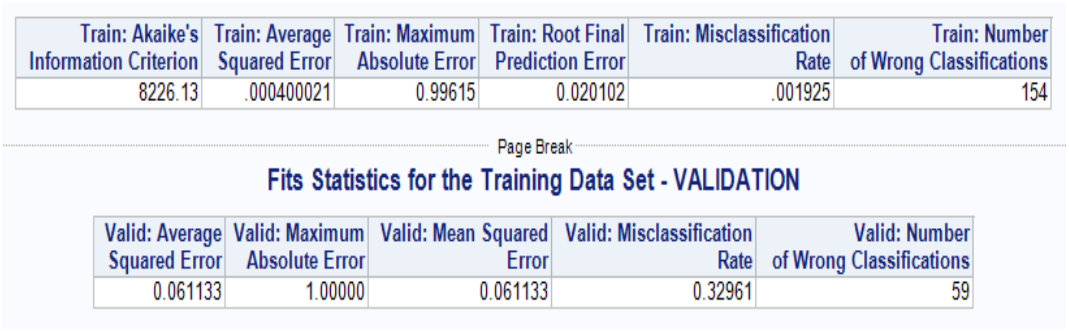
Page Break

Misclassification Table - VALIDATION

The FREQ Procedure

Table of F_patologia by I_patologia

		I_patologia(Into: patologia)							
		DIGEORGE	KABUKI	LEOPARD	NOONAN	NOONAN_LIKE	SOTOS	WILLIAMS	Total
F_patologia(From: patologia)									
CHARGE	Frequency	0	0	0	0	0	0	1	1
DIGEORGE	Frequency	64	1	4	5	6	0	5	85
KABUKI	Frequency	2	4	0	3	2	0	0	11
LEOPARD	Frequency	0	0	2	1	0	0	0	3
NOONAN	Frequency	1	0	1	21	5	0	1	29
NOONAN_LIKE	Frequency	0	2	7	3	5	0	1	18
SOTOS	Frequency	0	0	0	0	1	7	1	9
WILLIAMS	Frequency	4	0	0	1	1	0	17	23
Total	Frequency	71	7	14	34	20	7	26	179



Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Maximum Absolute Error	Train: Root Final Prediction Error	Train: Misclassification Rate	Train: Number of Wrong Classifications
8226.13	.000400021	0.99615	0.020102	.001925	154

Page Break

Fits Statistics for the Training Data Set - VALIDATION

Valid: Average Squared Error	Valid: Maximum Absolute Error	Valid: Mean Squared Error	Valid: Misclassification Rate	Valid: Number of Wrong Classifications
0.061133	1.00000	0.061133	0.32961	59

Immagine 11. Interfaccia del Sistema SAS con i risultati relativi alla seconda validazione del modello esperto sul gruppo a otto patologie (totalità pazienti reali).

L'analisi dei pazienti erroneamente identificati come affetti da una patologia diversa da quella di cui sono portatori, ha permesso di identificare due tipi di errore capaci di confondere il sistema esperto nella sua valutazione:

- in 23 pazienti su 53 (43% dei casi) le informazioni inserite nel database sono risultate troppo scarse o eccessivamente generiche (come ad esempio le voci: dismorfismi facciali, ritardo psico-motorio, anomalie del padiglione auricolare, ecc.);
- nei restanti 30 pazienti, invece, è stato evidenziato come l'elemento confondente per il sistema sia la discrepanza tra le prevalenze riportate in Letteratura e le prevalenze riscontrate nella popolazione di casi reali del database. Ad esempio, un paziente con Sindrome di Kabuki (il numero 5 nella Tabella 9) è stato identificato come probabilmente affetto da una delle patologie afferenti al gruppo delle Noonan-like: in questo soggetto la numerosità delle informazioni inserite dal centro clinico di riferimento è risultata adeguata (11 ID segno, 1 ID sintomo, 3 ID comorbidità) e il sistema ha correttamente attribuito la quasi totalità delle informazioni alla diagnosi corretta (Sindrome di Kabuki). Un unico ID segno è stato infatti attribuito alle patologie Noonan-like (il segno clinico: ipermetropia), che in Letteratura è solo raramente segnalato [69], mentre nella popolazione dei nostri pazienti, viene riportato nel 27% dei soggetti con Sindrome di Kabuki. Tale discrepanza tra quanto riportato in Letteratura e quanto constatato nella nostra popolazione, sembra determinare un maggior peso

nella scelta del sistema con conseguente errore di valutazione. La revisione dei casi errati ha mostrato molteplici episodi di discrepanza tra dati della Letteratura e dati della popolazione inclusa nello studio (Tabella 9).

	Malattia errata	Diagnosi corretta	segni1	segni2	segni3	segni4	segni5	segni6	segni7	segni8	segni9	segni10	segni11	segni12	segni13	sintomi1	sintomi2	com1	com2	com3	com4	com5	com6	com7	com8	com9	com10	men1	men2	men3	men4
1	KABUKI	DiGeorge	846													115		2525	5403	5625	8082	9345	9348	9371							
2	DIGEORGE	Kabuki	679	692	1034	1297	1545	1730	2322							94		3372	3955	4765	5196	9200	9340	14437							
3	DIGEORGE	Kabuki	531	647	679	762	799	1157	1545	1730	1838	2091	2322					4883	5196	9426	9468	9983									
4	NOONAN	Kabuki	1157	1174	1404	1838	2086	2169	2322									3356	3372	5437	5970	8870	9983	9984				111			
5	NOONAN_LIKE	Kabuki	782	844	956	1293	1545	1730	1838	1989	2027	2039	2086			206		5437	8915	9307											
6	NOONAN	Kabuki	347	370	1545	1554	2086	2117										5196	5639	9298	9306	9307	9339	9926	9983						
7	NOONAN_LIKE	Kabuki	767	956	1319	1877	2086	2322										3372													
8	NOONAN	Kabuki	518	647	808	838	1022	1030	1174	1849	2086							5196													
9	LEOPARD	NOONAN_LIKE	246	370	649	786	992	1033	1393	1838																					
10	LEOPARD	NOONAN_LIKE	373	444	806	1030	1174	1345	1844	1877								9339	9785												
11	LEOPARD	NOONAN_LIKE	782	806	1844	1877												6320	9331	9458											
12	NOONAN	NOONAN_LIKE	246	370	516	647	809	992	1024	1030	1292	1293	1368	1545	1989	89		5196	9306	9316	9983										
13	LEOPARD	NOONAN_LIKE	992	1174	1838	2086												9316	9983												
14	WILLIAMS	NOONAN_LIKE	806	1166	1404	1545	2159											3372	5074	5196	9307	9589	9983								
15	NOONAN	NOONAN_LIKE	256	761	992	1033	1166	1844	2118							89		4643	5062	9307	9983										
16	LEOPARD	NOONAN_LIKE	786	806	837	992	1166	1174	1545	1838								9316	9983												
17	NOONAN	NOONAN_LIKE	246	516	807	982	1174	1545	1730	2086						89		5061	9316	9983											
18	KABUKI	NOONAN_LIKE	370	373	761	786	806	1345	1838	1877	2086							9316													
19	KABUKI	NOONAN_LIKE	370	373	786	806	1166	1545	1838	2039	2194							5196	9316	9339											
20	LEOPARD	NOONAN_LIKE	246	991	2201													9307	9316												
21	LEOPARD	NOONAN_LIKE	246	370	740	786	806	1025	1174	2324								9204	9718												
22	NOONAN	WILLIAMS	786	2086														9316	9346												
23	DIGEORGE	WILLIAMS	444	786	1545	2086	2089											5196	9306	9307	9316										
24	DIGEORGE	WILLIAMS	846	1545												89	94	5031													
25	DIGEORGE	WILLIAMS	520	846	1545	1730												2430	2560	4878	9316	9468									
26	NOONAN_LIKE	WILLIAMS	256	434	435	647	820	846	1752	1989								6304													
27	WILLIAMS	CHARGE	264	465	1033	1128	1130	1166	1545	1730								5196	9307												
28	WILLIAMS	DiGeorge	516															8071	9145	9307	9324										
29	NOONAN	DiGeorge	692	957	1024	1033	1753	1844	2022	2086								7073	9306	9307	9348	9911	14409								
30	NOONAN_LIKE	DiGeorge	1751													170	192	3037	9306												
31	NOONAN_LIKE	NOONAN	846															9984													
32	LEOPARD	DiGeorge	246	808	1545	1838	2086	2124										9304													
33	WILLIAMS	DiGeorge	808	1545	2022	2041	2086											9304													
34	NOONAN_LIKE	NOONAN	846	992														9316	9984												
35	NOONAN	DiGeorge	846															5088	5196	9304	9316										
36	NOONAN_LIKE	NOONAN	846															9306	9984												
37	NOONAN_LIKE	DiGeorge	362	469	846																										
38	NOONAN_LIKE	DiGeorge	846	1545														5089	5405	9306	9307										
39	NOONAN	DiGeorge	246	260	692	786	992	1024	1085	1844	1965	2027	2086	2124	2129			9468	9984												
40	NOONAN_LIKE	DiGeorge	846	1545														9306	9345												
41	NOONAN	LEOPARD	806	978	979	1030	1174	1364	1838									5061	5071	9307	9468	9976	9984								
42	NOONAN_LIKE	NOONAN	846															9315	9316												
43	WILLIAMS	DiGeorge	246	474	622	745	800	1545	1844	2086								5196	9718												
44	NOONAN_LIKE	DiGeorge	846	1545														5196	9306	9307	9984										
45	DIGEORGE	NOONAN	846	1751														3358	3360	5196	9316										
46	NOONAN_LIKE	NOONAN	846															9984													
47	WILLIAMS	DiGeorge	1545															4883	9342												
48	WILLIAMS	DiGeorge	444	699	913	1034	1152	1546	1730	2022	2086							3037	3056	3372	9306	9762									
49	WILLIAMS	NOONAN	806	1024	1129	2086												5034	5196	7241	9202	9762	9984								
50	LEOPARD	NOONAN	246	829														2736	9316												
51	WILLIAMS	SOTOS	1545													391		4511	5196	5406	10003							67	112	233	
52	NOONAN_LIKE	SOTOS	704	761	838	956	1022	1033	1545	2174								3590	4000	4511	4547	5654	6371					235			
53	DIGEORGE	WILLIAMS	510	786	1545	2086	2089	3560										3268	9155	9983								78	113	231	235

Tabella 9. Descrizione dei 53 pazienti con diagnosi errata da parte del sistema esperto e dei codici identificativi descritti nel database (ID segno, ID sintomo, ID comorbidità e ID menonomazione). In grigio i 23 pazienti con informazioni di malattia scarse o aspecifiche.

Per minimizzare il rischio di errore, si è quindi proceduto ad eliminare dall'analisi i 23 soggetti "poco descritti" all'interno del database, lavorando con le informazioni dei restanti 156 pazienti.

La riformulazione dell'analisi ha mostrato (Immagine 12) un'ottima capacità di apprendimento del sistema (99.8%), come già in precedenza e un tasso d'errore pari al 23%.

Anche in questa occasione, 5/8 malattie sono state riconosciute correttamente:

- Sindrome di DiGeorge nel 82% dei casi,
- Sindrome Leopard nel 66.6% dei casi,
- Sindrome di Noonan nel 95% dei casi,
- Sindrome di Sotos nel 87.5% dei casi,

- Sindrome di Williams nel 94.4% dei casi.

Quelle con prestazioni meno valide sono risultate essere:

- Associazione CHARGE: 1 paziente identificato come Sindrome di Williams, come nell'analisi precedente,
- Sindrome di Kabuki, riconosciuta solo nel 40% dei casi;
- Sindromi Noonan-like riconosciute solo nel 31% dei casi.

Table of F_patologia by I_patologia

		I_patologia(Into: patologia)								
		CHARGE	DIGEORGE	KABUKI	LEOPARD	NOONAN	NOONAN_LIKE	SOTOS	WILLIAMS	Total
F_patologia(From: patologia)										
CHARGE	Frequency	9999	0	0	0	0	0	0	1	10000
DIGEORGE	Frequency	6	9979	1	5	9	0	0	0	10000
KABUKI	Frequency	0	0	10000	0	0	0	0	0	10000
LEOPARD	Frequency	0	0	0	9993	0	7	0	0	10000
NOONAN	Frequency	2	4	0	12	9956	21	0	5	10000
NOONAN_LIKE	Frequency	1	13	10	18	35	9923	0	0	10000
SOTOS	Frequency	0	0	0	0	0	1	9999	0	10000
WILLIAMS	Frequency	3	0	0	0	0	0	0	9997	10000
Total	Frequency	10011	9996	10011	10028	10000	9952	9999	10003	80000

Page Break

Misclassification Table - VALIDATION

The FREQ Procedure

Table of F_patologia by I_patologia

		I_patologia(Into: patologia)							
		DIGEORGE	KABUKI	LEOPARD	NOONAN	NOONAN_LIKE	SOTOS	WILLIAMS	Total
F_patologia(From: patologia)									
CHARGE	Frequency	0	0	0	0	0	0	1	1
DIGEORGE	Frequency	64	1	4	4	2	0	3	78
KABUKI	Frequency	2	4	0	2	2	0	0	10
LEOPARD	Frequency	0	0	2	1	0	0	0	3
NOONAN	Frequency	0	0	0	21	0	0	1	22
NOONAN_LIKE	Frequency	0	2	5	3	5	0	1	16
SOTOS	Frequency	0	0	0	0	1	7	0	8
WILLIAMS	Frequency	1	0	0	0	0	0	17	18
Total	Frequency	67	7	11	31	10	7	23	156

Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Maximum Absolute Error	Train: Root Final Prediction Error	Train: Misclassification Rate	Train: Number of Wrong Classifications
8226.13	.000400021	0.99615	0.020102	.001925	154

Page Break

Fits Statistics for the Training Data Set - VALIDATION

Valid: Average Squared Error	Valid: Maximum Absolute Error	Valid: Mean Squared Error	Valid: Misclassification Rate	Valid: Number of Wrong Classifications
0.045906	0.99999	0.045906	0.23077	36

Immagine 12. Interfaccia del Sistema SAS con i risultati relativi alla seconda validazione del modello esperto sul gruppo a otto patologie (popolazione di pazienti reali "senza il gruppo di pazienti poco descritti").

5. DISCUSSIONE

Nel mondo delle malattie rare si assiste ad un paradosso. Da più parti si afferma la scarsità di informazioni disponibili sulla patogenesi e la storia naturale di queste condizioni. Ciò corrisponde senza dubbio al vero per molte di queste patologie. Tuttavia, mai come in questo campo si assiste ad una vera e propria proliferazione di sistemi di raccolta di dati su queste malattie. Questi sistemi si sono sviluppati in maniera rapida e non articolata, per scopi diversi e quindi sono stati disegnati per rispondere alle diverse necessità di vari potenziali utilizzatori. Esiste anche un problema di visibilità delle malattie rare nei sistemi di classificazione generalmente utilizzati nei flussi sanitari. Nell'ICD-10 il numero di malattie rare aventi un codice specifico è di circa 500 [70]. Il miglioramento della codifica delle malattie rare è alla base del processo di revisione dell'ICD che esiterà nell'adozione della sua undicesima versione e rappresenta anche uno degli obiettivi della *Joint Action* europea sulle malattie rare *RD-Action*, attualmente in corso. La proposta a livello europeo è quella di utilizzare i codici orpha della classificazione Orphanet per identificare tutto ciò che è potenzialmente riconducibile alle specifiche entità rare corrispondenti, siano essi pazienti o episodi di cura. Questa scelta andrà certamente nella direzione di una minor frammentarietà dell'informazione relativa alle malattie rare.

La frammentazione dell'informazione è legata non solo alla minor visibilità nei sistemi informativi delle malattie rare rispetto ad altre malattie, ma anche all'eterogeneità degli strumenti di raccolta delle informazioni. Orphanet ha censito in Europa circa 690 registri che raccolgono dati su una o più malattie rare [71]. Essi differiscono per complessità, durata, *governance*, utilizzatori, strumenti utilizzati per la raccolta dei dati.

Esistono poi registri orientati non per patologia, ma per singolo trattamento disponibile. I limiti di questo tipo di approccio sono già stati evidenziati [72]. Lo scenario che si sta delineando con lo sviluppo delle reti di europee riferimento è al momento di difficile interpretazione. Da una parte tali reti, orientate per macro-gruppi di patologie, anche molto eterogenei al loro interno, costituiscono un'opportunità unica di condivisone di informazioni riferite ad un vasto numero di pazienti. D'altra parte lo sviluppo di piattaforme per la raccolta dei dati distinte, una per ciascuna ERN, potrebbe aumentare la frammentarietà, anziché ridurla. Alcuni gruppi ERN hanno manifestato l'interesse a sviluppare sistemi di raccolta dati del tutto nuovi, altri optano per non abbandonare quelli già in uso, preferendo orientarsi verso una loro possibile integrazione. Rimangono dei problemi aperti, legati alla effettiva possibilità di condivisione delle informazioni tra Centri siti in Paesi diversi, problemi di sicurezza del dato e problemi di tracciabilità dello stesso paziente tra ERN, soprattutto nelle aree cliniche di sovrapposizione. Entrambi gli approcci sopra descritti si gioverebbero comunque di una maggiore armonizzazione dei contenuti attorno ai quali questi sistemi di raccolta dati sono organizzati. Abbiamo già accennato alla codifica delle malattie rare. Ma il problema dell'eterogeneità che caratterizza il modo in cui l'informazione è raccolta riguarda anche il micro-dato, sia esso genetico o relativo al fenotipo. Nel presente lavoro ci si è focalizzati anzitutto su quest'ultimo aspetto. L'informazione sul fenotipo è presente e già raccolta in molti sistemi: registri, database, cartelle cliniche informatizzate. Tuttavia essa non è standardizzata. L'armonizzazione del modo in cui essa viene raccolta rappresenta un punto di svolta per disporre di dati utilizzabili a scopi di ricerca ed assistenziali. Tale attività è essenziale per interpretare meglio e più rapidamente il significato delle varianti del genoma identificate tramite le nuove tecniche di sequenziamento e rappresenta il prerequisito necessario, anche se non sufficiente, per rendere interoperabili sistemi informativi diversi. Nel presente lavoro si è voluto testare l'utilità di uno strumento applicabile per la raccolta di dati su una molteplicità di malattie rare, anche non correlate tra loro. Generalmente infatti i registri orientati su una singola malattia o su un gruppo di malattie simili, permettono la raccolta di dati di segni e

sintomi specifici di quelle condizioni, definiti a priori in base alle conoscenze già disponibili. Si raccoglie un'informazione di dettaglio, ma ci sono limiti alla raccolta di informazione non già prevista in fase di disegno del sistema. L'informazione aggiuntiva è registrata spesso come "altro", raccolta come testo libero, fatto che spesso ne pregiudica un'efficace elaborabilità a posteriori. Altri limiti di questo tipo di approccio sono la sostenibilità nel tempo di tali raccolte, slegate dal contesto assistenziale, riferite ad un limitato numero di casi e spesso quindi affette da distorsioni.

La cartella clinica generalizzata descritta nel presente lavoro rappresenta un esempio unico nel suo genere, permettendo potenzialmente la raccolta di dati su un ampio spettro di patologie, rare e più comuni. La cartella creata è risultata essere uno strumento:

- semplice da utilizzare, nonostante il considerevole numero di *records* contenuti (più di 30.000 voci e infinite possibilità di relazione tra entità principali e secondarie);
- adattabile alle singole realtà e implementabile in base alle richieste degli esperti coinvolti;
-sostenibile in termini di costi, tempo e risorse umane necessarie per la sua implementazione;
- affidabile relativamente alla tracciabilità del dato e alla sua protezione.

Il sistema esperto, formulato secondo la logica fuzzy e le reti neurali, è stato istruito attraverso l'inserimento di una popolazione di pazienti simulati affetta, in una prima analisi, da tre diverse patologie rare e, in un secondo momento da otto patologie, con caratteristiche in parte sovrapponibili tra loro. In entrambi i due tipi di analisi (quella con 3 patologie e quella con 8), il sistema è stato validato testando un gruppo di pazienti simulati a partire dalle caratteristiche generali sovrapponibili alla prima popolazione di pazienti simulati (prima validazione) e a partire dalla popolazione di pazienti rari con diagnosi formulata (seconda validazione). Entrambe le validazioni hanno dimostrato un'ottima capacità di apprendimento del sistema (98-100%) ed un errore di riconoscimento di malattia più basso nel gruppo ristretto di patologie (1,13%), più alto nel

gruppo di otto patologie (23%). In quest'ultimo tipo di analisi, la maggior fonte d'errore è risultata essere la scarsa accuratezza nella descrizione delle informazioni inserite nel database e la discrepanza con i dati di frequenza dei segni e dei sintomi riportati in Letteratura. Abbiamo infatti riscontrato una grande variabilità del dato di prevalenza riportato negi diversi lavori scientifici. Pur avendo attribuito in fase iniziale maggiore rilevanza ai lavori con campione più numeroso e più accuratamente descritto, in alcuni casi, tuttavia, si sono riscontrate maggiori analogie tra le prevalenze riportate nella nostra popolazione e quelle riportate in lavori aventi una casistica più limitata.

Dall'analisi delle fonti, è emerso molto evidente il problema dell'eterogeneità dei termini utilizzati nella descrizione delle informazioni cliniche. La complementarietà dell'informazione fenotipica rilevabile in fonti diverse è stata evidenziata anche in un recente lavoro, che ha indagato le differenze di espressione di uno stesso concetto, a seconda di dove l'informazione venga generata, ambito di ricerca versus ambito clinico, e di chi la interpreti e la registri [73]. Nessuna terminologia medica si è rivelata esaustiva. HPO contiene un'informazione di estremo dettaglio, ma poco fruibile in ambito assistenziale. L'opposto si può dire di SNOMED-CT. Il grande lavoro di definizione dei contenuti della cartella ha beneficiato della molteplicità di fonti analizzate.

Il sistema esperto ha dimostrato, a partire dai dati inseriti complessivamente una buona *performance*. Dopo opportuni accorgimenti per ridurre il rischio di errore, esso ha dato risultati molto promettenti, presentando un'ottimale capacità di apprendimento e un discreto tasso di identificazione di diagnosi (77%). Tra tutte le patologie oggetto di studio, le sindromi Noonan-like sono risultate quelle più difficilmente individuabili dal sistema, probabilmente a causa dell'eterogeneità dei profili fenotipici. Tale dato appare compatibile con quanto riportato in un recente lavoro scientifico [74]. L'implementazione del sistema con l'ampliamento ad altre patologie, permetterà, in un futuro prossimo, di aumentare la potenza dello stesso, consentendo inoltre l'analisi dei pazienti senza diagnosi.

La *performance* complessiva del sistema risente del seguente limite. È stato rilevato in questa prima fase il dato di prevalenza di un determinato

segno, singolarmente considerato. Non sono stati considerati i dati di frequenza di associazione tra segni che si riscontrano in una stessa patologia. *In primis* perché più difficilmente disponibili a partire dalla Letteratura. La definizione dei contenuti del modulo da utilizzare per la raccolta dei dati relativi al genotipo rappresenta uno sviluppo futuro previsto dello strumento. Si è parlato della necessità di lanciare, dopo il progetto Human Genome, il progetto *Human Phenome* [75]. A questo riguardo occorre sottolineare un aspetto determinante l'effettiva fruibilità, e quindi il successo, di qualsiasi strumento che si andrà a sviluppare in tal senso. Se il contesto è quello della ricerca è accettabile che lo strumento sia sviluppato e disponibile unicamente in lingua inglese, ma se uno strumento deve essere utilizzato nella pratica clinica, si pone il problema della sua traduzione in più lingue e dell'allineamento dei suoi contenuti. Questo aspetto riguarda ovviamente più la parte di raccolta dati sul fenotipo che quella sul genotipo.

In conclusione, la cartella clinica generalizzata si è dimostrata un valido strumento per la descrizione delle caratteristiche fenotipiche di pazienti con diverse malattie rare. Il sistema di raccolta, strutturato secondo una logica gerarchica e dotato di entità principali e secondarie, è risultato capace di descrivere condizioni molto differenti tra loro configurandosi come "sistema aperto", mantenendo nel contempo un grado accurato di dettaglio.

Le informazioni raccolte, per quanto numerose e complesse, sono risultate facilmente utilizzabili per elaborazioni, data la loro natura pre-codificata che crea omogeneità nel percorso conoscitivo, permettendone l'utilizzo da parte di clinici con diversi *background*. Il sistema inoltre ha dimostrato di poter essere implementato secondo le indicazioni dei singoli Centri accreditati partecipanti al progetto, ponendo le basi per progetti futuri di collaborazione e ampliamento della casistica. La condivisone di conoscenza e il confronto tra specialisti si è infatti dimostrato fondamentale per la corretta messa a punto del sistema.

La complessità del problema dei pazienti senza diagnosi può beneficiare della messa a punto di sistemi innovativi che indirizzino il clinico verso la

diagnosi di malattia attraverso la conoscenza e l'uso di algoritmi complessi come quelli inerenti le reti neurali e la logica fuzzy, capaci di autoapprendimento e funzioni filtro, come nel nostro studio. La messa a punto di questi sistemi esperti, istruiti con informazioni estratte dalla Letteratura scientifica e testati su una popolazione di pazienti reali, richiede tuttavia una descrizione analitica, sistematica ed accurata, delle informazioni immesse nel database al fine di predisporre il buon funzionamento di questi algoritmi. Da qui, come già sottolineato da più parti, l'importanza di sviluppare termini e definizioni condivisi ed integrabili tra loro, avvalendosi anche delle nuove ontologie sviluppate nell'ambito di malattie rare.

6. BIBLIOGRAFIA

1. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. *The next-generation sequencing revolution and its impact on genomics.* Cell. 2013;155(1):27-38.

2. Green ED, Watson JD, Collins FS. *Human Genome Project: Twenty-five years of big biology.* Nature. 2015;526(7571):29-31

3. Shen T, Lee A, Shen C, Lin CJ. *The long tail and rare disease research: the impact of next-generation sequencing for rare Mendelian disorders.* Genet Res (Camb). 2015;97:e15

4. McKusick VA. *The human genome through the eyes of Mercator and Vesalius.* Trans Am Clin Climatol Assoc 1981:92: 66–90

5. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR,Martin CL,McGuire AL, Nussbaum RL, O'Daniel JM, Ormond KE, Rehm HL, Watson MS, et al. 2013. *ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing.* Genet Med 15:565–574

6. Nachtomy O, Shavit A, Yakhini Z. 2007. *Gene expression and the concept of the phenotype.* Stud Hist Philos Biol Biomed Sci 38:238–254

7. Glaire MA, Brown M, Church DN, Tomlinson I. *Cancer predisposition syndromes: lessons for truly precision medicine.* J Pathol. 2016, Nov 9

8. London Dysmorphology Database, London Neurogenetics Database and Dysmorphology. M. Winter, M. Baraitser, Oxford University Press, ISBN 019851-780

9. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. *OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders.* Nucleic Acids Res. 2015;43: D789-98.

10. Hennekam RC1, Biesecker LG, Allanson JE, Hall JG, Opitz JM, Temple IK, Carey JC; *Elements of Morphology Consortium. Elements of morphology: general terms for congenital anomalies.* Am J Med Genet A. 2013;161A(11):2726-33.

11. Hunter A, Frias JL, Gillessen-Kaesbach G, Hughes H, Jones KL, Wilson L. *Elements of morphology: standard terminology for the ear.* Am J Med Genet A. 2009;149A(1):40-60.

12. Carey JC1, Cohen MM Jr, Curry CJ, Devriendt K, Holmes LB, Verloes A. *Elements of morphology: standard terminology for the lips, mouth, and oral region.* Am J Med Genet A. 2009;149A(1):77-92.

13. Hall BD, Graham JM Jr, Cassidy SB, Opitz JM. *Elements of morphology: standard terminology for the periorbital region.* Am J Med Genet A. 2009;149A(1):29-39.

14. Robinson PN1, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. *The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease.* Am J Hum Genet. 2008;83(5):610-5.

15. HPO : http://human-phenotype-ontology.github.io/

16. Groza T, Köhler S., Moldenhauer D., et al. *The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease.* Am J Hum Genet. 2015;97(1):111-24.

17. Lee D, de Keizer N, Lau F, Cornet R. *Literature review of SNOMED CT use.* J Am Med Inform Assoc. 2014 :21(e1):e11-9.

18. Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. *Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users.* Hum Mutat. 2012;33(5):803-8.

19. Liyanage H, Krause P, De Lusignan S. *Using ontologies to improve semantic interoperability in health data.* J Innov Health Inform. 2015 10;22(2):309-15.

20. Stelzer et Al. *VarElect: the phenotype-based variation prioritizer of the GeneCards Suite*. BMC Genomics 2016, 17(Suppl 2):444

21. Hamosh A, Sobreira N, Hoover-Fong J, Sutton VR, Boehm C, Schiettecatte F, Valle D. *PhenoDB: a new web-based tool for the collection, storage, and analysis of phenotypic features.* Hum Mutat. 2013;34(4):566-71.

22. Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chénier S, Chitayat D, Faghfoury H, Meyn MS, Ray PN, So J, Stavropoulos DJ, Brudno M. *PhenoTips: patient phenotyping software for clinical and research use.* Hum Mutat. 2013 Aug; 34(8):1057-65.

23. EU Commission recommendation on cross-border interoperability of electronic health record systems
http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=4214

24. McMurry JA, Köhler S, Washington NL. *Navigating the Phenotype Frontier: The Monarch Initiative.* Genetics. 2016;203(4):1491-5.

25. Mazzucato M, Visonà Dalla Pozza L, Manea S, Minichiello C, Facchin P. *A population-based registry as a source of health indicators for rare diseases: the ten-year experience of the Veneto Region's rare diseases registry.* Orphanet J Rare Dis. 2014;9:37.

26. Walker CE, Mahede T, Davis G, Miller LJ, Girschik J, Brameld K, Sun W, Rath A, Aymé S, Zubrick SR, Baynam GS, Molster C, Dawkins HJ, Weeramanthri TS. *The collective impact of rare diseases in Western Australia: an estimate using a population-based cohort.* Genet Med. 2016 Sep 22.

27. Graber ML, Kissam S, Payne VL, Meyer AN, Sorensen A, Lenfestey N, Tant E, Henriksen K, Labresh K, Singh H. *Cognitive interventions to reduce diagnostic error: a narrative review.* BMJ Qual Saf. 2012;21(7):535-57.

28. EurodisCare 2: survey of the delay in diagnosis for 8 rare diseases in Europe. http://www.eurordis.org/IMG/pdf/Fact_Sheet_Eurordiscare2.pdf.

29. Shoiania KG, Burton EC, McDonald KM, Goldman L. *Changes in rates of autopsy-detected diagnostic errors over time: a systemic review.* JAMA 2003; 289 (21): 2849-56.

30. Tiff CJ, Adams DR. *The National Institute of Health undiagnosed diseases program.* Curr Opin Pediatr 2014; 26(6):626-33.

31. Holtzer C, Meaney FJ, Andrewes J, Ciafaoni E, Fox DJ, James KA, Lu Z, Miller L, Pandya S, Ouvang L, Cunnif C. *Disparities in the diagnostic process of Duchenne and Becker muscular dystrophy.* Genet Med 2011; 13(11):942-7.

32. Huang L, Sadler L, O'Riordan MA, Robin NH. *Delay in diagnosis of Williams syndrome.* Clin Pediatr (Phila) 2002;41(4):257-61.

33. Seiber D, Hong CH, Takeuchi F, Olsen C, Hathaway O, Moss J, Darling TN. *Recognition of tuberous sclerosis in adult women: delayed presentation with life-threatening consequences*. Ann Intern Med 2011 Jun 21; 1 54 (12): 806-13; W-294

34. Svenstrup D, Jorgensen HL, Winther O. *Rare disease diagnosis: a review of web search, social media and large-scale data-mining approaches*. Rare Diseases, 2015, 3(1)e1083145.

35. Yang Y, Muzny DM, Reid JG, et al. *Clinical whole-exome sequencing for the diagnosis of mendelian disorders*. N Engl J Med. 2013;369(16):1502-11.

36. Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Bevan AP, Swaminathan GJ. *DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation.* Nucleic Acids Res. 2014;42(Database issue):D993-D1000.

37. Gahl WA et Al. *The National Institutes of Health Undiagnosed Disease Program: insight into rare diseases*. Genet Med 2012; 14(1):51-9.

38. Duan X, Markello T, Adams D, Toro C, Tifft C, Gahl WA, Boerkoel CF. *Cultural differences define diagnosis and genomic medicine practice: implications for undiagnosed diseases program in China*. Front Med 2013;7(3):389-94.

39. Tiff CJ, Adams DR. *The National Institute of Health undiagnosed diseases program.* Curr Opin Pediatr 2014; 26(6):626-33.

40. Brownstein CA, Holm IA, RAmoni R, Goldstein DB, Members of the Undiagnosed Diseases Network. *Data sharing in the Undiagnosed diseases network.* Hum Mutat 2015; 36(10):985-8.

41. Rizzi A. *Modelli statistici e reti neurali.* Statistica applicata. Vol 10, n. 4, 1998, pag. 595-608.

42. Mohammadpour RA, Abedi MA, Bagheri S, Ghaemian A. *Fuzzy rule-based classification system for assessing coronary artery disease.* Comput Math Methods Med. 2015:564867 Epub 2015Sep13.

43. Shaout A, Scharboeanu J. *Fuzzy logic based modification system for learning rate in backpropagation.* Comput Electr Eng 2000; 26:125-139.

44. Pagava K, Abesadze G, Uberi N, Korinteli I, Paghava I, Kvezereli-Kopadze M, Parulava T, Korinteli M, Kiseliova T. *Management options for rare diseases in children and adolescents in Georgia (experience of the country with transitional economy).* Georg Med News 2011; 193:8-11.

45. Bagirov A, Rubinov A, Yearwood J. *Using global optimization to improve classification for medicale diagnosis and prognosis.* Top Health Inf Manage, 2001:22(1):65-74.

46. Berks, G., Keyserling, D. Jantzeen, J. Ditoli, M., Axer, H., *Fuzzy clustering: A Versatile Mean to Explore Medical Database. ESIT 2000, Aachen, Germany, 2000, pp 453-457.*

47. Zadeh L.A., *Fuzzy Sets*. Information and Control 1965. 8: 338-353.

48. Bezdek JC. *Pattern Recognition with Fuzzy Objective Function Algoritm*, Plenum Press, 1981, New York.

49. Kruse R, Gebhardt JE, Klawon F. *Foundations of fuzzy systems*, 1994

50. Hellmann M. *Fuzzy logic introduction.* Epsilon Nought Radar Remote Sensing Tutorials, 2001.

51. Simpson PK. *Fuzzy Min-Max Neural Networks-Part 2: Clustering.* IEEE transactions on fuzzy systems, vol. 1, no. i , february 1993.

52. Jang JSR. *ANFIS: adaptive-network-based fuzzy inference system.* IEEE transactions on systems, man and cybernetics, vol. 23, no. 3, may-june 1993

53. Jang JSR, Sun CT, Mizutani E. *Neuro-fuzzy and soft Computing.* Prentice Hall, 1997.

54. Ekong VE, Onibere EA., Imianvan AA. *Fuzzy Cluster Means System for the Diagnosis of Liver Diseases.*, IJCST Vol. 2, Issue 3, September 2011.

55. Phuong NH, Kreinovich V. *Fuzzy logic and its applications in medicine.* International Journal of Medical Informatics 62 (2001) 165–173.

56. Adlassnig KP, *Fuzzy Set Theory in Medical Diagnosi.* IEEE transactions on systems, man, and cybernetics, vol. smc-16, no. 2, march/april 1986.

57. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.* Nat Med, 2001;7(6):673-679.

58. Haykin S. *Neural Network: a comprehensive foundation.* Prentice Hall, 1999.

59. Zhang GP. *Neural Networks for Classification: A Survey.* IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 30, no. 4, november 2000.

60. Acampora G, Kiseliova T, Pagava K, Vitiello A, *Towards Application of FML in Suspicion of Non-Common Diseases.* 2011 IEEE International Conference on Fuzzy Systems June 27-30, 2011, Taipei, Taiwan.

61. Senol C, Yildirim T. *Thyroid and Breast Cancer Disease Diagnosis using Fuzzy-Neural Networks*. Electrical and Electronics Engineering, 2009. ELECO 2009.

62. Singh S, Kumar A, Panneerselvam K, Vennila JJ. *Diagnosis of Arthritis Through Fuzzy Inference System.* J Med Syst. 2010 Oct 7.

63. Ubeyli E. D., Inan Guler. *Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems*. Computers in Biology and Medicine 35 (2005) 421–433.

64. Heden B, Ohlin H, Rittner R, Edenbrandt L. *Acute myocardial infarction detected in the 12-lead ECG by artificial neural networks.* Circulation, 1997;96:1798-1802.

65. Silipo R, Gori M, Taddei A, Varanini M, Marchesi C. *Classification of arrhythmc events in ambulatori electrocardiogram, using artificial neural networks*. Comput biomed Res 1995;28:305-318.

66. Ashizawa K, et Al. *Artificial neural networks in chest radiography: application to the differential diagnosis of interstitial lung disease*. Acad Radiol 1999;6:2-9.

67. Abdolmaleki P, et Al. *Neural network analysis o breast cancer from MRI findings*. Radiat Med 1997;15:283-293.

68. Festa I. *La Cartella Clinica Informatizzata nel percorso diagnostico-assistenziale del malato raro: sviluppo e implementazione di un sistema di raccolta e analisi dell'informazione clinica fenotipica e genotipica*. Tesi di Dottorato, 2015. http://paduaresearch.cab.unipd.it/9420/1/festa_ilaria_tesi.pdf.

69. Chaudhry IA, Shamsi FA, Alkuraya HS, Al-Sharif A. *Ocular manifestations in Kabuki syndrome: the first report from Saudi Arabia.* Int Ophthalmol, 2008;28:131–134.

70. Taruscio D, Gainotti S, Mollo E, Vitozzi L, Bianchi F, Ensini M, Posada M. *The current situation and needs of rare disease registries in Europe.* Public Helath Genomics 2013; 16(6):288-98.

71. Orphanet Report Series - Rare Disease Registries in Europe - January 2016 http://www.orpha.net/orphacom/cahiers/docs/GB/Registries.pdf

72. Hollak CE, Aerts JM, Aymé S, Manuel J. *Limitations of drug registries to evaluate orphan medicinal products for the treatment of lysosomal storage disorders.* Orphanet J Rare Dis. 2011;6:16.

73. Alnazzawi N, ThompsonP, Ananiadou S. *Mapping Phenotypic Information in Heterogeneous Textual Sources to a Domain Specific Terminological Resource.* PLoSONE, 2016;11 (9): e0162287.

74. Yang Y, Muzny DM, Reid JG, et al. *Clinical whole-exome sequencing for the diagnosis of mendelian disorders.* N Engl J Med. 2013; 369(16):1502-11.

75. Boycott KM1, Vanstone MR, Bulman DE, MacKenzie AE. *Rare-disease genetics in the era of next-generation sequencing: discovery to translation.* Nat Rev Genet. 2013;14(10):681-91.

A tutto il Dipartimento di Epidemiologia, a tutte le risate e le ore trascorse assieme,

A Ilaria, la mia "apri pista",

A mia madre e a tutta la mia Famiglia che, come me, ha sudato ogni parola di questa tesi,

e, naturalmente, a Giacomo che rassegnato mi aspetta a letto mentre faccio le "ore piccole" davanti al computer…

…un sincero GRAZIE!