

Indice

Introduzione	5
Sommario	7
Capitolo 1	
Il trauma grave	8
1.1 Considerazioni generali sul trauma grave	8
1.2 Considerazioni storiche: lo stile Utstein per gli arresti cardiocircolatori	9
1.3 Epidemiologia del trauma grave	10
1.4 La catena di soccorso	10
1.4.1 L'approccio preospedaliero	12
1.4.2 L'approccio ospedaliero	13
1.5 La costituzione di un Registro Traumi (RT)	13
Capitolo 2	
RITG : dettagli e specifiche.....	16
2.1 Linee guida per il registro traumi	16
2.2 Il sistema a punteggio.....	17
2.2.1 Panoramica sui parametri del RITG	18
2.2.2 I punteggi: le misure della gravità e degli esiti del trauma	22
2.2.3 Punteggi di gravità fisiologici.....	23
2.3 Criteri di inserimento.....	23
2.3.1 Criteri di Inclusione: l'utilizzo del NISS	23
2.3.2 Criteri di esclusione	24
2.4 L'importanza dei dati per la creazione di un predittore	24

Capitolo 3

Trasfusioni Massive	26
Introduzione.....	26
3.1 Trasfusioni nei traumi.....	26
3.1.1 Emorragie critiche e trasfusioni massive.....	27
3.1.2 Complicazioni legate alle trasfusioni.....	28
3.2 Parametri e punteggi per valutare la Massive Transfusion	31
3.3 Le trasfusioni massive sul campo di battaglia	34

Capitolo 4

Data mining in medicina clinica	36
Introduzione	36
4.1 Il Data Mining: considerazioni generali	36
4.2 Il Data Mining Predittivo.....	37
4.2.1 Principali Tecniche e Metodologie utilizzate nel data mining predittivo	37
4.3 I principali ambiti di utilizzo del data mining	42
4.3.1 Data mining e statistica	42
4.3.2 Data mining in Medicina: l'esperienza in genomica.....	43
4.4 Data Mining in medicina clinica.....	44
4.4.1 L'utilizzo dell'informazione pregressa o a priori	45
4.4.2 Tecniche del data mining clinico	45
4.4.3 Valutazione dei modelli predittivi: metriche e metodi di paragone	46
4.4.4 Implementazione del Data Mining e possibili evoluzioni	47
4.4.5 Data mining predittivo per l'analisi di pazienti con traumi gravi	48

Capitolo 5

Alberi di Decisione	50
Introduzione	50
5.1 Definizione.....	50

5.2 Costruzione dell'albero di decisione.....	51
5.2.1 L'algoritmo Divide and Conquer.....	52
5.2.3 La "potatura" dell'albero:il pruning	53
5.3 Il software Decision Tree : C5.0	55
5.3.1L'algoritmo: da ID 3 a C4.5 a C5.0	55
5.3.2 L'utilizzo di C5.0.....	58
5.3.3 I file utilizzati da C5.0	58
5.3.4 L'output di C5.0.....	59
5.3.5 Le opzioni	59
5.3.6 Utilizzi e limitazioni di C5.0.....	60

Capitolo 6

Analisi Sperimentale	61
Introduzione	61
6.1 L'Ospedale dell'Angelo di Mestre.....	61
6.1.2 Il pronto soccorso.....	61
6.2 La gestione dell'RITG a Mestre	62
6.3 Analisi Dataset	63
6.3.1 Obiettivi dell'analisi.....	64
6.3.2 Le tipologie di dato nel dataset.....	64
6.4 Calcolo TASH per la stima della probabilità di trasfusioni massive.	66
6.5 Costruzione di un predittore di Trasfusioni Massive con gli Alberi di Decisione	66
6.5.1 L'utilizzo di C5.0 : i test sul dataset proveniente dal RITG.....	67
6.5.2 Analisi dei task effettuati	68
6.5.3 L'algoritmo di boosting	70
6.5.4 La costruzione del Modello con gli attributi più predittivi	71
6.5.5 La costruzione del Modello con tutti gli attributi presenti nel registro	73
6.6 Pre-selezione Dati.....	73
6.6.1 L'analisi di Correlazione.....	73

6.6.2 Costruzione e ordinamento del Grafo disconnesso	76
6.7 Risultati finali	778
6.8 Conclusioni e sviluppi	81
Bibliografia	84

Introduzione

L'implementazione di modelli data mining in ambito clinico è una procedura che negli ultimi anni è stata sempre di più considerata nel supporto decisionale in diagnostica. In questo lavoro, l'analisi viene presentata in un contesto di emergenza, per la gestione del paziente traumatizzato grave nella predizione della necessità di terapia trasfusionale massiva.

Nei pazienti traumatizzati gravi il rischio di complicazioni legati a coagulopatie è frequente e non sempre facilmente individuabile, e richiede come soluzione terapeutica la trasfusione massiva o MT.

È necessario individuare il prima possibile e con certezza quali siano i pazienti che necessitino una trasfusione massiva, per allertare la banca del sangue in modo che le emazie siano disponibili il prima possibile a fronte di una situazione particolarmente critica di sopravvivenza e con possibili continui mutamenti, in cui la celerità con cui si effettua la terapia trasfusionale è determinante.

In questa tesi si sono studiati i diversi parametri contenuti in un dataset proveniente dal Registro Intraospedaliero multi-regionale Traumi Gravi al fine di arrivare ad un indice predittivo della trasfusione massiva utilizzando la tecnica degli alberi di decisione.

Il lavoro svolto ha portato alla costruzione di un modello predittivo, paragonabile prestazionalmente all'indice più utilizzato in pratica clinica.

Nel primo capitolo della tesi viene descritto l'ambito dell'analisi facendo una panoramica sulla gestione del trauma grave da parte delle strutture ospedaliere e delle principali problematiche legate alla gestione di pazienti in pronto soccorso.

Nel secondo capitolo si discute invece del registro traumi, lo strumento da cui sono stati estrapolati i dati analizzati in questo lavoro, esaminando standard di compilazione e tipo di dati registrati.

Nel terzo capitolo si analizza la problematica delle trasfusioni massive, esaminando prima la terapia trasfusionale in senso più generale e poi più specificatamente gli indici predittivi della trasfusione massiva e i contesti in cui essa viene applicata.

Nel quarto capitolo vengono introdotte le tecniche e metodologie utilizzate nello studio del predittore, facendo una panoramica sul data mining sia in senso generale che applicato alla medicina.

Nel quinto capitolo viene descritta la metodologia degli alberi di decisione, analizzando in particolare l'algoritmo C5.0 che verrà utilizzato per implementare il predittore.

Nel sesto capitolo si discutono le scelte dei vari dataset per la costruzione del modello e si analizzano i risultati delle implementazioni del predittore confrontando gli indici creati in un set di validazione con l'indice ottenuto attraverso il TASH, il più accurato indice di predizione MT in ambito clinico.

Sommario

Per trasfusione massiva si intende l'apporto di 10 o più unità di sangue intero e nell'ambito della terapia d'emergenza per i traumatizzati gravi; questa procedura rappresenta una soluzione per il recupero del paziente con emorragie critiche e gravi coagulopatie. Un terzo dei pazienti traumatizzati gravemente, muore a causa del dissanguamento: si è riscontrato in diversi casi che un intervento celere e specifico nel ristabilire le quantità ematiche è in grado di aiutare la rianimazione ed il recupero del paziente[13]. Come in ogni aspetto delle terapie nei reparti d'emergenza l'individuazione delle strategie di intervento più corrette deve essere fatto in maniera rapida ma accurata. Per cercare di prevedere il rischio di trasfusioni massive in pratica clinica si utilizzano degli indici basati su parametri fisiologici misurabili nel paziente nella prima ora di ingresso in ospedale.

Utilizzando un dataset proveniente dal Registro Intraospedaliero multiregionale Traumi Gravi dell'ospedale di Mestre è possibile utilizzare tecniche di machine learning e data mining per implementare un modello di predizione di trasfusioni massive e confrontare il predittore ottenuto con gli indici utilizzati maggiormente nella pratica clinica.

Il predittore con tecniche data mining è stato creato utilizzando un software per l'implementazione di alberi decisionali, utilizzando vari training set con parametri diversi allo scopo di trovare il classificatore più accurato. Dalle analisi svolte si è ottenuto un predittore comparabile prestazionalmente con il più usato indice di previsione MT in pratica clinica, in grado di individuare con la stessa accuratezza i casi di trasfusioni massive analizzate in un set di validazione .

Capitolo 1

Il trauma grave

1.1 Considerazioni generali sul trauma grave

Un **trauma grave** è il prodotto di un evento in grado di determinare lesioni mono- o multi-distrettuali tali da configurare un rischio immediato o potenziale di morte e/o di produrre severe inabilità immediate e future al paziente.

Già in questa affermazione si notano due momenti ben distinti: uno, immediato, sulla scena del trauma e l'altro, a distanza, in ambienti dedicati, solitamente strutture ospedaliere, ove avviene l'iter diagnostico terapeutico, differenziando una fase preospedaliera ed una ospedaliera.

L'obiettivo di ridurre la mortalità e la morbidità da trauma può essere raggiunto esclusivamente attraverso un complesso coordinamento di azioni, che vanno dalla fase di chiamata al centralino del Pronto Soccorso, all'invio dell'equipe di soccorso, fino alla fase di trattamento e riabilitazione, in cui ogni singolo anello della catena del percorso è fondamentale per garantire la migliore qualità del processo di cure al traumatizzato, osservando che un ottimo sistema intraospedaliero porterà pochi benefici, se i soccorsi preospedalieri sono scadenti.

L'organizzazione del traumatizzato grave è un problema di salute pubblica e prevede quindi una centralizzazione primaria o iniziale che gestisce la questione dal luogo dell'evento, in un tempo preospedaliero, verso il centro ospedaliero di accoglimento, organizzato e dedicato, ove si dà il via ad una gestione complessa e ad un iter diagnostico-terapeutico peculiare, con approccio in team multidisciplinare, affidando la gestione clinica ad un responsabile di riferimento, il tutto da affrontarsi in un tempo contenuto, la ben nota *golden hour.*, ossia quell'ora in cui se i pazienti gravemente feriti riescono a raggiungere la sala operatoria hanno una migliore prognosi, e quindi una più alta probabilità di sopravvivenza. Questa fase non sempre è immediata, perché si può

inserire un tempo intermedio presso una struttura ospedaliera secondaria ove avviene l'inquadramento del paziente, seguito poi dal trasporto mirato presso il centro dedicato.

Il tentativo di rendere il tutto più armonico e ponderato ha portato in successione nel tempo ad una serie di strumenti utili per coordinare i soccorsi e rendere il personale ed i mezzi in grado di affrontare le più diverse situazioni. Inizialmente si diede valore ai singoli operatori e alle singole realtà locali, ma questi sforzi non sempre miglioravano la qualità globale del servizio. Con queste esperienze, prima di fare investimenti inconcludenti, era necessario conoscere lo stato dell'arte in quel momento. L'organizzazione di tutto il sistema doveva essere testata per trovare i punti deboli e in quali settori agire per ottenere i miglioramenti, come confrontarsi con altre realtà adiacenti o distanti e quali erano i dati su cui basare il confronto.

1.2 Considerazioni storiche: lo stile Utstein per gli arresti cardiocircolatori

Nella gestione della cura degli arresti cardio circolatori si era vista una crescente collaborazione tra pronto soccorso e rianimazione, che ha portato nel tempo ad una migliore gestione delle risorse, sfruttando bagagli dati esistenti in ambito locale e sovra nazionale. All'inizio degli anni '90, vi era un prevalente interesse di tentare di impedire un arresto cardio circolatorio (ACR) facendo crescere una diffusa conoscenza dei sintomi prodromici e promuovendo piani di formazione delle Rianimazioni Cardio Polmonari (RCP) in ambito internazionale. Nell'isola di Mosteroy in Norvegia si incontrarono membri dell'European Resuscitation Council, dell'American Heart Association, dell'Australian Resuscitation Council, del Committees of the Heart and Stroke Canadese, del Resuscitation Council del Sud Africa e numerose altre rappresentanze di molti altri paesi per definire delle regole generali sulla raccolta dati delle unità RCP, effettuate in ambiente extraospedaliero. Questo complesso di regole prese il nome di Utstein Style dal nome dell'abbazia dove i invitati si riunirono per la prima volta nel Giugno del 1991.

Fu il primo passo, che stabilì gli standard oggi universalmente accettati non solo per la raccolta dati in ambiente extraospedaliero, ma anche, in seguito, per gli ACR occorsi in Ospedale. Fu definita una flow-chart a partire dalla quale era possibile strutturare una metodica di registrazione dei dati per la creazione di un database per un'accurata valutazione epidemiologica degli episodi ACR e delle risposte delle RCP.

Il database doveva raccogliere dati significativi relativi a tutti i tempi dall'inizio dell'intervento RCP, registrando le varie manovre terapeutiche, quali la prima defibrillazione, l'intubazione o la prima somministrazione di farmaco. Il *call response time* era stato registrato da tutte le realtà che ebbero intenzione di confrontarsi sul tema, e quindi fu ritenuto elemento fondamentale in questa raccolta dati, con una attenta valutazione degli esiti del trattamento,

segnando lo spontaneo ritorno alla normale circolazione, i tempi del ricovero, la mortalità a distanza di 24 ore, di tre mesi, di sei mesi e di un anno con la registrazione e valutazione degli esiti invalidanti possibili da eventi secondari intercorsi.

La organizzazione del registro degli episodi ACR così strutturata venne considerata come la pietra miliare da cui si può sviluppare qualsiasi ulteriore approfondimento e miglioramento nella catena assistenziale.

Appariva quindi fattibile una raccolta dati nell'ambito della traumatologia grave, utile per migliorare l'organizzazione delle strutture ad essa orientate, ma le variabili da introdurre diventavano sempre più numerose e spesso extra ospedaliere, rendendosi necessario sviscerare le dinamiche extraospedaliere e quelle all'interno delle strutture nosocomiali.

1.3 Epidemiologia del trauma grave

Se i prodromi di un evento cardiocircolatorio possono essere previsti, e i dati di Utstein avevano una loro valenza, nel caso di un trauma grave la fatalità appare spesso un elemento di peso. Tuttavia, dalle sempre più numerose esperienze, la probabilità che questi eventi accadano, appare più elevata se vi sono certe circostanze. Così negli incidenti stradali si è visto che vi sono strade più a rischio e tempi in cui vi è una frequenza maggiore nell'arco di un anno o della settimana o addirittura in un certo periodo del giorno in fasce orarie ben precise. Analogamente anche il luogo ove si è verificato l'evento assume un ruolo determinante, sia nel modo di operare, sia nella via da seguire e nella scelta del mezzo più idoneo alla prestazione di soccorso e quali vie percorrere, considerando che alla base di tutto rimane la rapidità per accedere alla scena dell'evento.

1.4 La catena di soccorso

Lo scopo principale per ridurre la mortalità e le conseguenze del trauma grave è quello di creare un complesso coordinato di azioni, che vanno dalla fase di allertamento del Sistema e all'invio delle equipe di soccorso, fino al trattamento definitivo: una vera catena in cui ogni singolo anello condiziona la solidità dell'intero insieme. Infatti, un ottimo ospedale porterà pochi benefici, se i soccorsi preospedalieri sono scadenti, con feriti non adeguatamente preparati, magari in ipossia, ipotesi e privi di una corretta immobilizzazione. Viceversa gli sforzi di soccorritori esperti e capaci sono vanificati da terapie ospedaliere non idonee.

La catena del soccorso del paziente traumatizzato è più complessa rispetto alla catena creata per gli arresti cardiocircolatori e presa a riferimento. La possibile presenza di più feriti e la difficoltà di ottenere informazioni cliniche attendibili comportano problemi rilevanti nei meccanismi di allarme del sistema di emergenza e nell'invio dei mezzi, mentre la pluralità delle lesioni riscontrate

rende più difficile la standardizzazione degli interventi sul terreno e dei criteri di indirizzamento dei pazienti.

Per trasferire i dati in modo schematico in un'analisi della situazione dell'emergenza, la catena di soccorso viene suddivisa in cinque anelli, cioè in cinque specifiche fasi organizzative.

- 1- Allarme e raccolta informazioni - Nei sistemi più avanzati, specifici criteri di interrogazione regolano i meccanismi del processo della chiamata di soccorso da parte degli operatori di centrale operativa, allo scopo di garantire l'invio immediato dell'equipe più adeguata sulla scena di un dato evento traumatico. L'arrivo dell'equipe sulla scena deve essere comunque subordinato alla valutazione della sicurezza ambientale e dei rischi potenziali.
- 2- Addestramento preliminare – Soprattutto nella traumatologia conseguente a incidenti del traffico, la presenza di più feriti è frequente; è pertanto essenziale che le equipe di soccorritori siano addestrate a riconoscere le lesioni che richiedono un intervento prioritario, allo scopo di stabilire quali feriti trattare e/o trasportare per primi. Il successo di questa fase dipende soprattutto dall'affiatamento dell'equipe e dall'anticipazione, ossia da quella fase di briefing preliminare all'arrivo sulla scena, nella quale avviene la preparazione dell'equipe di soccorso, il controllo dei materiali necessari e l'attribuzione dei compiti da parte del team leader, cioè la persona a cui si fa riferimento in queste emergenze e che coordina l'azione.
- 3- Trattamento preospedaliero – La patologia traumatica rappresenta un campo del tutto specifico, nel quale l'utilità di ogni singolo atto terapeutico deve essere valutata in un ideale rapporto fra i benefici derivanti in termini di prevenzione del danno secondario e miglioramento dell'outcome, a fronte dell'allungamento complessivo del tempo di soccorso e dei rischi di complicanze legati all'esecuzione di manovre terapeutiche complesse in condizioni disagiate. Ciascun operatore del soccorso, che si trovi ad agire sulla scena di un trauma, deve pertanto essere a conoscenza di un metodo di valutazione immediato nel trattamento del ferito, che, sulla base delle proprie competenze, gli consenta di ridurre gli errori e di offrire le massime garanzie per un buon esito dell'intervento di soccorso.
- 4- Indirizzamento dei feriti agli ospedali più idonei – Questo anello fondamentale esiste solo nei sistemi avanzati, regolati da precise strategie e normative o dove operino sul terreno equipe di soccorso avanzato in grado di stabilizzare adeguatamente i feriti, indirizzandoli direttamente a ospedali idonei a garantire il trattamento definitivo (strategia di centralizzazione traumi gravi). Si determina pertanto la necessità di una selezione dei feriti attraverso una serie di valutazioni che, come si vedrà, non sono solo di carattere clinico. Se per feriti gravemente instabili può essere opportuno l'invio ad un'idonea struttura

ospedaliera sita alla minima distanza possibile, un corretto indirizzamento presuppone spesso la necessità di allungare i tempi di trasporto del ferito all'ospedale per raggiungere destinazioni più lontane; risulta pertanto fondamentale che la fase di trasporto sia gestita con altrettanta competenza e supportata da adeguato monitoraggio.

- 5- Trattamento ospedaliero – La fase diagnostica e terapeutica intraospedaliera di emergenza riveste un'importanza cruciale nel soccorso al traumatizzato. Altrettanto importante è il collegamento fra la fase preospedaliera e quella ospedaliera, stadio estremamente critico che, nell'ottica di ottimizzare la tempistica, richiede un'adeguata e precoce preparazione della struttura ricevente (attivazione del Trauma team, allertamento della diagnostica, della chirurgia e delle terapie intensive, disponibilità eventuale di sangue universale). Una volta in ospedale è fondamentale un completo ed esaustivo passaggio di consegne fra le équipe.

E' chiaro che i risultati finali possono dipendere da molte variabili introdotte in questo percorso e sono di difficile valutazione in quanto coinvolge più di una struttura specialistica. Quindi queste problematiche non possono essere affrontate se non in modo marginale durante i corsi di aggiornamento sanitario, essendo questi percorsi formativi prevalentemente rivolti al soccorso preospedaliero. La conoscenza di ciò che avverrà dopo, rientra però nel patrimonio culturale dei soccorritori esperti e consente di disporre di strumenti indispensabili per valutare gli aspetti organizzativi del Sistema di Emergenza e contribuisce a migliorarne efficienza ed efficacia.

1.4.1 L'approccio preospedaliero

Da qualche anno in Italia è attivo il numero telefonico 118 per le emergenze su scala nazionale. Questa nuova apertura all'ambiente extraospedaliero e al territorio ha stimolato e consentito allo staff di infermieri e medici di ottenere nuove conoscenze e di occuparsi anche dell'aspetto organizzativo oltre che clinico di una fase dell'emergenza prima molto trascurata .

La valutazione del ferito grave costituisce l'anello fondamentale della catena di soccorso e questo avviene inizialmente al primo contatto in ambiente extra ospedaliero e con i mezzi che si hanno a disposizione, sul luogo dell'evento traumatico. Qui la definizione operativa di trauma grave deve essere rapida anche se in maniera meno precisa, considerando ed utilizzando un'opportuna lista di riscontro (*check list*), correlata con le principali casistiche legate al trauma, che prende in considerazione:

- La dinamica dell'evento definito ad alto rischio di trauma maggiore, come ad esempio la caduta dall'alto maggiore di 5 m, uno scontro automobilistico frontale, una eiezione di passeggero dall'abitacolo;
- La presenza di alterazione di parametri fisiologici. Qui si introduce un sistema a punteggio prendendo una scala specifica come la Revised Trauma Score;

- L'evidenza di lesioni traumatiche con alto rischio di vita e/o inabilità;

È da sottolineare che per i pazienti trovati vivi, la qualità dell'assistenza preospedaliera incide in modo determinante sugli esiti. Gli studi condotti negli ultimi 20 anni hanno documentato che esiste una percentuale elevata di feriti, che muore per cause potenzialmente risolvibili durante le diverse fasi di soccorso: la frequenza di queste morti potenzialmente evitabili varia dal 9 al 43 %.

1.4.2 L'approccio ospedaliero

L'altra fase di grosso peso nel trattamento del trauma grave è l'assistenza ospedaliera. Negli USA, in cui il modello assistenziale è diverso da quello europeo, una prima risposta organizzativa alle esigenze fu la creazione di strutture con elevati standard assistenziali e con un'organizzazione ad hoc dove concentrare questi pazienti, dette Trauma Center. In questo tipo di organizzazione, il ruolo importante è assunto dall'adeguata valutazione preospedaliera in grado di individuare subito i pazienti che necessitino della centralizzazione verso i Trauma Center, con l'adozione formale di procedure, che consentano il bypass degli ospedali più vicini, non adeguatamente attrezzati. Tuttavia la distribuzione territoriale di questi centri, rispetto alla restante rete ospedaliera, deve essere razionale e una loro mancanza può impedire di fatto l'attuazione delle procedure più adeguate e il ricorso all'immediata centralizzazione del traumatizzato.

In Europa e in Italia lo sforzo di organizzazione del soccorso ai traumatizzati è recente ed ha seguito principi diversi scegliendo di sviluppare le potenzialità specifiche di grandi ospedali polispecialistici, integrando gli stessi in un sistema organizzativo, che assicuri la continuità di tutto l'iter di soccorso, diagnosi e trattamento, includendo in questo processo anche la fase di riabilitazione. Analogamente a quanto avviene per i criteri di aggiornamento ospedaliero, si rende indispensabile nell'ambito della rete ospedaliera la condivisione di protocolli di accordo inter ed intra ospedaliero, che permettano di affrontare i traumi maggiori con procedure veloci, che riducano i tempi del trasferimento secondario.

1.5 La costituzione di un Registro Traumi (RT)

L'importanza di avere a disposizione dati attendibili a scopo di verifica della qualità dell'assistenza e di utilità nella scelta di strategie future in ambito interdisciplinare con coinvolgimento di ambienti extra, inter ed intra ospedalieri ha reso necessaria la stesura di una registrazione degli eventi in specifici data base: l'insieme di banche dati su cui basare i processi di revisione e miglioramento delle cure e della ricerca scientifica. Pertanto lo strumento insostituibile per garantire la corretta funzionalità di un Sistema Integrato di Soccorso viene considerata la creazione di un Registro Traumi.

Già nel 1966 negli Stati Uniti, col riscontro di un aumento degli episodi traumatici gravi, definito un'epidemia tragica e trascurata, l'Accademia Americana delle Scienze indicò le supposte

priorità per ridurre mortalità ed esiti invalidanti^[1]. Tra le indicazioni fu inserita l'istituzione dei Registri Ospedalieri dei Traumi, che avrebbero contribuito poi allo sviluppo del Trauma System, un approccio che ha permesso di abbassare drasticamente la media della mortalità e morbidità da trauma.

Queste registrazioni diffuse nel Nord America, con banche dati estese, in Europa sono utilizzate poco e limitatamente. I criteri di inclusione dei dati differiscono tra loro e sono disomogenei, talora includendo solo i pazienti ospedalizzati.

Le informazioni da raccogliere riguardano molti aspetti dell'evento tra cui la dinamica del trauma, la gravità delle lesioni anatomiche suddivise e catalogate secondo tabelle ed elenchi prestabiliti, i parametri fisiologici, le terapie eseguite o in atto, il trattamento praticato e gli esiti. Ovviamente l'estensione delle informazioni è variabile, per cui il tentativo di categorizzare si è reso necessario per una prima semplificazione delle indagini, proponendo una ulteriore suddivisione tra traumi chiusi e penetranti, e registrando la meccanica che ha portato alla lesione, proponendo differenze tra traumi da incidente stradale e quelli da caduta e segnalando l'eventuale intenzionalità dell'evento cioè se è avvenuto in modo accidentale, per autolesione o a seguito di un'aggressione, anche per fini medico legali.

In Italia l'esperienza dei registri è ancora agli albori. La regione Emilia Romagna ha sviluppato un registro traumi su base regionale, in fase di avvio. Nel 2004 il Ministero della Salute ha finanziato un progetto per la realizzazione di un primo registro multicentrico e multi regionale. Il progetto ha coinvolto tre grandi ospedali di aree diverse del paese ed ha consentito di raccogliere informazioni complete sulla gestione dei traumatizzati, riferite ad oltre 800 pazienti, anche se la raccolta dati è limitata ai pazienti giunti in ospedale e pertanto esclude le informazioni sui deceduti nel luogo dell'evento. Questo bypass impedisce un'analisi globale della catena di soccorso e limita la possibilità di verificare ed identificare tutti gli anelli nel percorso. Buoni ospedali possono essere inseriti in un sistema che non garantisce la qualità del soccorso preospedaliero e il corretto invio dei pazienti gravi ai centri in grado di trattarli adeguatamente. In questi casi i dati derivati dai registri ospedalieri potrebbero non evidenziare affatto l'esistenza di un problema, portando a ritenere che esista un buon livello di assistenza in un'area in cui la mortalità complessiva è invece elevata.

E' stato anche proposto di eseguire studi sulla popolazione per avere una migliore conoscenza dell'epidemiologia del trauma, registrando in un dato periodo tutti gli eventi traumatici accaduti entro un'area predefinita, ma sono studi difficili da realizzare perché estendono la ricerca dai sistemi di soccorso in tutti gli ospedali di una data area alle forze dell'ordine. Per questo motivo sono stati finora realizzati ben pochi studi di questo tipo. Gli studi europei pubblicati, benché realizzati in paesi diversi, dimostrano una sostanziale omogeneità nell'epidemiologia del trauma

grave a livello europeo, con una forte prevalenza di lesioni non penetranti. Gli incidenti stradali sono la causa del maggior numero di traumi gravi, seguito da caduti dall'alto. E' un problema di salute molto ampio e sentito nel nostro paese in cui gli incidenti traumatici gravi specialmente quelli da incidente stradale coinvolgono più frequentemente la popolazione giovane, ossia coloro che costituiscono la fascia di popolazione con la più alta aspettativa di vita e produttività.

Nel 2005 è stato ufficialmente proposto di dare il via ad un registro intraospedaliero multi regionale dei traumi, presentandolo sulla rivista *ACTA Anaesthesiologica Italica*. L'obiettivo è sempre quello di migliorare l'assistenza e l'efficacia dell'intervento eseguito in emergenza, valutandolo dai suoi prodromi alla conclusione.

Dai dati fin qui raccolti, in questi primi anni di sviluppo e di progressiva diffusione, il Registro Traumi Italiano sembra essere ben confrontabile con i riferimenti disponibili in letteratura. E' stato registrato finora un numero sostanziale di pazienti, con una buona affluenza anche dei dati pre-ospedalieri forniti, anche se alcune strutture non sono state in grado di fornire questo tipo di informazioni. Comunque negli ospedali partecipanti esiste una sostanziale omogeneità nelle caratteristiche generali dei dati raccolti sui pazienti registrati e questo lascia ben sperare che i Registri Traumi Italiani possono essere di riferimento per lo sviluppo e miglioramento dell'intervento e assistenza post-traumatica.

Capitolo 2

RITG : dettagli e specifiche

2.1 Linee guida per il registro traumi

Il Registro Traumi è l'elemento chiave per qualsiasi approccio in ambito di salute pubblica per la gestione delle gravi lesioni traumatologiche. Le nazioni europee affrontano le proprie difficoltà sfruttando la completezza e la qualità dei dati del registro, finanziandone da subito la gestione informatica.

Il valore di un dataset infatti accresce la sua potenzialità con l'aumentare del numero di dati registrati: un registro paneuropeo potrebbe permettere confronti più ampi tra le varie metodologie di trattamento di pazienti traumatizzati, tra i programmi di ricerca clinica, orientandone il finanziamento su larga scala.

A questo scopo la standardizzazione globale diventa una priorità alla luce dello sviluppo di un registro traumi internazionale, che permetta il confronto e l'omogeneità dei risultati tra le varie nazioni coinvolte: si tratta quindi di ripercorrere la strada della registrazione degli episodi di Arresto Cardio Circolatorio, che, attraverso lo stile Utstein, aveva trovato un accordo nella compilazione di un registro, che negli anni è diventato strumento fondamentale per le Rianimazioni Cardiopolmonari e per gli organismi di ricerca.

I primi segni di interesse nella creazione di un protocollo standard si hanno nel 1994, quando l'International Trauma Anesthesia and Critical Care Society (ITACCS) ha avviato una task force internazionale allo scopo di elaborare le linee guida per la segnalazione uniforme dei dati relativi ai traumi gravi, basate sullo stile Utstein in ambito cardiocircolatorio.

Questo tipo di linee guida da subito ha avuto grosso riscontro e grosso seguito da parte dei principali Trauma System internazionali desiderosi di trovare un strumento comune di valutazione e confronto nella gestione dei traumi ed anche nella valutazione della loro casistica, sebbene

richiedessero un notevole dispendio di risorse, necessarie all'implementazione del registro Traumi, e non si adeguassero alle possibilità dei Trauma System meno evoluti.

La scelta e la costruzione di linee guida condivise possono semplificare e ottimizzare l'utilizzo di queste risorse al fine di registrare dati che possano influenzare meglio e in maniera più accurata previsioni degli outcomes dei pazienti allorché venga utilizzata una determinata struttura ospedaliera.

Nel 1999 il gruppo ITACCS ^[35] pubblicò in via definitiva le linee guida basate sul modello di Utstein al fine di avere una segnalazione uniforme dei dati in seguito a traumi gravi. Il modello registra i dati della fase pre-ospedaliera, della prima fase ospedaliera e i dati sulla co-morbilità e sull'outcome clinico. In quell'occasione venne deciso che si debbano separare e catalogare i "dati chiave" relativamente ai campi obbligatori essenziali o "dati opzionali" riferendosi a campi supplementari. Nonostante l'iniziale buon riscontro da parte delle Organizzazioni Sanitarie, che vedevano realizzato uno strumento di grande semplificazione per gli studi sul trauma, in letteratura sono scarsi gli autori che hanno prestato attenzione a queste linee guida. Questo indicava la necessità di un ulteriore sviluppo dello stile Utstein ed in particolare mirato alla importante riduzione dei dati considerati "Chiave", ossia dei campi obbligatori, e all'aggiunta di definizioni più precise di questo tipo di dati, al fine di rendere meno pesante la compilazione.

Ma anche queste modifiche non sono state coronate da successo: sebbene si siano prodotti sforzi significativi da parte delle organizzazioni sanitarie europee nell'utilizzo di registri con le nuove linee guida, nessuno studio sulla cura del trauma e sull'outcome è stato fatto a livello continentale in maniera sistematica, principalmente perché non erano stati definiti i criteri di inclusione ed esclusione dei dati in maniera chiara e perché, data la varietà dei registri traumi coinvolti, le definizioni dei dati e i casi dei singoli pazienti risultavano poco comparabili.

Le linee guida quindi non hanno tentato di indicare un singolo strumento a tutti i membri, ma piuttosto hanno cercato di concordare un set di database comune a tutti, che può essere costruito per rispondere alle esigenze specifiche dei singoli stati.

Sulla base delle caratteristiche epidemiologiche del fenomeno trauma e dell'analisi della distribuzione della mortalità nelle varie fasi del soccorso sono stati sviluppati modelli organizzativi volti alla prevenzione degli errori e alla razionalizzazione dell'intero percorso clinico-assistenziale.

2.2 Il sistema a punteggio

I successivi sforzi dei gruppi che stabilirono le linee guida, puntarono quindi a stabilire criteri standardizzati di inclusione ed esclusione, una lista minima di dati chiave con precise definizioni e

una metodica robusta e consistente nel dare un punteggio ai vari scenari al fine di determinare uno score di valutazione.

A questo scopo e al fine di raggiungere un accordo europeo le varie organizzazioni deputate SCANTEM, TARN, DGU-TR e il Registro Intraospedaliero multiregionale dei Traumi Gravi (RITG) si riunirono nel 2007 per un simposio proprio all’Utstein Abbey in Norvegia per redigere “lo stile Utstein per un Report uniformato dei dati realtivi ai traumi gravi” ^[36]. In questa occasione esperti selezionati decisero criteri standard di inclusione ed esclusione dati, e definirono precisamente il gruppo di parametri chiave da memorizzare nel registro. Si cercò poi di creare un registro che fosse il più possibile compatibile con i principali registri Trauma in Europa e che aderisse al programme EuroTARN, per lo sviluppo di un Registro Traumi Europeo per promuovere lo sviluppo di un modello continentale per la predizione degli outcome e permettere un monitoraggio internazionale dei traumi gravi.

Nel 2008 alcuni autori ^[5] hanno effettuato un’ulteriore revisione dei criteri e dei parametri per la registrazione dei traumi. Tale revisione ha portato a linee guida che attualmente vengono seguite nella compilazione del registro interregionale per i traumi gravi anche per l'Italia.

Nella tabella sottostante vengono mostrati gli attributi principali registrati nel RITG secondo le linee guida delle più recenti revisioni.

2.2.1 Panoramica sui parametri del RITG:

Nella tabella sottostante viene descritto da sinistra a destra il nome del parametro; il tipo di dato ad esso associato ossia continuo se verranno registrati numeri reali, nominale se indicano un attributo categorizzato, ordinale se indicano un attributo appartenente ad un insieme finito discreto di numeri; le categorie ossia l’insieme di valori che possono appartenere ad un attributo, nel caso di numeri reali sarà indicato numero genericamente ; ed infine la definizione del tipo di parametro con la descrizione

Nome parametro	Tipo dato	Categorie del valore	Definizione del parametro
Età'	Continuo	Numeri	L'età del paziente
Genere	Nominale	1=femmina 2=maschio 3=sconosciuto	Il sesso del paziente
Tipo di Infortunio	Nominale	1=chiuso 2=penetrante 3=sconosciuto	Indica il tipo di infortunio prodotto dal trauma

Meccanismo di Infortunio	Nominale	1 =traffico, veicolo a motore (non 2 ruote) 2 =traffico motociclo 3 =traffico bicicletta 4 =traffico pedone 5 =traffico altro 6 =sparatorie 7 =colpito da lame varie 8 =colpito da corpo contundente generico 9 =caduta alta energia 10=-caduta bassa energia 11=altro 12=sconosciuto	Il meccanismo o fattore esterno che ha causato l'infortunio
Intenzionalità dell'infortunio	Nominale	1=accidentale 2=autoindotta 3=inferta da terzi (anche se sospettata e non provata) 4=altro 5=	Informazione sul ruolo intenzionale nell'occorrenza dell'infortunio primariamente determinato dall'incidente e non dal infortunio
Classificazione ASA-PS Pre-infortunio	Ordinale	1=Paziente Sano 2=Paziente con disturbi sistemici medi 3=Paziente con disturbi medici severi 4=Paziente con disturbi medici severi trattati a vita 5=Paziente moribondo non tollerante un'operazione 6=Paziente dichiarato morto con organi espianabili. 7=Sconosciuta	La comorbidità pre infortunio esistente prima dell'incidente
Arresto cardiaco prima dell'ospedale	Nominale	1 = No 2 = Sì 3 = Sconosciuto	Il paziente ha avuto un episodio i arresto cardiaco prima dell'ingresso in ospedale
Glasgow Coma Score all'arrivo del personale di emergenza sulla scena	Ordinale	Numero	Prima registrazione pre intervento del GCS all'arrivo sulla scena del personale medico addestrato all'assistenza

Componente motoria del GCS all'arrivo del personale di emergenza sulla scena	Ordinale	6 = Esegue gli ordini /risposta attesa al dolore 5 = Localizza il dolore 4 = Si ritira al dolore 3 = Flessione al dolore (decorticazione) 2 = Estensione al dolore (decerebrazione) 1 = Nessuna risposta motoria	Prima registrazione pre intervento del GCS motorio all'arrivo sulla scena del personale medico addetto all'assistenza
Glasgow Coma Score all'arrivo in Emergenza o in ospedale	Ordinale	Numero	Prima registrazione del GCS in ospedale
Componente motoria del GCS all'arrivo in Emergenza o in ospedale	Ordinale	6 =Esegue gli ordini /risposta attesa al dolore 5 = Localizza il dolore 4 = Si ritira al dolore 3 = Flessione al dolore (decorticazione) 2 = Estensione al dolore (decerebrazione) 1 = Nessuna risposta motoria	Prima registrazione del GCS motorio in ospedale
Pressione sistolica(SBP) all'arrivo del personale di emergenza sulla scena	Continuo	Numero	Prima SBP registrata all'arrivo sulla scena del personale medico
Pressione sistolica(SBP) - Categoria clinica - all'arrivo del personale di emergenza sulla scena	Ordinale	RTS 4 = >89 (polso radiale buono) RTS 3 = 76–89 (polso radiale debole) RTS 2 = 50–75 (polso femorale) RTS 1 = 1–49 (solo polso carotideo) RTS 0 = 0 (no polso carotideo)	Prima SBP registrata all'arrivo sulla scena del personale medico
Pressione sistolica(SBP) all'arrivo in Emergenza o in ospedale	Continuo	Numero	Prima SBP registrata all'arrivo in ospedale
Pressione sistolica(SBP) - Categoria clinica - all'arrivo in Emergenza o in ospedale	Ordinale	RTS 4 = >89 (polso radiale buono) RTS 3 = 76–89 (polso radiale debole) RTS 2 = 50–75 (polso femorale) RTS 1 = 1–49 (solo polso carotideo) RTS 0 = 0 (no polso carotideo)	Prima SBP registrata all'arrivo in ospedale
Frequenza respiratoria all'arrivo del personale di emergenza sulla scena a	Continuo	Numero	Prima Frequenza respiratoria registrata all'arrivo sulla scena del personale medico addetto all'assistenza

Frequenza respiratoria (categoria) all'arrivo del personale di emergenza sulla scena	Ordinale	RTS 4 = 10–29 (normale) RTS 3 = >29 (veloce) RTS 2 = 6–9 (lento) RTS 1 = 1–5 (affannoso) RTS 0 = 0 (no respiro)	Prima Frequenza respiratoria registrata all'arrivo sulla scena del personale medico addetto all'assistenza
Frequenza respiratoria all'arrivo in Emergenza o in ospedale	Continuo	Numero	Prima Frequenza respiratoria registrata all'arrivo in ospedale
Frequenza respiratoria(categoria) all'arrivo in Emergenza o in ospedale	Ordinale	RTS 4 = 10–29 (normale) RTS 3 = >29 (veloce) RTS 2 = 6–9 (lento) RTS 1 = 1–5 (affannoso) RTS 0 = 0 (no respiro)	Prima Frequenza respiratoria registrata all'arrivo in ospedale
Base Excess Arteriale	Continuo	Numero	Prima misura del Base Excess arteriale all'arrivo in ospedale
INR per coagulazione	Continuo	Numero	Prima misura di INR nella prime ore dopo l'arrivo in ospedale
Numero di giorni in ventilazione	Continuo	Numero	Il numero totale dei giorni in cui il paziente è rimasto collegato ad un respiratore meccanico
Durata della permanenza nel principale Ospedale in cui è stato ricoverato	Continuo	Numero	Calcolato dalla Data di dimissione meno la data di ingresso dal report ospedaliero
Destinazione a seguito delle dimissioni	Nominale	1 = Casa 2 = Re-Abilitazione 3 = Obitorio 4 = Altra Rianimazione, trattamento a più alto livello 5 = Altro reparto ospedaliero a minore livello assistenza 6 = Altro 7 = Sconosciuto	La destinazione del paziente a seguito del trattamento in Rianimazione

Glasgow Outcome Scale – alle dimissioni dall’ospedale	Ordinale	5 = Buon Recupero 4 = Moderata Disabilità: disabile ma indipendente 3 = Severa Disabilità: cosciente ma disabile; dipende da altri 2 = Persistente stato vegetativo: non risponde agli stimoli 1 = Morto 0 = Sconosciuto	Punteggio Glasgow Outcome Scale – alle dimissioni dall’ospedale
Stato di sopravvivenza	Nominale	1 = Deceduto 2 = Sopravvissuto 3 = Sconosciuto	Vivo o morto dopo 30 giorni dall’infortunio
Abbreviated Injury Scale (AIS)	Ordinale	Numero	Vedi sezione sotto

2.2.2 I punteggi: le misure della gravità e degli esiti del trauma

Oltre agli indici appena descritti, il registri solitamente riportano anche indici legati alla misura di gravità del trauma e della gravità quindi del paziente stesso. Avere dei riscontri oggettivi e statistici è quindi fondamentale per predisporre strumenti di valutazione il più possibile accurati e precisi.

Come visto dalla precedente tabella la scala di gravità anatomica utilizzata è la *Abbreviated Injury Scale* più conosciuta come AIS. Tale scala prevede un punteggio per tutte le lesioni possibili per ogni regione del corpo. I punteggi vanno da 1 (lesione lieve) a 6 (lesione incompatibile con la vita) e sono classificati secondo un manuale che contiene un elenco di lesioni con più di 1200 voci ordinate per regioni del corpo.

Una volta definito il trauma utilizzando le codifiche AIS, si calcola il punteggio ISS (Injury Severity Score), elevando al quadrato i tre punteggi AIS più alti in sei determinate regioni del corpo e sommandoli tra di loro. I valori dell’ISS sono compresi tra 0 e 75. Per convenzione quando è presente anche solo una lesione con valore AIS uguale a 6 il punteggio ISS diventa subito 75. IL calcolo del punteggio ISS è fondamentale perché permette di quantificare la gravità complessiva di un paziente traumatizzato. Anche la definizione internazionale standard di trauma grave è basata su questa scala e corrisponde ad una valore superiore a 15. Infatti la definizione di “trauma grave” è un fatto relativamente recente: nel 1999 le *“Recommendations for uniform reporting of data following major trauma”* codificano un traumatizzato grave il paziente in cui le lesioni nel complesso raggiungono il punteggio ISS > 15, con un rischio di mortalità compresa tra il 19% ed il 40%, mentre sotto questa soglia, la mortalità diventa molto bassa, salvo che per le età estreme. Questo valore tuttavia è solo di riferimento, in quanto si basa su dati talora meno raffinati, includendo nella registrazione statistica i decessi avvenuti sia in ospedale che in fase preospedaliera.

Anche se le scale AIS e ISS siano finora le più utilizzate, entrambe hanno inevitabilmente delle imperfezioni. Per questo sono stati proposti sistemi come l'*Anatomic Profile* o NISS come alternativa ai precedenti, ma finora nessuna scala si è dimostrata sicuramente la migliore.

La codifica degli stati morbosi più comunemente utilizzata è quella consigliata dell'*International Classification of Disease* (ICD), utilizzata dal nostro sistema sanitario nazionale con nomenclatore all'uopo.

Esistono sistemi per convertire questa classificazione in quella AIS/ISS con i relativi punteggi. Con la conversione si ha certo una perdita di accuratezza e di precisione che diventano accettabili solo quando la codifica delle diagnosi ICD è puntigliosa, con segnalazione del numero complessivo delle diagnosi riportate, che dovrebbe corrispondere a quello delle singole lesioni di un paziente, cosa che non sempre viene applicata e purtroppo succede raramente.

2.2.3 Punteggi di gravità fisiologici

Nel database italiano dell'RITG trovano spazio anche indici di gravità fisiologica, orientati alla capacità di un paziente di reagire all'evento traumatico, che varia secondo le sue condizioni cliniche. Queste vengono valutate con il *Revised Trauma Score* (RTS), dando un punteggio, estrapolato da una enorme casistica nord-americana, per capire quali siano i parametri fisiologici che più influenzano la prognosi nei pazienti traumatizzati. Il RTS tiene conto della frequenza respiratoria, della pressione arteriosa sistolica (PAS) e della *Glasgow Coma Score* (GCS). Il parametro rilevato viene categorizzato e attribuito ad un valore, che sarà successivamente moltiplicato per un coefficiente, allo scopo di pesare i parametri vitali in funzione della loro importanza prognostica.

Poiché dopo un trauma, le condizioni di un paziente sono in continua evoluzione, è importante specificare in quale momento questo punteggio venga calcolato, ben differenziandosi tra quello ottenuto sulla scena del trauma o all'arrivo in ospedale. La registrazione dei parametri al momento del primo approccio al paziente, permette di attribuire un punteggio anche quando vengono effettuate procedure invasive preospedaliere, che altrimenti ne impedirebbero la successiva determinazione.

2.3 Criteri di inserimento

2.3.1 Criteri di Inclusione: l'utilizzo del NISS

Il sistema NISS è una modifica del metodo ISS, che considera solo le regioni del corpo più severamente infortunate, e spesso crea problemi nel considerare le peggiori regioni. Il NISS invece è definito come la somma dei più severi infortuni, secondo la scala AIS, senza considerare la regione del corpo. Ultimamente vi è concordanza nel rimpiazzare ISS con il NISS, in quanto è più

semplicemente calcolabile e più predittivo delle probabilità di sopravvivenza. Il NISS infatti apporta un'indicazione di predittività della sopravvivenza per esiti uguali o migliori dell'ISS, per ogni genere di paziente ed in particolare per i pazienti con multiple fratture al capo. Pertanto è stato scelto considerare come traumatizzato grave, il paziente con NISS > 15, invece che ISS > 15 .

Cambiando il criterio di inclusione, vi è stato un aumento del numero dei pazienti considerati gravi: questo dovrebbe essere visto come un aumento della sensibilità del classificatore, piuttosto che una perdita di specificità. Uno sforzo ulteriore potrebbe essere fatto controllando se tutti i pazienti con NISS > 15 siano stati effettivamente inclusi, anche nel caso in cui prima o all'arrivo in ospedale non siano stati coinvolti dai team d'emergenza traumatica, o se non siano stati ricoverati in terapia intensiva.

2.3.2 Criteri di esclusione

Tuttavia anche usando il valore NISS > 15 come singolo criterio di inclusione, si rischia di includere pazienti che hanno alto rischio di rendere l'analisi dei dati più imprecisa e vaga. Per eliminare questo tipo di pazienti dalle analisi dei registri è stato creato un set di criteri di esclusione. Si ritiene utile infatti escludere a) quei pazienti la cui ammissione in ospedale è stata effettuata 24 ore dopo l'infortunio, b) pazienti dichiarati morti prima dell'arrivo in ospedale, c) quelli arrivati in ospedale privi di segni di vita, d) quelli morti al primo tentativo di rianimazione. Inoltre non si considerano traumi i casi di asfissia, affogamento e ustioni, che non sono traumi chiusi o penetranti e debbono essere separati. Talora il confronto tra ustioni e soffocamenti con gli altri generi di trauma è molto difficile, tanto che nel 2005 sono stati inseriti codici AIS a parte relativi a questo tipo di traumi. Anche in Italia esistono reparti per gli ustionati gravi: quindi i pazienti ustionati saranno esclusi dal registro traumi se l'ustione rappresenta l'infortunio predominante, o se un paziente è ricoverato in un reparto per gli ustionati

Le morti preospedaliere sono escluse per ragioni pratiche: solitamente i pazienti dichiarati morti vengono portati direttamente in obitorio, anche se in alcune regioni questi pazienti sono ammessi in ospedale e sono registrati tra i traumi gravi.

I pazienti invece che arrivano in emergenza con situazione circolatoria normale, dovrebbero essere inclusi anche se hanno avuto un arresto cardiocircolatorio temporaneo prima di entrare in emergenza o se sono morti nel reparto.

2.4 L'importanza dei dati per la creazione di un predittore

Le più recenti modifiche apportate alla quantità di dati da memorizzare nel registro traumi hanno principalmente lo scopo di ottenere migliorie nell'ambito di modelli predittivi degli esiti traumatici, per una valutazione di confronto, in linea con gli esiti delle tipologie di traumi più

comuni in Europa.

I modelli predittivi sono composti dall'insieme di pazienti e di variabili ad essi collegate per i traumi gravi che sono considerati importanti per una predizione degli outcome. Essi non sono determinativi ma danno una probabilità di outcome in relazione ad un singolo paziente: parametri complessi, come l'*Abbreviated Injury Scale* (AIS), derivati e soprattutto l'RTS sono spesso usati per creare modelli predittivi.

IL TRISS è l'indice predittivo attualmente più considerato ed è presente anche nel database del registro Traumi Italiano. Esso predice la probabilità di sopravvivenza ed è calcolato con indici pesati su età, *Injury Severity Score* e *Trauma Score* e modalità di infortunio. L'esperienza dei Registri Trauma di Germania e Gran Bretagna suggerisce che ci possano essere delle variabili migliori da includere in un modello predittivo che non il TRISS: in questi studi il modello predittivo si basa su parametri presi all'ingresso in Rianimazione e ne segua l'evoluzione nel paziente non oltre la prima ora di ricovero e soprattutto non considera le comorbidità pregresse del paziente.^{[5],[37]}

Attualmente si prova a considerare più variabili nello studio del modello predittivo nella ricerca di individuare le categorie "tipo" del traumatizzato con l'intento di raggiungere risultati più rilevanti che nel passato e soprattutto di ottenere una stima ottimale della probabilità di sopravvivenza.^[5]

I modelli predittivi sono quindi un utile strumento per la possibilità di prevedere quali possono essere le conseguenze di un incidente a livello di comorbidità, sopravvivenza e durata della degenza nei reparti di rianimazione o di terapia intensiva. In questo modo si otterrebbe una stima dei costi per il sistema sanitario sia in ambito locale che territoriale, nell'ottica in cui il sistema sanitario nazionale retribuisce le aziende ospedaliere attraverso il calcolo dei DRG basato su schemi di costo assicurativo correlato alla patologia. Ne consegue che un registro ordinato e ben compilato focalizzato ad un indice predittivo, porta ad una stima ottimale di quelle che saranno le risorse impiegate dai reparti nel trattamento dei traumatizzati e possa offrire indici per avallare ulteriormente con risultati auspicabili le decisioni cliniche, talvolta estremamente critiche e rischiose.

Capitolo 3

Trasfusioni Massive

Introduzione

Il trauma è la principale causa di morte per persone sotto i 40 anni e le emorragie critiche sono la maggior causa di mortalità nei traumi in ambito civile che militare. La morte a seguito di dissanguamento in genere arriva rapidamente, tipicamente intorno alle prime 6 o 12 ore dopo l'infortunio. Circa il 10 % di tutti i pazienti traumatizzati sono trasfusi con una o più unità di sangue e più del 30% di questi richiede l'intervento di una trasfusione massiva (MT), che è definita come la trasfusione di 10 o più unità di sangue intero nelle prime 24 ore.

Le trasfusioni possono essere necessarie per migliorare la riorganizzazione dei tessuti: diversi studi ^{[6][7][8]} hanno tuttavia mostrato che le trasfusioni di sangue sono associate a basse percentuali di outcome positivo e un aumentato rischio di infezione, edemi polmonari acuti e sindromi da distress polmonare, sospensione della funzionalità organica in più sedi sistemiche e morte.

La capacità di decidere in tempi brevi della necessità in un paziente di effettuare trasfusioni massive è la base per permettere una rapida correzione dei più comuni problemi ematici indotti nei traumi e relativi alla coagulopatia, acidosi e ipotermia. La rapidità permette una valida mobilitazione dalle banche del sangue, anche in luoghi dove le strutture sanitarie sono meno sviluppate, con una preparazione dell'approvvigionamento ematico in tempi brevi.

3.1 Trasfusioni nei traumi

La trasfusione è la trasmissione di sangue da un organismo detto donatore ad un altro detto ricevente. Deve essere considerata una terapia atta a sostituire il sangue perduto in toto o in alcune sue componenti, in attesa che venga risolta la situazione patologica che ha portato alla perdita.

Sotto certi aspetti può essere considerata come un trapianto e come tale può dare fenomeni gravi di incompatibilità conosciuti come reazioni trasfusionali. I casi in cui si pratica più frequentemente sono:

- emorragie acute post traumatiche;
- ustioni;
- interventi chirurgici;
- trapianti d'organo;
- parto;
- emorragie di tipo organico (anemie, talassemia, leucemie, linfomi, neoplasie, emofilia, ecc.);
- carenze di componenti del plasma (albumina, fibrinogeno, fattori della coagulazione o altri fattori plasmatici).

Nel caso di trauma grave, in cui sia in corso una forte emorragia, la terapia trasfusionale è necessaria per la sopravvivenza dei pazienti: in America il 15 % dei quasi 12.000.000 di sacche trasfuse ogni anno viene utilizzato per il trattamento di pazienti traumatizzati^[10].

Tra le principali cause di morte collaterali al trauma stesso in un episodio di infortunio traumatico vi è infatti il dissanguamento: di fatto le forti emorragie rappresentano una possibile complicazione nel soccorso del paziente, vista anche la rapidità con cui le condizioni peggiorano, richiedendo la massima celerità nell'intervento.

Particolare attenzione deve avere il personale di soccorso nell'individuare il tipo di emorragia, l'origine di questa e le possibili complicazioni ad essa legate: a) quelle esterne sono ben visibili, perciò facilmente correlabili all'origine della perdita ematica e al quantitativo di sangue perso; b) quelle interne sono di più difficile inquadramento, per le diverse cause non sempre direttamente correlabili al trauma, ma a quelle patologie intercorse a seguito dell'infortunio, di difficile rilevamento in situazioni di emergenza.

La gestione di un sanguinamento critico si basa quindi sulla diagnosi precoce, sul possibile rapido controllo della fonte di sanguinamento e sul tentativo di reintegro del volume di sangue circolante.

3.1.1 Emorragie critiche e trasfusioni massive

Il termine di emorragia grave, o critica, è indicativo di uno scenario clinico tale da aumentare il rischio di alta morbidità e mortalità del paziente. Generalmente si caratterizza per il volume di sangue coinvolto, per la vasta area del corpo interessata, o, per aree limitate, in una sede corporea estremamente critica o difficilmente individuabile.

A seguito di un trauma le misure di emergenza durante il soccorso possono aumentare i disordini coagulativi a causa della diluizione ematica e della riduzione delle piastrine, con un ulteriore rischio di ipotermia, rendendo indispensabile recuperare il volume ematico con rapide trasfusioni di grandi quantitativi di sangue.

Viene definita trasfusione massiva l'apporto di 1 volume di sangue nelle 24 ore e quindi 10 U di sangue intero in un uomo adulto di 70 kg. Altre definizioni indicano il ripristino di metà dell'intero volume ematico nelle 4 ore seguenti ad una perdita di 150 ml a minuto.

Una volta determinata la necessità di effettuare una trasfusione massiva, verrà dato il via ad un protocollo di procedure urgenti basate sulla stima del sangue perso e delle condizioni ematiche del paziente da parte del medico soccorritore, attivando i laboratori trasfusionali ospedalieri per ottenere la trasfusione delle necessarie unità di emazie e di plasma, dopo l'adeguato controllo del sangue.

3.1.2 Complicazioni legate alle trasfusioni

Vari rischi possono essere associati alla terapia trasfusionale, che in situazioni di emergenza deve essere attentamente soppesata, data l'urgenza ed il poco tempo a disposizione, considerando danni e benefici in momenti in cui è in gioco la sopravvivenza stessa del paziente.

COMPLICANZE IMMEDIATE

È importante monitorare il paziente, attentamente In occasione di ogni trasfusione di sangue specialmente durante i primi 10-15 minuti di infusione, proprio per rilevare tempestivamente e precocemente i segni clinici di reazione. Molte reazioni, che in alcuni casi possono mettere a repentaglio la vita del paziente, richiedono un intervento tempestivo adeguato, anche se in reparti di terapia intensiva o di rianimazione il paziente è sempre strettamente monitorato. Le principali sono le seguenti:

- *Emolisi intravascolare acuta.* Si tratta di una complicanza rara, ma potenzialmente letale, dovuta alla reazione fra antigeni presenti sui globuli rossi trasfusi e anticorpi del ricevente, dovuta a somministrazione di sangue ABO incompatibile. La lisi delle emazie ABO incompatibili, conduce a shock, coagulazione intravasale diffusa (CID) e danno renale con emoglobinuria. All'origine di queste reazioni è quasi sempre un errore umano: il campione di sangue utilizzato per i test non proviene dal paziente sottoposto a trasfusione, il sangue viene trasfuso al paziente sbagliato, la tipizzazione del campione è errata. Poiché la reazione può esordire con un semplice rialzo termico, risulta talora difficile distinguere tempestivamente una reazione trasfusionale di lieve entità da una reazione emolitica intravascolare ben più temibile.

- *Reazioni febbrili.* La febbre può presentarsi in molti tipi di reazioni trasfusionali e può essere il primo segno di una grave reazione emolitica. Le reazioni febbrili, tuttavia, sono di solito dovute ad anticorpi diretti contro antigeni leucocitari o piastrinici.
- *Contaminazione batterica.* La reazione febbrile può essere causata anche da contaminazione batterica della sacca. Tale reazione, dovuta per lo più a endotossine batteriche, è caratterizzata da febbre alta, shock, emoglobinuria, CID e insufficienza renale, spesso associati a crampi addominali, diarrea, vomito, dolori muscolari.
- *Reazione allergica.* Le reazioni allergiche sono causate da anticorpi diretti contro proteine plasmatiche. I segni ed i sintomi possono variare da manifestazioni cutanee localizzate come orticaria, pomfi e prurito, fino a reazioni anafilattiche sistemiche con broncospasmo.
- *Sovraccarico circolatorio.* Il sovraccarico circolatorio, o ipervolemia, si presenta quando si somministra una eccessiva quantità di emocomponenti, o se la trasfusione è troppo rapida. Si manifesta con mal di testa, difficoltà respiratoria, insufficienza cardiaca congestizia, incremento della pressione arteriosa di 50mm Hg e cianosi. I sintomi di solito migliorano se si sospende la trasfusione e se al paziente, vengono somministrati ossigeno e diuretici per rimuovere l'eccesso di liquidi. Nella terapia trasfusionale post traumatica, e in caso di elevate quantità di sangue trasfuso, è una delle complicazioni più considerate.
- *Transfusion Related Acute Lung Injury (TRALI).* Con la trasfusione vi può essere la somministrazione passiva di anticorpi anti-leucocitari (leucoagglutinine) diretti contro i leucociti del ricevente provocando reazioni polmonari come l'edema polmonare acuto. I sintomi sono dispnea, ipossia, ipotensione, febbre ed edema polmonare bilaterale, durante la trasfusione o entro poche ore dal suo inizio.
- *Emolisi non immune.* Un'emolisi meccanica degli eritrociti trasfusi, dovuta ad un eccessivo traumatismo, si può avere con valvole cardiache di tipo meccanico o con la circolazione extra-corporea. Raramente la sintomatologia si manifesta in modo così critico come nell'emolisi immune, anche se si può riscontrare emoglobinuria.
- *Reazioni avverse da Immunoglobuline.* Si presenta come grave shock anafilattico in seguito all'infusione di sostanze proteiche e di IgA, in soggetti con deficit congenito di IgA.

COMPLICANZE TARDIVE

Gli effetti sfavorevoli della trasfusione possono presentarsi in un arco di tempo che va da alcuni giorni a diversi anni; è importante perciò essere consapevoli degli effetti tardivi della terapia trasfusionale, per poterli poi riconoscere ed eventualmente evitare.

- *Alloimmunizzazione post-trasfusionale.* Pregresse trasfusioni possono determinare immunizzazione, ovvero comparsa di anticorpi rivolti verso gli antigeni eritrocitari o leuco-

piastrinici o verso gli antigeni delle proteine plasmatiche e delle Immunoglobuline. Gli anticorpi eritrocitari compaiono, in genere, dopo alcune settimane o mesi dalla trasfusione di emazie, e raramente emolizzano le emazie trasfuse ancora circolanti. Se non identificati, possono creare problemi nelle successive trasfusioni.

- *Emolisi extravascolare.* La presenza di anticorpi specifici, in pazienti sensibilizzati da trasfusioni avute in passato o da precedenti gravidanze, può portare alla distruzione delle nuove emazie trasfuse. L'emolisi, che ha luogo principalmente nella milza o nel fegato, è accompagnata da febbre e ittero.
- *Malattia da trapianto contro l'ospite (GVHD).* La GVHD è una complicanza rara ma quasi sempre fatale, riscontrabile in pazienti gravemente immunodepressi in seguito all'attecchimento dei linfociti trasfusi che reagiscono contro i tessuti del ricevente. Si verifica generalmente dopo 10-12 gg, è caratterizzata da febbre, eruzioni cutanee, diarrea ed epatite. Talora si associa ad aplasia midollare per evolvere verso un esito fatale, per infezioni incontrollabili. Per evitare la GVHD è necessario irradiare a dosi elevate queste unità di emocomponenti.
- *Sovraccarico di ferro.* Una unità di globuli rossi contiene approssimativamente 250 mg di ferro. Se vi è un supporto trasfusionale continuato, si può provocare un accumulo di ferro, con tossicità a carico soprattutto del cuore, fegato e ghiandole endocrine, cui si può ovviare con una terapia ferro-chelante.
- *Rischi infettivi.* Talora con la trasfusione di sangue si possono essere trasmettere infezioni virali come quella dell'epatite B, la più frequente, anche se la più temuta, dall'opinione pubblica sembra essere l'infezione da HIV. Attualmente tutte le unità di sangue da donatore sono testate per l'epatite B, l'epatite C, l'HIV 1/2, sia con test sierologici che con metodiche di dimostrazione di acidi nucleici virali specifici. Permane il rischio di trasmissione di altre infezioni virali come da Citomegalovirus, Parvovirus B19, Virus del Nilo Occidentale ed altre, o da prioni (Creutzfeld-Jacob o della mucca pazza) o da batteri.

3.1.2.1 Complicazioni tipiche legate alle trasfusioni massive

Quando un paziente riceve trasfusioni in quantitativo elevato, il sangue personale può subire una diluizione, tale da determinare la permanenza anche di solo 1/3 del sangue originale, provocando una emodiluizione. Se non intercorrono una prolungata ipotensione o una CID, la più probabile complicanza è una trombocitopenia da diluizione. Poiché il sangue conservato non contiene piastrine pienamente funzionali, si possono manifestare sanguinamenti microvascolari come stillicidio anomalo e/o continuo da superfici di tagli profondi. Una complicanza simile è causata

più da piastrine disfunzionali che dalla piastrinopenia. Poiché i fattori coagulativi non sono significativamente ridotti, non è utile eseguire i test di coagulazione.

Una rapida trasfusione di notevoli quantità di sangue freddo può causare ipotermia che a sua volta porta aritmia o arresto cardiaco. L'ipotermia è evitata usando un dispositivo termico EV specificamente ideato per riscaldare delicatamente il sangue. Altri mezzi per riscaldare il sangue sono controindicati a causa del possibile danneggiamento dei Globuli Rossi e della possibile emolisi da calore.

3.2 Parametri e punteggi per valutare la Massive Transfusion

Attualmente si stima che un terzo dei traumatizzati gravi arriva nei reparti di Emergenza con gravi coagulopatie^[18], rendendo necessario individuare il prima possibile i pazienti a rischio di gravi emorragie e bisognosi di trasfusioni massive. Il tutto è finalizzato allo sviluppo di strategie terapeutiche per prevenire le complicanze trasfusive accennate e migliorare l'outcome clinico generale del paziente.

Ad oggi, moderne e sofisticate tecniche di imaging e di laboratorio permettono di individuare emorragie persistenti e problemi ematici difficilmente visibili se non durante un intervento chirurgico, come in caso di emorragie da infortuni toracici o retroperitoneali, in cui spesso non si riesce ad individuare l'origine dell'emorragia interna. Non sempre è possibile utilizzare la CT (Tomografia Computerizzata) o fare un angiografia durante le prime fasi della rianimazione perché porterebbero via tempo prezioso alla stabilizzazione del paziente. Analogamente, anche i test di coagulazione eseguiti in laboratori specifici sono disponibili mezz'ora o un'ora dopo l'ingresso in ospedale, anche se non sempre rendono la reale capacità coagulativa di quel momento, perché non si è ancora presentata la diluizione ematica.

È perciò in quest'ottica che si sono sviluppate metodologie di stima o veri propri modelli predittivi, basati su parametri fisiologici e sulle informazioni dell'infortunio e del paziente, per determinare il rischio di un'emorragia grave o ancor prima dell'arrivo in ospedale o almeno nella prima mezz'ora dall'ingresso nel reparto di emergenza.

Presentiamo qui sotto i principali indici:

TASH

Una semplice metodologia per stimare il rischio di trasfusione massiva è il punteggio TASH, ossia Trauma Associated Severe Hemorrhage, che può essere facilmente e velocemente calcolato basandosi su pochi parametri fisiologici misurabili anche prima dell'entrata in emergenza:

essi sono la pressione sanguigna, sesso, il tasso di emoglobina, la necessità di sonografia addominale, i battiti cardiaci, l'alcalosi o base excess e le fratture alla pelvi o alle gambe inferiori. Le variabili sono pesate secondo l'importanza nel provocare coagulopatie critiche e i pesi sono sommate tra di loro. Nella tabella allegata sono riportati i punteggi associati ai vari parametri ed il loro calcolo: il punteggio TASH varia da un minimo di 0 ad un massimo di 28. Il punteggio TASH così calcolato si usa per derivare un indice di probabilità di rischio di trasfusione massiva la cui formula viene riportata qui sotto:

Variable	Value	Points
Hemoglobin (mg/dL)	<7	8
	<9	6
	<10	4
	<11	3
	<12	2
Base excess (mmol/L)	<-10	4
	<-6	3
	<-2	1
Systolic blood pressure (mm Hg)	<100	4
	<120	1
Heart rate (beats/min)	>120	2
Free intraabdominal fluid (e.g. by FAST)		
Extremities		3
Clinically instable pelvic fracture		6
Clinically femur fracture open/dislocated		3
Male patient		1

FAST, focused assessment sonography in trauma.

$$p = 1/[1+\exp(4.9 - 0.3 * \text{TASH})]$$

Dagli studi presenti in letteratura ^[13] si è visto più del 80 % dei pazienti con probabilità di trasfusione massiva superiore al 50 % (quindi con punteggio TASH superiore a 16), vengono effettivamente trasfusi e soprattutto tutti quelli aventi un TASH score superiore a 27 subiscono una MT. Questi buoni risultati, uniti a semplicità di calcolo e rapidità con cui l'indice viene determinato rispetto all'ingresso in ospedale, hanno fatto sì che il TASH diventasse l'indice clinico di riferimento per il calcolo delle probabilità di trasfusioni massive. Si sottolinea però che dai parametri clinici utilizzati per determinare il TASH mancano alcune variabili importanti: in particolare la temperatura e il pH sono due variabili che influenzano considerevolmente la coagulopatia e che non sono considerate nel punteggio in uso, in parte perché spesso non vengono rilevate o non vengono riportate nei registri. Una situazione simile si verifica anche per il parametro del lattato, che è però presente solo in alcuni registri.

La principale critica che viene fatta al TASH in termini di immediatezza di calcolo è quello di basarsi anche su parametri legati agli "Injury Score", che spesso sono difficili da stimare nelle prime ore di intervento.

McLaughlin Score

Questo tipo di punteggio consiste di 4 componenti dicotomici, che richiedono sia misure fisiologiche, sia risultati di laboratorio. Le componenti non sono pesate e sono semplici da identificare tramite un sì e un no: il battito cardiaco > 105 bpm (HR), la pressione sistolica < 110 mm Hg (SBP), il pH < 7.25 (Ph) e l'ematocrito < 32% (Hct) sono le variabili di riferimento, se una di queste è

presente ci sarà una probabilità del 20 % di trasfusione massiva, se sono presenti tutte la probabilità sarà dell'80%.

Questo score richiede l'utilizzo di un laboratorio e dipende dai suoi tempi di elaborazione, come si intuisce dalla presenza di un dato ematico quale l'ematocrito.

La formula di calcolo delle probabilità di Trasfusione Massiva è la seguente:

$$\log (p/[1-p]) = 1.576 + (0.825 * SBP) + (0.826 * HR) + (1.044 * Hct) + (0.462 * pH)$$

ABC score

Il punteggio Assessment of Blood Consumption (ABC) è composto da 4 componenti dicotomiche, che sono derivabili direttamente dal paziente acutamente infortunato "senza muoverlo dal letto" anche in una fase di prima assistenza. La presenza di ciascuna variabile determina un punto sullo score totale, con un range possibile di punteggio da 0 a 4. L'infortunio di tipo penetrante, la pressione sistolica inferiore a 90 mm Hg, il battito cardiaco superiore a 120 bpm e un'analisi FAST positiva sono le variabili prese in considerazione. Di creazione relativamente recente, l'ABC cerca di essere immediato, facile da ricordare e da calcolare come il punteggio McLaughlin, ma accurato e preciso come il TASH score. Alcuni studi ^[9] hanno effettivamente dimostrato una comparabilità statistica tra il TASH e l'ABC, ma ad oggi rimane solo un'alternativa al primo, considerata soprattutto per la sua semplicità e per il fatto che può essere calcolato direttamente sui dati immediati del paziente.

Base Deficit

In fisiologia, l'alcalosi o base excess, o l'acidosi o base deficit (BD), si riferiscono ad un eccesso o ad un deficit, rispettivamente, della quantità di basi presenti nel sangue. È stato dimostrato^[15] che il base deficit può essere indicativo di shock, traumi addominali, dell'efficacia terapeutica nella rianimazione e quindi possa essere usato come predittore di trasfusioni massive e mortalità nelle prime 24 ore di ospedalizzazione. Non esistono vere e proprie metodologie di utilizzo del BD come predittore di MT, ma può essere utilizzato per avvalorare ulteriormente la scelta clinica della trasfusioni o in condizioni estremamente critiche risultare un rapido indice di stima prognostica.

Solitamente il BD è diviso in 4 categorie, che indicano la situazione del paziente: per valori < -10, ossia nella categoria considerata più critica, la percentuale di trasfusioni massive ed il rischio di mortalità salgono drasticamente ^[14].

Altri indici

Numerosi sistemi di valutazione fisiologici sono stati sviluppati negli ultimi decenni per calcolare in maniera oggettiva lo stato del paziente traumatizzato. Diversi studi hanno cercato di determinare se effettivamente con alcuni di questi indici come il Glasgow Coma Score o l'Injury Severity Score si fosse in grado di stimare la necessità di una trasfusione massiva: i risultati hanno dimostrato che può essere un valore soglia sia della GCS che della ISS sopra ai quali vi sono grosse percentuali di trasfusioni.

Altro parametro studiato nell'ottica della ricerca di un predittore di Massive Transfusion è l'ipotermia. È stata dimostrata la correlazione tra basse temperature corporee e coagulopatie; una temperatura inferiore ai 35 gradi, a seguito di una emorragia acuta, implica una trasfusione massiva in alte percentuali di pazienti.

Questi indici hanno il pregio della semplicità nello sviluppo della predizione delle MT ma spesso si basano su metodologie non oggettive e spesso non testate su larga scala ma basate su considerazioni empiriche più che scientifiche.

3.3 Le trasfusioni massive sul campo di battaglia

Lo sviluppo di metodologie di predizione dell'utilizzo di risorse mediche e degli outcome dei feriti durante le operazioni di guerra è materia di grande interesse soprattutto nell'ultimo decennio in cui i conflitti in Iraq e Afghanistan hanno causato numerose vittime civili e militari a seguito di traumi legati in particolare a corpi penetranti. La logistica delle risorse nei teatri di guerra è condizionata da diversi fattori quali la quantità di personale e strutture mediche, la capacità nella gestione dei traumi, il tempo delle operazioni e il numero di vittime coinvolte. Vi sono poi strutture e approvvigionamenti che devono essere imprescindibili in un ambiente prossimo ad uno scenario di guerra, come le sale operatorie e la disponibilità di sacche di sangue per le trasfusioni, viste le condizioni in cui si presta assistenza, spesso con risorse da centellinare e ottimizzare, in ospedali da campo e in strutture di fortuna.

L'emotrasfusione è una delle più importanti innovazioni mediche del precedente secolo ed ha aiutato a salvare molte vite sui campi di battaglia dalla prima Guerra Mondiale. Nei successivi conflitti le tecniche di trasfusione si sono raffinate e si è iniziato a registrare i volumi di sangue trasfuso. Questi dati sul volume trasfusionale furono principalmente usati per la pianificazione logistica delle risorse mediche da utilizzare per supplire alla possibile domanda di emotrasfusioni in altre possibili situazioni di belligeranza.

Negli ultimi anni si è cercato di predire l'utilizzo di risorse mediche più critiche come le sacche di sangue per esempio, o i possibili outcome clinici dei pazienti in base ai parametri fisiologici e clinici dei feriti: ogni scenario di guerra è caso a sé e diventa difficile perciò stimare l'utilizzo di

strutture e mezzi sanitari in base ai dati pregressi. Il potenziale di un predittore sta nell'adattarsi alla molteplicità casistiche dei singoli pazienti che non ad una stima globale.

Diversi predittori di trasfusioni massive e diversi metodi di stima degli outcome sono stati creati sulla base di dati nell'ambito civile, ma risulta necessario uno studio solo sui dati di tipo militare, in quanto le tipologie traumatiche sono tra loro simili, quasi tutte di tipo penetrante a differenza dei registri in ambito civile in cui i traumi penetranti rappresentano solo il 20 %. Inoltre la tipologia fisiologica e di comorbidità pregressa dei soggetti sono molto diverse da quelle registrate in ambito civile. Talvolta infatti l'uso di questi predittori, come il GCS o l'ISS, ha portato a previsioni meno accurate di quelle date dagli stessi utilizzati in ambito civile.^{[16],[17],[11]} .

La creazione di sistemi ad-hoc di valutazione e predizione dello stato clinico in ambito militare, allo scopo di ottimizzare le risorse mediche, da utilizzare sui campi di battaglia e rendere più rapido l'approvvigionamento, è uno degli obiettivi futuri degli studi sia in campo sanitario sia civile che militare.

Capitolo 4

Data mining in medicina clinica

Introduzione

Con *data mining* si intendono tutte quelle metodologie e strumenti atti all'analisi di grandi moli di dati il data mining è stato applicato con successo in diversi campi come l'economia e il marketing oltre che le scienze naturali e l'Ingegneria in tutte le sue branchie. Attualmente in letteratura^[20] si è cominciato a vedere un interesse anche da parte della medicina clinica alla metodologie di creazione di predittori, nell'affrontare problemi tipici: un approccio con tecniche di data mining permette di costruire infatti predittori, che sfruttano le conoscenze disponibili in ambito clinico e che spiegano le scelte effettuate una volta che i modelli vengono utilizzati per supporto a decisioni mediche. Quindi l'obiettivo di un predittore data mining in medicina clinica sarà di derivare un modello che possa utilizzare informazioni relative al paziente per predire l'outcome di interesse e per supportare ulteriormente le decisioni cliniche: esso viene infatti utilizzato per la costruzione di modelli decisionali per procedure di prognostica, diagnostica e pianificazione del trattamento.

La sempre più diffusa disponibilità di metodologie e strumenti computazionali per l'analisi e la creazione di modelli predittivi richiede ai medici ricercatori e agli esperti informatici di ricercare tra queste le strategie più adatte per far fronte a problemi di previsione clinica.

4.1 Il Data Mining: considerazioni generali

Per data mining si intendono i processi di selezione, di esplorazione e di modellizzazione di un largo numero di dati allo scopo di individuare classi o relazioni che possano portare a risultati chiari e utilizzabili per chi li analizza. Coniato negli anni '90 il termine indica lo scavar tra i dati con chiara metafora del gold mining, ossia la ricerca dell'oro. Oggi è sinonimo di ricerca di informazione in vasti database, ed enfatizza il processo di analisi all'interno dei dati in alternativa all'uso di

specifici metodi di analisi. I problemi di data mining sono spesso risolti utilizzando una varietà di approcci diversi derivati dall'Informatica, come database multi dimensionali, machine learning, soft computing e tecniche di visualizzazione dati, o derivati dalla Statistica, come i test di ipotesi, il clustering, le tecniche di classificazione e l'analisi di regressione.

La forza del data mining dipende proprio dalla corretta scelta e combinazione di queste metodologie allo scopo di ottenere due obiettivi principali sui dati: la descrizione e la predizione. Con descrizione si intende la ricerca di classi e associazioni interpretabili secondo quella che potrebbe essere una classificazione intellegibile anche da un essere umano, con predizione si definisce la stima di un possibile outcome di interesse. Sebbene le metodologie utilizzate per giungere a questi risultati spesso sono analoghe, le differenze tra il classificatore e lo stimatore predittivo stanno sostanzialmente nella preparazione dei dati: nel secondo caso infatti bisognerà includere nel test in esame le variabili di interesse particolare che il modello dovrà predire.

4.2 Il Data Mining Predittivo

Lo scopo di tecniche di data mining predittivo, come detto, è quello di individuare modelli statistici e modelli che possono essere utilizzati per predire la risposta di interesse. Questi modelli sono costruiti utilizzando dati per cui il valore della variabile di risposta è già noto.

4.2.1 Principali Tecniche e Metodologie utilizzate nel data mining predittivo

Le principali metodologie di data mining predittivo hanno origine dai differenti campi di ricerca e spesso utilizzano approcci di modellizzazione molti diversi.

Sono presenti diverse tecniche che possono essere messe a confronto sulla base di

- gestione di dati mancanti e rumorosi
- modo in cui vengono gestiti i diversi tipi di parametri (continui, ordinali, categorici)
- presentazione di modelli di classificazione facilmente interpretabili da personale esperto
- riduzione del numero di parametri per arrivare a conclusioni
- costi computazionali per la costruzione e l'uso di modelli di classificazione
- nei modelli utilizzati nel supporto decisionale capacità di questi di giustificare determinate scelte
- la duttilità nell'adattarsi ai diversi casi testabili.

Le principali tecniche di data mining più comunemente usate sono le seguenti:

Alberi di decisione

I Decision Tree o Alberi di decisione usano un partizionamento ricorsivo dei dati ed inducono una classificazione robusta e trasparente, sebbene le performance soffrono della caratteristica

segmentazione dei dati: ogni foglia dell'albero può arrivare a contenere un numero elevato di istanze, il che non permette talvolta di arrivare a predizioni utilizzabili. La complessità computazionale dell'algoritmo di induzione è bassa grazie alle potenti basi euristiche.

Le più utilizzate suite di programmi per il data mining contengono varianti di C4.5 e See 5 e Decision Tree CART: questi software ed in generale i decision tree saranno esposti in dettaglio nel prossimo capitolo.

Regole Decisionali

Questa metodologia può essere definita come l'equivalente verbale di un decision tree; un set di Regole Decisionali, o più comunemente definito con il termine inglese decision rules, specifica l'appartenenza di un istanza ad una classe, basandosi su una sequenza gerarchica di decisioni contingenti. Ogni regola nell'insieme delle regole decisionali generalmente prende la forma di un clausola di Horn (in logica nel calcolo proposizionale è una disgiunzione di letterali in cui al massimo uno dei letterali è positivo) l'appartenenza alla classe è dedotta da una serie di osservazioni contingenti. Le Decision Rules sono quindi nella forma

IF condizione basata sui valori degli attributi **THEN** valore di out-come
e possono essere costruite da alberi decisionali come le regole derivate dal C4.5 (il programma di decision tree) o possono essere derivate direttamente dai dati come nel caso degli algoritmi AQ e CN2 [22]. Questo tipo di algoritmi sono molto simili per certe caratteristiche ai decision tree, ma rispetto a quest'ultimi sono molto più costosi dal punto di vista computazionale

Regressione Logistica

La regressione logistica è una metodologia potente e già particolarmente utilizzata in ambito statistico. È un caso particolare di modello lineare generalizzato, un'estensione della regressione ordinaria, e si applica ai casi in cui la variabile dipendente y sia di tipo dicotomico, come in un classificatore che stabilisca l'occorrenza o la non occorrenza di un determinato evento. Il modello viene descritto quindi da una formula di regressione che determina la funzione legata alla probabilità che un evento y accada; la stima della probabilità poi avviene effettuando la stima dei parametri del modello attraverso il metodo di massima verosimiglianza.

Reti neurali

Una rete neurale artificiale è un modello di calcolo basato su alcuni principi delle reti neurali biologiche. Tale modello è costituito dall'unità base detta neurone artificiale che può avere una o più connessioni in entrata, proveniente da altri neuroni della rete, ed ha una sola uscita diretta eventualmente, verso altri neuroni o verso un output. Nella maggior parte dei casi una rete neurale artificiale è un sistema adattivo che cambia la sua struttura in base alle informazioni che

attraversano la rete durante la fase di apprendimento. Esse possono essere utilizzate per simulare relazioni complesse tra ingressi e uscite che altre funzioni analitiche non riescono a rappresentare. Una rete neurale artificiale riceve segnali esterni su uno strato di nodi (unità di elaborazione) d'ingresso, ciascuno dei quali è collegato con numerosi nodi interni, organizzati in più livelli. Ogni nodo elabora i segnali ricevuti e trasmette il risultato a nodi successivi. Le reti neurali sono state finora tra le principali tecniche utilizzate per il data mining in ambito di medicina clinica^[21]. Principalmente tutto ciò è dovuto alle buone performance di questo tipo di metodica nella creazione di modelli predittivi, nonostante una serie di difetti che includono alta sensibilità ai parametri del metodo, alto costo computazionale nell'addestramento della rete, e la difficoltà interpretativa del modello dedotto per un personale esperto. Le reti neurali sono tuttavia strumento capace di modellare complesse relazioni non lineari, ed in questo sta indubbiamente il grande vantaggio delle reti rispetto ad altri metodi di modellistica più semplici.

Support Vector Machine

Le macchine a vettori di supporto (SVM, dall'inglese Support Vector Machines), o macchine kernel, sono un insieme di metodi di apprendimento supervisionato per la classificazione di pattern: sono anche note come classificatori a massimo margine, poiché allo stesso tempo minimizzano l'errore empirico di classificazione e cercano di classificare due istanze massimizzando il margine geometrico di distanza tra due punti in un insieme. Rappresentano attualmente ^[23] lo strumento forse più potente di classificazione in termini di accuratezza di predizione, essendo basate su forti basi matematiche e sulla teoria di apprendimento statistico. Centrale a questo tipo di metodica è la procedura di ricerca dell'iperpiano che separa gli esempi di due diversi esiti. Essendo stato creato principalmente per problemi a due classi, le SVM trovano un iperpiano, detto iperpiano ottimo, con la distanza massima al punto più vicino tra le due classi. Il set di attributi con la distanza minima dall'iperpiano ottimo è chiamato support vector. La ricerca dell'iperpiano porta ad un classificatore lineare. Al di là di queste basi lineari, le support vector machine sono usate più spesso in modelli non lineari, in cui lo spazio dei parametri viene trasformato in un nuovo spazio, a più dimensioni in cui il classificatore lineare viene inferito. Il concetto di base è che ogni punto del piano degli attributi, sia rappresentato da un set di funzioni matematiche conosciute come kernel, permettendo all'algoritmo di ricavare l'iperpiano a massimo margine nel nuovo spazio dei parametri trasformati. La trasformazione può essere non lineare e lo spazio trasformato può essere multidimensionale. Per questo il classificatore può essere un iperpiano nello spazio dei parametri multidimensionali, e può essere non lineare nello spazio originale degli input. Le più utilizzate funzioni di kernel sono polinomiali a base radiale o funzioni sigmoidali. La scelta del più appropriato kernel dipende ovviamente dalle proprietà del dataset e del dominio del problema. Nel caso reale però spesso non

esiste un iperpiano che separa chiaramente tutti i dati in due classi. Per bypassare questo inconveniente, viene proposto un metodo a margine ridotto dove il risultante iperpiano divide l'insieme dei dati in 2 set: un dato sarà classificato in una determinata classe, se sta nel semipiano con maggioranza di dati in quella classe. Le SVM hanno cominciato ad acquistare grande popolarità in ambito medico-informatico solo in tempi recenti, in particolare grazie alla bioinformatica^[20], dal momento che i formalismi teorici possono essere facilmente interpretabili anche dagli esperti clinici e non solo dai data miners. In termini di accuratezza di previsione le SVM si avvicinano alle reti neurali, ma sono più robuste e sono meno dipendenti dalla scelta degli attributi del dataset.

Classificatore naive bayesiano

Il classificatore naive bayesiano è un semplice classificatore basato sull'applicazione del teorema di Bayes, con forti assunzioni di indipendenza tra le variabili predittive. Questo metodo si fonda sull'assunzione che la presenza o assenza di un determinata caratteristica di una classe sia scorrelata dalla presenza o assenza di qualsiasi altro attributo: per esempio, una mela può essere definita dagli attributi rossa, rotonda e di 8 centimetri di diametro. Anche se questi dipendono l'uno dall'altro o dall'esistenza di un altro parametro, il classificatore naive li considera essere contributi indipendenti alla probabilità di classificare un oggetto istanza. Dipendendo dalla precisa natura del modello predittivo, il classificatore naive Bayesiano può essere addestrato in maniera molto efficiente con un approccio di apprendimento supervisionato. In diverse applicazioni pratiche, la stima dei parametri del modello probabilistico Bayesiano utilizza il metodo della massima verosimiglianza, perciò tutti i parametri di probabilità Bayesiana devono essere a disposizione. Pur essendo un modello relativamente semplice, è estremamente performante e comparabile con altri sistemi, come gli alberi decisionali. Il suo grande vantaggio sta nel richiedere un piccolo gruppo di dati da utilizzare come training set al fine di stimare i parametri necessari alla classificazione, necessitando solo della varianza delle variabili per ciascuna classe e non dell'intera matrice di covarianza.

Reti Bayesiane

Una rete bayesiana è un modello grafico probabilistico capace di esprimere chiaramente una distribuzione di probabilità congiunta sopra un numero di variabili attraverso un set di distribuzione di probabilità condizionata. Matematicamente una rete bayesiana è un grafo aciclico, dove ogni nodo rappresenta una variabile aleatoria e gli archi rappresentano le relazioni di dipendenza statistica tra le variabili e le distribuzioni locali di probabilità dei nodi figli rispetto ai valori dei nodi padre. Ogni variabile è assunta indipendente dai suoi non discendenti ed ha un insieme di genitori detto $pa(X_i)$. Sotto questa ipotesi di partenza, detta di Markov, la probabilità condizionata di tutte la

variabili x può essere scritta:

$$p(S|x) = \frac{p(x|S)p(S)}{p(x)} \propto p(x|S)p(S)$$

La rete è completamente descritta dall'insieme di distribuzioni di probabilità condizionate, che quantificano le relazioni qualitative tra le variabili espresse dal grafo. Queste distribuzioni di probabilità descrivono il set di parametri Θ . Così come i valori della tabella delle probabilità condizionali descrivono variabili discrete, la media e la varianza della distribuzione gaussiana danno le variabili continue. Le reti sono state tradizionalmente utilizzate in medicina come strumento per ragionamenti probabilistici, grazie alla loro affidabilità. Possono facilmente essere usate per analisi probabilistiche di classificazione, dove sono inquadrare come una generalizzazione della metodica del naive Bayesiano. L'algoritmo di apprendimento dai dati è basato sulla struttura del modello di selezione. L'obiettivo è costruire la struttura S con la migliore distribuzione di probabilità a posteriori, dato un dataset x : questa probabilità a posteriori è proporzionale a due termini, la verosimiglianza marginale, che indica come dovrebbe essere il modello rispetto ai dati, e la probabilità a priori della struttura S . Il loro utilizzo è sempre più frequente sia nel data mining predittivo che in bioinformatica. I principali difetti di questa metodica risiedono nella costruzione della struttura del grafo, che necessita di dataset molto grandi, e nell'interpretazione dei casi inseriti nel grafo.

K-nearest neighbors

È l'algoritmo più semplice fra quelli utilizzati nell'apprendimento automatico, utilizzati nel data mining, basandosi per il riconoscimento di pattern da classificare sulle caratteristiche degli oggetti vicini a quello considerato. Un oggetto è classificato in base alla maggioranza dei "voti" sulla classe dei suoi k vicini, in cui k è un intero positivo, tipicamente non molto grande. Se $k=1$ allora l'oggetto viene assegnato alla classe del suo vicino. In un contesto binario, in cui sono presenti esclusivamente due classi, è opportuno scegliere un valore di k dispari, per evitare di ritrovarsi in situazioni di parità. Questo metodo può essere utilizzato per la tecnica di regressione assegnando all'oggetto la media dei valori dei k oggetti suoi vicini. Se si considerano solo i "voti", ossia la classe di appartenenza dei k oggetti vicini, vi è il rischio dell'influenza dovuta alla predominanza delle classi con più oggetti e quindi può risultare utile pesare i contributi dei vicini, in modo da dare, nel calcolo della media, maggior importanza, in base alla distanza dall'oggetto considerato. La scelta di k dipende dalle caratteristiche dei dati. Generalmente all'aumentare di k si riduce il rumore che compromette la classificazione, ma il criterio di scelta per la classe diventa più labile. La scelta di k quindi può esser presa attraverso tecniche euristiche, come ad esempio la cross-validation. L'algoritmo è composto da a) una fase di training, in cui si partiziona lo spazio in regioni secondo le

posizioni e le caratteristiche degli oggetti; b) una fase di calcolo della distanza in cui gli oggetti sono rappresentati attraverso vettori di posizione in uno spazio multidimensionale, in cui si misura la distanza euclidea; c) una fase di classificazione in cui un punto è assegnato alla classe C, se questa è la più frequente fra i k esempi più vicini all'oggetto sotto esame. La vicinanza si misura in base alla distanza fra punti. I vicini sono presi da un insieme di oggetti per cui è nota la classificazione corretta. La ricerca della classe di appartenenza può essere lunga e richiede l'addestramento piuttosto elaborato, con un set simile per caratteristiche al validation set dell'insieme di training, per avere una classificazione accurata.

4.3 I principali ambiti di utilizzo del data mining

4.3.1 Data mining e statistica

I concetti di data mining e Statistica (il campo della matematica applicata connesso con l'analisi dei dati) possono essere entrambi definiti come "estrazione di informazione utile da insiemi di dati" ^[24]. La differenza che caratterizza in modo saliente i metodi del primo è però la possibilità di prendere un numero estremamente elevato di dati, anche di diversa tipologia permettendo una modellizzazione molto flessibile.

L'iniziale successo riscontrato da queste tecniche in aree come l'analisi del paniere ed il text mining, incoraggiò gli statistici e la loro comunità scientifica nell'utilizzo e sviluppo del data mining. Mentre suite di programmi di data mining sono diventate parte dei pacchetti software più importanti utilizzati dagli statistici ed i principali libri sull'argomento sono stati scritti da studiosi di statistica, i più recenti sviluppi nel data mining si sono focalizzati nel creare un ponte tra statistica e rappresentazione analitica dei dati dei diversi campi dell'Ingegneria.

Una distinzione estremamente enfatizzata tra statistica pura e data mining viene fatta sulla base della mole di dati da analizzare, in cui il secondo si considera essere più performante per grandi quantità di dati piuttosto che la statistica che nelle sue analisi risulta più efficace con dati meno ampi. Il data mining ha poi la grande possibilità di mostrare le deduzioni ottenute dai dati in modo esplicito e comunicabili ad uno gruppo di esperti interessati, permettendo di spiegare le scelte per ulteriori casistiche nuove, schematizzando ed utilizzando il dominio delle conoscenze carpite dall'analisi dei dati.

Altra differenza tra data mining e statistica sta nel modo di approcciare le non linearità e i disturbi nei dati. I dataset in generale sono soggetti a differenti sorgenti di rumore, che possono essere legate a diversi fattori talvolta prevedibili, talvolta causate da una sostanziale mancanza di dati compilati e governate quindi da un comportamento non lineare. La statistica moderna e il data

mining sono a sufficienza potenti da essere utilizzabili per gestire la maggior parte dei casi, a prescindere dalla sorgente di rumore. Tuttavia vi è un differente approccio nell'analisi dei dati tra questi due tipologie di studio: in statistica si crea una modellizzazione "a priori" delle non linearità o delle relazioni tra le caratteristiche, nel data mining si cercano invece le tecniche per automatizzare delle induzioni costruttive, come la scoperta dell'interazione tra i vari parametri e la gestione delle non linearità di modello, analizzando i dati e lo spazio degli attributi. Per gli algoritmi di data mining, i dati sono solo la forma principale di informazione, a cui qualsiasi conoscenza aggiuntiva dedotta da questa, può essere utilizzata in modo complementare.

4.3.2 Data mining in Medicina: l'esperienza in genomica

Negli ultimi anni, il data mining predittivo ha avuto grande impulso dalla ricerca in biologia molecolare. I metodi data mining, come il clustering gerarchico o le support vector machine, sono usati di routine nell'analisi di dati altamente complessi ottenuti dai DNA-microarrays o dalla stereospettrometria di massa.

E' stato evidenziato che il potenziale del data mining predittivo per inferire modelli clinicamente ^[25] rilevanti da misure molecolari e per supportare scelte operative in medicina genomica. Ad oggi tre differenti tipologie di misure molecolari sono utilizzate per l'analisi clinica: a) i dati sul genotipo, la variazione di sequenza di DNA, che si ha quando un singolo nucleotide nella sequenza genomica è alterato, b) i dati sull'espressione genica, che possono essere misurati con i DNA microarrays, per ottenere un'istantanea dell'attività dei geni in un tessuto ad un determinato istante temporale, c) l'espressione proteomica, che include un set completo di profili di proteine ottenute con lo spettrometro di massa. Gli utilizzi dei metodi data mining predittivo nella ricerca riguardano principalmente l'analisi dell'espressione genica contenuta nei microarrays, che consiste nella valutazione di migliaia di geni per ogni paziente, con lo scopo di diagnosticare la tipologia di patologia ad essi legata, ottenere una prognosi, e individuare le opportune scelte terapeutiche. Si dimostra l'efficacia dell'utilizzo dei Dna microarrays nel determinare l'outcome predittivo nella ricerca contro il cancro ^[26].

Per migliorare l'accuratezza del classificatore e la rilevanza clinica dei modelli prognostici, è stata proposta ^[27] l'integrazione di dati clinici e dell'espressione genica: un approccio con algoritmi decision-tree è stato condotto raggruppando i geni mutageni, unendoli ai dati clinici, al fine di prevedere la sopravvivenza del paziente. Questo utilizzo per ricavare la lista dei geni utili per la predizione dell'outcome non è esente da difetti. Infatti nei diversi studi in letteratura su predittori di geni con grandi performance in fatto di qualità di classificazione, si è osservato come la lista dei geni predittivi vari radicalmente tra i diversi esperimenti ^[28]. Al fine di aumentare la robustezza e l'impatto pratico di questi studi, si cerca sempre più di integrare conoscenze funzionali dei geni e dei

processi biomedici, con dati di espressione genica e dati clinici, uniti al confronto di dati provenienti da altri studi. Una considerazione infine meritano anche le ricerche sui profili proteomici, derivati dalle tecniche di spettrometria di massa, per inferire modelli predittivi. In quest'ambito di ricerca sono stati utilizzati per predire l'outcome di pazienti affetti da cancro ovarico e prostatico.

Se si considerano i limiti dei metodi precedenti, nel futuro gli informatici e i bioingegneri avranno il compito di creare strumenti e di inventare euristiche efficienti per oltrepassare l'ostacolo della proibitiva complessità nella ricerca delle interazioni tra geni: una sfida nel design di software, che possa portare ad un'interfaccia di analisi interattiva ed esplorativa, che porti gli utenti non esperti di informatica a scoprire e formulare nuove ipotesi dai dati grezzi.

4.4 Data Mining in medicina clinica

A differenza del data mining in economia finanziaria o nel marketing, l'applicazione del data mining in campo medico ha alcune caratteristiche particolari, in quanto si opera in un contesto in cui sicurezza e accuratezza nelle decisioni sono estremamente critiche. Per questo motivo il valore di ciascun dato può essere più importante che in altri contesti, anche se i dati clinici sono soggetti a diverse fonti di rumore ed incertezza, legate ad errori di misura, dati mancanti, dovuti ad imprecisioni nella codifica sui registri digitali delle informazioni dei report testuali. Il data mining può superare questi problemi con un'attenta applicazione di metodi di selezione di variabili e modelli adeguati, mediante un'interpretazione delle conoscenze pregresse nell'analisi di dati.

I modelli predittivi in medicina clinica sono "modelli di aiuto decisionale che combinano due o più attributi dei dati del paziente per arrivare a predizioni di tipo clinico."^[38]

Il data mining può contribuire allo sviluppo di modelli predittivi utili grazie ad alcuni importanti aspetti correlati: a) un approccio all'analisi di dati comprensivo e propositivo che coinvolge l'applicazione di metodi da diverse aree scientifiche; b) la capacità esplicativa di tali modelli; c) la capacità di usare un background di conoscenze pregresse nel processo di analisi dei dati.

Il data mining si basa su processi di modellizzazione, che beneficiano di un costante sviluppo e miglioramento di standard e di metodiche, vista la loro importanza in diversi ambiti in continua evoluzione. Il vantaggio ma anche la difficoltà del data mining sta nell'essere uno strumento, che integra diverse metodologie, assunte da diverse discipline scientifiche.

Vista la buona esperienza nell'ambito di medicina genomica anche per l'incombente necessità di integrare dati clinici e dati molecolari, il data mining, come altre conoscenze computazionali avanzate, è diventato estremamente richiesto, per raggiungere lo stato dell'arte sia in ricerca, che nelle applicazioni della vita reale in ambito medico.

4.4.1 L'utilizzo dell'informazione pregressa o a priori

L'uso del data mining può avere un duplice scopo: può essere usato per derivare una particolare classificazione o per capire quale informazione sia contenuta nei dati. Nell'ottica di scoprire nuove informazioni è importante riuscire a sfruttare anche le conoscenze pregresse (a priori) sui dati in analisi per la costruzione del modello. Per conoscenze a priori si intendono quella quantità di informazione utile a capire la situazione o il problema in esame, e in un modello quelle conoscenze composte da dati o da assunti teorici che non servono direttamente per la sua implementazione ma per descrivere l'insieme di utilizzo, i limiti di derivazione e contribuire al calcolo del modello. Nella costruzione di un predittore è importante, specialmente in ambito medico, utilizzare le conoscenze pregresse per inserire informazioni note, che non dovrebbero essere riscoperte dai dati. Le conoscenze a priori possono essere espresse mediante diversi strumenti matematici come i modelli bayesiani, la logica fuzzy e i concetti gerarchici. Nel metodo bayesiano, la probabilità a priori inserita nel calcolo della probabilità condizionata permette alle conoscenze pregresse di essere considerate nel modello. Le probabilità a priori permettono ad un modello di essere derivabile, anche quando l'informazione dei dati è povera, e può aiutare nel correggere modelli costruiti, seguendo i dati troppo "alla lettera", con il rischio di fallire nella classificazione di casi nuovi, non esaminati in fase di training.

4.4.2 Tecniche del data mining clinico

Il data mining applica molto spesso diverse tecniche, prese da vari ambiti scientifici con lo scopo di scoprire elementi e aspetti interessanti dai dati analizzati: vista la moltitudine di metodiche attualmente utilizzabili proveniente da diversi campi di studio non stupisce che il data mining sia visto come materia ardua e ostica da trattare ed applicare. Allo scopo di semplificare l'approccio del personale esperto nel mondo del data mining, sono state create diverse suite, come SPSS e SAS, in ambito statistico informatico ed ingegneristico con l'obiettivo di una standardizzazione globale nelle procedure. Il data mining predittivo si basa sull'analisi di dataset composti da istanze e in medicina da pazienti singoli da esaminare o dai casi clinici, ove ciascuna istanza è caratterizzata da un numero di attributi, alcuni chiamati variabili di outcome, riferibili in ambito clinico all'esito di una terapia o di una serie di eventi e casistiche. Queste variabili possono essere dipendenti da altri parametri, riferibili ad una classe ed in genere, rappresentano l'attributo più importante da analizzare e nei modelli predittivi l'attributo da predire.

Al contrario del data mining standard, i dataset per il data mining medico sono più piccoli: tipicamente il numero di istanze ed il numero di attributi variano da diverse decine a diverse migliaia. L'obiettivo del data mining predittivo nella medicina clinica è costruire un modello predittivo, che aiuti i medici a migliorare le prognosi, le diagnosi e la pianificazione delle procedure

di trattamento. In termini di analisi di dati, ci sono diverse problematiche a cui il data mining è chiamato a rispondere, valutando se la quantità di dati e parametri predittivi sia sufficiente a costruire un modello: questo è comprensibile grazie ad osservazioni empiriche sulla validità del modello, con l'aiuto di esperti o valutando la capacità degli algoritmi di costruire i modelli, a seconda della quantità di dati. E' necessario trovare l'attributo maggiormente predittivo da inserire nel modello, per comprendere la relazione tra l'attributo e l'outcome da predire. Trovare combinazioni di interesse tra gli attributi o i fattori intermedi degli attributi originali, può potenziare le performance del modello predittivo e indicare fenomeni particolari.

Per il data mining i dati clinici più frequentemente arrivano da database dedicati che sono stati creati allo scopo di studiare una particolare situazione clinica. Gli algoritmi più usati richiedono di mettere i dati in forma tabulare, includendo fattori predittivi e outcome. I dataset dei modelli sono compilati attraverso query su tabelle da database più ampi per estrapolare le informazioni più significative.

I modelli predittivi non devono essere testati sullo stesso set di dati usato per la costruzione. Per questo motivo, sono utilizzate tecniche come la cross validation, anche se talvolta risulta più conveniente dividere i dati in due insiemi: a) il primo, indicato come training set, è usato per comparare diversi algoritmi di data mining, trovando il miglior insieme per ordinare e selezionare i parametri, usando metriche statistiche per la valutazione delle loro performance, ed infine selezionando i modelli che funzionano meglio, b) il secondo, detto validation set, per testare il modello finale completo dal training set sviluppato, usando queste metodiche.

Separare training e validation set è necessario per valutare oggettivamente le performance predittive. I modelli data mining possono essere complessi e in casi estremi possono "ricordare" ciascuna istanza dei dati da cui hanno appreso. Questo tipo di modelli, infatti, si comporta perfettamente su dati, che sono stati usati per la fase di apprendimento. E' indispensabile la valutazione del predittore per ogni nuovo caso, che non rappresenti le istanze dei dati del dataset di apprendimento.

4.4.3 Valutazione dei modelli predittivi: metriche e metodi di paragone

Dopo aver categorizzato e selezionato gli attributi e dopo aver indotto una predizione di outcome, diverse misure statistiche possono essere usate per stimare la qualità del modello derivato:

- Accuratezza di classificazione*: misura la proporzione tra test classificati correttamente e test con stima sbagliata, stimando quindi una probabilità di classificazione corretta.

- Sensibilità e Specificità*: misura l'abilità del modello nel riconoscere i pazienti di un certo gruppo. Se decidiamo per esempio di osservare la predizione di un outcome , come la

sopravvivenza o meno di un paziente, la sensibilità è la probabilità che un paziente, che sia sopravvissuto, sia predetto come sopravvissuto; la specificità è la probabilità che un paziente non sopravvissuto, sia classificato come non sopravvissuto

-*Area sotto la curva ROC (aROC)*: è basato sul test non parametrico del segno statistico e stima la probabilità, continuando con l'esempio fatto sopra, che la predizione di sopravvivenza per un individuo estratto a caso dal gruppo dei sopravvissuti sia superiore a quello estratto a caso dal gruppo dei non sopravvissuti.

-*Probabilità media di assegnamento alla classe corretta*: è collegato all'accuratezza di classificazione, ma dà informazioni addizionali sulla coerenza con i dati delle scelte del classificatore.

Le metriche e statistiche sopra descritte sono effettuate in genere attraverso la ten-fold cross-validation, in cui si divide l'insieme dei dati delle istanze da osservare in dieci insiemi, che hanno approssimativamente la stessa grandezza e uguale distribuzione di outcome. In ciascun esperimento, un singolo insieme è usato per testare il classificatore che sarà sviluppato dai rimanenti nove insiemi. Le statistiche per ciascun metodo sono poi estrapolate da una media tra i dieci esperimenti. Questi metodi stimano la qualità di un singolo classificatore: tuttavia per comparare i classificatori viene usato il test di McNemar, test migliore e più corretto per questo scopo. Esso compara due classificatori contando i casi, che sono classificati correttamente dal primo classificatore, ma non dal secondo (n_{10}) e viceversa (n_{01}). L'insieme di test e di training è usato per entrambi i metodi di induzione, e i risultati possono essere sommati per tutti i dieci esperimenti di cross validazione. L'indice viene calcolato come:

$$D = \frac{(|n_{01} - n_{10}| - 1)}{n_{01} + n_{10}}$$

che è distribuito approssimativamente con una distribuzione χ^2 con un grado di libertà.

4.4.4 Implementazione del Data Mining e possibili evoluzioni

Il data mining in ambito clinico dovrebbe essere anche considerato per le relazioni scoperte con la sua applicazione. La scoperta di una relazione interessante tra due attributi può cambiare l'attuale pratica medica? Il modello predittivo costruito può essere utilizzato per un supporto decisionale in maniera più pragmatica? Una volta che il modello è derivato, la qualità della assistenza sanitaria migliora o i costi derivati diminuiscono? Sono queste le varie domande di base a livello clinico applicativo. Tuttavia le analisi dell'utilità del modello costruito dal data mining predittivo sono piuttosto rare, tanto più per un modello in ambiente clinico. Gli strumenti del data mining sono spesso abbastanza complessi, in quanto le suite di software sono estremamente costose e costruite per specialisti, e ottimizzate più per lo sviluppo del modello, che per usare specifiche interfacce per l'utilizzo del modello nel contesto specifico. Non sempre dispongono di ambienti appropriati per

l'utilizzo come supporto decisionale, che è uno degli aspetti più importanti di utilizzo del data mining in ambito clinico.

In tempi recenti sono comparsi diversi strumenti utili ad affrontare questo problema e agevolare l'unione tra data mining e supporto decisionale: si pensi allo standard introdotto dal linguaggio PMML, linguaggio di programmazione per la creazione di modelli predittivi. Con l'evoluzione di questo standard o di altri simili, i principali strumenti di data mining potrebbero essere affiancati a shell di supporto decisionale, utili ad implementare uno specifico modello predittivo e ad offrire un'interfaccia appropriata anche per un personale esperto operante nell'ambito clinico.

L'idea base di questi o di simili approcci è di collegare data mining e supporto decisionale da una parte e dall'altra di rendere indipendente la gestione per utenti come i medici e il personale sanitario, utilizzando shell decisionali, facili da usare. La capacità di estrapolare in maniera chiara il modello di data mining predittivo, può portare significativi vantaggi nella comunicazione dei risultati, nella valutazione e nei test dei progetti data mining.

I modelli di data mining predittivo possono essere anche usati come strumenti per giustificare e comparare risultati medici basati sulle evidenze sperimentali con l'outcome ottenuto dalla pratica clinica. La disponibilità dei dati registrati nelle istituzioni cliniche su processi specifici di gestione di problemi di salute, permette di integrare strategie comprovate e di fare delle osservazioni prognostiche, mediante le informazioni provenienti dai dati nella routine clinica. La fusione della medicina evidence-based con l'esperienza clinica specifica può essere vista come un particolare problema di data mining dove il background di conoscenza può essere ottenuto da studi clinici: in questo caso le conoscenze pregresse portano informazioni da integrare nel processo di apprendimento degli algoritmi.

L'obiettivo dell'analisi dei dati è quindi collegato ad una migliore comprensione delle informazioni che sono contenute nei dataset d'analisi, evidenziando casi che non confermino conoscenze stabilite o problemi nelle procedure di registrazione dei dati. L'obiettivo per un centro clinico è analizzare ogni caso specifico nell'ottica di tessere in maniera migliore linee guida prognostiche. Il data mining appare quindi come parte dell'infrastruttura informatica delle istituzioni mediche.

4.4.5 Data mining predittivo per l'analisi di pazienti con traumi gravi

Il data mining clinico può sposare perfettamente i cosiddetti problemi 'bed-side', ossia il genere di problematiche di diagnostica su un paziente ospedalizzato osservandone i sintomi, con modelli che prevedano la predizione dell'outcome del paziente. Un supporto decisionale, che usi un particolare modello predittivo, potrebbe tener conto di diversi aspetti nel predire o meno un certo

outcome e fornire dei costi per determinate scelte.

I recenti progressi del machine learning e dell'analisi intelligente dei dati, combinati con l'evoluzione del data mining clinico, hanno portato ad una crescente utilità di queste metodiche, utilizzando dati retrospettivi nella fase prognostica: tutto ciò è diventato possibile grazie anche ai miglioramenti delle tecniche di archiviazione dati e del sistema informatico degli ospedali, che hanno permesso di ottenere volumi di dati più consistenti e più accurati. Questi miglioramenti riguardano anche i reparti di emergenza e le strutture ad essi collegati.

Recentemente si sono potuti studiare modelli predittivi anche per traumatizzati gravi in occasione del primo approccio chirurgico, allorché si devono affrontare complessi problemi gestionali e decisionali, per pazienti con gravissime situazioni cliniche. Nelle prime fasi di trattamento ospedaliero è necessario il non sempre agevole recupero della instabilità emodinamica così da stabilizzare il paziente tener sotto controllo gli esiti immediati dell'infortunio, allo scopo di evitare procedure chirurgiche estese e importanti su pazienti ancora scompensati. Questa procedura richiede investimenti importanti di personale, di strutture e di risorse orientati verso un numero di casi relativamente esiguo tra gli infortunati gravemente, che tuttavia hanno una bassa probabilità di sopravvivenza, perché dopo l'intervento, il decesso è legato a deficit respiratori, alle coagulopatie o ai distress degli organi interni. Nella scelta della distribuzione delle risorse per ottimizzarne l'uso, si ritiene necessaria la presenza di un modello prognostico, che prenda in considerazione soprattutto il periodo che intercorre tra il trauma e la prima chirurgia.

Per modelli prognostici si è dimostrato che un set sperimentale di pochi dati è sufficiente per costruire un modello predittivo, ma la stessa scarsa quantità di dati presente su particolari tipologie di traumatizzati, rende difficile attualmente validare modelli data mining^[29]

Capitolo 5

Alberi di Decisione

Introduzione

Nella costruzione di un classificatore, l'obiettivo principale non è limitato alla sola costruzione di un predittore, ma anche alla modellizzazione intellegibile e comprensibile da un osservatore umano. Gli alberi di decisione o decision tree riescono a soddisfare questi requisiti con una limitata richiesta computazionale rispetto ad altri classificatori molto usati nel data mining, come le reti neurali o le support vector machine.

Tra i software più usati per l'implementazione dei decision tree, particolare importanza riveste il progetto di Ross Quinlan, C4.5 e la sua evoluzione C5.0: i campi di utilizzo di questo software sono molteplici e vanno dalla bioinformatica all'informatica forense passando per la statistica e la biologia.

I risultati rilevanti ottenuti in questi diversi ambiti dal classificatore, fanno riflettere sulla duttilità del programma e ci consentono di pensare ad un utilizzo anche in ambito medico clinico, dove i decision tree si adatterebbero all'esigenze di confrontabilità e facilità d'osservazione da parte del personale medico.

5.1 Definizione

Nel machine learning un albero di decisione è un modello predittivo, dove ogni nodo interno rappresenta una variabile, un arco verso un nodo figlio rappresenta un possibile valore per quella proprietà, ed è detto ramo o ramificazione, e una foglia indica il valore predetto per la variabile obiettivo a partire dai valori delle altre proprietà. Un albero di decisione viene costruito utilizzando tecniche di apprendimento a partire dall'insieme dei dati iniziali (dataset), il quale come già spiegato

viene diviso in due sottoinsiemi: il training set sulla base del quali si crea la struttura dell'albero ed il validation set, che viene utilizzato per testare l'accuratezza del modello predittivo così creato.

Nel data mining l'albero di decisione viene utilizzato per classificare le istanze dei dati: in questo caso descrive una struttura ad albero dove i nodi foglia rappresentano le classificazioni e le ramificazioni l'insieme delle proprietà che portano a quelle classificazioni. Di conseguenza ogni nodo interno risulta essere una macro-classe costituita dall'unione delle classi associate ai suoi nodi figli.

Ogni ramo dell'albero di decisione rappresenta un test su una singola variabile che segmenta lo spazio in una o più parti. Ogni ramificazione deve essere necessariamente binaria.

Il predicato, che si associa ad ogni nodo interno e sulla base del quale avviene la ripartizione dei dati, è chiamato condizione di split.

Esiste poi un criterio di arresto (halting), o anche criterio di potatura (pruning), che ha il fine di determinarne la profondità massima. Questo perché il crescere della profondità di un albero, ovvero della sua dimensione, non influisce direttamente sulla bontà del modello, ma anzi, una crescita eccessiva della dimensione dell'albero potrebbe portare solo ad un aumento del tempo di calcolo sproporzionato, rispetto ai benefici riguardanti l'accuratezza delle previsioni/classificazioni.

Gli alberi di decisione in genere sono di due tipologie principali.

Alberi di classificazione: assegnano un'etichetta ai record e quindi alle classi appropriate; possono anche indicare il livello di confidenza sulla classificazione effettuata. In questo caso l'albero di classificazione fornisce la probabilità della classe.

Alberi di regressione: stimano il valore di una variabile target che assume valori numerici. Un albero di regressione potrebbe calcolare per esempio il contributo di un sostenitore o il numero di sinistri, quindi, i rimborsi di un assicurato.

Esprimere i concetti degli alberi decisionali, con tabelle e numeri è difficile e, anche se formalmente una tabella è corretta, può lasciare perplesso e confuso chi la legge, perché non appare immediata la giustificazione di un esito. Una rappresentazione grafica è di aiuto, permettendo di raffigurare con maggiore leggibilità le stesse informazioni, ed evidenziando il ramo che è stato percorso, per arrivare a determinare la scelta o la valutazione. La tecnica dell'albero di decisione quindi è utile per identificare una strategia decisionale o per perseguire un obiettivo predittivo, creando un modello il cui grafo orienta immediatamente la lettura del risultato.

5.2 Costruzione dell'albero di decisione

La costruzione di un albero di decisione segue un processo detto di partizionamento ricorsivo. Si tratta di un processo iterativo di divisione dei dati in partizioni destinate ad essere a loro volta suddivise in altre sottopartizioni. All'inizio tutti i record del set di addestramento si trovano in

un unico gruppo e l'algoritmo tenta di dividere i dati, utilizzando ogni possibile suddivisione binaria per ogni campo, scegliendo la segmentazione che massimizza un criterio euristico di partizionamento. Questo processo di creazione di partizioni viene poi applicato ad ogni nuova casella, fin che non si trovano più segmentazioni utili.

5.2.1 L'algoritmo Divide and Conquer

Lo scheletro della procedura di costruzione di un decision tree è basato sul così detto algoritmo "Divide and Conquer", che fornì le idee principali per le prime sperimentazioni in ambito alberi di decisione. Considerando l'insieme S di casi di addestramento e le possibili classi di appartenenza degli attributi descritte da $\{C_1, C_2, \dots, C_K\}$, abbiamo due possibilità:

- S contiene uno o più casi, tutti appartenenti ad una classe singola C_j : il decision tree per T è una foglia identificante la classe C_j .
- S contiene casi che appartengono ad una varietà di classi: in questa situazione l'idea è di affinare S in subset di casi che sono il più possibile appartenenti ad una singola classe che contenga diversi casi. Si sceglie un insieme di test B , con outcome $b_1, b_2, b_3, \dots, b_t$ che produce una partizione non banale di S , e si denoti con S_i i casi che hanno outcome b_i di B . L'albero di decisione è partizionato in sottoinsiemi T_1, T_2, \dots, T_n dove T_i contiene tutti i casi in T che hanno outcome b_i , ottenuto dal test. L'albero di decisione per T è composto da un nodo decisionale, che identifica il test e un ramo per ogni possibile risultato. Lo stesso meccanismo di costruzione dell'albero è applicato ricorsivamente per ogni subset di casi di addestramento, così che il ramo i-esimo porti ad un decision tree costruito dal subset T_i di casi di addestramento.

Nell'algoritmo divide and conquer la scelta dell'insieme dei test B che partiziona in maniera non banale S porta ad un albero di decisione, ma diversi insiemi di test B_s portano ad alberi di decisione diversi. L'algoritmo seleziona i test candidati a formare l'albero tra i test che massimizzano un criterio euristico di partizione.

Per gli algoritmi di costruzione degli alberi sono utilizzati i criteri di guadagno d'informazione e di rapporto di guadagno. Sia $RF(C_j, S)$ che indica le frequenze dei casi che appartengono alla classe C_j in S . Il contenuto di Informazione del messaggio che identifica la classe di un caso in S sarà poi

$$I(S) = \sum_{j=1}^x RF(C_j, S) \log(RF(C_j, S))$$

Dopo che S è stato partizionato in sottoinsiemi S_1, S_2, \dots, S_t dall'insieme test B , l'informazione avrà poi in

$$G(S, B) = I(S) - \sum_{i=1}^t \frac{|S_i|}{|S|} I(S_i)$$

il criterio di guadagno sceglie l'insieme di test B che massimizza $G(S,B)$

Un problema di questo criterio è che favorisce test con outcome numerosi, per esempio $G(S,B)$ è massimizzato da un test in cui ogni S_i contiene un caso singolo. Il criterio di rapporto di guadagno bypassa questo problema tenendo conto della potenziale informazione della partizione stessa:

$$P(S,B) = - \sum_{i=1}^t \frac{|S_i|}{|S|} \log\left(\frac{|S_i|}{|S|}\right)$$

Il criterio gain ratio massimizza il rapporto tra il guadagno $G(S,B)$ e l'informazione della partizione $P(S,B)$

5.2.2 La gestione dei test con esito sconosciuti

L'algoritmo precedentemente descritto contiene un'assunzione nascosta data per implicita: che i risultati di un test per ogni caso siano determinati. Talvolta ci si trova a dover gestire istanze i cui attributi sono sconosciuti o addirittura i risultati dei test sperimentali sono indeterminabili. Gli algoritmi legati agli alberi di decisione più comuni assumono che i test con outcome sconosciuto, siano distribuiti probabilisticamente in accordo con le relative frequenze di outcome conosciuto. I test con esito ignoto sono consapevolmente "sparpagliati" in insiemi di test con outcome conosciuto proporzionalmente alle frequenze relative di questi, con il risultato che un caso ignoto può seguire diverse strade nell'albero. Tutto questo si applica allo stesso modo quando i casi di training sono divisi durante la costruzione dell'albero e quando l'albero è usato per classificare casi diversi.

5.2.3 La "potatura" dell'albero: il pruning

Gli algoritmi usati negli alberi di decisione effettuano la segmentazione migliore all'altezza della radice, dove il numero di record è maggiore. Ogni segmentazione successiva ha una popolazione più piccola o meno rappresentativa da elaborare. Verso la fine, le idiosincrasie dei record di training di un nodo specifico evidenziano pattern specifici solo per quei record. Questi pattern sono irrilevanti e pericolosi per le previsioni.

Esistono situazioni in cui la segmentazione serve solo a separare due record e non aiuta una classificazione: l'albero trova quella che può essere una distinzione effettiva nel set di addestramento ma che non può essere altrettanto caratteristica in senso generale. Si presenta quindi un problema di overfitting, cioè sovradattamento del modello al set di addestramento ed è difficile stabilire esattamente l'aumento dell'errore attribuibile a questa problematica.

Per evitarlo esistono due diversi approcci: uno detto tecnica di stopping e l'altra detta tecnica di sfoltimento o pruning.

La tecnica stopping ha lo scopo di bloccare la crescita dell'albero prima che raggiunga profondità eccessiva. Il funzionamento si basa sull'applicazione di vari test ad ogni nodo per stabilire

se sia utile o meno effettuare un'altra segmentazione. Il test può essere semplice quanto richiedere la presenza di un minimo di record ad ogni nodo, oppure più complesso, comportando l'esecuzione di un test statistico sulla significatività della segmentazione proposta. L'esperienza insegna tuttavia che stabilendo dimensioni minime per il nodo, l'albero cresce in modo corretto. Questo perché quando i nodi contengono molti record costituiscono campioni più rappresentativi.

L'operazione di pruning consiste nell'eliminazione di foglie e rami per migliorare le prestazioni dell'albero di decisione, diminuendo l'errore complessivo. In effetti l'albero sfoltito è un sottogruppo dell'albero decisionale completo.

Nella sua procedura di sfoltimento è possibile usare un algoritmo basato sul tasso di errore aggiustato, indicato dalla formula sottostante:

$$A_t = E_t + \alpha \lambda_t$$

dove l'indice t indica il particolare sottoalbero, E_t l'errore associato ad un sottoalbero t , λ_t è il numero di foglie nel sottoalbero t . Il tasso di errore aggiustato tiene conto anche del peso del numero di foglie (peso indicato con α); esso permette inoltre di effettuare confronti tra i sottoalberi e l'albero completo. Se $t = 0$ si ha l'albero completo e λ_0 è il numero di foglie dell'albero, A_0 è l'errore aggiustato dell'albero completo. Esso coincide con E_0 solo se $\alpha = 0$ ed in generale se $\alpha = 0$ l'errore aggiustato di un qualunque sottoalbero coincide con E_t . L'algoritmo di pruning prevede le seguenti fasi:

1. fissato α si calcola A_t per ogni sottoalbero contenente la radice;
2. si aumenta α da 0 a 1 fino ad ottenere un sottoalbero t_1 per cui risulta $A_{t_1} \leq A_0$ e si eliminano tutti i rami che non sono parte di t_1 ;
3. si procede allo stesso modo fino ad arrivare alla radice ottenendo in tal modo un set di sottoalberi $\{t_1, t_2, \dots\}$;
4. si calcola l'errore di ogni sottoalbero utilizzando un altro insieme di dati (test-set) e si sceglie il sottoalbero con l'errore minore.

Questa soluzione algoritmica è abbastanza efficace, in quanto procede allo sfoltimento sulla base delle prestazioni dell'albero misurandole su un set indipendente di dati preclassificati. Con un singolo set di test, l'algoritmo può portare l'albero fino a quel sottoalbero che minimizza l'errore sul set di test. Con set di test multipli, invece possiamo anche affrontare direttamente la questione della generalità del modello selezionando l'albero che fornisce le prestazioni migliori su molti insiemi di test.

5.3 Il software Decision Tree : C5.0

C5.0 è l'evoluzione di C4.5, un software di data mining largamente utilizzato per la creazione e lo sviluppo di alberi di decisione, che a sua volta è stato l'aggiornamento di ID3, un altro software molto popolare in quest'ambito. Tutti questi programmi sono stati progettati da Ross Quinlan, uno scienziato americano specialista in data mining e machine learning.

Il codice del programma è disponibile sul sito della società del suo inventore, la "Rulequest", e può essere utilizzato sia su macchine UNIX che su dispositivi Windows.

5.3.1L'algoritmo: da ID 3 a C4.5 a C5.0

La prima incarnazione ufficiale dell'algoritmo C5.0 si ha con il primo programma creato da Quinlan per implementare i decision tree: l'Iterative Dichotomizer (ID3).

I passaggi chiave dell'algoritmo sono:

1. Prendere tutti gli attributi non ancora utilizzati relativamente ai campioni del test da esaminare e calcolare la loro entropia
2. Scegliere l'attributo A la cui entropia è minima, o equivalentemente il cui guadagno di informazione è massimo
3. Creare un nodo radice che contiene quell'attributo A
4. Costruire l'albero di decisione, con rami da questo nodo radice per ogni possibile valore dell'attributo
5. Assegnare ogni possibile caso con valore corrispondente agli attributi dei rami
6. Partizionare il dataset seguendo il valore di ogni ramo, in modo che siano presenti, solo casi con il valore corrispondente dell'attributo del ramo e torna al passo 1.

L'algoritmo della parte di costruzione dell'albero di decisione in pseudocodice assumerà poi questa formulazione:

ID3 (Esempi, Attributi_Target, Attributi)

- Crea un nodo radice per l'albero
- **IF** <tutti gli Esempi sono della stessa classe + >,
 THEN return Radice Albero nodo singolo con etichetta = +
- **IF** <tutti gli Esempi sono della stessa classe - >
 THEN return Radice Albero nodo singolo con etichetta = -
- **IF** <il numero di attributi predittivi è vuoto>
 THEN return Radice albero nodo singolo, con etichetta = valore più comune di attributi target
- **ELSE:**
 - A = L' Attributo ottimo con la migliore entropia o guadagno di informazione in Esempi appartenente ad Attributi.
 - Attributo Decision Tree come Radice = A.
 - **for** <ogni possibile valore v_i , di A>

- Aggiungi nuovi rami sottostante la Radice, corrispondente ai test in cui $A = v_i$.
- $Esempi(v_i)$ = il subset di esempi che hanno il valore v_i di A
- **return** l'albero con il nodo radice A e gli archi etichettati con v_1, v_2, \dots, v_i vanno agli archi con sottoalberi rispettivamente $ID3(Esempi(v_1), Attributi_Target, Attributi - \{A\}), ID3(Esempi(v_2), Attributi_Target, Attributi - \{A\}) \dots ID3(Esempi(v_i), Attributi_Target, Attributi - \{A\})$
- **END**
- **Return** Root

L'attributo ottimo A è considerato, come visto quello che massimizza l'entropia o il guadagno di informazione nel set S , insieme dei casi.

Il concetto di Entropia è formalizzato dalla seguente formula:

$$E(S) = - \sum_{j=1}^n f_S(j) \log_2 f_S(j)$$

dove:

- $E(S)$ è l'entropia dell'informazione dell'insieme S dei test;
- n_i è il numero di valori diversi degli attributi in S (entropia viene calcolata per un determinato attributo alla volta)
- $f_S(j)$ è la frequenza del valore j nel set S

Relativamente all'algoritmo in pseudocodice sopra descritto essa viene quindi calcolata come:

$$E(S) = -p_- \log_2(p_-) - p_+ \log_2(p_+)$$

Dove $p_-(p_+)$ è la proporzione di esempi negativi (positivi) nell'insieme S .

Un'entropia di 0 identifica un set perfettamente classificato e come mostrato nell'algoritmo l'entropia è utilizzata per determinare quale nodo verrà splittato al passo successivo: più alta è l'entropia, più alto è il potenziale di migliorare il classificatore.

Un modo per quantificare i miglioramenti d'entropia del set a seguito dei vari split dell'albero, è il calcolo del guadagno d'informazione o Information gain, precedentemente introdotto per l'algoritmo "Divide and Conquer", formalizzato dalla seguente formula:

$$G(S, A) = E(S) - \sum_{i=1}^m f_S(A_i) E(S_{A_i})$$

Dove :

- $G(S,A)$ è il guadagno del set S dopo lo split per un attributo A
- $E(S)$ è l'entropia dell'informazione del set S ;
- m è il numero di valori diversi dell' attributo A in S
- $f_S(A_i)$ è la frequenza dei casi che hanno A_i come valore per A nell'insieme S
- A_i è il valore i -esimo di A
- S_{A_i} è il sottoinsieme di S contenente tutti gli oggetti dove il valore di A è A_i

ID-3 è algoritmo estremamente performante nella velocità di creazione degli alberi e contiene già un algoritmo di “pruning”, per evitare possibili overfitting. Tuttavia il rischio di overfitting si presenta comunque con piccoli dataset e solo un attributo alla volta può essere testato per modelli decisionali: negli anni 80, quando venne progettato ID3 ,questi svantaggi implicavano, uniti alle difficoltà di classificazione continua e a causa degli alti costi computazionali, che gli esami data mining di questo algoritmo fossero estremamente dispendiosi in termini di tempo.

Alla luce di questi limiti diversi anni più tardi Quinlan studiò un evoluzione del suo algoritmo: nella nuova incarnazione l'algoritmo chiamato C4.5, parte dagli stessi assunti del suo predecessore, ossia un data set di addestramento $S = s_1, s_2$ di dati già classificati, ogni suo campione $s_i = x_1, x_2, \dots$, un vettore dove x_1, x_2, \dots rappresentano gli attributi dei parametri dei campioni ed i dati di addestramento sono un vettore $C = C_1, C_2, \dots$ dove C_1, C_2, \dots, C_n sono la classe a cui ciascun campione appartiene. L'algoritmo poi ad ogni nodo dell'albero sceglie un attributo dei dati, che più marcatamente divide il suo set di campioni in sottoinsiemi, appartenenti all'una o all'altra classe, equivalentemente al suo predecessore. Per questa scelta di attributo viene prese l'attributo con il guadagno di informazione più alto, procedendo ricorsivamente fino a sottoinsiemi più piccoli.

C4.5 poi si interrompe se:

- L'albero T contiene uno o più esempi, tutti appartenenti alla medesima classe → singola foglia con etichetta la classe
- L'albero T non contiene nessun esempio (insieme vuoto) → singola foglia con etichetta la classe più frequente nell'insieme padre
- Nessun test riesce a generare almeno due insiemi con un numero minimo di esempi → singola foglia con etichetta la classe più frequente (alcuni esempi saranno mal classificati)

C4.5 ha fatto diversi miglioramenti rispetto ad ID3 permettendo di gestire sia attributi continui che discreti, gestendo anche dataset con attributi mancanti, permettendo di dare un costo/peso ad

alcuni attributi rispetto ad altri e perfezionando l'algoritmo di pruning, con una potatura dopo la creazione albero.

A seguito della nascita della sua società Rulequest, Quinlan decise di sviluppare un'aggiornamento con importanti migliorie di C4.5, con l'obiettivo di commercializzare un software basato sull'algoritmo. Questo aggiornamento prende il nome di C5.0 e mostra significativi miglioramenti in termini di accuratezza e velocità rispetto al suo predecessore, puntando alla costruzione di alberi di decisione più piccoli e ad una gestione della memoria più efficiente di C4.5. Altra importante novità è l'aggiunta di 3 nuove opzioni: l'inserimento del file cost, che permette di far pesare diversamente classificazioni errate a seconda della classe, il setacciamento (winnowing) - un filtraggio automatico dei dati considerati inutili- e il boosting, cioè la creazione congiunta di diversi alberi di decisione allo scopo di migliorare l'accuratezza.

5.3.2 L'utilizzo di C5.0

Come detto precedentemente lo scopo di C5.0 è trovare il modo di predire le classi di appartenenza di determinati casi in esame dal valore degli attributi dei casi di un training set: l'algoritmo fa questo creando un classificatore che fornisca delle previsioni. In particolare il software di Quinlan può costruire classificatori espressi come decision tree o come regole decisionali. Per ottenere una classificazione è necessario fornire alla funzione di elaborazione dell'albero un dataset ordinato ed elaborato secondo un particolare ordine ed un particolare formato, arrivando poi ad avere un file di output rappresentante l'albero di decisione ed una schermata con le informazioni sulla classificazione, ove viene rappresentato graficamente l'albero decisionale e il dataset di partenza classificato.

5.3.3 I file utilizzati da C5.0

Ogni task di C5.0 è identificato da un breve nome identificativo chiamato filestem; tutti i dataset e alberi letti e scritti dal sistema sono nella forma filestem.extension, dove extension caratterizza il tipo di informazione inserita.

Il file fondamentale per ogni task di classificazione è il file con estensione name che descrive gli attributi e le classi di questi. Tra questi ce ne è uno più evidenziato detto attributo target, che indica all'algoritmo la classe da predire. I nomi di ogni attributo esplicitamente definito in questo file possono essere arbitrari, mentre la descrizione dei valori che essi possono avere, deve appartenere necessariamente a sei tipi di classe che saranno poi i formati che C5.0 è in grado di gestire. Le tipologie di formato che il programma può prendere in input sono i numeri continui, i numeri discreti, il formato data, il formato temporale sia nella forma ora, minuti, secondi che nelle forma data, ora, minuti, secondi, l'ordine ignore (che ignora l'attributo nel dataset), ed il formato label, ossia stringhe che descrivono le classi possibili

Il secondo file essenziale è il data file che contiene il training set da cui saranno costruiti il decision tree o le regole decisionali. Qui l'entry di ciascun caso è composta da una o più righe, che danno il valore per tutti gli attributi esplicitamente definiti. Altri file utilizzati sono i filestem.test che permette di creare un validation set con nuovi casi test ed ha lo stesso formato del data, ed il file di costo, implementato nell'ultima versione, chiamato costs che è opzionale e indica il costo di eventuali errori di classificazione.

5.3.4 L'output di C5.0

Una volta disposti i file names e data, è possibile avviare il programma ottenendo quindi la costruzione di un albero di decisione. L'output del software presenterà una schermata, in cui verrà presentata una rappresentazione dell'albero mediante una serie di condizioni if concatenate. Assieme all'output grafico il software fornisce anche un file con estensione .tree che può essere utilizzato per testare il classificatore su di un programma interactive interpreter, che utilizza il decision tree all'interno del file, per testarlo su un validation set, creato come i file .data utilizzati nella costruzione del classificatore.

La schermata della shell di output permette di verificare quanti sono stati gli errori globali e di vedere quanti soggetti sono stati classificati in una determinata classe, confrontandoli con la classe di effettiva appartenenza.

Gli alberi di decisionae possono essere abbastanza difficili da capire: è quindi anche possibile poi ottenere un insieme di regole decisionali, invece di un albero come output dell'algoritmo

5.3.5 Le opzioni

Per ottenere alberi più accurati o meglio interpretabili talvolta è necessario unire la funzione di calcolo del classificatore con alcune opzioni nell'algoritmo, che permettono di implementare alcune specifiche particolari.

Le opzioni principali, utilizzabili nella costruzione dell'albero, sono il *winnowing* (setacciamento), il *boosting*, il *softening threshold*, l'*advanced pruning*.

Il *setacciamento* viene utilizzato nelle applicazioni in cui C5.0 usa un numero importante di attributi, rendendo la classificazione meno accurata: l'abilità di prendere e scegliere tra i parametri, quelli che predicono meglio, è un aspetto chiave della modellizzazione degli alberi di decisione. L'opzione winnow effettua una ricerca analizzando il training set prima di costruire il decision tree, ed estrapola gli attributi più importanti per la classificazione elencandoli in ordine di importanza: C5.0 stima il fattore per cui il tasso di errore o il costo della classificazione errata sarebbe incrementato, se quel determinato attributo non fosse considerato nella costruzione dell'albero.

Il *boosting* è una delle innovazioni più recenti di C5.0 ed è basato sugli studi di Rob Schapire e Yoav Freund [34]: si basa sull'idea di creare diversi classificatori invece che uno solo, in cui ad ogni

nuovo caso esaminato, ognuno di questi classificatori esprimerà un voto, ed il conteggio dei voti di tutti i classificatori determinerà la classe finale. L'algoritmo verrà descritto più in dettaglio nel prossimo capitolo.

La *softening threshold*, permette di gestire meglio i valori nell'intorno di una soglia di classificazione, o di separazione dei rami dell'albero, creando soglie meno nette di separazione, gestendo l'intorno di valori vicini alla soglia in maniera accorta e scegliendo i lower ed upper bound della stessa alla luce di un'analisi di accuratezza del classificatore.

L' *advanced pruning* è un'opzione estremamente sofisticata, che viene utilizzata raramente nei test di classificazione, ma che può tuttavia incrementare le prestazioni dell'albero di decisione se usata in maniera accorta. C5.0 costruisce l'albero in due fasi: prima cercando di adattarsi al meglio ai dati e poi rimuovendo (pruning) le parti che hanno un tasso di errore estremamente alto. Questa potatura è applicata prima ad ogni sottoalbero poi globalmente. Le opzioni di potatura avanzata permettono di evitare questo secondo tipo di pruning, ottenendo così alberi più grandi, e di controllare il modo in cui viene stimato il tasso d'errore e quindi il grado di pruning

5.3.6 Utilizzi e limitazioni di C5.0

C5.0 è stato usato in diversi ambiti scientifici al fine di fare ricerche di previsioni accurate e di regole di decisione facilmente intelleggibili da personale esperto. ^{[30][33]}

I più recenti utilizzi si sono concentrati in ambito bioinformatico ed ingegneristico ed hanno portato a diverse pubblicazioni. Anche il mondo della statistica pura e dell'informatica forense ha sempre considerato questo genere di classificatori in maniera molto positiva.

Tuttavia, i decision tree utilizzati dall'algoritmo di Quinlan si sposano solo ad alcune tipologie di classificazione. C5.0 infatti può gestire solo classi discrete e non ordinate, ma non le classi continue e non il caso delle classificazioni delle quantità numeriche. L'algoritmo nella sua ultima versione è in grado di gestire diverse tipologie di attributi, tra cui anche gli ordinali, anche se non può gestire attributi invece strutturati, che hanno una gerarchia di possibili valori, come la gerarchia dei colori, che si dipana dai colori primari fino ad ogni possibile sfumatura di ciascun colore. Queste limitazioni comunque serviranno ad indicare la strada per possibili nuove implementazioni aggiuntive per un algoritmo, che si è rinnovato, migliorando di volta in volta in ogni sua successiva incarnazione e che è diventato uno standard tra le più moderne tecnologie in ambito data mining

Capitolo 6

Analisi Sperimentale

Introduzione

In questo capitolo verrà descritto il setup sperimentale e la serie di test effettuati al fine di creare un indice di probabilità di trasfusioni massive, con tecniche di data mining, in particolare usando gli alberi decisionali con l'algoritmo C5.0. Si è poi effettuato un confronto con il relativo standard clinico attualmente utilizzato, ossia l'indice TASH.

6.1 L'Ospedale dell'Angelo di Mestre

I dati del Ritg che verranno utilizzati per le successive analisi sono stati concessi dal reparto di rianimazione dell'Ospedale dell'Angelo di Mestre, che è il nuovo ospedale di Mestre e sostituisce l'Umberto I situato in centro. L'espansione e la modernizzazione di alcune zone della città Veneta hanno portato via via lo spostamento e l'ampliamento dei servizi anche in zone più periferiche rispetto il vecchio centro cittadino. Di fondazione estremamente recente, è stato aperto al pubblico nel 2008 e fornisce 680 posti letto.

6.1.2 Il pronto soccorso

I traumatizzati gravi sono pazienti gestiti solitamente dalle strutture di urgenza dell'ospedale: sono infatti soccorsi dal pronto intervento e la prima ospedalizzazione avviene attraverso il pronto soccorso, per poi, a seconda delle loro condizioni, essere trasferiti in rianimazione ed anestesia, il reparto che gestisce la stabilizzazione dei pazienti in situazioni critiche, prima che siano eventualmente spostati in reparti specifici. L'archiviazione dei dati nel registro traumi viene effettuata solitamente dal personale medico anestesista della rianimazione, o dai medici del pronto soccorso e del pronto intervento.

In particolare consideriamo il Servizio di Pronto Soccorso dell'Ospedale dell'Angelo, da cui proviene il nostro dataset, che è una delle venti strutture più utilizzate d'Italia. La piena entrata a

regime dell'ospedale dell'Angelo ha aumentato infatti la sua capacità attrattiva: l'ospedale è punto di riferimento provinciale non solo per i casi di trauma grave, ma anche per infarto, patologie neurochirurgiche e cardiocirurgiche. Rispetto alle altre strutture lagunari è possibile che la maggiore modernità e l'offerta di attrezzature tecnologiche possano aver spostato il bacino d'utenza verso questo ospedale: il Pronto Soccorso dell'Angelo è ad oggi il punto di principale di arrivo e smistamento per pazienti di tutta la provincia di Venezia, in particolare per i traumatizzati grazie a una gestione dei soccorsi più ottimizzata rispetto agli altri ospedali veneziani, data la sua posizione strategica a livello provinciale e per le strutture più moderne e tecnologicamente avanzate nella rianimazione.

Uno dei problemi maggiori di un pronto soccorso sono i tempi d'attesa: nella vita quotidiana di tutti i presidi ospedalieri di emergenza e urgenza, il 50% degli accessi riguarda i codici bianchi, ovvero patologie non gravi, soggette al pagamento del ticket e che sono all'ultimo posto nella scala di priorità nell'intervento dei sanitari. Tuttavia per un Servizio di Pronto Soccorso è importante essere chiamati a dare risposte rapide a chi è in pericolo di vita come i traumatizzati gravi: attualmente in questa realtà il tempo di attesa per i codici rossi, cioè i casi gravissimi, è stato portato a 9 minuti, per i gialli a 12 minuti, essendo quindi in grado di intervenire efficacemente nel primo quarto d'ora, una porzione temporale cruciale, in cui davvero si decide la sorte di un malato o di un ferito.

6.2 La gestione dell'RITG a Mestre

Attualmente la gestione dei dati del Registro Intraospedaliero multiregionale Traumi Gravi per l'ospedale di Mestre è gestita da una società di consulenza informatica specializzata in applicativi di gestione database in ambito informativo, che attraverso il portale Ritg.net permette al personale sanitario preposto di inserire i dati attraverso un'applicazione web based lato server.

La Rianimazione Generale dell'ospedale partecipa attivamente alla costituzione del database che permette di effettuare un percorso di miglioramento continuo della qualità attraverso il confronto e lo scambio di esperienze con i più importanti Trauma Center italiani.

L'applicazione è realizzata con tecnologia "Active-Server-Pages" (ASP.net), e funziona direttamente da browser internet senza alcuna installazione software dal lato dell'utilizzatore (lato client). Questa soluzione aggira ogni problema di compatibilità e dialogo tra ospedali. In più evita tutti i problemi derivanti dai software installati su macchina singola e garantisce al personale la comodità e flessibilità di immissione dati giacché consente di utilizzare qualsiasi computer connesso a Internet. Allo stesso tempo, garantisce anche la massima indipendenza tra i "data-set" dei singoli ospedali, grazie ad un sistema di password specifiche. Il software offre infine una serie di piccole

differenze a seconda dell'ospedale connesso, così da fornire la massima adattabilità alle specifiche esigenze di ciascun centro. Il fatto che i dati sensibili non siano rintracciabili su Internet rispetta le norme sulla privacy. Il sistema informatico comprende un sito web dove possono essere pubblicati aspetti del progetto, con multipli e progressivi livelli di restrizione, come, ad esempio, indicatori generati automaticamente dal programma in tempo reale relativi a numero di pazienti ammessi, gravità, mortalità, ecc. Il programma è in grado di inviare un file contenente una selezione dei record, generata in tempo reale in base al profilo utente connesso (ad esempio tutti i casi di un determinato ospedale), a qualsiasi indirizzo di posta elettronica.

I dati presenti nelle schede del software sono divisi in 5 principali categorie:

1. Dati demografici.
2. Dati preospedalieri essenziali: livello di assistenza, parametri vitali sulla scena, ora e tipo di manovre diagnostiche e terapeutiche.
3. Dati intraospedalieri nelle prime 24 ore: parametri vitali all'arrivo, ora e tipo di manovre diagnostiche e terapeutiche, intervalli, reparto di ricovero definitivo.
4. Gravità: Abbreviated Injury Scale (AIS), ISS, RTS, TRISS.
5. Esito: durata del ricovero in ospedale e in terapia intensiva, complicazioni, interventi chirurgici successivi, luogo e ora dell'eventuale decesso, disabilità a 6 mesi (Glasgow Outcome Score, questionario EQ 5D).

Nel 2009 è iniziato uno studio europeo basato su queste variabili, che ha tra gli obiettivi, analogamente al famoso *Major Trauma Outcome Study* americano, lo sviluppo di un modello predittivo di outcome tarato sulla realtà europea con lo scopo quindi di arrivare a predittori più accurati di quelli esistenti.

6.3 Analisi Dataset

Sono stati presi in esame i dati relativi al database RITG dell'Ospedale di Mestre relativi a 320 pazienti registrati dall'entrata in funzione nel 2008 del database nel reparto di Pronto Soccorso e Rianimazione. All'interno del database sono stati memorizzati diversi parametri sia di carattere clinico, come pressione arteriosa o indici di gravità, che di carattere logistico, seguendo l'ordine cronologico di assistenza del paziente infortunato con campi dedicati per ogni fase del soccorso. Tuttavia diversi campi spesso non sono stati compilati e i dati relativi ad alcuni pazienti erano estremamente scarsi. Sono stati quindi scartati dall'analisi 15 soggetti, che avevano meno del 20 % dei parametri compilati, in modo da rimuovere una sicura fonte di errore.

6.3.1 Obiettivi dell'analisi

I primi sviluppi dell'analisi sono stati improntati nella costruzione di un dataset, che potesse essere funzionale alla creazione di un training set per derivare un classificatore decision tree. Sono stati eliminati quindi tutti i parametri non riconducibili a categorie, a numeri reali o a numeri discreti, come per esempio la dettagliata descrizione del tipo di terapia o dell'iter clinico. Si sono anche individuati i parametri che consentissero di calcolare l'indice TASH, con l'obiettivo di confrontarlo con un indice di trasfusione massiva, creato con tecniche di data mining, in particolare con gli alberi decisionali: il confronto ha consentito di verificare l'affidabilità e l'accuratezza delle tecniche dei decision tree rispetto allo standard clinico attualmente più usato. Dei parametri analizzati per il data mining, dapprima si sono utilizzati i principali attributi clinici, e successivamente è stata fatta una selezione più sofisticata, come verrà spiegato in seguito.

6.3.2 Le tipologie di dato nel dataset

Nella pagina seguente è presentata la tabella (tabella 6.1) con i principali dati e la tipologia.

Per attributi *numericali razionali* si intendono gli attributi legati a dati con misure precise, i cui intervalli tra valori, hanno significato con un valore minimo possibile.

Per attributi *nominali* si intendono gli attributi che non hanno un ordine, ma che contengono un nome o un'etichetta di categoria: si dividono poi in *dicotomici*, che possono assumere solo due valori, e *multicotomici* che hanno più di due valori possibili.

Per attributi *ordinali* si intendono i dati che hanno un ordine, ma non è significativo l'intervallo tra due possibili valori.

Alcuni di questi attributi sono stati registrati in diverse fasi del soccorso al paziente, e compaiono più volte, chiaramente con valori diversi. I dati con registrazione multiple sono: la pressione sistolica, il Glasgow Coma Score, la categoria della pressione sistolica al polso secondo il Revised Trauma Score (categoria polso), la frequenza respiratoria, la SpO2 registrate al momento del soccorso e all'ingresso in ospedale.

Le fasi di soccorso in cui sono stati registrati i dati sono: il primo intervento sul paziente, l'ingresso in pronto soccorso, la fase di rianimazione, la fase di interventi di emergenza e le eventuali fasi di reparto, seguendo quindi l'iter della catena di soccorso.

Tra i soggetti registrati è possibile individuare i trasfusi massivi dai dati relativi alle emazie e alle unità di plasma trasfuse: la loro somma viene considerata come la quantità di sangue intero trasfuso. Nell'attuale data set i pazienti con sangue intero trasfuso ≥ 10 sono stati classificati come trasfusi massivi e rappresentano il 7 % del numero totale. Tuttavia sono stati esclusi diversi pazienti, i cui dati nel registro non erano ben compilati e risultavano non utilizzabili ai fini di un'analisi corretta, riducendo l'insieme dei dati a 289 elementi.

Attributi di tipo razionale	Attributi di tipo nominale	Attributi di tipo ordinale
età	Sedazione	Gravità AIS Testa
SBP arrivo ambulanza	Intubazione Tracheale	Gravità AIS Collo
FrequenzaRespiratoriaAtti arrivo ambulanza	Ventilazione Artificiale	Gravità AIS Faccia
SpO2Specifica arrivo ambulanza	Incannulamento Vena Periferica	Gravità AIS Rachide Cervicale
SBP in pronto soccorso	Incannulamento Vena Centrale	Gravità AIS Rachide Dorsale
INR in pronto soccorso	Incannulamento Arteria	Gravità AIS RachideLombare
PO2 in pronto soccorso	Infusione Sangue	Gravità AIS _Addome
Ph in pronto soccorso	Raggi X Torace	Gravità AIS Torace
Numero Totale Unità Plasma trasfuse	Raggi X Bacino	Categoria pressione Polso in pronto soccorso
Emazie trasfuse	Raggi X Rachide Cervicale	GlasgowComaScore all'arrivo dell'ambulanza
TempoChiamataArrivo	Eco Addome	Glasgow Coma Score in pronto soccorso
FrequenzaRespiratoriaAtti in pronto soccorso	Rianimazione Cardiopolmonare	RevisedTraumaScore
PCO2 in pronto soccorso	Tipo_trauma	InjurySeverityScore (ISS)
HCO3 in pronto soccorso	Air Way	TRaumaISS
Tempo arrivo ambulanza	Decompressione Toracica	NewISS
Lattati misurati in pronto soccorso	Drenaggio Toracico	TRaumaNewISS
Indice Emoglobina misurato in pronto soccorso	Intubazione preospedaliera	Categoria pressione Polso arrivo in ambulanza
Tempo normalizzazione base excess	Sesso	GravitàAIS arti inferiori/bacino
baseexcess		Gr AiS arti superiori
SpO2 in pronto soccorso		Gravità AIS termica
Cannule		
Fluidi endovena		

TABELLA 6.1 – La tabella illustra i parametri e la tipologia degli attributi memorizzati nel registro

Accanto al training set, per poter testare il predittore creato è stato utilizzato un validation set con 21 pazienti, provenienti dall'RITG dell'Ospedale di Mestre con infortuni gravi registrati nell'ultimo semestre 2010, con analoghe caratteristiche e attributi memorizzati del training set.

6.4 Calcolo TASH per la stima della probabilità di trasfusioni massive

Il TASH, ossia Trauma Associated Severe Hemorrhage, è un indice che misura la probabilità che sia necessaria una trasfusione massiva in un paziente traumatizzato: può essere calcolata dalla somma pesata dei parametri fisiologici di pressione sanguigna, sesso, emoglobina, presenza di FAST (focused assessment sonography in trauma), battito cardiaco, alcalosi o base excess e fratture alla pelvi o alle gambe inferiori. Grazie alla sua semplicità e alla sua buona accuratezza rispetto agli altri indici è considerato uno degli standard clinici nella previsione di trasfusioni massive. È stato calcolato anche l'indice di probabilità trasfusionale secondo la formula

$$p = 1/[1 + \exp(4.9 - 0.3 * \text{TASH})]$$

sul set dati, in modo da avere un test clinico retrospettivo su una procedura, la cui robustezza e accuratezza fosse ben testata e dimostrata in letteratura^[13], utile a creare un confronto tra lo standard clinico ed il predittore implementato con le tecniche data mining. Si sono classificati come casi necessitanti di trasfusione massiva i pazienti con un indice di probabilità di MT superiore al 40 %, valore di soglia con cui si ottiene la massima accuratezza del predittore^[13]: il TASH sul nostro dataset è in grado di classificare correttamente il 70 % di pazienti effettivamente trasfusi ed il 98 % dei non trasfusi. Si è poi utilizzato l'indice TASH per valutare i casi del validation set, utili per il confronto con il predittore creato con tecniche data mining. La tabella con l'esito di classificazione TASH è riportata nella pagina seguente (tabella 6.2) dove nelle caselle, i veri positivi e i casi di massive transfusion sono indicati in giallo, in rosso i falsi positivi, e in arancione i veri negativi.

Dal validation set non è stato possibile comprendere se alcuni pazienti (3 casi) avessero subito trasfusioni massive, a causa della mancanza o della cattiva compilazione dei dati.

6.5 Costruzione di un predittore di Trasfusioni Massive con gli Alberi di

Decisione

Il data mining, diventa utile specie quando vi sono una grossa quantità di attributi a disposizione ed un rilevante numero di pazienti come quelli considerati nell'analisi e nel dataset ricavato dal registro traumi.

A tal proposito si sono considerati gli algoritmi per la costruzione di alberi decisionali, metodologia che fornisce risultati facilmente interpretabili e, alla luce dell'ambiente di studio, quindi più facilmente comunicabili al personale medico, per un eventuale utilizzo pratico; l'esigenza di semplicità è una necessità in ambienti di emergenza dove gli altri indici di previsione di trasfusione massiva utilizzati puntano su rapidità, immediatezza e semplicità in primis.

indice Hb	indice be	indice SBP	indice sex	indice FIA	indice artiinfabacino	Unità Plasma	Emazie	Sangue intero	MT	TASH	probabilità TASH
0	1	1	1	0	0	0	0	0	F	3	1.798620996
8	4	4	1	3	6	16	4	20	T	23	88.0797078
0	0	0	1	0	0	0	0	0	F	1	0.995180187
0	0	0	1	0	0	0	0	0	F	1	0.995180187
0	0	0	1	0	0	0	0	0	F	1	0.995180187
4	4	4	1	0	0	7	7	14	T	13	26.89414214
0	1	4	1	0	0	0	0	0	F	6	4.310725494
0	0	0	1	0	3	4	3	7	F	4	2.412702142
0	1	0	1	0	0	0	0	0	F	2	1.338691783
6	1	0	0	0	3	1	3	4	F	10	13.01084744
0	0	0	1	0	0	0	0	0	F	1	0.995180187
8	4	0	1	0	3	4	6	10	T	16	47.50208125
0	0	4	1	0	3	0	0	0	F	8	7.585818002
0	0	0	0	0	0	0	0	0	F	0	0.739154134
0	3	0	0	0	3	0	?	0	F	6	4.310725494
2	1	1	1	0	3	5	9	14	T	8	7.585818002
8	4	0	1	0	3	0	0	0	F	16	47.50208125

TABELLA 6.2 – Calcolo del TASH per i pazienti del validation set e relativi indici per il calcolo

Uno degli standard attualmente più considerati tra gli algoritmi data mining per la costruzione di decision tree, è il software C5.0 di Ross Quinlan, che è stato dimostrato avere esiti estremamente accurati in esperimenti di classificazione, gestendo anche un numero estremamente consistente di parametri, anche in ambito clinico e bioinformatico^[31].

6.5.1 L'utilizzo di C5.0 : i test sul dataset proveniente dal RITG

Per l'addestramento ed implementazione dei decision tree, si sono considerati i dati del dataset proveniente dal registro RITG. Si è testato poi l'albero di decisione, così creato, su un validation set, ottenuto dai dati del registro traumi archiviati più di recente, per vedere l'effettiva capacità del predittore di individuare le trasfusioni massive anche in nuovi casi, su cui non è stato addestrato l'algoritmo.

Per ogni test è stato necessario preparare il dataset e memorizzare i parametri e i dati secondo le convenzioni richieste dal programma C5.0. Una volta costruito il predittore, viene fornita dal software una schermata che indica l'accuratezza e gli errori di classificazione dell'algoritmo e mostra una rappresentazione grafica dell'albero di decisione.

In fase di validazione si è utilizzato invece un programma interactive interpreter associato a C5.0, che utilizza un file con il test da validare ed un file con i parametri, come quelli usati per creare il classificatore, ma a differenza dell'algoritmo per l'apprendimento utilizza l'albero di decisione appena costruito per effettuare la predizione sui nuovi dati.

Gli esiti del validation set saranno utilizzati per il confronto tra i vari test di classificazione, per verificare la scelta migliore dei parametri dal dataset, da utilizzare per implementare l'albero.

6.5.2 Analisi dei task effettuati

Le prove sperimentali successive sono state effettuate considerando un numero di parametri crescente con lo scopo di migliorare l'accuratezza del classificatore, prima selezionando i principali parametri in base alle conoscenze mediche a priori, poi cercando di utilizzare strategie per massimizzare la quantità di informazione utili per creare il classificatore, accettando tutti gli attributi disponibili nel dataset che non fossero tra loro correlati e quindi ridondanti.

Nella prima prova (**test 1**) si sono considerati i soli attributi utilizzati per il calcolo del TASH allo scopo di creare una prova su attributi che dall'esperienza clinica si considerano essere molto rilevanti nella predizione delle trasfusioni massive, e vedere se una metodologia machine learning, potesse portare alle stesse osservazioni e pesare allo stesso modo i parametri di una metodologia basata solo su considerazioni medico-cliniche.

Il classificatore ottenuto ha costruito un albero, utilizzando solo tre parametri tra quelli disponibili, ritenendo il più importante, la radice dell'albero, la pressione sistolica seguita dal sesso e dalla presenza di fluidi intraddominali, pesando quindi in maniera radicalmente differente i dati rispetto all'indice TASH. Il predittore sul training set ha il 2.5 % di errore, margine di errore causato però dalla mancata predizione di trasfusione massiva tra gli effettivamente trasfusi massivi, dove l'errore è del 30 %, paragonabile a quello ottenuto dall'indice clinico. Nel validation set l'indice di predittività si avvicina al 25%, valore inferiore a quello ottenuto con il TASH (vedi tabella 6.8 in Risultati Finali).

Successivamente per creare il training set è stato considerato un numero maggiore di parametri del dataset, sulla base di considerazioni cliniche (**test 2**), quindi prendendo in esame attributi legati a parametri ematici (come PH, HCO₃, INR, Base excess) , vitali (come la frequenza respiratoria, la pressione sistolica, l'SpO₂) e indici di gravità (come gli Abbreviated Injury Score locali, e i severity score), che sono anche i principali attributi utilizzati in letteratura nelle predizioni di outcome o di severe coagulopatie^{[9] [11] [14] [13]}

In questo test in particolare si sono selezionati i dati relativi agli attributi di tipo numerico perché più semplici da implementare nell'algoritmo e perché possono indurre un risultato più performante rispetto altri attributi: i dati numerici ad un'analisi dei parametri relativi ai trasfusi massivi, sembrano caratterizzare meglio i soggetti rispetto tipi di dati, alla luce della possibilità di essere usati come valori su cui ramificare l'albero. Tra gli attributi compilati più volte in diverse fasi si sono selezionati quelli maggiormente compilati a seconda della fase, e considerando anche i parametri relativi agli indici di gravità che possono essere ottenuti solo dopo un tempo consistente di intervento sul paziente, trascurando per ora la necessità di un modello che possa essere costruito con i soli parametri relativi alle primissime fasi di soccorso, e che quindi sia utilizzabile per una

immediata predizione di trasfusione massiva. Lo scopo di questa scelta di parametri è dovuto in questa fase all'elaborazione un predittore prima di tutto accurato .

In fase di apprendimento del classificatore si è usata la tecnica dei “boosted tree” descritta nella sezione seguente, in cui vengono utilizzati più alberi decisionali “in parallelo” . Questa tecnica ha permesso di aumentare sensibilmente la predittività, tuttavia rendendo il modello meno semplice da comprendere da un personale esperto in quanto il predittore con questa opzione diventa un insieme di alberi piuttosto che un albero singolo, più facilmente implementabile come modello decisionale. Quest’approccio risulta necessario per migliorare l’accuratezza di predizione rispetto al classificatore costruito con un albero di decisione singolo: l’errore totale sul training set passa infatti dal 12 % al 2.5% nel test con il “boost” , in cui non vi sono falsi negativi ed i veri positivi si attestano sul 80 %. Nella tabella allegata alla sezione (tabella 6.3), sono descritti i parametri utilizzati dal training set e quelli poi effettivamente utilizzati nell’albero di decisione per classificare le massive transfusion: le percentuali indicano con che frequenza questi siano stati considerati nel classificare i casi. I risultati sul validation set sono migliori rispetto al precedente test 1, come si può vedere dalla tabella nel paragrafo Risultati finali, mostrando una predittività del 25 % e un azzeramento dei falsi negativi, migliorando i risultati del TASH dove l’errore di questo genere era del 5%.

Si è poi proceduto ad ingrandire il set di dati, andando ad inserire non solo parametri di tipo numerico razionale o ordinale utilizzati nel precedente test 2, ma anche di tipo dicotomico e multicotomico, scegliendo dati che avessero una possibile correlazione medica o che potessero descrivere in qualche modo lo scenario clinico per un paziente trasfuso massivamente, in base ad informazioni a priori (**test 3**) .Si sono quindi considerati anche campi in cui venivano registrate presenza o meno di particolari procedure come la decompressione toracica,o la presenza di incannulamenti e intubazioni (ad esempio per fornire ossigeno puro),o la tipologia di trauma e il tipo di esami radiografici fatti nelle prime ore, ritenendo che questi parametri potessero contenere ulteriori informazioni caratterizzanti il trasfuso massivo.

L’albero di decisione così creato ha una predittività del 60 %, ma sul validation set si comporta in maniera analoga al quello creato nel test 2 come è possibile vedere dalle tabelle presentate per la conclusione delle analisi (tabella 6.8).

ATTRIBUTI UTILIZZATI PER IL TRAINING SET	
Sesso	
Età	
Tipo Trauma	
Tempo chiamata arrivo	
Pressione sistolica	
SpO2	
Frequenza respiratoria atti	
Polso	
fluidi Endo Venosi	
INR	
PO2	
PCO2	
Indice emoglobina	
RTS	
NISS	
TRNISS	
Gravità AIS pelvi arti inferiori	
Gravità AIS addominali	
Baseexcess	
Gravità AIS testa	
Gravità AIS collo	
Gravità AIS faccia	
Gravità AIS torace	
Gravità AIS rachide cervicale	
Gravità AIS rachide dorsale	
Gravità AIS rachide lombare	
Gravità AIS arti superiori	
Gravità AIS esterno-lesione termica	
Gravità AIS altre lesioni	

ATTRIBUTI DEL DECISION TREE	
100% Età	
99% NISS	
97 Tempo chiamata arrivo	
96% Gravità AIS faccia	
96% Gravità AIS rachide dorsale	
94% RTS	
93% Gravità AIS arti sup	
93% Tipo Trauma	
91% sesso	
84% Pressione sistolica	
82% fluidi EndoVena	
78% Gravità AIS torace	
76% Gravità AIS pelvi arti inferiori	
72% Gravità AIS collo	
58% Indice emoglobina	
57% Gravità AIS addominali	
53% Frequenza respiratoria atti:	
52% base excess	
51% PO2	
46% Gravità AIS rachidecervi	

TABELLA 6.3 Attributi considerati nel test 2 per la creazione del training set ed attributi utilizzati da C5.0 per costruire l'albero con percentuali di utilizzo nella classificazione

6.5.3 L'algoritmo di boosting

C5.0 annovera tra le sue opzioni una funzione di boost: l'utilizzo di questa caratteristica può incrementare notevolmente le prestazioni in fatto di accuratezza del predittore MT ed è stata utilizzata in tutte le prove sperimentali effettuate, vista la grande efficienza e rapidità di elaborazione, ad eccezione del test 1 dove la scarsa accuratezza del classificatore iniziale ne impediva l'implementazione.

Il boosting implementato in C5.0 è simile ad ADABOOST un algoritmo elaborato da Freund e Shapirò^[34] per studi di machine learning ed intelligenza artificiale: con boost si indica uno strumento algoritmico generico e testato per la produzione di accurate e robuste regole di predizione combinando tra di loro, classificatori a regole decisionali più grezzi e semplici.

Il pseudocodice di ADABOOST, il più considerato metodo di boosting è riportato in figura sottostante (figura 1)^[34].

ADABOOST si è visto essere capace di migliorare le prestazioni di numerosi algoritmi di apprendimento, nel caso di classificatori semplici le migliorie sono notevoli, e si può pensare di considerare quest'algoritmo come parte integrante dell'addestramento

del classificatore stesso: non stupisce quindi che il boost sia diventata una delle aggiunte di C5.0 più elogiate. L'idea base dietro il boosting è di creare diversi classificatori invece che uno solo: ciascuno di essi fornirà una predizione sulla classe di appartenenza, e le diverse predizioni saranno contate per determinare la classe finale. Il primo passo, in C5.0, è stato quindi costruire un albero decisione utilizzando il completo dataset; questo classificatore molto probabilmente commetterà degli errori su alcuni casi nei dati.

Given: $(x_1, y_1), \dots, (x_m, y_m)$ where $x_i \in X$ and $y_i \in Y = \{-1, +1\}$
Initialize $D_1(i) = 1/m$
For $t = 1, \dots, T$:

- Train weak learner using distribution D_t
- Get weak hypothesis $h_t : X \rightarrow \{-1, +1\}$ with error $\epsilon_t = \Pr_{D_t}[h_t(x_i) \neq y_i]$
- Choose $\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$
- Update
$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases}$$
$$= \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$
where Z_t is a normalization factor chosen so that D_{t+1} will be a distribution.

Output of the final hypothesis:

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right)$$

FIGURA 6.1 – Il pseudocodice di ADABOOST

Quando sarà costruito il secondo classificatore sarà riposta su questi casi più attenzione nel tentativo di correggere gli errori. In genere il secondo classificatore è diverso dal primo, sbaglierà tuttavia in alcuni casi e questi diventeranno l'obiettivo da migliorare per il classificatore costruito successivamente.

Il numero di classificatori che si possono costruire nelle fasi di boost è detto boost-trial è può essere gestito dall'utente che ne deciderà la quantità.

Si è cercato in questa analisi di vedere quale fosse il numero di prove, che ottimizzasse l'accuratezza nella costruzione del classificatore finale a fronte della complessità computazionale e del tempo di elaborazione: il valore ottimale di trial è risultato essere intorno al 10, come avvalorato anche da uno studio in letteratura [33] che dimostra che oltre un certo numero di boost trial le prestazioni in accuratezza dell'algoritmo non migliorano.

6.5.4 La costruzione del Modello con gli attributi più predittivi

Uno dei punti di forza del TASH, come detto, è il suo essere un indice calcolabile su parametri rilevabili nella prima ora dopo l'ingresso in ospedale, riuscendo ad anticipare diagnosi di emorragie gravi ottenibili solo con TAC o esami di laboratorio accurati. Tuttavia questo modello si basa anche su attributi clinici, come la gravità dell'infortunio addominale e dell'infortunio pelvi-arti inferiori, che necessitano di una stima di danni fisici per i quali difficilmente, si riesce ad arrivare ad una misura non arbitraria e precisa nelle prime ore dell'emergenza. Con lo scopo di costruire un predittore con alberi di decisione che potesse utilizzare dei parametri ottenibili solo nei primi minuti

o nella prima ora di soccorso e che potesse quindi “anticipare” l’esito di un eventuale predizione TASH, si è implementato il classificatore utilizzando un training set composto da soggetti in cui venivano considerati solo i parametri relativi alle prime fasi di soccorso (**test 4**). Gli attributi considerati sono mostrati nella tabella sottostante (tabella 6.4). L’esito della fase di addestramento e costruzione dell’albero tuttavia porta ad un predittore inaccurato nel classificare soprattutto i casi di trasfusione massiva, con un tasso di errore del 5.6% e con il 65% di trasfusi massivi classificati erroneamente. L’algoritmo non è in grado di iniziare la fase di boosting a causa degli errori negli alberi creati, in quanto la classificazione di ciascun albero di boost, avendo un alto tasso di errore è considerata inaffidabile per esprimere una predizione, e la somma di predizioni inaccurate rischia di peggiorare ulteriormente l’accuratezza finale. L’impossibilità di utilizzare la funzione boost è probabilmente determinante nella non accuratezza del predittore finale ottenuto: l’albero costruito è basato sui soli parametri di Pressione sistolica al momento del soccorso, sesso e base excess. Nel validation set, come è possibile vedere dalla tabella 6.8 nella parte relativa ai risultati finali, il predittore non è in grado di individuare alcuna delle trasfusioni massive risultando peggiore in quest’ottica rispetto al TASH.

PARAMETRI CONSIDERATI NELLA COSTRUZIONE DEL PREDITTORE
Sesso
Età
Tipo Trauma
GlasgowComaScore all’arrivo dell’ambulanza
Pressione sistolica all’arrivo dell’ambulanza
Frequenza Respiratoria Atti all’arrivo dell’ambulanza
Sp O2Specifica all’arrivo dell’ambulanza
Tempo Chiamata Arrivo in pronto soccorso
AirWay1 in pronto soccorso
IntubazionePreospedaliera
FrequenzaRespiratoriaAtti in pronto soccorso
SpO2Specifica in pronto soccorso
Decompressione Toracica
Pressione Sistolica in pronto soccorso
Polso in pronto soccorso
FluidiEndoVenosi in pronto soccorso
Effettuata Sedazione in pronto soccorso
GlasgowComaScore in pronto soccorso
INR
Ph
PO2
PCO2
HCO3
BaseExcess
Lattati
Indice EmoGlobina

TABELLA 6.4 – Parametri considerati nel test 4

6.5.5 La costruzione del Modello con tutti gli attributi presenti nel registro

L'utilizzo di un numero maggior di parametri ha permesso di arrivare a dei margini d'errore totale di classificazione nel training set paragonabili all'indice TASH, sebbene i falsi positivi siano ancora in numero consistente. Si è provato quindi di utilizzare tutti gli attributi disponibili nel database RITG indipendentemente dalla loro importanza clinica correlata al rischio di trasfusioni massive, con l'obiettivo di aumentare l'informazione da fornire al classificatore, per aumentarne l'accuratezza, ma con il rischio di incrementare l'entropia di informazione: infatti l'algoritmo fallisce nel tentativo di creare l'albero di decisione a causa dei troppi errori in fase di addestramento.

Si è deciso pertanto di applicare delle metodiche per "filtrare" i dati ridondanti e limitarsi a dati indipendenti e scorrelati. Queste metodiche saranno descritte in dettaglio nel capitolo successivo. Con i dati ottenuti dalla procedura di pre-elaborazione dati possiamo utilizzare un numero cospicuo di parametri, rappresentanti tutti i dati del dataset del RITG ritenuti indipendenti e scorrelati, per costruire l'albero di decisione. Questo consente di sfruttare tutta l'informazione disponibile nel registro minimizzando le fonti di ridondanza, per ottenere una maggiore predittività del classificatore.

6.6 Pre-selezione Dati

Come visto precedentemente il database RITG contiene dati di varia tipologia, presi in diversi momenti durante la prestazione d'emergenza: ci sono quindi attributi uguali presi in fase di primo soccorso ed in tempi successivi. Tuttavia i dati non sono sempre compilati, e spesso hanno lo stesso valore. Si è quindi cercato di limitare la ridondanza dei dati, tenendo conto per l'analisi, degli attributi che avevano più campi compilati in tempi diversi. Le coppie omologhe eliminate sono state la *Glasgow coma score*, la frequenza respiratoria registrata in pronto soccorso e la pressione sistolica registrata al momento dell'intervento.

6.6.1 L'analisi di Correlazione

Si è poi proceduto all'analisi di correlazione tra i vari attributi presenti: questa analisi ha portato alla creazione di un grafo connesso con tutti gli attributi più strettamente correlati. L'obiettivo di questa selezione di parametri è la creazione di un dataset il più possibile non ridondante, con un numero comunque importante d'attributi ai fini della costruzione del classificatore. L'eliminazione di questi parametri è un contributo alla diminuzione del rumore dovuto alla ridondanza dei parametri stessi, e quindi è una possibile limitazione dell'entropia più che una perdita di informazione. Per individuare i parametri non indipendenti si è cercato di determinare quali campi fossero correlati in base al calcolo dell'indice di correlazione, un indice statistico per determinare quanto fortemente possano essere relazionate due variabili causali.

Per il calcolo delle correlazione si è proceduto ad analizzare i dati in R, un ambiente di sviluppo per l'analisi statistica, software che permette di gestire dataset con campi non compilati o segnati da rumore statistico. Per i dati di ciascun attributo si è studiata la correlazione con i dati di ciascun altro attributo del dataset calcolandone l'indice. Avendo dati di tipo numerico razionale, numerico ordinale, e dicotomici è stato necessario calcolare indici diversi a seconda del tipo di dato correlato.

Vengono presentati i vari indici di correlazione calcolati:

Tra due attributi quantitativi o razionali: Pearson r

Il coefficiente di Pearson è il più comunemente studiato e analizzato indice di correlazione per dati statistici, tuttavia non si adatta ai soli dati di tipo quantitativo nella sua formulazione originale.

Date due variabili aleatorie X e Y, indice di correlazione di Pearson è definito come la loro covarianza divisa per il prodotto delle deviazioni standard delle due variabili:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

dove

σ_{xy} , è la covarianza tra X e Y

σ_x, σ_y , sono le due deviazioni standard

Il coefficiente assume sempre valori compresi tra -1 e 1:

$$-1 \leq \rho_{xy} \leq 1$$

Tra due attributi ordinali: Spearman ρ

Non è infrequente nei nostri dati trovarsi di fronte a tipologie di dati diverse da quelle razionale o quantitativa, come i dati di tipo ordinale: si pensi alle scale GCS o agli indici di gravità. Per questi dati si calcola l'indice di correlazione di Spearman secondo la formula:

$$\rho_s = \frac{\sum_i (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_i (r_i - \bar{r})^2} \sqrt{\sum_i (s_i - \bar{s})^2}}$$

Dove gli n punteggi delle variabili ordinali X e Y, sono convertiti in ranghi r ed s rispettivamente, ed r_i ed s_i sono i ranghi all'i-esima osservazione.

Il coefficiente ρ è semplicemente un caso particolare del coefficiente di correlazione di Pearson dove i valori vengono convertiti in ranghi prima di calcolare il coefficiente, anche se solitamente si segue un calcolo più semplice, in quanto si calcola la differenza D tra i ranghi delle due misure di un'osservazione, ottenendo così

$$\rho_s = 1 - \frac{6 \sum_i D_i^2}{N(N^2 - 1)}$$

dove

$D_i = r_i - s_i$ è la differenza dei ranghi, essendo r_i e s_i rispettivamente il rango della prima variabile e della seconda variabile della i -esima osservazione

N il numero complessivo di osservazioni

Tra attributi quantitativo razionale o numerico e booleano: *Point biserial*

r_{pb} è il coefficiente di correlazione tra un attributo nominale booleano ed uno numerico ed è un caso particolare dell'indice di Pearson, in formule. Detta X la variabile con attributi booleani e Y quella con attributi numerici

$$r_{pb} = \frac{M_1 - M_0}{s_y} \sqrt{pq},$$

dove M_0 e M_1 sono i punteggi medi dei valori di Y associati ai valori di X con un punteggio di 0 o 1 rispettivamente, $q = 1 - p$ e p sono le proporzioni di dati con punteggio in X di 0 o 1, rispettivamente, e s_y è la deviazione standard per la popolazione di dati y .

Tra due attributi di tipo booleano: *Phi*

Se entrambe le variabili sono di tipo booleano allora l'indice di correlazione può essere calcolato in maniera semplificata dall'indice di Pearson, utilizzando la tabella di contingenza, una tabella bidimensionale con le frequenze per ogni categoria

$Y \backslash X$	0	1	Totali
1	A	B	A + B
0	C	D	C + D
Totale	A + B	C + D	N

La formula è :

$$\phi = (BC - AD) / \sqrt{(A+B)(C+D)(A+C)(B+D)}.$$

che non è altro che la formula di Pearson modificata e semplificata per un caso completamente booleano: a livello computazionale quindi la formulazione non cambia e può essere utilizzata la formula di Pearson implementata precedentemente.

Tra attributi di tipo booleano e ordinale: *D Somers*

Assieme al coefficiente rank-biserial è l'indice di correlazione usato tra dati booleani e di tipo ordinale: sia D l'indice di correlazione relativamente a due variabili C ed R , $C|R$ indica che la variabile C è considerata la variabile indipendente, mentre la R si considerata

dipendente, con n numero osservazioni, e P-Q che misura l'eccesso di coppie discordanti rispetto le coppie concordi,

$$D(C|R) = \frac{P-Q}{w_r}$$

considerando

$$w_r = n^2 = \sum_i n_i^2$$

Tra attributi di tipo ordinale e numerico:

Si utilizza in questo caso in coefficiente di correlazione studiato per questa particolare situazione, molto frequente nello studio di questi dati ma meno considerata in ambito statistico.

Le equazioni che descrivono l'indice considerando che per una variabile X esistono N ranghi che ed R è la somma di questi i ranghi, sono le seguenti

$$\sum_i^N R_i = N(N+1)$$

Indicando X la variabile ordinale ed Y la variabile quantitativa:

$$\sum_i^k X_i Y_i = \sum_i^k R_i Y_i$$

Il cui indice di correlazione sarà

$$r_{ri} = \frac{\sum_i^k R_i Y_i / N - [(N+1) \sum_i^k Y_i]}{\sqrt{(N^2 - 1) / 12 (s_y^2)}}$$

6.6.2 Costruzione e ordinamento del Grafo disconnesso

A seguito del calcolo degli indici di correlazione, si è cercato di trovare una soglia oltre cui due valori potessero ritenersi associati. Si sono utilizzati i grafici relativi ad ogni correlazione, in cui sull'asse delle x era indicato un indice numerico tra 0 e 1 e sull'asse delle y il numero di coppie ritenute correlate utilizzando come soglia quell'indice. Si è cercato quindi il valore per cui la curva della funzione del numero di correlati tendesse ad aver un cambio di derivata consistente per ciascuna correlazione tra tipologie dati diverse. Dopo avere identificato le coppie con indice di correlazione superiore a questa soglia come coppie di valori correlati, si è proceduto a creare un Grafo delle Correlazioni con queste coppie. I nodi del grafo rappresentano parametri e gli archi rappresentano le correlazioni, è possibile che il grafo sia costituito da più sottografi sconnessi.

L'obiettivo è tenere il parametro o i parametri contenenti più informazione, ossia quelli con grado massimo all'interno dei vari Grafi di correlazione che si saranno creati, procedendo ad eliminare ad ogni passo i nodi collegati a quelli di grado massimo, in una sorta di algoritmo di ordinamento topologico. Nel caso di grado uguale si eliminerà il nodo il cui attributo è meno compilato nel dataset. L'obiettivo della procedura di ordinamento e eliminazione è ottenere un grafo finale che dovrà essere a) privo di archi; b) tale che ciascun nodo eliminato sia connesso ad almeno uno dei suoi nodi; c) tra tutti i grafi ottenibili da quello di partenza e che rispettano a) e b) deve avere il minimo numero di nodi. In seguito a questa analisi i parametri filtrati in quanto considerati ridondanti da questa procedura sono i seguenti:

HCO3
<u>RianimazioneCardiopulmonare</u>
PressioneSistolica in pronto soccorso
<u>DecompressioneToracica</u>
<u>TRISS</u>
FrequenzaRespiratoriaAtti all'arrivo dell'ambulanza
GlasgowComaScore_ all'arrivo dell'ambulanza
VentilazioneArtificiale
RaggiX Torace
Raggi X Rachide Cervicale
Incannulamento Arteria
Eco Addome
baseexcess

TABELLA 6.5 - Parametri eliminati dall'analisi

Tra i parametri correlati sono stati considerati quelli più importanti e con più contenuto informativo:

Incannulamento Vena Centale
Rx Bacino
Drenaggio Toracico
Ph
IntubazioneTracheale
Sedazione

TABELLA 6.6 – Parametri a più alto contenuto informativo tra i correlati

Riassumendo in questo test sono stati scelti i seguenti parametri per creare il training set:

età
Sesso
SpO2specifica nel
INR
PO2
Ph
TempoChiamataArrivo
PCO2
BaseExces
Lattati
Indice Emoglobina
Sedazione
IntubazioneTracheale
IncannulamentoVenaPeriferica
IncannulamentoVenaCentrale
RXBacino
Tipo_trauma
AirWay1
DrenaggioToracico
Gravità AIS Testa
Gravità AIS Collo
Gravità AIS Faccia
Gravità AIS Rachide Cervicale
Gravità AIS Rachide Dorsale
Intubazione pre ospedaliera
Tempo normalizzazione base excess
Gravità AIS Rachide Lombare
Gravità AIS addominale
Gravità AIS Arti Superiori
Gravità AIS estremo-lesioni termiche
RTS
ISS
NISS
TRNISS
Fluidi endovena
Pressione al Polso in pronto soccorso
Cannule
Calibro Massimo
Infusione sangue
Gravità AIS arti inferiori e pelvi

TABELLA 6.7-Parametri considerati nel test 5

6.7 Risultati finali

Completata la scelta dei parametri, si è proceduto all'addestramento e costruzione dell'albero. I risultati in fase di training sono ottimi: il decision tree ha lo 0.7 % di tasso di errore sul totale del set di addestramento, e il 90 % di indice di predittività sulle trasfusioni massive. Anche il comportamento con il validation set dà risultati interessanti e migliori delle prestazioni dell'indice TASH: è accurato al 92 % anche se presenta 50 % di falsi positivi come il TASH.

Nelle pagine seguenti viene presentata la tabella (tabella 6.8) con il confronto tra i vari esperimenti fatti costruendo decision tree con diverso tipo e numero di attributi. Da sinistra vengono indicati: i soggetti esaminati, la presenza di trasfusione massiva(t o f a seconda che il paziente rispettivamente sia stato trasfuso o meno) e gli esiti dei test effettuati con i predittori, in

blu le caselle con i veri negativi, in giallo i veri positivi, in rosso i falsi negativi ed in viola i falsi positivi.

I primi 3 test migliorano progressivamente il loro comportamento nel validation set, fino al terzo task ed arriva a risultati migliori del TASH sul validation set, non avendo però altrettanta accuratezza in fase di addestramento, dove risulta meno performante nel classificare i soggetti effettivamente trasfusi.

Il **test 4** basato invece sui dati solo sulle prime fasi di soccorso non è accurato quanto le altre prove di classificazione, né in fase di addestramento né in fase di validazione: questo a testimoniare una difficoltà notevole dell'algoritmo C5.0 nel costruire un albero con pochi campi rispetto ai test 3 e 5, in cui questi attributi siano compilati scarsamente: i parametri legati alle prime fasi di soccorso sono spesso incompleti o mal compilati, al contrario dei parametri di gravità o di indice clinico che sono meglio registrati, e sono relativi a fasi più avanzate di assistenza al paziente. Nella costruzione di un albero di decisione quindi è preferibile puntare ad usare più attributi e scegliendo per il training set un insieme Y con i parametri meglio compilati. Tuttavia questo task rappresenta un tentativo di utilizzare dati che permetterebbero se utilizzati nella pratica clinica di anticipare notevolmente rispetto al TASH la predizione di una possibile terapia con trasfusione massiva, in quanto l'indice utilizzato nella pratica clinica sfrutta anche parametri legati agli indici di gravità che non sono considerati in questo modello, e che sono relativi ad una fase di soccorso più avanzata.

L'ultimo test (**test 5**) svolto permette di costruire un predittore con albero di decisione accurato e con comportamento analogo all'indice clinico TASH. Facendo un confronto tra la predizione di quest'ultimo e i classificatori elaborati con C5.0, per tutti i casi del validation set compresi quelli di cui non vi sono informazioni sulle trasfusioni, è possibile vedere come gli esiti predittivi il più delle volte si equivalgano. Tuttavia è necessario incrementare il numero di veri positivi per tutti gli indici, per avere una metodologia che possa essere più affidabile da un punto di vista clinico. Questi test hanno permesso comunque di offrire un'indicazione su quali possano essere i parametri più importanti da utilizzare per la costruzione di un modello predittivo di trasfusione.

In particolare nell'ultimo test (**test 5**) effettuato l'albero di decisione è costruito utilizzando i parametri di tabella 6.9, illustrati con le relative percentuali di utilizzo nella costruzione e classificazione da parte dell'algoritmo: per questo task di C5.0 aventi un numero consistente di attributi può essere utile saper come individualmente questi attributi contribuiscono al classificatore. Per ogni attributo è indicata la percentuale di quanto il valore di tale attributo è usato nel predire una classe, quindi nel nostro albero sappiamo che l'età è utilizzata quando si classificano il 100 % dei casi di training

Soggetto	Trasfusione massiva	probabilità TASH	Test 1	Test 2	Test 3	Test 4	Test 5
Paziente1	f	13.01084744	f	f	f	f	f
Paziente 2	t	7.585818002	f	f	t	f	t
Paziente 3	t	47.50208125	f	f	f	f	f
Paziente 4	t	26.89414214	f	f	f	f	f
Paziente 5	f	2.412702142	f	f	f	f	f
Paziente 6	f	0.739154134	f	f	f	f	f
Paziente 7	f	4.310725494	f	f	f	f	f
Paziente 8	f	0.995180187	f	f	f	f	f
Paziente 9	t	88.0797078	t	t	t	f	t
Paziente 10	f	47.50208125	f	f	f	f	f
Paziente 11	f	0.995180187	f	f	f	f	f
Paziente 12	f	0.995180187	f	f	f	f	f
Paziente 13	f	1.338691783	f	f	f	f	f
Paziente 14	f	4.310725494	f	f	f	f	f
Paziente 15	f	0.995180187	f	f	f	f	f
Paziente 16	f	7.585818002	f	f	f	f	f
Paziente 17	f	1.798620996	f	f	?	f	f

TABELLA 6.8- Confronto tra i vari test effettuati e l'esito del TASH nel validation set

100% età
98 % sesso
98% tipo trauma
96% gr_testa
96% gr_rachidecervi
96% gr addominale
96% gr_artisup
96% gravità arti inferiori e pelvi
95% ISS
95% NISS
86% gravità rachidedorsale
83% gravità torace
82% TRNISS
73% fluidi endovena
68% Incannulamento vena
65% tempo chiamata
65% Spo2
63% RTS
62% Infusione sangue
55% intubazione preospedaliera
47% PO2
43% indice emoglobina
41% PS CalibroMax
38% PS cannule
33% PS polso
25% gravità rachidelombare
25% Rx bacino
23% airway
16% lattati

Tabella 6.9 – Parametri utilizzati nel test 5 e relative percentuali di utilizzo nel classificatore

Tra gli attributi più utilizzati nell'albero di decisione (che anche in questo caso è in realtà un boosted tree) costruito dall'algoritmo vi sono alcuni dei parametri utilizzati dal TASH, ossia indice gravità addominale e degli arti inferiori e pelvi, ed alcuni degli attributi clinici, come ISS e NISS, che sono considerati i più importanti nella stima di outcome o di coagulopatie. Tuttavia parametri verificati essere fortemente predittivi come il ph e il base excess non sono stati considerati nell'analisi perché eliminati dalla preselezione in quanto contenenti ridondanza di informazione. Queste defezioni non si riflettono sull'accuratezza del predittore finale che è anzi maggiore per il test in cui queste caratteristiche non sono state considerate tra gli attributi su cui costruire il dataset. Anche il tasso di emoglobina, che è uno dei valori che influenzano maggiormente il calcolo dell'indice TASH pesando fortemente per valori critici sulla probabilità di trasfusione massiva, viene poco considerato dall'albero ed in ramificazioni profonde in questo ultimo test.

6.8 Conclusioni e sviluppi

L'utilizzo di una tecnica di data mining semplice, ma comunque sofisticata, come l'albero di decisione, ci ha permesso di creare un indice predittivo di trasfusioni massive paragonabile come prestazione all'indice clinico più accurato e considerato dal personale medico esperto: è tuttavia una metodologia non esente da errori che, in una ipotetica applicazione in ambito clinico "bed-side", risulterebbero estremamente rilevanti. Va in ogni caso evidenziato come il set di dati sia ristretto ad un numero di casi piuttosto esiguo rispetto al numero che C5.0 è in grado di elaborare: il software permette infatti risultati migliori su dataset con numeri più importanti di istanze, per scongiurare i problemi di overfitting, tuttavia una grossa mole di dati potrebbe influire sui costi computazionali che in questa analisi non sono mai stati considerati in quanto il tempo di elaborazione dei predittori è stato sempre inferiore ai due secondi. Il riscontro di un non trascurabile numero di parametri mal compilato o non compilato del tutto, rappresenta una possibile fonte di rumore nel training set atto a costruire il predittore, e questo rumore si riflette pesantemente sul classificatore creato. La scarsa completezza di dati del registro è dovuta principalmente alla grandezza stessa del database che rende difficile per il personale addetto l'accurata compilazione di tutti i campi; considerando che il predittore migliore utilizza solo alcuni di questi parametri e che l'albero di decisione costruito sfrutta una quantità di attributi ancora minore, la selezione di questi può fornire comunque al personale sanitario le informazioni necessarie per stabilire quali siano i campi chiave più importanti nella compilazione del registro.

Tecniche data mining snelle e computazionalmente poco onerose, come gli alberi di decisione, possono essere implementate anche direttamente nelle applicazioni che gestiscono il

database, per ottenere più rapidamente una predizione immediatamente successiva o al momento della registrazione dei dati. L'attuale database con il registro traumi è comunque compilato a seguito del trattamento del paziente, quindi i risultati ottenibili con i dati RITG potranno essere unicamente di tipo retrospettivo; è invece più interessante studiare l'implementazione di modelli predittivi all'interno di strutture dati costruite su parametri rilevati in tempo reale durante le prime fasi di soccorso, al fine di realizzare un predittore con effettiva utilità clinica. Attualmente i parametri nelle prime fasi di soccorso, se rilevati, sono registrati a mano; con l'evoluzione della telemedicina e dei dispositivi di registrazione e rilevazione dei parametri vitali, è possibile che i dati vengano direttamente immessi in un registro o in un database al momento della misurazione.

Il grosso vantaggio del TASH rispetto ai predittori calcolati dai decision tree sta soprattutto nella semplicità di calcolo e di derivazione anche "bed-side", sfruttando parametri ottenibili nei primi minuti dopo l'arrivo in ospedale. L'albero di decisione si basa anche su parametri che possono essere misurati in anticipo rispetto alcuni campi del TASH; tuttavia questi, in numero limitato e spesso mal compilati, non permettono da soli la costruzione di un albero accurato.

Un albero di decisione è una soluzione estremamente semplice e facilmente riproducibile ed interpretabile dal personale medico esperto; tuttavia per poter avere risultati accurati utilizzando questi dati, è sempre necessario l'ausilio di implementazioni più complesse come il boosting, che rendono il modello predittivo più sofisticato e quindi più difficilmente utilizzabile da parte del personale. Vista la difficoltà di creare alberi decisionali semplici, che offrano anche predizioni sufficientemente accurate, si pensa non sia possibile superare, in fatto d'immediatezza, l'indice basato su sole considerazioni cliniche, al momento dell'utilizzo in pratica medica. Sviluppando ulteriormente il modello predittivo con tecniche data mining sarà più facile puntare ad aver risultati più accurati, che non ad ottenere un indice più pratico, semplice e facile da implementare del TASH, che si basa su una semplice formula pesata e che all'atto pratico è affidabile al 70 %. Pertanto, i modelli predittivi data mining sono visti come soluzione di supporto piuttosto che come strumento per la diagnosi diretta, soprattutto in condizioni di emergenza in cui si preferisce grande celerità e praticità piuttosto che strumenti accurati che richiedano però tecnologie elaborate e ancora poco testate in ambito clinico. Sul piano dell'accuratezza però l'indice data mining utilizzato in questa analisi deve migliorare per essere avvalorato per qualsiasi ambito clinico di predizione delle MT, cercando in particolare di aumentare il numero di veri positivi partendo comunque dalle buone premesse date dai risultati in fase di apprendimento.

In questa tesi si è cercato di implementare un predittore per individuare i pazienti necessitanti di trasfusione massiva, non il numero unità di sangue trasfuso. Un'analisi successiva potrebbe essere fatta nell'ottica di predire la quantità di sangue trasfuso allo scopo di individuare le

moli di emazie o di plasma che verranno consumate, e quindi misurare preventivamente il dispendio di risorse critiche soprattutto in ambito militare: a tal proposito è necessario effettuare un'analisi su registri con dati più numerosi e meglio compilati, soprattutto nei campi legati alle trasfusioni, e considerare tecniche data mining più evolute nelle classificazioni di interi.

A meno di una registrazione informatizzata dei dati al momento del primo soccorso e del primo intervento, da cui un sistema automatizzato possa applicare un modello data mining affidabile e determinare una predizione, è difficile un utilizzo a scopo predittivo, che possa anticipare e risultare più pratico rispetto ad un indice basato sull'esperienza clinica, anche se non si può escludere il possibile miglioramento delle strutture informative ospedaliere e del pronto soccorso. Nel futuro sarebbe auspicabile lavorare non più su dati retrospettivi, ma su quelli provenienti in real time e ottenuti all'interno delle strutture di primo soccorso, fin dalle prime fasi dell'evento per un'analisi finalizzata all'ottimizzazione delle risorse cliniche e logistiche, tenendo d'occhio i costi.

Bibliografia

- [1] Sanson G, Nardi G, De Blasio E, Di Bartolomeo S, Moroni C, Serantoni C, “*La Prehospital Trauma Care - Approccio e trattamento al traumatizzato in fase preospedaliera e nella prima fase intraospedaliera*”
- [2] Bartolomeo Stefano; Nardi Giuseppe; Sanson, Gianfranco; Gordini, Giovanni; Michelutto, Vanni; Ciminello, Michela, Giugni Aimone; Cingolani, Emiliano; Cancellieri, Francesco, *The first Italian trauma registry of national relevance: methodology and initial results, European Journal of Emergency Medicine*: August 2006 - Volume 13
- [3] Hans Morten Lossius, Audun Langhelle, Eldar Søreide, Johan Pillgram-Larsen, Tine Askvik Lossius, Petter Laake, Petter Andreas Steen “*Reporting data following major trauma and analysing factors associated with outcome using the new Utstein style recommendations*”, *Resuscitation*, Volume 63, Issue 2, November 2004,
- [4] Ian Jacobs; Vinay Nadkarni and the ILCOR “*Task Force on Cardiac Arrest and Cardiopulmonary Resuscitation Outcomes, Update and Simplification of the Utstein Templates for Resuscitation*” Registries, Cardiac Arrest and Cardiopulmonary Resuscitation Outcome Reports
- [5] Kjetil G, Ringdal,, Timothy J, Coats, Rolf Lefering, Stefano Di Bartolomeo, Petter Andreas Steen, Olav Roise, Lauri Handolin, Hans Morten Lossius¹ and Utstein TCD expert panel “*The Utstein template for uniform reporting of data following major trauma: A joint revision by SCANTEM, TARN, DGU-TR and RITG*”, *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 2008
- [6] Cotton , Gunter OL, Isbell J, et al. “*Damage control hematology: the impact of a trauma exsanguination protocol on survival and blood product utilization*” *J Trauma*. 2008;64:1177–1182; discussion 1182– 1183.
- [7]. Tieu , Holcomb , Schreiber . “*Coagulopathy: its pathophysiology and treatment in the injured patient*”. *World J Surg*. 2007;31:1055–1064.
- [8] Malone DL, Hess JR, Fingerhut “*A. Massive transfusion practices around the globe and a suggestion for a common massive transfusion protocol*”. *J Trauma*. 2006;60:S91–S96.

[9] Timothy Nunez, Igor V. Voskresensky, Lesly . Dossett, MPH, Ricky Shinall, , William D. Dutton, Bryan A. Cotton, *Early Prediction of Massive Transfusion in Trauma: Simple as ABC (Assessment of Blood Consumption)?* J Trauma, 2009

[10] Bethesda, “*Comprehensive report on blood collection and transfusion in the United States in 2001*”: National Blood Data Resources Center, 2002.

[11] Brian J. Eastridge, Debra Malone, John B. Holcomb, FACS, “*Early Predictors of Transfusion and Mortality After Injury: A Review of the Data-Based Literature*”, J Trauma. 2006;60:S20–S25.

[12] *Patient Blood Management Guidelines: Module 1 | Critical Bleeding/Massive Transfusion*, Patient Blood Management Guidelines

[13] Nedim Yucel, Rolf Lefering, Marc Maegele, Matthias Vorweg, Thorsten Tjardes, Steffen Ruchholtz, Edmund. Neugebauer, Frank Wappler, Bertil Bouillon, Dieter Rixen, “*Polytrauma Study Group*” of the German Trauma Society Trauma Associated Severe Hemorrhage (TASH)-Score: Probability of Mass Transfusion as Surrogate for Life Threatening Hemorrhage after Multiple Trauma”, J Trauma. 2009;66:346–352.

[14] , Davis, James W, Parks, Steven N, Kaups, Krista L,; Gladen, Herbert E., O'Donnell-Nicol, Sheila, “*Admission Base Deficit Predicts Transfusion Requirements and Risk of Complications*”, The Journal of Trauma: Injury, Infection, and Critical Care

[15] Davis JW, Shackford SR, Hollbrook TL: “*Base deficit as a sensitive indicator of compensated shock and tissue oxygen utilization*” Surg Gynecol Obstet 173:473, 1991

[16] , Martin A Schreiber, Jeremy Perkins, Laszlo Kiraly, Samantha Underwood, Charles Wade, John B Holcomb, “*Early Predictors of Massive Transfusion in Combat Casualties*”

[17] Daniel F. McLaughlin, Sarah E. Niles, , Jose Salinas, Jeremy G. Perkins, E. Darrin Cox, Charles E. Wade, COL John B. Holcomb, “*A Predictive Model for Massive Transfusion in Combat Casualty Patients*”, J Trauma. 2008;64:S57–S63.

[18] Como JJ, Dutton RP, Scalea TM, Edelman BB, Hess JR. “*Blood transfusion rates in the care of acute trauma*”, Transfusion. 2004; 44:809 – 813.

[19] Michael J. A. Berry, Gordon S. Linoff, “*Data Mining*”, Apogeo, 2001

[20] Riccardo Bellazzia, Blaz Zupan “*Predictive data mining in clinical medicine: Current issues and guidelines, international journal of medical informatics*” 77 (2008) 81–97

- [21] G.Schwarzer,W.Vach,M.Schumacher, “*On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology*”, Stat. Med. 19 (2000) 541–561.
- [22] P.Clark,T.Niblett, “*The CN2 Induction Algorithm*”, Mach. Learn. 3 (1989) 261–283
- [23] N.Cristianini,J.Shawe-Taylor, “*An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*”, Cambridge University Press, Cambridge, UK, New York, 2000.
- [24] D.J. Hand, “*Data mining: statistics and more?*”, Am. Statist. 52 (1998) 112–118.
- [25] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, P. Tarczy-Hornoch, “*Data integration and genomic medicine*”, J. Biomed. Inform. 40 (2007) 5–16.
- [26] L.J. van’t Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, et al., “*Gene expression profiling predicts clinical outcome of breast cancer*”, Nature 415 (2002) 530–536.
- [27] Z. Hu, C. Fan, D.S. Oh, J.S. Marron, X. He, B.F. Qaqish, C. Livasy, L.A. Carey, et al., “*The molecular portraits of breast tumors are conserved across microarray platforms*, BMC” Genom. 7 (2006) 96.
- [28] J.D. Brenton, L.A. Carey, A.A. Ahmed, C. Caldas, “*Molecular classification and molecular forecasting of breast cancer: ready for clinical application?*” J. Clin. Oncol. 23 (2005) 7350–7360.
- [29] Janez Demsčar, Blaž Zupan, Noriaki Aoki, Matthew J. Wall, Thomas H. Granchi e, J. Robert Beck, “*Feature mining and predictive model construction from severe trauma patient’s data*, International Journal of Medical Informatics” 63 (2001) 41–50
- [30] J. R. Quinlan, C4.5, “*Programs for Machine Learning*”, Morgan Kaufmann Publishers, 1993.
- [31] S.B. Kotsiantis, Supervised Machine Learning, “*A Review of Classification Techniques*”, Informatica 31(2007) 249-268, 2007
- [32] J. R. Quinlan, “*Improved use of continuous attributes in c4.5*”, Journal of Artificial Intelligence Research, 4:77-90, 1996.
- [33] David Arditi, Thaveeporn Pulket, “*Predicting the Outcome of Construction Litigation*

[34] Freund Y, Schapire, “*A short introduction to boosting.*” J. Jpn. Soc. Artif. Intell., 14 5 , 771–780. 1999.

[35] Dick WF, Baskett PJ, “*Recommendations for uniform reporting of data following major trauma – the Utstein style. A report of a working party of the International Trauma Anaesthesia and Critical Care Society (ITACCS)*” . Resuscitation 1999, 42:81

[36] *Italian National Registry of Major Injuries: Registro Intraospedaliero Multiregionale Traumi Gravi.*

[37] Boyd CR, Tolson MA, Copes WS, *Evaluating trauma care: the TRISS method. Trauma Score and the Injury Severity Score.* J Trauma 1987, 27:370-378.

[37] R.J. Wherry S, “*Contributions to correlational analysis*”, Academic Press 1984

[38] C.J. Wyatt, D.G. Altman, “*Prognostic models: clinically useful or quickly forgotten?*” BMJ (1995) 311.