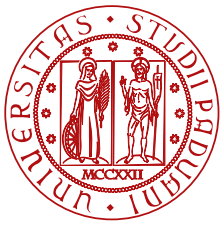

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

Sviluppo di una rete bayesiana dinamica per lo studio della progressione della sclerosi multipla

Relatore
Prof. Tavazzi Erica

Laureando
Rinaldi Matteo

Correlatore
Prof. Di Camillo Barbara

ANNO ACCADEMICO 2023-2024

Data di laurea 11/04/2024

Alla mia famiglia. Ai miei amici.

Abstract

Multiple sclerosis is a chronic inflammatory disease of the central nervous system that affects the brain and spinal cord and is caused by an over-response of the immune system that causes inflammation that damages myelin and neurons. It is the most common cause of neurological disability in young adults, with progressive problems (in motor, visual, sensory, and autonomic systems) that can be lifelong. The underlying causes of multiple sclerosis are varied and include genetic factors, environmental factors, and possible exposure to infectious agents. The prognosis of this disease is highly variable, resulting in high phenotypic heterogeneity.

In order to model the progression of multiple sclerosis, in this thesis a model was developed based on dynamic Bayesian networks, a methodology that allows the analysis of probabilistic relationships between variables monitored over time. To this end, demographic and clinical data from 1792 subjects with multiple sclerosis, made available within the European BRAINTEASER project, were used. After an initial phase of data preprocessing, the model was trained, obtaining a graph that allows visualizing how variables influence each other over time. From the network, interesting relationships between the patients' place of residence and the value of EDSS scale scores (a scale for assessing disability in people with multiple sclerosis) are revealed, highlighting the impact that air pollution has on the disease; a result that is reflected in the literature. From this network, two others were created by introducing variables regarding magnetic resonance imaging and drug therapies. Finally, in light of the evidence found in the first network, a fourth DBN was sought to be created by introducing data on environmental readings of two pollutants (PM10 and NO2), with the aim of further investigating their relationships with the disease.

After an extensive analysis of the relationships emerged in the networks, possible future developments of this methodology are discussed, introducing how, for example, starting with patient-specific data, the model could be used to predict the probabilistic evolution of the disease, towards an increasingly personalized medicine.

Abstract

La sclerosi multipla è una malattia cronica infiammatoria del sistema nervoso centrale che colpisce il cervello e il midollo spinale ed è causata da una risposta eccessiva del sistema immunitario che provoca un'infiammazione che danneggia la mielina e i neuroni. È la causa più comune di disabilità neurologica nei giovani adulti, con problemi progressivi (nei sistemi motori, visivi, sensoriali ed autonomi) che possono essere portati avanti tutta la vita. Le cause all'origine della sclerosi multipla sono varie e comprendono fattori genetici, ambientali, e la possibile esposizione ad agenti infettivi. La prognosi di questa malattia è molto variabile, risultando in un'elevata eterogeneità fenotipica.

Allo scopo di modellizzare la progressione della sclerosi multipla, in questa tesi si è sviluppato un modello basato su reti bayesiane dinamiche, una metodologia che permette di analizzare le relazioni probabilistiche tra variabili monitorate nel tempo. A tale fine, si sono utilizzati i dati demografici e clinici di 1792 soggetti affetti da sclerosi multipla, messi a disposizione all'interno del progetto europeo BRAINTEASER. Dopo una fase iniziale di preprocessing dei dati si è allenato il modello, ottenendo un grafo che permette di visualizzare come le variabili si influenzino tra di loro nel tempo. Dalla rete si evidenziano relazioni interessanti tra il luogo di residenza dei pazienti ed il valore dei punteggi della scala EDSS (una scala per la valutazione della disabilità nelle persone con sclerosi multipla), evidenziando l'impatto che l'inquinamento atmosferico ha sulla malattia; un risultato che trova riscontro in letteratura. A partire da questa rete ne sono state create altre due introducendo variabili riguardanti la risonanza magnetica e le terapie farmacologiche. Infine, alla luce delle evidenze riscontrate nella prima rete si è voluto creare una quarta DBN introducendo dei dati sulle rilevazioni ambientali di due inquinanti (PM10 e NO2), con l'obiettivo di indagare ulteriormente le loro relazioni con la malattia.

Dopo un'ampia analisi delle relazioni emerse nelle reti, vengono discussi i possibili sviluppi futuri di questa metodologia, introducendo come, ad esempio, partendo da dati specifici del paziente, il modello potrebbe essere utilizzato per prevedere l'evoluzione probabilistica della malattia, in direzione di una medicina sempre più personalizzata.

Indice

1 Contesto biologico **3**

1.1 Sclerosi multipla . . . 3
1.2 Cause . . . 3
1.2.1 Effetto dell'inquinamento sulla malattia . . . 5
1.3 Patogenesi . . . 5
1.4 Fisiopatolgia . . . 6
1.5 Manifestazioni cliniche . . . 7
1.6 Diagnosi . . . 9
1.7 Gestione e cura . . . 11
1.8 Scala EDSS . . . 12

2 Reti Bayesiane Dinamiche **15**

2.1 Teoria dei grafi . . . 15
2.2 Teoria delle probabilità . . . 16
2.3 Rete bayesiana . . . 17
2.4 Inferenza . . . 18
2.5 Apprendimento . . . 20
2.5.1 Algoritmi per l'apprendimento strutturale nel pacchetto R bnstruct . . . 22
2.5.2 Funzioni costo nel pacchetto R bnstruct . . . 25
2.6 Reti bayesiane dinamiche . . . 25

3 Dati e preprocessing **27**

3.1 Dati . . . 28
3.1.1 PATIENTS_DATA_AND_SYMPTOMS (1792x21) . . . 28
3.1.2 EDSS (25289x13) . . . 28
3.1.3 EVOKED_POTENTIALS (6212x6) . . . 29
3.1.4 MRI (7077x11) . . . 29
3.1.5 MS_TYPE (3646x4) . . . 29

3.1.6 RELAPSES (6197x16) . . . 29
3.1.7 THERAPEUTIC_PROCEDURES (13035x8) . . . 30
3.1.8 ENVIRONMENT (242878x68) . . . 30
3.2 Pulizia dei dati . . . 30
3.2.1 Preprocessing scheda EDSS . . . 31
3.2.2 Preprocessing scheda MRI . . . 31
3.2.3 Preprocessing scheda therapeutic_procedures . . . 32
3.2.4 Preprocessing scheda relapses . . . 34
3.2.5 Preprocessing scheda ms_type . . . 34
3.2.6 Preprocessing scheda evoked_potentials . . . 34
3.2.7 Preprocessing scheda environment . . . 35

4 Sviluppo dei modelli **37**
4.1 Selezione ed organizzazione dei dati per la prima DBN . . . 37
4.2 Selezione ed organizzazione dei dati per la seconda DBN . . . 39
4.3 Selezione ed organizzazione dei dati per la terza DBN . . . 41
4.4 Selezione ed organizzazione dei dati per la quarta DBN . . . 42
4.5 Quantizzazione . . . 43
4.6 Design e apprendimento delle DBN . . . 43
4.6.1 Implementazione della prima DBN . . . 43
4.6.2 Implementazione della seconda DBN . . . 46
4.6.3 Implementazione della terza DBN . . . 46
4.6.4 Implementazione della quarta DBN . . . 47

5 Risultati **49**
5.1 Prima DBN . . . 49
5.2 Seconda DBN . . . 51
5.3 Terza DBN . . . 53
5.4 Quarta DBN . . . 56

6 Conclusioni **57**
6.1 Sviluppi futuri . . . 59

Bibliografia **61**

Elenco delle figure

1.1 Numero di persone con SM. Prevalenza su 100.000 persone [3] 4
1.2 Sclerosi multipla recidivante-remittente (RR) [11] 7
1.3 Sclerosi multipla progressiva secondaria (SP) [11] 8
1.4 Sclerosi multipla primariamente progressiva (PP) [11] 8
1.5 Criteri di McDonald 2017 [12] 10
1.6 Confronto tra terapie immunomodulanti/immunosoppressive e immunoricostituenti [10] 12

2.1 Esempio di grafo aciclico diretto [16] 16
2.2 Esempio di una DBN [19] 18
2.3 Esempio di Markov blanket [21] 19

3.1 Primi 10 record della scheda EDSS dopo il preprocessing 31
3.2 Primi 10 record della scheda MRI dopo il preprocessing 33
3.3 Riassunto imputazione *end_date* nella scheda therapeutic_procedures 33
3.4 Primi 10 record della scheda therapeutic_procedures dopo il preprocessing . . . 34
3.5 Primi 10 record della scheda environment dopo il preprocessing 35

4.1 Primi 10 record di window_edss 38
4.2 Primi 10 record di final_table_net1 39
4.3 Primi 10 record di window_mri_stat 40
4.4 Primi 10 record di window_therapeutic 41
4.5 Primi 10 record di window_environment 42

5.1 Plot prima DBN 50
5.2 Plot seconda DBN 52
5.3 Plot terza DBN 54
5.4 Plot quarta DBN 55

Elenco delle tabelle

1.1 Scala EDSS 14

2.1 Quattro casi di apprendimento per una BN 20

3.1 Nuove variabili della tabella MRI. Per ciascuna variabile derivata è riportata la percentuale di FALSE, TRUE e NA nel dataset. BS = Brain Stem (tronco encefalico); SC = Spinal Cord (midollo spinale); CSC = Cervical Spinal Cord (tratto cervicale del midollo spinale); TSC = Thoracic Spinal Cord (tratto toracico del midollo spinale) 32

3.2 Percentuali delle tipologie di SM nella scheda ms_type 34

4.1 Variabili Rete 1 39

4.2 Variabili aggiunte alla Rete 2 40

4.3 Variabili aggiunte alla Rete 3 41

4.4 Variabili aggiunte alla Rete 4 42

4.5 Quantizzazione delle variabili incluse nelle quattro reti 44

4.6 Matrice della layer structure per la prima rete 46

4.7 Matrice della layer structure per la quarta rete 48

Introduzione

La sclerosi multipla (SM) è una malattia autoimmune cronica del sistema nervoso centrale in cui l'infiammazione causa la demielinizzazione e la perdita assonale, provocando sintomi e decorso potenzialmente molto variabili. Nelle aree del cervello e del midollo spinale colpite dalla SM, i segnali trasmessi attraverso i nervi vengono rallentati o bloccati, causando sintomi neurologici che possono comportare una diminuzione della qualità della vita e disabilità. I sintomi che si posso verificare sono diversi ed includono ad esempio problemi visivi come la neurite ottica, difficoltà motorie e di equilibrio, debolezza o alterazioni della sensibilità, disturbi della memoria o del pensiero ed un rischio maggiore di soffrire di depressione o ansia. Per valutare la disabilità si usa la Expanded Disability Status Scale (EDSS), ovvero una scala di valutazione funzionale che permette di valutare i livelli di disabilità, con punteggi da 0 a 10, nelle persone affette da sclerosi multipla. L'esordio dei sintomi di solito avviene tra i 20 e i 30 anni con un'incidenza aumentata nel tempo, fatto probabilmente dovuto ad una progressiva esposizione ai fattori ambientali, i quali concorrono allo sviluppo della malattia insieme ai fattori genetici. Considerando anche il fatto che con una prevalenza di 113 casi ogni 100 mila abitanti, l'Italia è considerata un'area ad alto rischio rispetto agli altri paesi europei [1], risulta quanto mai più importante comprendere al meglio la malattia cercando di individuarla preventivamente, di identificare i fattori correlati alla sua insorgenza e prognosi, e di massimizzare il trattamento attraverso una personalizzazione delle terapie.

Questo lavoro di tesi, grazie all'implementazione di una rete bayesiana dinamica (Dynamic Bayesian Network, DBN), si pone lo scopo di esplorare in modo approfondito la sclerosi multipla, cercando di contribuire al raggiungimento di tali difficili obiettivi. Infatti, le reti bayesiane sono dei modelli grafici probabilistici che rappresentano un insieme di variabili stocastiche, con le loro dipendenze condizionali, attraverso l'uso di un grafo aciclico diretto. Introducendo poi il fattore tempo, le DBN risultano strumenti molto utili per la rappresentazione di sistemi dinamici codificando, specificatamente, come le variabili si influenzino probabilisticamente nel tempo. In questo lavoro di tesi, le DBN sono state impiegate per identificare e quantificare i fattori che influenzano maggiormente la sclerosi multipla e, quindi, comprendere nel tempo, le relazioni probabilistiche tra le variabili.

A tal proposito si sono utilizzati i dati di 1792 pazienti provenienti dal database europeo del progetto BRAINTEASER [2]. I dati utilizzati comprendono sia informazioni di tipo demografico che clinico e dopo un accurato preprocessing, sono stati: aggregati, quantizzati ed elaborati al fine di costruire quattro reti bayesiane grazie al pacchetto bnstruct che consente anche una gestione efficace dei missing values, molto comuni in ambito clinico. Partendo dalla prima rete dove si consideravano solo alcuni dati statici del paziente e i valori EDSS, si sono costruite tre nuove reti, aggiungendo altre variabili come: la presenza di eventuali lesioni rilevate con la risonanza magnetica, le procedure terapeutiche e i dati di inquinamento ambientale dell'aria.

Analizzando i grafi ottenuti, dopo l'addestramento, tra le relazioni più rilevanti, si possono notare: l'influenza dell'età di esordio della malattia sulla presenza dei sintomi all'esordio, come la presenza di una lesione comporti una maggiore probabilità nella prescrizione di farmaci immunoattivi e soprattutto la dipendenza del valore EDSS dalla residenza del paziente a seconda che questo abiti in un'area industrializzata o poco densamente popolata a causa degli elevati valori degli inquinanti ambientali PM10 ed NO2. Si fa notare comunque che, rispetto alla letteratura, alcune dipendenze non sono state colte dalla rete.

Si traggono infine le conclusioni sul lavoro svolto e si fa una panoramica sui possibili sviluppi futuri, in particolare, di come le DBN possono aiutare a prevedere l'andamento futuro della malattia in un individuo o in una popolazione.

La tesi si svilupperà come segue:

Nel **Capitolo 1** si tratta il contesto biologico e la scala di valutazione EDSS, dando così un quadro generale della malattia.

Nel **Capitolo 2** si presentano le reti bayesiane con i relativi metodi di inferenza e di apprendimento con un maggior focus su quello strutturale e sui 5 metodi presenti in bnstruct.

Nel **Capitolo 3** vengono descritti i dati utilizzati e il preprocessing effettuato.

Nel **Capitolo 4** si descrive la selezione ed organizzazione dei dati per ciascuna DBN, spiegando poi la quantizzazione e l'implementazione (comprensiva di layering) utilizzate.

Nel **Capitolo 5** vengono presentati e discussi i risultati ottenuti per ognuna delle quattro reti.

Nel **Capitolo 6**, infine, si traggono le conclusioni sul lavoro svolto e vengono presentati dei possibili sviluppi futuri.

Capitolo 1

Contesto biologico

1.1 Sclerosi multipla

La sclerosi multipla è la malattia neurologica invalidante non traumatica più comune che affligge i giovani adulti. È principalmente una malattia infiammatoria del cervello e del midollo spinale, in cui un'infiltrazione linfocitaria focale porta al danneggiamento della mielina (demielinizzazione) e degli assoni.

Due fenomeni clinici caratteristici della malattia sono il sintomo di Lhermitte (una sensazione elettrica che corre lungo la colonna vertebrale o gli arti durante la flessione del collo) e il fenomeno di Uhthoff (un peggioramento temporaneo, di breve durata, dei sintomi, in risposta ad un aumento della temperatura, che si verifica ad esempio: dopo l'esercizio fisico, per il calore della stagione estiva o per la febbre). Le manifestazioni più comuni includono neurite ottica, sindromi del tronco cerebrale e del midollo spinale. Tuttavia, sono presenti altri fenomeni clinici, non specifici della malattia, nei sistemi motori, sensoriali, visivi ed autonomi. Ad esempio, per quanto riguarda il telencefalo, si verifica un deterioramento cognitivo con conseguente deficit dell'attenzione, del ragionamento e demenza nelle fasi più avanzate, oppure depressione come sintomo a livello affettivo. Il danneggiamento della mielina nel tronco encefalico comporta problemi visivi, vertigini, compromissione della deglutizione, della parola ed instabilità emotiva. Altri sintomi che si possono riscontrare sono: tremore, scarso equilibrio, stanchezza, rigidità e spasmi dolorosi, disfunzioni della vescica, disfunzione erettile, dolore e senso di affaticamento.

1.2 Cause

La sclerosi multipla è causata sia da fattori ambientali che genetici. Analizzando più nello specifico i primi, in generale, la malattia aumenta allontanandosi dall'equatore, con alcune zone che fanno eccezione, come mostrato in Figura 1.1.

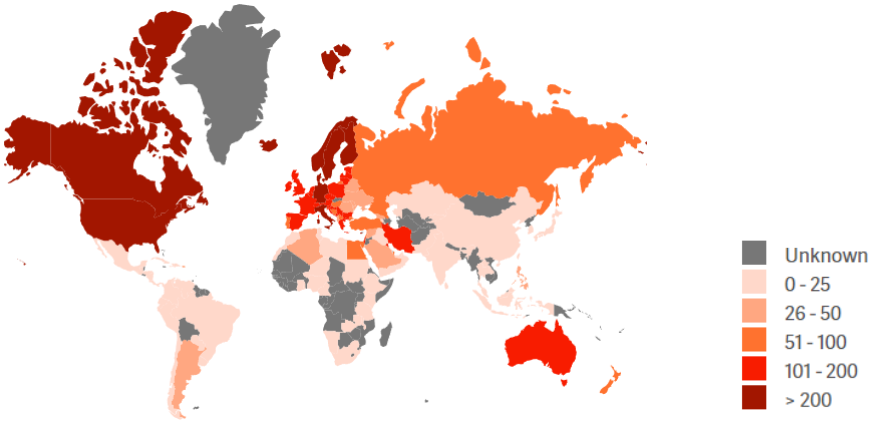

Figura 1.1: Numero di persone con SM. Prevalenza su 100.000 persone [3]

Ovviamente la distribuzione è fortemente influenzata dalle migrazioni e gli studi di quest'ultime indicano come l'ambiente prevalga sulla genetica, con un tempo di 10-20 anni tra l'esposizione ai fattori di rischio ambientali e l'insorgenza della malattia. Infatti, gli immigrati adulti provenienti da paesi a basso rischio, come le Indie occidentali, e trasferitisi in Europa hanno un basso rischio di sviluppare la SM; tuttavia, i figli nati da migranti in Europa sono a rischio elevato [4]. Di particolare rilevanza sembra essere il virus Epstein-Barr: uno studio afferma che un'infezione dovuta a questo virus, nei primi anni di vita, provocherebbe un rischio fino a 2/3 volte maggiore, supportando così "l'ipotesi dell'igiene" [4], [5]. Tale ipotesi suggerisce che bambini e adulti sviluppino risposte anomale alle infezioni per la mancata esposizione ai microbi necessari a costruire le difese immunitarie nei primi anni di vita. Tuttavia altri fattori ambientali sono stati proposti come possibili cause, tra cui: carenza di vitamina D data dalla poca esposizione alla luce ultravioletta B e l'obesità (confermati come fattori di rischio indipendenti da studi di randomizzazione mendeliana), gli inquinanti atmosferici e il fumo che aumenta il rischio di circa il 50%. [4], [5]

Per quanto riguarda i fattori genetici, la sclerosi multipla ha una ricorrenza familiare di circa il 20%, col rischio che si abbassa a seconda del grado di parentela, dimostrando che c'è un'influenza genetica sulla suscettibilità individuale alla malattia. Studi discordanti riguardano la zigosità dei gemelli in merito ai tassi di concordanza clinica. I recettori del complesso maggiore di istocompatibilità (MHC) identificati, che sono associati alla sclerosi multipla, sono DR15 e DQ6 e sono codificati dal complesso dell'antigene leucocitario umano (HLA) del cromosoma 6. Questa associazione è presente in tutte le popolazioni, (ad eccezione di alcuni gruppi mediterranei dove è associata a DR4), ma maggiormente in quelle nord europee. [5]

1.2.1 Effetto dell'inquinamento sulla malattia

Rilevante risulta essere anche il luogo di residenza in quanto vari studi indicano una correlazione positiva tra l'esposizione al PM10 e il rischio di sviluppare la SM e la sua ricaduta e attività. La pessima qualità dell'aria e quindi la presenza di inquinanti atmosferici potrebbero esercitare un grave impatto sul sistema nervoso centrale attraverso vari processi. Anche se i meccanismi non sono ancora del tutto chiari, è stato scoperto che l'esposizione cronica all'inquinamento aumenta il livello di alcuni marcatori neuroinfiammatori e proinfiammatori nel cervello, con conseguenti eventi a cascata che porterebbero allo sviluppo della malattia [6]. Ciò viene avvalorato da un ampio studio retrospettivo condotto in Lombardia, la regione in cui l'inquinamento atmosferico ha l'impatto maggiore, che ha riportato un aumento del 42% dei ricoveri ospedalieri dovuti a riacutizzazioni quando i livelli di PM10 erano nel quartile più alto nella settimana precedente [7]. Inoltre Elgasbi et al., oltre a riportare come il livello degli inquinanti differisca a seconda della stagione (più alto in inverno e più basso in estate), sottolinea una correlazione tra PM10 e NO2, individuando un'associazione positiva nei risultati tra le ricadute della SM e l'esposizione ambientale a questi due inquinanti [8].

1.3 Patogenesi

La sclerosi multipla è una patologia infiammatoria del sistema nervoso centrale (SNC) la cui patogenesi è attribuita ad un'infiammazione autoimmune in cui il sistema immunitario attacca erroneamente la mielina danneggiandola e in alcuni casi provocando danni alla fibra nervosa. A causa dei difetti regolatori, che permettono a queste cellule di innescare una risposta immunitaria all'interno del cervello, avviene la transizione da una sorveglianza fisiologica a un processo patologico. È ben nota la correlazione tra la suscettibilità alla malattia e l'aplotipo DR2 esteso del complesso maggiore di istocompatibilità. I recenti studi di sequenziamento dell'intero genoma umano hanno confermato tale associazione e identificatone altre con i polimorfismi a singolo nucleotide nei geni per il recettore dell'interleuchina 2 e dell'interleuchina 7. Il pensiero corrente è che questi polimorfismi dei recettori delle citochine possano influenzare l'equilibrio tra i linfociti T a effetto patogeno e i linfociti T a effetto protettivo di regolazione. [9] L'innesco della reazione autoimmune è ancora oggetto di ipotesi e potrebbe originare da fenomeni di molecular mimicry o dal difettoso funzionamento della tolleranza immunologica (Treg cells). [10] Ruolo centrale in questo processo lo hanno i linfociti T autoreattivi del tipo Th1 e Th17 che attraversano la barriera emato-encefalica ed entrano nel SNC inducendo un'infiammazione locale.

Sono coinvolte nel processo anche altre cellule del sistema immunitario e cellule gliali, come gli astrociti e le microglia. Infatti i linfociti Th1 secernono IFN-γ (interferone gamma), che

attiva i macrofagi che costituiscono l'infiltrato nelle placche e nelle regioni cerebrali circostanti; mentre i linfociti Th17 promuovono il reclutamento dei leucociti che con i loro prodotti lesivi causano la demielinizzazione [9]. Recentemente ai linfociti B è stata attribuita una funzione più complessa, includente la capacità di operare da cellule presentanti l'antigene e di agire mediante il rilascio di molecole co-stimolatorie per i linfociti T [10]. Per quanto riguarda la rimielinizzazione, essa avviene sia durante il processo infiammatorio acuto, in cui è più attiva, che durante la fase progressiva. In circa il 20% delle persone le placche vengono rimielinate con successo dai precursori degli oligodendrociti [5].

1.4 Fisiopatolgia

Il processo caratteristico della SM include infiammazione, perdita della mielina (demielinizzazione) con successivo tentativo del corpo di riparare il tessuto danneggiato (rimielinizzazione), esaurimento di oligodendrociti e astrocitosi. Nel frattempo, gli assoni e le cellule nervose possono subire danni e degenerare, culminando infine nella formazione della caratteristica placca sclerotica associata alla SM. La mielina è sintetizzata da oligodendrociti maturi e avvolge più di un assone nel sistema nervoso centrale, mentre i processi di sviluppo dei precursori degli oligodendrociti in cellule mielinizzanti sono regolati da fattori di crescita definiti. Le cellule gliali che formano la mielina si dispongono lungo tutta la lunghezza dell'assone coprendone l'intera estensione, escludendo il monticolo assonico, le terminazioni sinaptiche e lasciando parzialmente scoperte delle zone chiamate nodi di Ranvier che corrispondono al punto di contatto tra due cellule gliali adiacenti; qui, inoltre, la resistenza elettrica è bassa quindi è favorita la depolarizzazione, la generazione di una corrente elettrica e di conseguenza l'innesco della conduzione saltatoria.

I sintomi associati alla fatica fisiologica derivano dal fatto che gli assoni parzialmente demielinizzati non possono trasmettere treni rapidi di impulsi. La depolarizzazione, che attraverserebbe la lesione ad una velocità ridotta, spiegherebbe il ritardo caratteristico dei potenziali evocati. Le sensazioni possono essere alterate dagli assoni parzialmente demielinizzati che scaricano spontaneamente. La maggiore sensibilità meccanica porta poi a sintomi indotti dal movimento. Questi includono bagliori di luce causati dai movimenti oculari e contrazioni muscolari involontarie piuttosto ampie (miochimia). I sintomi peggiorano quando il corpo è sottoposto a temperature elevate per la mancanza di mielina (sintomo di Uhthoff). Al contrario, la rimielinizzazione contribuisce alla ripresa dai sintomi, infatti gli assoni possono nuovamente condurre l'impulso nervoso e ripristinare la loro funzione.

1.5 Manifestazioni cliniche

All'esordio, circa l'85% dei pazienti presenta un episodio acuto che colpisce un sito (o occasionalmente più di uno), classificato come sindrome clinicamente isolata (Clinically Isolated Syndrome, CIS). La sclerosi multipla evolve in diversi fenotipi (ovvero l'insieme di tutte le caratteristiche manifestate da un organismo vivente), infatti se a seguito del primo episodio questo è accompagnato da anomalie nella sostanza bianca, rilevate tramite risonanza magnetica (Magnetic Resonance Imaging, MRI), in siti clinicamente non colpiti, la probabilità di un secondo attacco di demielinizzazione aumenta dal 50% a 2 anni all'82% a 20 anni. In questo caso si classifica il fenotipo come recidivante-remittente (RR), sempre qualora i criteri diagnostici per la sclerosi multipla fossero soddisfatti. I nuovi episodi si verificano in modo irregolare, ma il tasso raramente supera l'1,5 all'anno. Nel tempo, come si osserva da Figura 1.2, il recupero da ciascun episodio risulta via via sempre più incompleto e i sintomi persistenti si accumulano. Più nello specifico, le ricadute si sviluppano in poche ore o giorni, raggiungono un plateau che dura alcune settimane e poi si riprendono gradualmente, con la maggior parte delle ricadute che, però, lascia dietro di sé qualche danno.

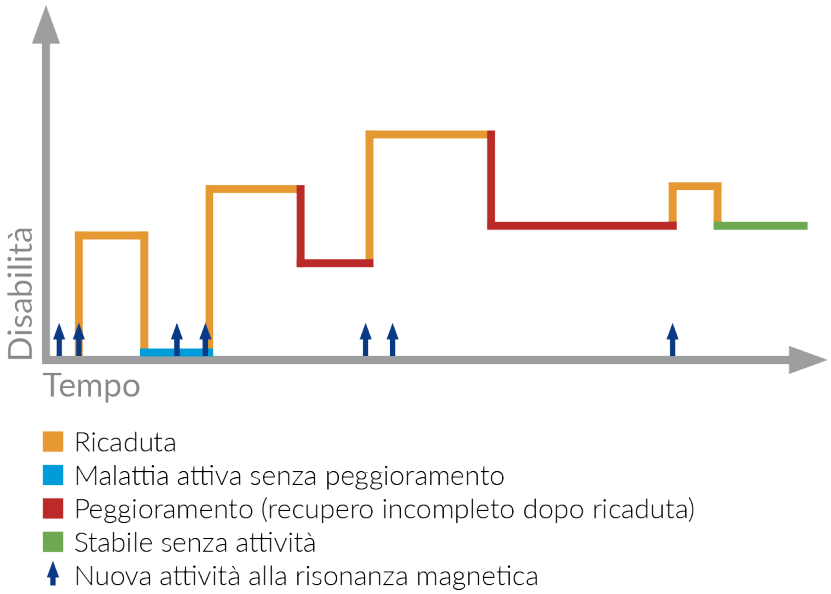

Figura 1.2: Sclerosi multipla recidivante-remittente (RR) [11]

In una seconda fase, circa il 65% dei pazienti passa alla fase progressiva secondaria (SP), caratterizzata da una disabilità persistente che progredisce gradualmente nel tempo, come illustra il grafico in Figura 1.3. Questa forma si sviluppa generalmente 10-15 anni dopo l'inizio della forma recidivante-remittente.

Circa nel 15% dei casi, come si può notare in Figura 1.4, la malattia è progressiva fin dall'inizio (PP), con l'assenza di ricadute o remissioni; tipicamente si verifica, però, un accumu-

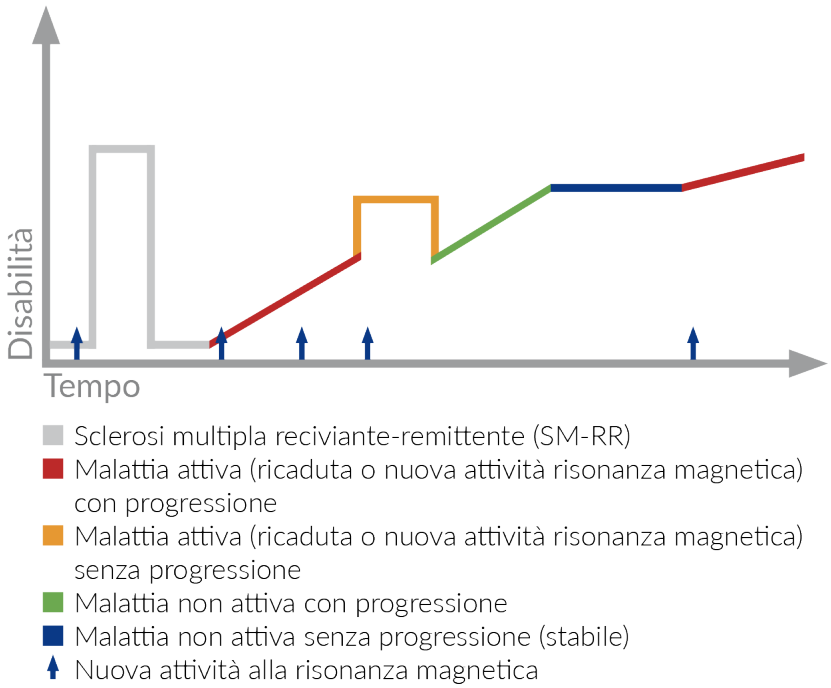

Figura 1.3: Sclerosi multipla progressiva secondaria (SP) [11]

lo graduale di disabilità progressiva coinvolgente un sistema neurale dominante. In entrambe queste situazioni, la progressione inizia intorno ai 40 anni.

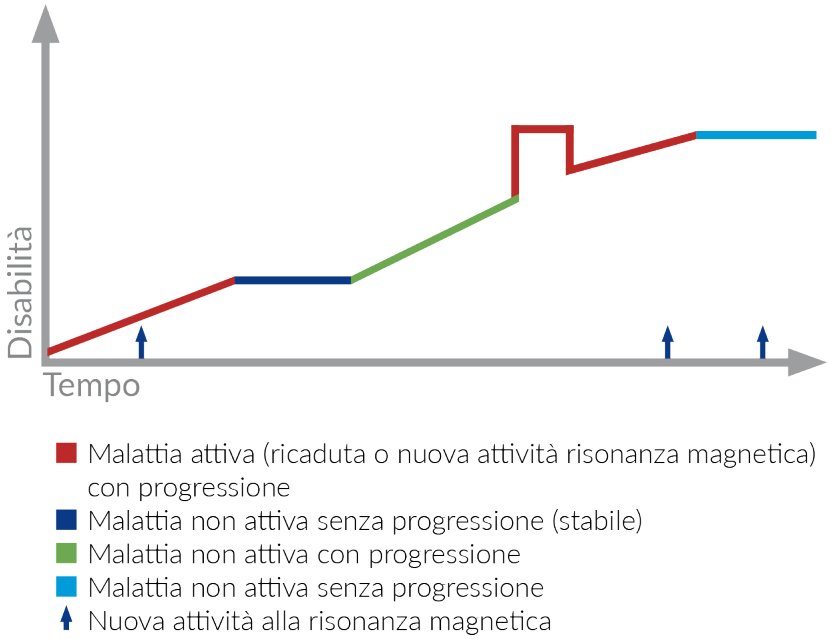

Figura 1.4: Sclerosi multipla primariamente progressiva (PP) [11]

Una piccola percentuale di individui viene diagnosticata con una forma relativamente ra-

ra di sclerosi multipla nota come sclerosi multipla a evoluzione progressiva-recidivante (PR). Questo tipo di SM peggiora costantemente dall'esordio dei primi sintomi, indipendentemente da ricadute o periodi di remissione che comunque possono essere sporadiche. In alcuni soggetti si riscontra la sindrome radiologicamente isolata (RIS), ovvero la presenza di lesioni caratteristiche della SM riscontrate tramite risonanza magnetica, pur in assenza di sintomi clinici.

L'andamento della malattia evolve durante il corso di decenni e anche a causa delle infezioni che essa porta (dovute alle disabilità neurologiche) è responsabile della morte in due terzi dei casi, con una riduzione dell'aspettativa di vita di 5-10 anni. Per quanto riguarda il tasso di recidiva esso diminuisce durante la gravidanza, ma raddoppia dopo l'esposizione virale ad infezioni delle vie respiratorie e gastrointestinali, invece sembra che la promozione delle cellule T regolatorie, data da un'infezione parassitaria persistente, protegga dalla malattia [4], [5].

La progressione della malattia nella sclerosi multipla dipende dalla degenerazione accumulata degli assoni, è quindi importante studiare anche questo aspetto per capire il legame che ha con l'infiammazione. Secondo Compston and Coles la progressione nella sclerosi multipla è dovuta alla perdita cumulativa degli assoni, iniziata e mantenuta da complesse risposte infiammatorie che agiscono su individui intrinsecamente suscettibili alla neurodegenerazione e cambiano man mano che aumenta il danno tissutale. Affermano, quindi, che esiste un'interazione tra neurodegenerazione ed infiammazione e che quest'ultima non perde del tutto il suo effetto, anche se dovesse diminuirne la quantità assoluta, poiché c'è una crescente suscettibilità degli assoni danneggiati dal trauma infiammatorio residuo [5]. Anche Dobson e Giovannoni riportano come il danno agli oligodendrociti e la demielinizzazione si verifichino a causa dell'infiammazione, con gli assoni che con il progredire della malattia sviluppano un danno assonale irreversibile. Esiste, poi, una distinzione clinica tra la forma RR, dove c'è una profonda infiammazione linfocitaria e le forme SP e PP, dove le lesioni tendono ad avere un nucleo di lesione inattivo circondato da una stretta fascia di microglia e macrofagi attivati. [4]

1.6 Diagnosi

Al fine di confermare la diagnosi di SM ed escludere le malattie con sintomi affini, in passato si effettuava l'analisi del liquido cerebrospinale (alla ricerca di bande oligoclonali, che esprimono la presenza di un'attività immune anomala). Ad oggi questa tecnica è stata sostituita dalla risonanza magnetica tramite la quale è possibile riscontrare la presenza di lesioni. In genere mediante MRI si osservano circa 10 lesioni "asintomatiche" per ogni attacco clinico [4]. Il complesso di sintomi risultanti è una combinazione di posizione e dimensioni: una lesione, anche se piccola, in un'area significativa, è probabile che causi sintomi. Molte lesioni possono

essere osservate a livello microscopico e ancora di più nella sostanza grigia profonda e corticale, mentre quelle macroscopiche, visibili in MRI, sono solo una piccola parte. [4]

La SM viene diagnosticata in presenza di sintomi e di segni di interessamento del sistema nervoso centrale in più zone (disseminazione spaziale), con comparsa di lesioni in tempi successivi (disseminazione temporale), con la diagnosi che può arrivare anche dopo anni dai primi sintomi. Per accelerare questo processo di identificazione è stato ideato il criterio di McDonald (revisione del 2017) riportato in Figura 1.5 [12].

	NUMERO DI LESIONI CON EVIDENZA CLINICA OBIETTIVA	DATI ADDIZIONALI NECESSARI PER LA DIAGNOSI DI SM
≥ 2 attacchi clinici	≥ 2	Nessuno
≥ 2 attacchi clinici	1 (così come l'evidenza di un precedente attacco coinvolgente un sito anatomico distinto)	Nessuno
≥ 2 attacchi clinici	1	DIS (attacco clinico addizionale in un sito anatomico distinto)
1 attacco clinico	≥ 2	DIT (attacco clinico addizionale, RM o bande oligoclonali)
1 attacco clinico	1	DIS e DIT

Figura 1.5: Criteri di McDonald 2017 [12]

Una critica che viene spesso mossa al criterio di McDonald è che non sia abbastanza approfondito poiché non include una diagnosi di "SM asintomatica", escludendo di fatto dal trattamento i pazienti diagnosticati con RIS. Questo rappresenta nella pratica un problema, dal momento che una proporzione di questi soggetti ha già danni cerebrali e compromissione cognitiva. Circa il 30%, inoltre, statisticamente svilupperà la sclerosi multipla entro 5 anni: per questi soggetti, con interventi precoci potrebbe essere possibile prevenire lo sviluppo di malattie neurologiche clinicamente evidenti. [4]

Tradizionalmente, l'accumulo/conteggio delle lesioni insieme alle lesioni "attive" (con contrasto al gadolinio) è stato utilizzato per stimare l'attività della malattia; tuttavia, la correlazione con gli esiti a lungo termine non è limitata. Al contrario ha una buona correlazione se viene preso in considerazione insieme all'atrofia cerebrale. Le nuove tecniche di MRI, tra cui l'imaging a trasferimento di magnetizzazione, l'imaging a tensori di diffusione e l'imaging funzionale a risonanza magnetica, forniscono approfondimenti sulla malattia con diffusi disturbi al di fuori dello sviluppo di lesioni focali; tuttavia, queste tecniche non sono ancora nella pratica clinica di routine. [5]

Altri esami che possono aiutare a stabilire la disseminazione spaziale sono invece i potenziali evocati visivi, uditivi e sensoriali e i tempi di conduzione motoria centrale. Questa classe di test

è utilizzata per registrare e misurare la velocità e l'integrità con cui i segnali nervosi raggiungono il cervello. Può quindi essere utile per dimostrare una conduzione rallentata in pazienti con segnali clinici e apparizioni nell'MRI equivoci. Tuttavia, queste informazioni, potrebbero non aggiungere molto valore clinico, per questo per avere una diagnosi efficace è necessaria la combinazione dei sintomi del paziente, dei risultati degli esami fisici oggettivi e dei risultati delle neuroimmagini.

1.7 Gestione e cura

Il trattamento della sclerosi multipla può essere suddiviso in terapie modificanti la malattia, specifiche per questa condizione, e terapie sintomatiche, spesso utilizzate per trattare i sintomi derivanti da disfunzioni neurologiche.

Per quanto riguarda la prima categoria, le strategie di trattamento attuali sono due: quella immunomodulante/immunosoppressiva e quella immunoricostituente. Il primo modello si basa sulla somministrazione continua di farmaci con due diversi meccanismi d'azione. Questi farmaci possono essere classificati in immunomodulanti (o immunoattivi), come gli interferoni e il natalizumab, o immunosoppressivi (o cortisonici) come il prednisone e il metilprednisolone in base al loro effetto linfopenizzante. Entrambi i farmaci devono essere somministrati in modo persistente per raggiungere il controllo della risposta autoimmune e di conseguenza una remissione della malattia. I cortisonici agiscono principalmente riducendo l'infiammazione sopprimendo la risposta immunitaria e riducendo la produzione di sostanze chimiche infiammatorie e sono usati per il trattamento a breve termine di infiammazioni acute, come avviene durante una ricaduta. Gli immunoattivi, invece, possono modulare la risposta immunitaria a lungo termine e prevenire le ricadute, inoltre non sopprimono completamente le cellule immunitarie, ma ne alterano il funzionamento.

Il secondo modello con trattamento immunoricostituente, al contrario, si distingue per l'uso limitato nel tempo di farmaci o di trattamenti (tra cui alemtuzumab e cladribine) con drastico effetto linfopenizzante in grado di determinare un reset della risposta immunitaria con riequilibrio del repertorio di specificità antigenica dei linfociti T e B. Anche dopo la loro sospensione, queste terapie possono causare una remissione prolungata della malattia, anche se vengono somministrate in cicli brevi ed intermittenti.

Come riportato in Figura 1.6, le terapie immunoricostituenti sono altamente efficaci nel trattamento della SM e il loro vantaggio è di indurre un reset del sistema immunitario con il ripristino di una condizione di self-tolerance e un cambiamento del repertorio linfocitario. [10]

Il secondo macrotipo di terapia riguarda le terapie sintomatiche le quali mirano ai sintomi derivanti da danni al sistema nervoso centrale. Comprendono quindi diversi farmaci a seconda di

	TERAPIE IMMUNOMODULANTI/ IMMUNOSOPPRESSIVE	TERAPIE IMMUNORICOSTITUENTI
Modo di somministrazione	Continuo	Intermittente o a brevi cicli
Problemi di aderenza	Possibili	Rari
Efficacia	Da modesta a molto elevata	Da elevata a molto elevata
Reversibilità degli effetti sul sistema immunitario	Si	No
Rischio di eventi avversi	Basso inizialmente (ma aumenta nel tempo)	Elevato all'inizio (ma si riduce nel tempo)
Effetto *rebound* alla sospensione	Molto probabile	Meno probabile
Remissione prolungata dopo sospensione	Improbabile	Probabile

Figura 1.6: Confronto tra terapie immunomodulanti/immunosoppressive e immunoricostituenti [10]

quale e dove è situata la disfunzione; tali terapie includono ad esempio il sativex per la spasticità e la fampridina per le difficoltà deambulatorie.

1.8 Scala EDSS

La "Expanded Disability Status Scale" (EDSS) è una scala utilizzata in clinica per valutare e quantificare la disabilità nella sclerosi multipla e quindi per monitorare nel tempo i cambiamenti nel livello di disabilità. Ha range da 0 (nessuna disabilità) a 10 (morte causata dalla sclerosi multipla) con incrementi di 0,5 unità [13].

Le misure di compromissione in otto sistemi funzionali fanno da base per il calcolo del punteggio EDSS. Un sistema funzionale (Functional System, FS) rappresenta una rete di neuroni nel cervello con responsabilità per compiti specifici e ognuno viene valutato su una scala da 0 (nessuna disabilità) a 5 o 6 (disabilità più grave). Di seguito vengono riportati gli otto FS coinvolti:

1. sistema piramidale: debolezza muscolare o difficoltà nel movimento degli arti
2. sistema cerebellare: atassia, perdita dell'equilibrio, coordinazione o tremore
3. funzionalità del tronco encefalico - problemi di linguaggio, deglutizione e nistagmo
4. sistema sensoriale - intorpidimento o perdita di sensazioni
5. funzionalità intestinale e vescicale.

6. funzionalità visiva – problemi di vista
7. funzionalità cerebrali – problemi di pensiero e memoria
8. altro

La Tabella 1.1 riporta la scala EDSS. I punteggi da 1 a 4,5 si riferiscono a persone con SM in grado di camminare senza alcun aiuto, mentre quelli da 5 a 9 sono definiti in base alla difficoltà nel camminare. Tuttavia proprio per questa dipendenza, forte dalle abilità legate alla sola deambulazione, la scala EDSS viene talvolta criticata. Nonostante tenga conto della disabilità associata ad un grado avanzato della malattia è difficile che le persone raggiungano un punteggio di 7 o superiore [14], [15].

Punteggio	Descrizione
0	Esame neurologico normale, nessuna disabilità in nessun sistema funzionale (FS).
1.0	Nessuna disabilità, segni minimi in un FS.
1.5	Nessuna disabilità, segni minimi in più di un FS.
2.0	Disabilità minima in un FS.
2.5	Disabilità lieve in un FS o disabilità minima in due FS.
3.0	Disabilità moderata in un FS o disabilità lieve in tre o quattro FS. Nessun impedimento al camminare.
3.5	Disabilità moderata in un FS e più di una disabilità minima in diversi altri FS. Nessun impedimento al camminare.
4.0	Disabilità significativa, ma autosufficiente e in movimento circa 12 ore al giorno. In grado di camminare senza ausilio o riposo per 500 metri.
4.5	Disabilità significativa, ma in movimento gran parte del giorno, in grado di lavorare una giornata intera, potrebbe avere alcune limitazioni o richiedere assistenza minima. In grado di camminare senza ausilio o riposo per 300 metri.
5.0	Disabilità sufficientemente grave da compromettere le attività quotidiane e la capacità di lavorare una giornata intera senza provvidenze speciali. In grado di camminare senza ausilio o riposo per 200 metri.
5.5	Disabilità sufficientemente grave da escludere le attività quotidiane complete. In grado di camminare senza ausilio o riposo per 100 metri.
6.0	Richiede un supporto per camminare, come un bastone, una stampella, ecc., per percorrere circa 100 metri con o senza riposo.
6.5	Richiede due supporti per camminare, come una coppia di bastoni, stampelle, ecc., per percorrere circa 20 metri senza riposo.
7.0	Incapace di camminare oltre circa 5 metri anche con aiuto. Essenzialmente limitato alla sedia a rotelle; se ne spinge una standard e si trasferisce da solo. In movimento con la sedia a rotelle circa 12 ore al giorno.
7.5	Incapace di fare più di pochi passi. Limitato alla sedia a rotelle e potrebbe aver bisogno di assistenza nel trasferirsi. Può spingersi, ma non può continuare per un'intera giornata in una sedia a rotelle standard e potrebbe richiedere una sedia a rotelle motorizzata.
8.0	Essenzialmente limitato a letto o in sedia o spinto in sedia a rotelle. Potrebbe essere fuori dal letto gran parte del giorno. Conserva molte funzioni di auto-cura. Generalmente ha un uso efficace delle braccia.
8.5	Essenzialmente limitato a letto gran parte del giorno. Ha un uso efficace delle braccia e conserva alcune funzioni di auto-cura.
9.0	Confinato a letto. Può comunque comunicare e mangiare.
9.5	Confinato a letto e totalmente dipendente. Incapace di comunicare efficacemente o mangiare/deglutire.
10.0	Morte causata dalla SM.

Tabella 1.1: Scala EDSS

Capitolo 2

Reti Bayesiane Dinamiche

Le reti bayesiane dinamiche (Dynamic Bayesian Networks, DBN) sono una classe di modelli probabilistici che estendono le reti bayesiane tradizionali per modellare relazioni probabilistiche e dinamiche nel tempo. Sono costituite da nodi che rappresentano le variabili del sistema e archi che rappresentano le relazioni probabilistiche tra di esse. Le DBN trovano applicazioni in una vasta gamma di settori, tra cui l'ingegneria, le scienze ambientali, la finanza, la medicina e l'informatica. In particolare, un'applicazione si trova nei sistemi diagnostici medici, ad esempio, una rete bayesiana potrebbe rappresentare le relazioni probabilistiche tra malattie e sintomi. La rete può essere utilizzata per calcolare le probabilità della presenza di diverse malattie, dati i sintomi a priori. Più in generale queste reti possono essere usate per la modellazione delle relazioni causali, la rappresentazione della conoscenza incerta, l'inferenza probabilistica e la risposta a query probabilistiche. In pratica, sono modelli probabilistici che combinano la teoria dei grafi e la teoria delle probabilità.

2.1 Teoria dei grafi

I grafi sono strutture matematiche discrete che rivestono interesse sia per la matematica che per un'ampia gamma di campi applicativi essendo alla base di modelli di sistemi e processi. La definizione afferma che un grafo $G = (V, E)$ è composto da un insieme finito di nodi (o vertici) V e da un insieme di archi $E \subset V \times V$ che connettono coppie di nodi. Due nodi connessi da un arco sono detti adiacenti. Due vertici u, v connessi da un arco e prendono il nome di estremi dell'arco, il quale viene anche identificato dai suoi estremi (u, v). Se E è una relazione simmetrica allora si dice che il grafo è non orientato (o indiretto), altrimenti si dice che è orientato (o diretto). Un percorso di lunghezza n in G è dato da una sequenza di vertici $v_0, v_1, ..., v_n$ (non necessariamente tutti distinti) e da una sequenza di archi che li collegano $(v_0, v_1), (v_1, v_2), ..., (v_{n-1}, v_n)$. I vertici v_0 e v_n si dicono estremi del percorso. Un percorso con

gli archi a due a due distinti tra loro prende il nome di cammino. Un cammino chiuso $(v_0 = v_n)$ senza archi ripetuti viene detto circuito. Nella teoria dei grafi, un grafo aciclico diretto (Directed Acyclic Graph, DAG) è un grafo diretto che non ha circuiti, come si può notare nell'esempio in Figura 2.1. Infatti, comunque scegliamo un vertice del grafo, non possiamo tornare ad esso percorrendo gli archi del grafo; in più, una visita in profondità non presenterà archi all'indietro.

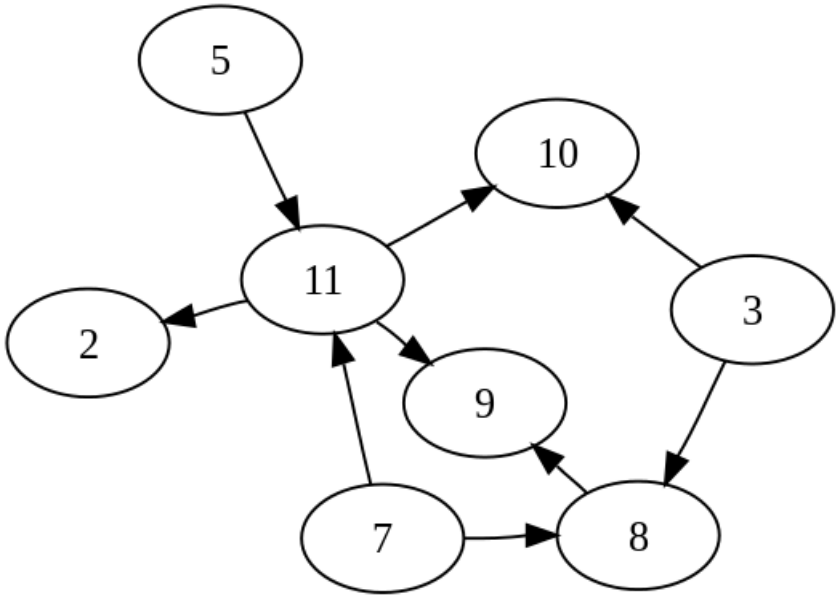

Figura 2.1: Esempio di grafo aciclico diretto [16]

2.2 Teoria delle probabilità

La definizione di probabilità condizionata è: siano A e B due eventi, la probabilità condizionale di A condizionata da B è definita da: $P(A|B) = P(B)P(A, B)$. Da qui si ottiene il Teorema di Bayes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{con} \quad P(A), P(B) \neq 0$$

Dove:

- $P(A)$ è la probabilità a priori dell'evento A
- $P(A|B)$ è la probabilità condizionata di A, noto B, chiamata anche probabilità a posteriori
- $P(B|A)$ è la probabilità a posteriori dell'evento B, noto A
- $P(B)$ è la probabilità a priori dell'evento B

2.3 Rete bayesiana

Una rete bayesiana (Bayesian Network, BN) è un modello grafico probabilistico che rappresenta un insieme di variabili stocastiche con le loro dipendenze condizionali attraverso l'uso di un grafo aciclico diretto. Nello specifico, una BN consiste di:

1. Un insieme di variabili casuali $X_1, \ldots X_n$

2. Un grafo aciclico diretto (DAG) in cui ogni variabile appare una sola volta ed è graficamente rappresentata come nodo del grafo. Per ogni variabile X_i, è definita una probabilità condizionata $p(x_i|Parents(X_i))$ dove $Parents(X_i)$ corrispondono ai genitori della variabile X_i nel DAG. L'insieme delle variabili casuali è completamente determinato dalla distribuzione di probabilità congiunta. Sotto l'ipotesi di Markov, cioè l'ipotesi che ciascun x_i sia condizionalmente indipendente dai suoi non-discendenti dato i suoi genitori, questa distribuzione di probabilità congiunta può essere determinata dalla scomposizione attraverso:
$$p(X_1 = x_1, \ldots, x_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i|Parents(X_i))$$

3. Una rappresentazione delle probabilità condizionate richieste

Al fine di soddisfare quest'ultimo punto è conveniente usare le tabelle di probabilità condizionata (Conditional Probability Table, CPT) che esprimono la probabilità associata ad ogni valore che la variabile può assumere, in funzione di ogni configurazione possibile dei suoi nodi genitore. Per nodi senza genitori vengono definite tabelle di probabilità a priori. È possibile calcolare l'output del sistema utilizzando la condizione di input rilevante (riga) nella CPT per ciascun nodo, generando un "1" con la probabilità di output specificata per quella condizione. Quindi, così facendo, possiamo valutare gli output dei nodi che ricevono i valori appena generati come input [17], [18]. Per comprendere al meglio quanto appena detto, di seguito, in Figura 2.2, viene riportato un esempio riguardante diagnosi mediche dove, ad esempio, è possibile calcolare la probabilità di avere mancanza di respiro essendo un fumatore $P(Shortness\ of\ Breath = T|Smokes = T)$.

Allo scopo di capire meglio quanto riportato in precedenza al punto 2 è utile definire due importanti concetti:

- Il primo è l'indipendenza condizionale: due insiemi di nodi A e B sono detti condizionalmente indipendenti se tutti i percorsi tra i nodi in A e B sono separati da un nodo in un terzo insieme C. In modo più specifico, due variabili casuali X e Y sono condizionalmente indipendenti dato un'altra variabile casuale Z se $P(X|Z) = P(X|Y, Z)$. L'indipendenza

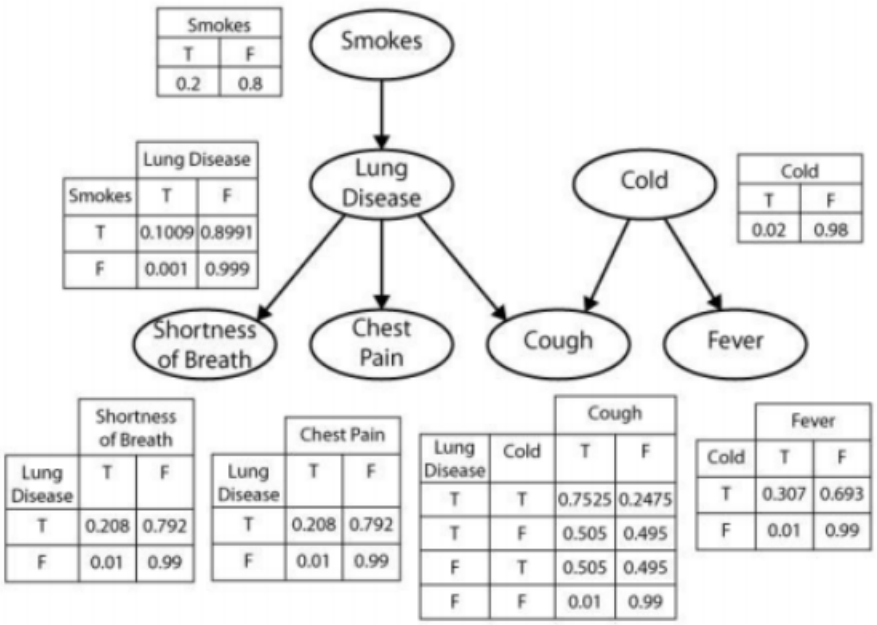

Figura 2.2: Esempio di una DBN [19]

definita per le reti bayesiane deve tener conto della direzionalità degli archi. Un arco da un nodo A ad un nodo B può essere interpretato come A causa B.

- Il secondo invece è la D-separazione: Sia V l'insieme dei nodi. Due variabili A e B in una rete bayesiana sono d-separate da $X \subseteq V$ se tutti i percorsi tra A e B sono bloccati da X [20]. Questa regola è alla base della definizione della Markov blanket (uno stato di indipendenza locale) la quale afferma che data una rete bayesiana e un nodo X, quest'ultimo è indipendente da tutti gli altri nodi della rete, dati i suoi nodi genitori, i suoi nodi figli e i nodi genitori dei suoi nodi figli. Questo insieme di nodi è detto Markov blanket del nodo X e qualsiasi variazione alla rete che non comprenda la Markov blanket non influirà sul suo valore. In Figura 2.3 ne è riportato un esempio.

2.4 Inferenza

Come già menzionato la rete bayesiana può essere utilizzata per rispondere ad interrogazioni probabilistiche sulle sue variabili e relazioni. Per farlo usa un processo chiamato inferenza probabilistica che consiste nel calcolo della distribuzione posteriore delle variabili basandosi sulle evidenze. Ad esempio, grazie ad essa, è possibile aggiornare la conoscenza dello stato di un sottoinsieme di variabili quando le variabili di evidenza sono osservate: questo aggiornamento

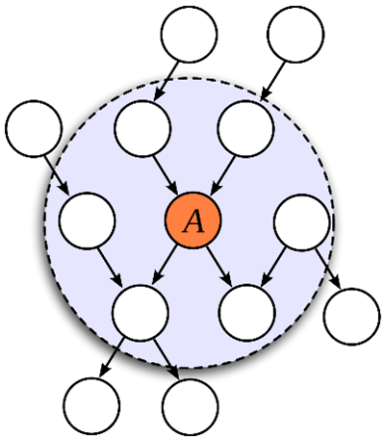

Figura 2.3: Esempio di Markov blanket [21]

prende il nome di propagazione dell'evidenza. In grandi reti calcolare l'inferenza è molto dispendioso dal punto di vista computazionale, infatti per variabili discrete si è dimostrato essere NP-hard; in generale, vengono quindi usati dei metodi per ridurre la quantità di calcoli. Questi metodi nelle reti bayesiane si dividono in due tipi, di inferenza esatta e di inferenza approssimata.

I primi sfruttano la struttura di indipendenza contenuta nella rete per propagare efficacemente l'incertezza; i più comuni sono: variable elimination, che elimina le variabili osservate e non di query riorganizzando le sommatorie in modo che vengano utilizzati solo i fattori che coinvolgono una data variabile nella marginalizzazione di tale variabile [20]. L'algoritmo junction tree elimination mira ad eliminare i cicli (che sono cammini non vuoti, in cui tutti gli archi sono distinti, dove solo il primo e l'ultimo vertice sono uguali) raggruppandoli in nodi singoli, riesce quindi poi a memorizzare i calcoli in modo che diverse classi estese di interrogazioni possono essere elaborate contemporaneamente in strutture dati più ampie e nuove evidenze propagate rapidamente. In ultima l'algoritmo recursive conditioning che offre un trade-off tra spazio e tempo e spiega una relazione quantitativa tra queste due risorse. [22]

Per reti ancora più grandi i metodi esatti non sono sufficienti e bisogna ricorrere a quelli per l'inferenza approssimata dato che forniscono stime di probabilità che richiedono significativamente meno calcoli. Uno di questi è il metodo Markov chain Monte Carlo (MCMC) che estrae campioni da una distribuzione di probabilità costruendo prima una Markov chain. Un altro algoritmo è il loopy belief propagation, utile quando la struttura del grafo include cicli e che itera al fine di ottenere le stime delle probabilità marginali fino a quando si ha la convergenza o una certa condizione di arresto viene soddisfatta. Tuttavia, potrebbe non convergere sempre, ma spesso produce risultati ragionevoli. Altri algoritmi sono il mini-bucket elimination e il variational Bayesian learning, che è un'approssimazione al Bayesian learning, ed è sviluppato per approssimare la densità a posteriori con l'obiettivo di minimizzare il "misfit" tra di essi,

oppure per derivare un limite inferiore per la probabilità marginale dei dati osservati, al fine di selezionare un modello. [20]

2.5 Apprendimento

Nel caso in cui una rete bayesiana non sia specificata da un esperto e sia troppo complessa da definire, bisogna effettuare, dai dati, un apprendimento (learning) della struttura e dei parametri della funzione di densità di probabilità congiunta. In generale possono esserci 4 casi di apprendimento, osservabili in Tabella 2.1; per osservabilità parziale si intende quando i nodi sono nascosti o mancano dei dati (missing values).

Caso	Struttura BN	Osservabilità	Metodo di learning proposto
1	Conosciuto	Piena	Stima di massima verosimiglianza
2	Conosciuto	Parziale	EM, MCMC
3	Sconosciuto	Piena	Ricerca nello spazio del modello
4	Sconosciuto	Parziale	EM + ricerca nello spazio del modello

Tabella 2.1: Quattro casi di apprendimento per una BN

Nel primo caso, il più facile, si ha struttura conosciuta e osservabilità piena e l'apprendimento è mirato a trovare i valori dei parametri, in ogni distribuzione di probabilità condizionata, che massimizzano la verosimiglianza logaritmica (maximum-likelihood) del set di dati. Quest'ultimo contiene m casi che spesso vengono considerati come indipendenti. Dato il set di dati $D = x_1, ..., x_m$, dove $x_l = (x_{l1}, ..., x_{ln})^T$, e il set di parametri $\Theta = (\theta_1, ..., \theta_n)$, dove θ_i è il vettore di parametri per la distribuzione condizionale della variabile X_i (rappresentata da un nodo nel grafo), il logaritmo della verosimiglianza del set di dati di addestramento è la somma di termini, uno per ciascun nodo:

$$\log L(\Theta|\Sigma) = \sum_m \sum_n \log P(x_{li}|\pi_i, \theta_i)$$

Nel secondo caso dove l'osservabilità è parziale si può utilizzare l'algoritmo expectation-maximization (EM) per ottenere una stima localmente ottimale della massima verosimiglianza dei parametri [23]. Definiamo come X i dati osservati, Z quelli non osservati o missing values e

θ come un vettore di parametri incogniti con $L(\theta; X, Z) = p(X, Z|\theta)$ funzione di verosimiglianza. L'algoritmo è iterativo ed alterna l'esecuzione di due passi: il primo chiamato expectation, che crea una funzione per il valore atteso della verosimiglianza logaritmica, calcolato usando la stima dei parametri corrente θ^t. Viene definito come:

$$Q(\theta|\theta^{(t)}) = E_{z|x,\theta^{(t)}}[\log L(\theta; X, Z)]$$

L'altro step è detto maximization, e calcola nuove stime dei parametri (che possono essere usate al passo successivo) massimizzando la funzione di verosimiglianza logaritmica attesa trovata al passo precedente. Viene definito come:

$$\theta^{(t+1)} = arg \max_{\theta} Q(\theta|\theta^{(t)})$$

Per quanto riguarda i casi 3 e 4, questi si basano sull'apprendimento della struttura, che va ricercata nello spazio esponenziale di tutti i possibili DAG ed è quindi più difficile, computazionalmente dispendiosa (NP-hard) [23]. Al fine di rendere questo compito più semplice e veloce si usano degli algoritmi euristici che trovano una soluzione approssimata. Questi algoritmi si possono dividere in due classi: independence analysis-based e score-based.

I metodi basati sull'independence analysis effettuano dei test condizionali sui dati e successivamente limitano ad un solo elemento il numero di possibili strutture che sono coerenti con i risultati dei test [20]. Ci sono approcci globali come PC e PC-stable che cercano di apprendere in una sola volta la struttura del grafo basandosi su decisioni di indipendenza condizionale, oppure esistono anche approcci di scoperta locale come Grow-Shrink e IAMB che cercano di apprendere prima lo scheletro del grafo, e poi di orientarlo. Anche se molto efficienti su set di dati con un grande numero di variabili, questi tipi di algoritmi rimangono dipendenti dalla scelta della soglia di significatività per il test statistico.

Invece, l'approccio basato sullo score (punteggio) consiste in una ricerca euristica, nello spazio di tutte le strutture, di una rete che si adatti al meglio al set di dati, ma solo dopo aver mappato ogni possibile struttura ad un punteggio. Si ricerca quindi, quella col punteggio maggiore utilizzando un metodo di ottimizzazione stocastica. L'algoritmo più usato è il greedy Hill-Climbing (HC), che partendo da una struttura vuota, esplora lo spazio dei possibili DAG aggiungendo, eliminando o invertendo un arco. Ad ogni passo, il DAG viene valutato tramite un punteggio di idoneità (fitness score) come, ad esempio, il Bayesian Information Criterion (BIC), l'Akaike Information Criterion (AIC), il Quotient Normalized Maximum Likelihood (qNML) o il Bayesian-Dirichlet equivalent uniform (BDeu). Verrà quindi selezionato il DAG, G, che massimizza il fit dei dati D:

$$argmax\ score(G, D)$$

con G che appartiene al set dei DAGs generati. Al fine di non rimanere bloccati in un massimo locale e avvicinarsi al massimo globale si perturbano i dati (esempio guardando subset incompleti) o si eseguono dei riavvi casuali. [24]

2.5.1 Algoritmi per l'apprendimento strutturale nel pacchetto R bnstruct

In questa tesi si è fatto uso del pacchetto bnstruct, implementato in R, che consente una gestione efficace dei missing values, molto comuni in ambito clinico. Il pacchetto dispone di 5 algoritmi che implementano l'apprendimento della struttura della rete (tramite il parametro algo) [25]. Sono quindi riportati qui di seguito, più nel dettaglio gli algoritmi:

- Silander-Myllymaki
- Max-Min Parent-and-Children
- Hill Climbing
- Max-Min Hill-Climbing
- Structural Expectation-Maximization

Silander-Myllymaki (SM)

È un algoritmo di search-and-score esatto (fornisce la soluzione globale ottima per il problema) che valuta l'intero spazio di ricerca con l'obbiettivo di trovare la miglior rete possibile. Si basa su 5 step logici:

1. Calcolare lo score locale per tutte le $n2^{n-1}$ differenti variabili
2. Usando il punteggio locale trovare i migliori genitori per tutte le $n2^{n-1}$ variabili
3. Trovare il miglior sink per tutti i 2^n insiemi di variabili. Ogni DAG ha almeno un nodo senza archi uscenti, quindi almeno un nodo, chiamato sink, non è genitore di nessun altro nodo.
4. Usando i risultati dello step 3, trovare il miglior ordinamento delle variabili
5. Trovare il miglior network usando i risultati trovati agli step 2 e 4

La sua applicazione è ristretta a reti di piccole dimensioni, infatti per reti con più di 25-30 nodi risulta inutilizzabile, poiché l'algoritmo tra i suoi difetti, oltre all'elevata memoria necessaria, richiede anche molto tempo per essere applicato. [26]

Max-Min Parent-and-Children (MMPC)

Questo algoritmo ha un approccio constraint-based, ovvero basato sui vincoli. La parte Max-Min del nome si riferisce al tipo di euristica che usa, mentre la parte Parent-and-Children all'output che produce. Si può dimostrare che l'insieme di genitori e figli di T è unico tra tutte le reti bayesiane aventi stessa distribuzione, possiamo quindi scriverlo come PC_T. Quindi, avendo un grafo fedele alla distribuzione dei dati e dei test statistici che producano risultati affidabili, MMPC restituisce PC_T, indicando che esistono degli archi entranti e uscenti da T (ma senza dirci l'orientamento). Usando come variabile target T ogni variabile della rete è possibile identificarne lo scheletro. L'euristica Max-Min seleziona la variabile che massimizza l'associazione minima con T relativa ai candidati per l'insieme PC_T. Grazie alla formula 2.1 che descrive un test di indipendenza statistica e misura la forza di associazione tra coppie di variabili è possibile identificare l'insieme dei candidati genitori e figli:

$$G^2 = 2 \sum_{abc} S_{ijk}^{abc} \ln \frac{S_{ijk}^{abc} S_k^c}{S_{ik}^{ac} S_{jk}^{bc}} \tag{2.1}$$

Con S_{ijk}^{abc} che rappresenta il numero di volte nei dati che $X_i = a$, $X_j = b$, $X_k = c$. Vengono definiti in modo simile S_k^c, S_{ik}^{ac}, S_{jk}^{bc}. La statistica G^2 è asintoticamente distribuita come χ^2 con i gradi di libertà che sono:

$$df = (|D(X_i)| - 1)(|D(X_j)| - 1) \prod_{X_l \in X_k} |D(X_l)|$$

dove $D(X)$ è il numero di valori distinti di una variabile X. Il test χ^2 ritorna un p-value che se è minore del livello α di significatività (di solito 0.05) rifiuta l'ipotesi nulla di mantenimento dell'indipendenza condizionata. [27]

Hill-Climbing (HC)

Hill-Climbing è un algoritmo di ricerca locale, o più specificatamente di ottimizzazione matematica, basato su iterazioni che partendo da un grafo vuoto prova a trovare una soluzione migliore. Il nome deriva dal fatto che l'algoritmo cerca di "scalare" gli archi verso quelli con valori maggiori, ovvero viene eseguita (ricorsivamente) l'aggiunta, la cancellazione o l'inversione dell'arco che massimizza una funziona obiettivo $f(x)$, dove x è un vettore di valori continui e/o discreti. Quindi, ogni qualvolta si ha un miglioramento nel valore di $f(x)$

si accetta il cambiamento e x viene considerato "localmente ottimale". La ricerca è vincolata a considerare solo l'aggiunta di un arco se è stato scoperto dall'algoritmo MMPC nella prima fase. Pur essendo un algoritmo veloce, perché non analizza tutti i nodi, ha la problematica di poter bloccarsi in un massimo locale trascurando così eventuali massimi globali presenti. Al fine di evitare questa criticità viene utilizzata una lista TABU che conserva le ultime 100 strutture esplorate e viene eseguita la migliore modifica locale che porta a una struttura non presente nella lista. Questa modifica potrebbe però ridurre il punteggio e quando si verificano 15 modifiche senza un aumento del punteggio massimo mai incontrato durante la ricerca, l'algoritmo termina e viene restituita la struttura complessiva con il miglior punteggio [27].

Max-Min Hill-Climbing (MMHC)

Questo algoritmo ha un approccio che combina metodi basati sull'apprendimento locale, di constraint-based e di search-and-score in modo efficace. Prima ricostruisce lo scheletro di una rete bayesiana e successivamente esegue una ricerca greedy basata su un punteggio bayesiano per orientare gli archi. In pratica è una combinazione dei due algoritmi precedenti appena descritti: infatti, prima esegue un'accurata selezione statistica dello spazio di ricerca grazie a MMPC e poi fa una valutazione di tipo greedy usando HC.

Structural Expectation-Maximization (SEM)

Il structural Expectation-Maximization risulta molto utile qualora si volesse apprendere una rete da un set di dati con valori mancanti. Essenzialmente si compone di due step, che vengono compiuti ad ogni iterazione, fino alla convergenza, che avviene quando non ci sono ulteriori miglioramenti nello score obiettivo (oppure quando viene raggiunto il numero massimo di iterazioni):

1. Stima i valori mancanti nei dati grazie all'uso dell'algoritmo generalizzato di Expectation-Maximization. L'algoritmo cerca di migliorare il modello calcolando e massimizzando lo score atteso, su un piccolo subset di modelli, invece del loro punteggio effettivo.

2. Utilizza MMHC per apprendere la struttura

Il guadagno che si ottiene, ed è stato dimostrato, è che così stiamo facendo una scelta migliore, in termini del punteggio marginale della rete. Un aspetto problematico potrebbe invece essere che SEM converga verso un modello sub-ottimale, ma si potrebbe eseguire l'algoritmo da diversi punti di partenza per ottenere una stima migliore. [28]

2.5.2 Funzioni costo nel pacchetto R bnstruct

I metodi di search-and-score necessitano di una funzione di valutazione per calcolare una misura stimata di ciascuna configurazione di nodi. Di seguito sono elencate le 3 funzioni costo implementate in bnstruct:

- Il criterio AIC è basato sul concetto di entropia come misura di informazione ed è definito come:

$$AIC = 2k - 2\ln L$$

  dove k è il numero di parametri nel modello statistico e L è il valore massimizzato della funzione di verosimiglianza del modello stimato.

- Il criterio BIC risulta strettamente correlato a quello AIC, ma con una penalizzazione più forte. E' definito come:

$$BIC = -2\ln(L) + k\ln(n)$$

  dove k è il numero di parametri nel modello statistico, n il numero di osservazioni e L è il valore massimizzato della funzione di verosimiglianza del modello stimato.

- Le proprietà chiave del Bayesian-Dirichlet equivalent uniform (BDeu) derivano dalla sua distribuzione uniforme sui parametri di ciascuna distribuzione locale nella rete, il che rende il learning della struttura computazionalmente efficiente. E' stato proposto da Buntine nel 1991 come caso particolare dello score BDe, la formula è:

$$BDeu(B,T) = \log(P(B)) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}(\log(\frac{\Gamma(\frac{N'}{q_i})}{\Gamma(N_{ij}+\frac{N'}{q_i})}) + \sum_{k=1}^{r_i}\log(\frac{\Gamma(N_{ijk}+\frac{N'}{r_i q_i})}{\Gamma(\frac{N'}{r_i q_i})}))$$

  Che succede quando

$$P(X_i = x_{ik}, \prod X_i = w_{ij}|G) = \frac{1}{r_i q_i}$$

  Il punteggio BDeu dipende solo da un parametro, la dimensione campionaria equivalente N', però non ci sono regole per determinarne il valore quindi ne vanno provati vari per capire quale è il migliore. [29]

2.6 Reti bayesiane dinamiche

La rappresentazione del modello di una rete bayesiana dinamica estende i modelli statici delle reti bayesiane a modelli di previsione dinamici più generali integrando e raffinando iterativamen-

te le dipendenze temporali rilevanti in differenti istanti di tempo, catturando come le variabili evolvono e come possono influenzarsi a vicenda. Viene però fatta l'assunzione che, sebbene le variazioni nelle forze esterne non modellate influenzino la forza delle dipendenze, cambiamenti nell'ambiente esogeno del modello non ne introducono di nuove o annullano quelle esistenti. [30] Il modello assume anche una proprietà di Markov per cui lo stato futuro dipende solo dallo stato corrente e non dalle osservazioni passate, in pratica tutta l'informazione necessaria per predire lo stato al tempo t è contenuta nella descrizione dello stato al tempo $t-1$. Poiché le DBN sono solo una sottoclasse delle reti bayesiane, i relativi algoritmi basati sulla struttura e sviluppati per esse, possono essere immediatamente applicati al ragionamento con le reti bayesiane dinamiche. [31]

Capitolo 3

Dati e preprocessing

I dati di pazienti affetti da sclerosi multipla che sono stati analizzati in questo lavoro di tesi, in ambiente di lavoro R, provengono dal database europeo del progetto BRAINTEASER (BRinging Artificial INTelligencE home for a better cAre of amyotrophic lateral sclerosis and multiple SclERosis [2]) e sono stati raccolti in diversi centri partner europei. In totale sono presenti 1792 pazienti, ognuno identificato da un ID univoco anonimizzato, il cui uso è fondamentale al fine di preservare la privacy dei pazienti. I dati consistono in un totale di 63878 record e 67 variabili diverse e sono suddivisi in 8 file CSV così organizzati:

- patients_data_and_symptoms: racchiude i dati demografici e d'esordio della malattia (unico file contenente dati di tipo statico)
- edss: riporta le valutazioni EDSS effettuate dal clinico
- evoked_potentials: include le informazioni riguardanti i potenziali evocati
- mri: contiene i dati relativi agli esami di risonanza magnetica
- ms_type: riporta i fenotipi della malattia
- relapses: include i dati riguardanti le ricadute
- therapeutic_procedures: contiene i trattamenti terapeutici prescritti
- environment: racchiude i dati sull'inquinamento ambientale

In ogni scheda è presente l'ID del paziente, la data di osservazione degli eventi in essa contenuti e il centro medico presso cui è stata effettuata la visita. Di seguito, affianco al nome di ciascuna scheda, viene riportata tra parentesi la dimensione della tabella (numero righe x numero colonne) ed una breve descrizione dei dati in essa contenuti.

3.1 Dati

3.1.1 PATIENTS_DATA_AND_SYMPTOMS (1792x21)

È l'unico file che contiene dati di tipo statico, ovvero il cui valore al tempo t non dipende dal suo valore al tempo $t-1$ o la cui variazione avviene in modo deterministico (ad esempio l'età). *Patients_data_and_symptoms* comprende una raccolta di informazioni demografiche come il sesso (*sex*), la residenza (*residence* e *residence_classification*), etnia (*ethnicity*), anno di nascita (*birth_year*), data di morte (*death_date*); ma anche informazioni cliniche relative alla prima visita come: la presenza o meno di sclerosi multipla in età pediatrica (*ms_in_pediatric_age*), la data della prima visita nel centro (*date_first_visit_in_the_centre*), la data ed età dell'esordio della malattia (*onset_date* ed *age_at_onset*), la data della diagnosi (*diagnosis_date*), il ritardo della diagnosi, in mesi, (*diagnostic_delay*) calcolato come la differenza tra la data dell'esordio della malattia e la data di diagnosi. È presente anche il criterio con cui è stata effettuata la diagnosi e il risultato di essa (*diagnosis_criteria_detail*, *diagnosis_criteria*) ed infine se sono stati rilevati o meno questi tipi di sintomi all'esordio: nel midollo spinale, nel tronco encefalico, alle vie ottiche, sopratentoriali o di altro genere (*spinal_cord_symptom*, *brainstem_symptom*, *eye_symptom*, *supratentorial_symptom*, *other_symptoms*).

3.1.2 EDSS (25289x13)

Il file contiene i punteggi della scala EDSS per ogni parte anatomica (vedi sezione 1.8), nello specifico per: il sistema piramidale, cerebellare, del tronco encefalico, intestinale e vescicale, sensoriale, funzione visiva, funzione cerebrale (*pyramidal*, *cerebellar*, *brainstem*, *bowel_and_bladder*, *sensory*, *visual_function*, *cerebral_functions* rispettivamente); tutti questi punteggi sono espressi in un range, che a seconda della variabile, va da 0 a 5 o da 0 a 6. La difficoltà nella deambulazione (*ambulation*) invece è espressa da 0 a 15 partendo da valori che esprimono una deambulazione totale e aumentano man mano che è richiesta sempre più assistenza al paziente che compie meno metri in autonomia, fino ad arrivare alla costrizione a letto e all'impossibilità di comunicare o deglutire. Infine, sono presenti il punteggio EDSS totale (*total_edss*) ottenuto come semplice somma degli score delle singole parti anatomiche, e il punteggio EDSS valutato dal clinico (*edss_as_evaluated_by_clinician*) ovvero calcolato tramite l'algoritmo della scala EDSS [13]. Quest'ultimo dato risulta fondamentale per il nostro studio poiché ci offre un criterio che cerca di essere il più oggettivo possibile per la comparazione dello stato di salute dei vari pazienti.

3.1.3 EVOKED_POTENTIALS (6212x6)

Sono descritti i dati delle visite relative ai potenziali evocati alterati (*altered_potential*) che possono essere di 4 tipi: visivo, somatosensoriale, motorio, uditivo. La variabile *potential_value* ci dice se il test ha osservato o meno un'alterazione del potenziale evocato e *location* ne indica la posizione.

3.1.4 MRI (7077x11)

Qui sono riportati i dati relativi alle visite di risonanza magnetica e con *mri_area_label* è indicata la zona su cui è stato effettuato l'esame, che può essere: tronco encefalico (Brain Stem), midollo spinale (Spinal Cord), tratto cervicale del midollo spinale (Cervical Spinal Cord), tratto toracico del midollo spinale (Thoracic Spinal Cord). È indicato poi se si sono osservate delle lesioni in T1 (*lesions_T1*), T1 con l'uso del mezzo di contrasto gadolinio (*lesions_T1_gadolinium*), col relativo numero di lesioni (*number_of_lesions_T1_gadolinium*) e T2 (*lesions_T2*) con il relativo numero di lesioni nuove o allargate dall'ultima risonanza magnetica (*number_of_new_or_enlarged_lesions_T2*) e numero di lesioni totali (*number_of_total_lesions_T2*).

3.1.5 MS_TYPE (3646x4)

In questo file sono riportate le visite fatte al fine di diagnosticare il tipo (*multiple_sclerosis_type*) di sclerosi multipla. Si contano 5 diverse forme: la sindrome clinicamente isolata (CIS), la primariamente progressiva (PP), la progressiva-recidivante (PR), la recidivante-remittente (RR), la progressiva secondaria (SP).

3.1.6 RELAPSES (6197x16)

In RELAPSES sono riportate le riacutizzazioni con tanto di data di inizio (*relapse_start_date*) e fine (*relapse_end_date*). Per ogni sistema funzionale è indicato se è stato colpito o meno e quelli presi in considerazione sono: sensoriale (*sensory_relapse*), piramidale (*pyramidal_relapse*), tronco encefalico (*brainstem_relapse*), sfinterico (*sphincter_relapse*), visivo (*vision_relapse*), cervelletto (*cerebellum_relapse*), psichico (*psychic_relapse*) o di altro tipo (*other_relapses*). Inoltre, è riportata anche la durata della ricaduta (*relapse_length*) ricavata dalle date, se c'è stato un trattamento cortisonico (*cortisone_treatment*), un ricovero clinico (*clinical_admission*) o un ricovero ospedaliero (*hospital_admission*).

3.1.7 THERAPEUTIC_PROCEDURES (13035x8)

Per ogni trattamento terapeutico prescritto, oltre alla data di inizio (*start_date*) e di fine (*end_date*) è specificata la categoria (*treatment_category*) che può essere: cortisonica, immunoattiva, un trattamento sintomatico della SM, un trattamento degli effetti collaterali o altri tipi di medicazione. Inoltre, è riportata l'indicazione che riguarda la sospensione del trattamento o meno (*has_been_suspended*), la ragione per la conclusione del trattamento (*end_reason*) e il codice ATC (Anatomical Therapeutic Chemical code, *ATC_code*) che è un codice unico assegnato a un farmaco in base all'organo o al sistema su cui agisce e al modo in cui agisce; il sistema di classificazione è gestito dall'Organizzazione Mondiale della Sanità (World Health Organization, WHO) [32].

3.1.8 ENVIRONMENT (242878x68)

Per quanto riguarda invece i dati ambientali, questi includono diverse rilevazioni per misurare la qualità dell'aria, e sono stati raccolti per 814 pazienti. La scheda riporta le date di rilevazione (che iniziano dal primo gennaio 2013 e finiscono il 6 dicembre 2021, in finestre di una settimana ciascuna), i dati su vari tipi di inquinanti e sulla temperatura, per un totale di 68 variabili. Ai fini dello studio però, proprio per le evidenze trovate in letteratura, abbiamo incluso nell'analisi solo: la data di inizio e fine della rilevazione che corrisponde sempre ad un arco di tempo di 7 giorni (*start_date* ed *end_date*), i valori di *PM10_mean*, ovvero particolato fine costituito da particelle inquinanti, di diametro inferiore a $10\mu m$, presenti nell'aria che respiriamo e i valori di *NO2_mean*, ovvero biossido di azoto che è sempre un inquinante dell'aria che viene normalmente generato a seguito di processi di combustione. Entrambi questi valori sono calcolati come la media delle rilevazioni effettuate durante la settimana. È stata mantenuta anche la variabile *season* che indica (con valori numerici da 1 a 4) in quale stagione è avvenuta la rilevazione.

3.2 Pulizia dei dati

Al fine di fornire i migliori dati possibili in input alla rete bayesiana e proprio per le caratteristiche di essa (che verranno descritte in seguito) è necessario un lavoro accurato di preprocessing dei dati. A grandi linee sono stati eliminati i dati e le variabili considerati superflui con un eventuale riadattamento delle singole schede ed unificazione di tutti i dati selezionati in un'unica tabella finale.

3.2.1 Preprocessing scheda EDSS

Più nel dettaglio, si è partiti dalla scheda EDSS in cui sono state eliminate tutte le variabili, tranne l'ID, la data della visita e *edss_as_evaluated_by_clinician*. Essendo la variabile riportante la valutazione dei punteggi EDSS centrale nel nostro studio, sono state rimosse tutte le righe il cui valore era valorizzata NA (290); data la piccola quantità è stato deciso di non imputare questi valori, anche perché l'algoritmo per il calcolo dell'EDSS non risulta banale 3.1. In particolare *centre* è stata cancellata da tutte le tabelle perché non considerata rilevante per l'analisi. Dopo l'elaborazione la scheda risulta come in Figura 3.1.

	patient_id	edss_date	edss_as_evaluated_by_clinician
1	10025147959814041565030309928029864407 7	2006-10-19	2.0
2	10025147959814041565030309928029864407 7	2007-01-22	2.0
3	10025147959814041565030309928029864407 7	2007-05-10	1.5
4	10038199677222038202107097495517621823 1	1997-01-14	1.0
5	10038199677222038202107097495517621823 1	1997-05-19	1.5
6	10038199677222038202107097495517621823 1	1997-09-08	1.0
7	10038199677222038202107097495517621823 1	1997-11-10	1.5
8	10038199677222038202107097495517621823 1	1998-02-02	1.0
9	10038199677222038202107097495517621823 1	1998-05-18	1.0
10	10038199677222038202107097495517621823 1	1998-09-07	1.0

Figura 3.1: Primi 10 record della scheda EDSS dopo il preprocessing

3.2.2 Preprocessing scheda MRI

Per quanto riguarda i dati di risonanza magnetica sono state rimosse le variabili: *number_of_lesions_T1_gadolinium*, *new_or_enlarged_lesions_T2*, *number_of_new_or_enlarged_lesions_T2*, *number_of_total_lesions_T2* perché o con elevato numero di NA o perché considerate non rilevanti per lo studio. In seguito, per ogni valore di *mri_area_label*, ovvero le 4 zone anatomiche su cui è stata eseguita l'MRI, è stata creata una colonna di T1, T1_gadolinium e T2 (esempio *BS_T1* indica la presenza o meno di una lesione di tipo T1 nel Brain Stem), portando così alla formazione di 12 nuove variabili dove per ognuna è stata calcolata la percentuale di TRUE, FALSE o NA, come è possibile osservare in Tabella 3.1.

Dalla sua analisi si può evincere che la presenza di NA risulta molto elevata, sopra al 70%, per tutte le variabili tranne che per *BS_T1g*, ovvero lesioni T1 con gadolinio nel tronco encefalico, motivo per cui è stato deciso di tenere solamente questa variabile e scartare le altre, trasformando così le dimensioni del dataframe MRI in 5522x3. La Figura 3.2 riporta la scheda

Nome	FALSE	TRUE	NA
BS_T1	0.10	0.16	0.74
BS_T1g	0.61	0.18	0.21
BS_T2	0.01	0.26	0.73
SC_T1	0.04	0.03	0.93
SC_T1g	0.20	0.04	0.76
SC_T2	0.02	0.04	0.94
CSC_T1	0.047	0.001	0.952
CSC_T1g	0.08	0.01	0.91
CSC_T2	0.01	0.03	0.96
TSC_T1	0.0187	0.0003	0.981
TSC_T1g	0.028	0.002	0.97
TSC_T2	0.005	0.008	0.987

Tabella 3.1: Nuove variabili della tabella MRI. Per ciascuna variabile derivata è riportata la percentuale di FALSE, TRUE e NA nel dataset. BS = Brain Stem (tronco encefalico); SC = Spinal Cord (midollo spinale); CSC = Cervical Spinal Cord (tratto cervicale del midollo spinale); TSC = Thoracic Spinal Cord (tratto toracico del midollo spinale)

MRI dopo il preprocessing, che ora contiene: l'ID del paziente, la data della visita e la variabile appena creata che indica quanto appena descritto.

3.2.3 Preprocessing scheda therapeutic_procedures

Rispetto alla tabella iniziale in therapeutic_procedures sono state eliminate tutte quelle righe con valori "–" e anche le colonne relative al codice ATC e alla motivazione di fine del trattamento dato che quest'ultime risultano informazioni troppo dettagliate per lo studio. Siamo più interessati a capire se è stato prescritto almeno un trattamento, in un certo periodo di tempo, la durata di esso ed il tipo. Per questo motivo è stato importante individuare i record con valore NA presenti in *end_date* (ovvero il 13% circa), di questi si è cercato poi quali avessero

	patient_id	mri_date	BS_T1g
1	10025147959814041565030309928029864407 7	2006-03-15	FALSE
2	10025147959814041565030309928029864407 7	2006-08-01	FALSE
3	10025147959814041565030309928029864407 7	2007-03-27	FALSE
4	10038199677222038202107097495517621823 1	1997-01-14	*NA*
5	10038199677222038202107097495517621823 1	2010-07-08	FALSE
6	10038199677222038202107097495517621823 1	2011-05-04	FALSE
7	10038199677222038202107097495517621823 1	2012-11-19	FALSE
8	10038199677222038202107097495517621823 1	2013-10-28	*NA*
9	10038199677222038202107097495517621823 1	2015-02-12	FALSE
10	10038199677222038202107097495517621823 1	2015-10-13	FALSE

Figura 3.2: Primi 10 record della scheda MRI dopo il preprocessing

anche *has_been_suspended* uguale a FALSE per imputare i valori in modo logico. Infatti alle righe che avessero come *treatment_category* uguale ad "Immunoactive medications" o "Symptomatic treatment of MS" è stata assegnata una *end_date = 2022-01-01*, giorno considerato come l'ultimo per l'osservazione dei pazienti nel database; per chi, invece, avesse *treatment_category* uguale a "Cortisone for relapses" è stata assegnata *end_date = start_date* dato che come spiegato in precedenza, alla sezione 1.7, possiamo assumere la somministrazione di questi farmaci come "one-shot" 3.3. Uno schema riassuntivo è illustrato in Figura 3.3.

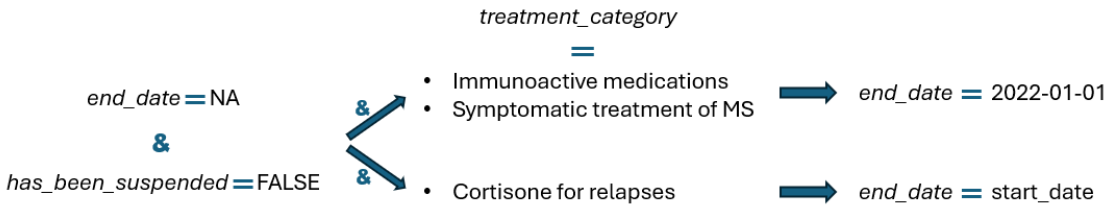

Figura 3.3: Riassunto imputazione *end_date* nella scheda therapeutic_procedures

Successivamente, riprendendo quanto fatto per la scheda MRI, sono state create tre nuove colonne: la prima denominata "cort_side_other" che accorpa le categorie riguardanti il cortisone, gli effetti collaterali e le altre medicazioni; la seconda riguarda i trattamenti sintomatici e la terza le medicazioni immunoattive. Per ognuna è stato indicato con TRUE o FALSE se fosse stato prescritto quel trattamento o meno. Alla fine di questa fase di preprocessing quindi therapeutic_procedures risulta come in Figura 3.4

	patient_id	start_date	end_date	cort_side_other	symptomatic_treatment	immunoactive_medications
1	10025147959814041565030309928029864407 7	2006-10-19	2006-11-05	TRUE	FALSE	FALSE
2	10025147959814041565030309928029864407 7	2007-01-22	2007-02-08	TRUE	FALSE	FALSE
3	10038199677222038202107097495517621823 1	1997-01-18	1997-01-23	TRUE	FALSE	FALSE
4	10038199677222038202107097495517621823 1	2013-03-14	2022-01-01	FALSE	TRUE	FALSE
5	10038199677222038202107097495517621823 1	2014-01-07	2022-01-01	TRUE	TRUE	FALSE
6	10038199677222038202107097495517621823 1	2015-03-16	2016-06-06	FALSE	FALSE	TRUE
7	10047040381791688904269018632283378295	1995-07-01	1995-07-05	TRUE	FALSE	FALSE
8	10047040381791688904269018632283378295	1996-04-01	1996-04-14	TRUE	FALSE	FALSE
9	10047040381791688904269018632283378295	1997-02-03	2001-03-20	FALSE	TRUE	FALSE
10	10047040381791688904269018632283378295	1997-02-03	2022-01-01	FALSE	TRUE	FALSE

Figura 3.4: Primi 10 record della scheda therapeutic_procedures dopo il preprocessing

3.2.4 Preprocessing scheda relapses

Per la scheda relapses sono stati rimossi: *relapse_end_date*, *relapse_length*, *clinical_admission* e *hospital_admission* per eccessiva presenza di NA (tutti sopra al 70%), mentre *cortisone_treatment* non è stata osservata congruente rispetto alla corrispettiva variabile nella tabella relativa alle procedure terapeutiche ed è quindi stata rimossa.

3.2.5 Preprocessing scheda ms_type

La scheda che raccoglie le informazioni sui tipi di sclerosi multipla è restata pressoché uguale, con le percentuali elencate in Tabella 3.2 che sono:

CIS	PP	PR	RR	SP
47.12%	1.67%	0.96%	46.15%	4.1%

Tabella 3.2: Percentuali delle tipologie di SM nella scheda ms_type

Al fine delle analisi implementate in questa tesi, si è deciso di non considerare i fenotipi, che risultavano dinamicamente poco valorizzati e, come atteso, per la maggioranza CIS e RR.

3.2.6 Preprocessing scheda evoked_potentials

Per quanto riguarda la tabella evoked_potentials si è deciso di eliminarla completamente visto che circa il 60% dei pazienti non ha nessun dato che riguarda i potenziali evocati.

3.2.7 Preprocessing scheda environment

Infine, per il dataframe riportante i dati ambientali, sono state rimosse le rilevazioni effettuate prima dell'esordio del paziente e posti uguali ad NA i valori di *PM10_mean_week* e *NO2_mean_week* valorizzati erroneamente come negativi. In Figura 3.5 è riportata la scheda environment dopo l'elaborazione, che contiene: l'ID del paziente, la stagione e la data di inizio e fine della rilevazione, i valori di PM10 e NO2 medi della settimana.

	patient_id	season	start_date	end_date	PM10_mean_week	NO2_mean_week
1	1409576303922404155996665112598820799 08	4	2013-01-01	2013-01-07	62.67143	49.21906
2	1409576303922404155996665112598820799 08	4	2013-01-08	2013-01-14	50.04286	43.52467
3	1409576303922404155996665112598820799 08	4	2013-01-15	2013-01-21	37.75000	61.03050
4	1409576303922404155996665112598820799 08	4	2013-01-22	2013-01-28	51.57143	59.91809
5	1409576303922404155996665112598820799 08	4	2013-01-29	2013-02-04	36.76429	38.52466
6	1409576303922404155996665112598820799 08	4	2013-02-05	2013-02-11	34.57857	38.79005
7	1409576303922404155996665112598820799 08	4	2013-02-12	2013-02-18	83.97857	55.40290
8	1409576303922404155996665112598820799 08	4	2013-02-19	2013-02-25	48.52857	32.42055
9	1409576303922404155996665112598820799 08	4	2013-02-26	2013-03-04	59.90000	34.20328
10	1409576303922404155996665112598820799 08	1	2013-03-05	2013-03-11	35.92143	29.44783

Figura 3.5: Primi 10 record della scheda environment dopo il preprocessing

Capitolo 4

Sviluppo dei modelli

Presi in considerazione i dati elaborati come descritto nel Capitolo 3 si è deciso di costruire quattro reti bayesiane, in cui, a partire dalla prima basata sulle sole variabili statiche e dall'informazione derivata dalla scheda EDSS, si sono via via aggiunte ulteriori variabili statiche e dinamiche. Si descrive di seguito la creazione di queste quattro reti.

4.1 Selezione ed organizzazione dei dati per la prima DBN

Si è passati quindi alla selezione delle variabili statiche che più sembravano significative da dare in input a tutte le reti, e sono:

- *sex*
- *residence_classification*
- *age_at_onset*
- *diagnostic_delay*
- *spinal_cord_symptom*
- *brainstem_symptom*
- *eye_symptom*
- *supratentorial_symptom*

Oltre a queste variabili statiche, la prima rete comprende come uniche variabili dinamiche informazioni derivate dalla scheda EDSS preprocessata. Specificatamente, si è fatta la scelta di utilizzare una finestratura annuale, di cui segue la spiegazione.

Dopo il processo di pulizia dei dati si è scelto di seguire un approccio atto a tracciare l'andamento dei valori di *edss_as_evaluated_by_clinician* nel tempo. Si è quindi come prima cosa introdotta una variabile chiamata "time since onset" (TSO) calcolata come la differenza tra la data in cui viene effettuata la valutazione EDSS e l'esordio della malattia (quest'ultimo corrispondente al giorno 0). Per fare ciò, si sono inizialmente individuati i pazienti con TSO negativo e quelli che hanno meno di 3 visite EDSS (numero minimo scelto per avere un andamento): questi risultavano essere 268 e sono stati eliminati da tutte le schede, di conseguenza il numero di pazienti totali rimasti si è abbassato a 1524.

Dopo un'analisi dell'andamento dei valori EDSS, si è notato che questo fosse prettamente lineare, con picchi, che teoricamente dovevano indicarci la presenza di una ricaduta, solo in una bassa percentuale di pazienti, motivo per cui si è deciso di eliminare dall'analisi le informazioni riguardanti i relapses contenute nella scheda omonima. Il passo successivo è stato interpolare i dati con un punto ogni mese, e creare finestre con lunghezza di 1 anno (essendo una malattia a lungo termine si è ritenuto fosse il periodo di tempo più adatto), introducendo la variabile *window_anni*, e per ognuna di esse calcolare l'EDSS medio (*edss_mean*) e l'EDSS massimo (*edss_max*). Dopo questa fase di manipolazione dei dati, si è ottenuto un dataframe chiamato window_edss che viene illustrato in Figura 4.1, contenente: l'ID del paziente, la variabile *window_anni* che racchiude l'informazione relativa alla finestra temporale di riferimento e i valori medi e massimi dell'EDSS.

	patient_id	window_anni	edss_mean	edss_max
1	10025147959814041565030309928029864407 7	1	1.9566358	2.0000000
2	10025147959814041565030309928029864407 7	2	1.5783179	1.6566358
3	10038199677222038202107097495517621823 1	20	1.2436119	1.4872239
4	10038199677222038202107097495517621823 1	21	1.1624452	1.4731254
5	10038199677222038202107097495517621823 1	22	1.0000000	1.0000000
6	10038199677222038202107097495517621823 1	23	1.0805005	1.3898042
7	10038199677222038202107097495517621823 1	24	1.1247233	1.4886727
8	10038199677222038202107097495517621823 1	25	1.3355838	1.5000000
9	10038199677222038202107097495517621823 1	26	1.1921865	1.3635540
10	10038199677222038202107097495517621823 1	27	1.5660793	1.7374468

Figura 4.1: Primi 10 record di window_edss

La DBN, per ogni variabile al tempo t, vuole in input anche la corrispettiva variabile all'istante successivo $t+1$, che quindi creeremo per ogni variabile dinamica utilizzata nelle quattro DBN. Come illustrato in Tabella 4.1 le variabili utilizzate, per la costruzione della prima DBN,

sono: le variabili statiche già selezionate, le due variabili EDSS appena ricavate (intese al tempo t) e la loro versione al tempo $t + 1$.

Variabili statiche	variabili statiche selezionate precedentemente
Variabili dinamiche	• edss_mean • edss_max • edss_mean_t1 • edss_max_t1

Tabella 4.1: Variabili Rete 1

Dall'unione dei dati statici selezionati e dei dati dinamici relativi all'EDSS così ottenuti, si è ricavata una prima tabella denominata final_table_net1 che contiene i dati da utilizzare per lo sviluppo della prima rete bayesiana. Come risulta visibile da Figura 4.2, in cui sono stati riportati: l'ID, i dati statici selezionati, la finestra temporale e i valori EDSS medi e massimi ricavati, nella struttura final_table_net1 i dati sono stati organizzati considerando una riga per ciascun soggetto e *window_anni*, risultando così in una collezione di più righe per soggetto (una per ogni finestra annuale disponibile), ciascuna con appesi i dati statici del soggetto stesso.

	id	sex	residence_classification	age_at_onset	diagnostic_delay	spinal_cord_symptom	brainstem_symptom	eye_symptom	supratentorial_symptom	window_(anni)	edss_mean	edss_max
1	10025147959814041565030309928029864407	female	Towns	25	7	TRUE	FALSE	FALSE	FALSE	1	1.9566358	2.0000000
2	10025147959814041565030309928029864407	female	Towns	25	7	TRUE	FALSE	FALSE	FALSE	2	1.5783179	1.6566358
3	10038199677222038202107097495517621823	female	Towns	32	1	FALSE	TRUE	FALSE	FALSE	20	1.2436119	1.4872239
4	10038199677222038202107097495517621823	female	Towns	32	1	FALSE	TRUE	FALSE	FALSE	21	1.1624452	1.4731254
5	10038199677222038202107097495517621823	female	Towns	32	1	FALSE	TRUE	FALSE	FALSE	22	1.0000000	1.0000000
6	10038199677222038202107097495517621823	female	Towns	32	1	FALSE	TRUE	FALSE	FALSE	23	1.0805005	1.3898042
7	10038199677222038202107097495517621823	female	Towns	32	1	FALSE	TRUE	FALSE	FALSE	24	1.1247233	1.4886727
8	10038199677222038202107097495517621823	female	Towns	32	1	FALSE	TRUE	FALSE	FALSE	25	1.3355838	1.5000000
9	10038199677222038202107097495517621823	female	Towns	32	1	FALSE	TRUE	FALSE	FALSE	26	1.1921865	1.3635540
10	10038199677222038202107097495517621823	female	Towns	32	1	FALSE	TRUE	FALSE	FALSE	27	1.5660793	1.7374468

Figura 4.2: Primi 10 record di final_table_net1

4.2 Selezione ed organizzazione dei dati per la seconda DBN

Per la creazione della seconda rete, ai dati della prima, è stata aggiunta come variabile statica *BS_T1g*, cioè l'unica variabile non scartata in precedenza riguardante la risonanza magnetica. La scelta di utilizzare i dati MRI in modo statico è figlia dell'alto numero di NA che ci sarebbero stati in caso fossero stati usati in modo dinamico. La Tabella 4.2 riporta quanto appena descritto.

Variabili statiche aggiunte	• BS_T1g

Tabella 4.2: Variabili aggiunte alla Rete 2

Specificatamente, volendo applicare una finestratura sempre di 1 anno, si è selezionata l'informazione relativa alla variabile *BS_T1g* nell'intorno dell'esordio. Nel dettaglio: per i soggetti con una sola rilevazione nel primo anno dopo l'esordio (TSO compreso tra 0 e 1 anno), si è considerato il valore di quest'ultima; per i soggetti con più rilevazioni nel primo anno dopo l'esordio è stato tenuto solamente, se presente, il valore TRUE, altrimenti sempre se presente, il valore FALSE e in assenza di entrambi è stato valorizzato ad NA; per i soggetti senza rilevazioni nel primo anno dopo l'esordio, si è cercata una rilevazione nell'anno precedente all'esordio. Per tutti i soggetti per cui non è stato possibile ricavare informazioni sui dati, è stato posto il valore di NA alla variabile statica. Viene quindi creata window_mri_stat che raccoglie solamente le informazioni relative a *BS_T1g* in un intorno dell'esordio, come si può notare in Figura 4.3.

	patient_id	window_anni	BS_T1g
1	1002514795981404156503030992802986440 77	1	FALSE
2	1004704038179168890426901863228337829 5	1	*NA*
3	1006192561890673867704844509606321242 11	1	*NA*
4	1016003339614271151252663455218264075 39	1	TRUE
5	1023093220042290792517917607964032964 11	1	TRUE
6	1031760040771522298933689685229145540 78	1	TRUE
7	1032231502703920583523703391533146747 92	1	TRUE
8	1058324027627549257699346269626659759 29	1	TRUE
9	1058520515486396157484869522021767111 48	1	TRUE
10	1076825523416347651864932410285461302 07	1	FALSE

Figura 4.3: Primi 10 record di window_mri_stat

Anche qui si sono uniti i dati in un'unica tabella finale che risulta come final_table_net1, ma con l'aggiunta della variabile *BS_T1g*. Qualora il dato MRI non sia presente per il paziente, la variabile statica *BS_T1g* appesa al dataset per la seconda rete risulta pari ad NA per tutte le righe relative a quel soggetto.

4.3 Selezione ed organizzazione dei dati per la terza DBN

Successivamente, abbiamo aggiunto ai dati della precedente DBN, quelli relativi alle terapie.

Sulla scheda preprocessata therapeutic_procedures sono state svolte delle elaborazioni dei dati per quanto riguarda il TSO (rispetto a *start_date*) sempre al fine di ottenere una finestratura di 1 anno. Qui, però, sono state considerate solo le variabili dei trattamenti sintomatici e dei farmaci immunoattivi, mentre la categoria "cort_side_other", viene scartata dato che non può essere usata per una rappresentazione dinamica della variabile. Per entrambe le variabili, per ogni anno, è stato tenuto il valore TRUE, se durante quell'anno era stato prescritto un medicamento di quella categoria per almeno un giorno, altrimenti è stato indicizzato a FALSE. Si è costruita quindi la tabella window_therapeutic illustrata in Figura 4.4.

	patient_id	window_anni	symptomatic_treatment	immunoactive_medications
1	10038199677222038202107097495517621823 1	36	TRUE	FALSE
2	10038199677222038202107097495517621823 1	37	TRUE	FALSE
3	10038199677222038202107097495517621823 1	38	TRUE	TRUE
4	10038199677222038202107097495517621823 1	39	TRUE	TRUE
5	10038199677222038202107097495517621823 1	40	TRUE	TRUE
6	10038199677222038202107097495517621823 1	41	TRUE	FALSE
7	10038199677222038202107097495517621823 1	42	TRUE	FALSE
8	10038199677222038202107097495517621823 1	43	TRUE	FALSE
9	10038199677222038202107097495517621823 1	44	TRUE	FALSE
10	10038199677222038202107097495517621823 1	45	TRUE	FALSE

Figura 4.4: Primi 10 record di window_therapeutic

Quindi rispetto alle variabili della seconda rete, sono state aggiunte per la costruzione della terza DBN, le variabili in Tabella 4.3. La tabella finale con i dati aggregati risulta uguale alla precedente con l'aggiunta delle variabili *symptomatic_treatment*, *immunoactive_medications*. Qualora una di queste due variabili non sia valorizzata in un anno specifico, risulta pari ad NA.

Variabili dinamiche aggiunte	• symptomatic_treatment • immunoactive_medications • symptomatic_treatment_t1 • immunoactive_medications_t1

Tabella 4.3: Variabili aggiunte alla Rete 3

4.4 Selezione ed organizzazione dei dati per la quarta DBN

Infine, per la quarta rete, si sono aggiunti, ai dati della terza, quelli inerenti alle rilevazioni ambientali.

Anche environment è stato finestrato e analogalmente ad EDSS sono state create 4 variabili: una rappresentante la media e una il massimo, rispetto ai valori assunti nell'anno da PM10; medesima cosa per NO2. Onde evitare che le variabili codificanti il valore massimo ed il valore medio degli inquinanti durante l'anno soffrissero di bias dovuti alla mancanza di dati per lunghi intervalli, è stato verificato che fosse presente almeno una rilevazione dell'inquinante per stagione nella finestra considerata pari ad un anno. In caso contrario, per quell'anno il valore dell'inquinante è stato posto pari a NA. Il dataframe creato risulta come in Figura 4.5, con la successiva unione di queste 4 variabili alla tabella finale della terza DBN.

	patient_id	window_anni	PM10_mean	PM10_max	NO2_mean	NO2_max
1	1409576303922404155996665112598820799 08	17	NA	NA	NA	NA
2	1409576303922404155996665112598820799 08	18	33.36042	78.17857	33.38599	63.63544
3	1409576303922404155996665112598820799 08	19	36.89205	89.21429	34.20832	70.30550
4	1409576303922404155996665112598820799 08	20	34.03950	89.15309	32.97540	76.12627
5	1409576303922404155996665112598820799 08	21	34.92131	81.29398	34.00925	73.90446
6	1409576303922404155996665112598820799 08	22	NA	NA	NA	NA
7	6739665461258937008362309240781076669 3	5	NA	NA	NA	NA
8	6739665461258937008362309240781076669 3	6	35.89855	76.35714	30.68186	55.35275
9	6739665461258937008362309240781076669 3	7	33.87748	86.04286	38.36846	108.66862
10	6739665461258937008362309240781076669 3	8	33.11550	94.44286	22.90646	50.48391

Figura 4.5: Primi 10 record di window_environment

La Tabella 4.4 illustra le variabili dinamiche aggiunte alla quarta rete.

Variabili dinamiche aggiunte	• PM10_mean • PM10_max • NO2_mean • NO2_max • PM10_mean_t1 • PM10_max_t1 • NO2_mean_t1 • NO2_max_t1

Tabella 4.4: Variabili aggiunte alla Rete 4

4.5 Quantizzazione

Il passaggio successivo è stato la quantizzazione dei dati contenuti nelle quattro tabelle finali secondo le regole per la costruzione di una DBN con il pacchetto bnstruct. Infatti, il dataframe deve contenere i dati così organizzati:

- no ID
- tutte le colonne quantizzate e numeriche
- ciascuna riga già organizzata per includere, oltre ai dati statici, le variabili dinamiche raccolte in due istanti consecutivi: dati statici, dati dinamici al tempo t, dati dinamici al tempo $t + 1$

I criteri adottati per la discretizzazione delle variabili sono principalmente due: per le variabili qualitative sono stati assegnati valori da 1 fino alla cardinalità della variabile, invece per le variabili quantitative, che contengono un range di valori numerici, sono state utilizzate soglie di cut-off basate sui terzili. In Tabella 4.5 vengono mostrati i relativi valori di quantizzazione.

Rispettando quanto detto sopra è stato tolto l'*id*, e per ogni variabile dinamica utilizzata viene creato un altro nodo che le rappresenta al tempo successivo $t + 1$; queste variabili nell'ultimo record di ogni paziente avranno valore NA. Gli id sono stati salvati come nome della riga puramente per un aspetto comprensivo.

4.6 Design e apprendimento delle DBN

Parte fondamentale della costruzione di una rete bayesiana sono le decisioni da prendere su: (i) gli archi obbligatori, dove si forzano due nodi ad essere collegati, (ii) il layering, tramite il quale le variabili vengono raggruppate in layer numerati, (iii) la layer structure, tramite la quale si specificano le relazioni tra i layer. Tramite questi parametri, sostanzialmente, si fornisce conoscenza a priori alla rete.

4.6.1 Implementazione della prima DBN

Gli archi obbligatori sono rappresentati da una matrice di zeri di dimensione $n \times n$, con n uguale al numero di nodi, con però il valore 1 in posizione (i, j) nel caso ci dovesse essere un arco dal nodo i al nodo j. Per le reti sviluppate in questa tesi, è stato inserito un arco obbligatorio tra la variabile *window_anni* e tutte le variabili al tempo $t + 1$, in particolare per la prima rete tra:

- *window_anni* e *edss_mean_t1*

Variabile	Livello di quantizzazione	Valore di quantizzazione
sex	$male$	1
	$female$	2
residence_classfication	$Cities$	1
	$RuralArea$	2
	$Towns$	3
age_at_onset	≤ 25	1
	$]25, 33]$	2
	> 33	3
diagnostic_delay	≤ 6	1
	$]6, 40]$	2
	> 40	3
spinal_cord_symptom	$FALSE$	1
	$TRUE$	2
brainstem_symptom	$FALSE$	1
	$TRUE$	2
eye_symptom	$FALSE$	1
	$TRUE$	2
supratentorial_symptom	$FALSE$	1
	$TRUE$	2
window_anni	≤ 7	1
	$]7, 15]$	2
	> 15	3
edss_mean	≤ 1.665	1
	$]1.665, 3.025]$	2
	> 3.025	3
edss_max	≤ 1.992	1
	$]1.992, 3.468]$	2
	> 3.468	3
BS_T1g	$FALSE$	1
	$TRUE$	2
symptomatic_treatment	$FALSE$	1
	$TRUE$	2
immunoactive_medications	$FALSE$	1
	$TRUE$	2
PM10_mean	≤ 31.121	1
	$]31.121, 36.17]$	2
	> 36.17	3
PM10_max	≤ 72.342	1
	$]72.342, 88.714]$	2
	> 88.714	3
NO2_mean	≤ 27.122	1
	$]27.122, 40.03]$	2
	> 40.03	3
NO2_max	≤ 55.618	1
	$]55.618, 78.699]$	2
	> 78.699	3

Tabella 4.5: Quantizzazione delle variabili incluse nelle quattro reti

- *window_anni* e *edss_max_t1*

Le relazioni tra i layer, invece, sono rappresentate da una matrice che segue le regole di quella per gli archi obbligatori, con la differenza che la presenza dell'arco che collega i due nodi è possibile e non mandatorio. Di default, il primo layer contiene variabili senza genitori, inoltre variabili in un layer possono avere genitori solo nello stesso layer o precedenti, a meno di relazioni sovraimposte nella layer structure.

La suddivisione delle variabili nei layer e le relazioni tra i layer stessi sono stati definiti come segue:

- Layer 1: *sex*, *residence_classification*, cioè le informazioni demografiche.
- Layer 2: *age_at_onset*. L'età d'esordio può essere influenzata dalle variabili al layer 1.
- Layer 3: *spinal_cord_symptom*, *brainstem_symptom*, *eye_symptom*, *supratentorial_symptom*. Il layer rappresenta le variabili statiche di tipo clinico. Per questo layer è stato rispettato il default, infatti dipende dai due layer precedenti e dalle variabili interne allo stesso layer.Ad esempio è ragionevole pensare che sesso e residenza influenzino la presenza o meno di sintomi all'esordio e che magari alcuni di essi siano correlati.
- Layer 4: *diagnostic_delay*. Il ritardo diagnostico dipende da tutti e tre i layer precedenti. Infatti questa variabile è temporalmente conseguente all'esordio.
- Layer 5: *edss_mean*, *edss_max*. Questo layer rappresenta le variabili al tempo t, le quali non vogliamo vengano influenzate dai layer precedenti né tra loro stesse. Ci interessa vedere solo che effetti queste abbiano sulle variabili al tempo $t + 1$.
- Layer 6: *edss_mean_t1*, *edss_max_t1*. Questo layer rappresenta le variabili al tempo $t + 1$ che possono avere archi provenienti da tutti gli strati precedenti, ma non da variabili interne allo stesso layer. Ci interessa vedere come queste variabili sono influenzate da sè stesse al tempo precedente (layer 5) e dalle variabili statiche. Inoltre è concesso anche un arco dal layer 7, essendo che contiene la variabile relativa al tempo.
- Layer 7: *window_anni*. La finestratura temporale non dipende da nulla, appunto perchè si tratta della variabile che riguarda il tempo.

La matrice della layer structure per la rete 1, illustrata in Tabella 4.6, risulterà quindi essere composta come:

Viene quindi creata la rete bayesiana usando come algoritmo strutturale Hill-Climbing e come funzione costo BIC. Come menzionato nella Sezione 2.5 per essere sicuri di ottenere un

layer	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
[1,]	0	1	1	1	0	1	0
[2,]	0	0	1	1	0	1	0
[3,]	0	0	1	1	0	1	0
[4,]	0	0	0	0	0	1	0
[5,]	0	0	0	0	0	1	0
[6,]	0	0	0	0	0	0	0
[7,]	0	0	0	0	0	1	0

Tabella 4.6: Matrice della layer structure per la prima rete

grafo "solido" e di non finire in un massimo locale, essendo Hill-Climbing un algoritmo deterministico, come rete di partenza per la ricerca della struttura, abbiamo fornito una rete ottenuta attraverso un processo di cross-validazione (CV) a 10 fold sulla totalità dei dati. In particolare, a ogni iterazione, è stata appresa una DBN su un nuovo sottoinsieme ottenuto unendo i dati di 9/10 fold, partendo da una rete nulla. I 10 diversi DAG risultanti da questa procedura sono stati combinati insieme in un Weighted Partially Directed Acyclic Graph (WPDAG), la cui matrice di adiacenza in posizione i, j tiene conto del numero di occorrenze dell'arco che va dal nodo i al nodo j nei 10 DAG così ottenuti. Un tale numero di fold ci permette di avere un tempo di calcolo ragionevole, inoltre sono state fatte prove con diversi valori, ma i risultati del WPDAG non cambiano. Infine, è stata ricavata tramite sogliatura una rete con solo gli archi più affidabili, cioè quelli inclusi in almeno l'80% dei DAG addestrati nella CV. Dopo aver visualmente verificato che l'ipotesi di aciclicità fosse soddisfatta, il DAG risultante è stato utilizzato come rete iniziale nell'apprendimento della prima rete bayesiana.

4.6.2 Implementazione della seconda DBN

La precedente rete viene fornita come rete di partenza per l'addestramento di questa DBN. Avendo inserito una sola variabile statica in più, rispetto alla prima rete, non ci sono vincoli aggiuntivi sugli archi obbligatori, ma viene solo aggiunta *BS_T1g* al layer 3 che quindi ora sarà:

- Layer 3: *spinal_cord_symptom*, *brainstem_symptom*, *eye_symptom*, *supratentorial_symptom*, *BS_T1g*

Non ci sono comunque variazioni nelle relazioni tra i layer.

4.6.3 Implementazione della terza DBN

La seconda DBN viene fornita come rete di partenza per l'addestramento di questa rete. In questo scenario si erano aggiunte alle variabili della seconda rete le due variabili dinamiche *symp-*

tomatic_treatment e *immunoactive_medications*. Di conseguenza si inseriscono i due vincoli aggiuntivi per gli archi obbligatori tra:

- *window_anni* e *symptomatic_treatment_t1*
- *window_anni* e *immunoactive_medications_t1*

Le due variabili al tempo t vengono aggiunte, insieme alle altre al tempo t, al layer 5 e le rispettive variabili al tempo $t + 1$ nel layer 6:

- Layer 5: *edss_mean*, *edss_max*, *symptomatic_treatment*, *immunoactive_medications*
- Layer 6: *edss_mean_t1*, *edss_max_t1*, *symptomatic_treatment_t1*, *immunoactive_medications_t1*

Anche qui non ci sono variazioni tra le relazioni dei layer.

4.6.4 Implementazione della quarta DBN

La terza rete viene fornita come rete di partenza per l'addestramento di quest'ultima DBN. Con l'introduzione delle variabili ambientali è stato modificato il layering per soddisfare la coerenza logica tra l'influenza delle variabili cliniche ed ambientali (le variabili ambientali possono avere un effetto sulle variabili cliniche nella finestra temporale successiva, come riportato dalla letteratura 1.2.1, ma non viceversa). Innanzitutto sono stati introdotti i vincoli tra:

- *window_anni* e *PM10_mean_t1*
- *window_anni* e *PM10_max_t1*
- *window_anni* e *NO2_mean_t1*
- *window_anni* e *NO2_max_t1*

Mentre il nuovo layering risulta:

- Layer 1: *sex*
- Layer 2: *residence_classification*
- Layer 3: *age_at_onset*
- Layer 4: *spinal_cord_symptom*, *brainstem_symptom*, *eye_symptom*, *supratentorial_symptom*, *BS_T1g*
- Layer 5: *diagnostic_delay*

- Layer 6: *edss_mean*, *edss_max*, *symptomatic_treatment*, *immunoactive_medications*
- Layer 7: *PM10_mean*, *PM10_max*, *NO2_mean*, *NO2_max*
- Layer 8: *edss_mean_t1*, *edss_max_t1*, *symptomatic_treatment_t1*, *immunoactive_medications_t1*
- Layer 9: *PM10_mean_t1*, *PM10_max_t1*, *NO2_mean_t1*, *NO2_max_t1*
- Layer 10: *window_anni* che non dipende da nulla, appunto perchè si tratta della variabile che riguarda il tempo.

Queste modifiche sono state necessarie in quanto è possibile che la residenza, ma non il sesso, influenzi i livelli di PM10 e NO2; inoltre, le variabili cliniche sono state divise da quelle ambientali poiché quest'ultime sono indipendenti dallo stato di salute del paziente e dipendono da fattori esterni non presi in considerazione nell'analisi. La matrice dei layer modificata per la quarta DBN risulta quindi essere come in Tabella 4.7:

layer	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7],	[,8]	[,9]	[,10]
[1,]	0	0	1	1	1	0	0	1	0	0
[2,]	0	0	1	1	1	0	0	1	1	0
[3,]	0	0	0	1	1	0	0	1	0	0
[4,]	0	0	0	1	1	0	0	1	0	0
[5,]	0	0	0	0	0	0	0	1	0	0
[6,]	0	0	0	0	0	0	0	1	0	0
[7,]	0	0	0	0	0	0	0	1	1	0
[8,]	0	0	0	0	0	0	0	0	0	0
[9,]	0	0	0	0	0	0	0	0	0	0
[10,]	0	0	0	0	0	0	0	1	1	0

Tabella 4.7: Matrice della layer structure per la quarta rete

Capitolo 5

Risultati

5.1 Prima DBN

In Figura 5.1 è rappresentata la rete bayesiana dinamica nella sua prima versione, in cui le variabili dinamiche sono solo *edss_mean* e *edss_max*. La presenza di una freccia entrante nella variabile z e proveniente dalla variabile x e di un'altra freccia sempre entrante in z, ma proveniente dalla variabile y, indica che "la variabile z dipende in maniera probabilistica e congiunta dalla combinazione delle variabili genitori, in questo caso x e y ". Analizzando le relazioni che emergono nella rete ottenuta, possiamo affermare che nel nostro caso, *age_at_onset* dipende in maniera probabilistica e congiunta dalla combinazione di *sex* e *residence_classification*. Questa dipendenza, infatti, viene anche confermata in letteratura con le donne che presentano un esordio prematuro della malattia rispetto agli uomini [33].

Essendo che, logicamente, abitare in un conglomerato urbano affollato come una città comporta un'esposizione all'inquinamento atmosferico maggiore che abitare in un'area rurale con una bassa densità abitativa, vista l'associazione riscontrata in letteratura tra inquinanti e SM, si può affermare che vivere in aree industrializzate e nelle vicinanze a strade principali potrebbe comportare un aumento del rischio di incorrere in un peggioramento della SM, oltre ad un rischio aumentato di sviluppare la malattia. Per quanto appena affermato risulta, quindi, notevole la relazione che si riscontra nella rete, in cui *residence_classification* influenza *edss_mean_t1* e *edss_max_t1*, congiuntamente ad altri parents (*edss_mean*, *edss_max*, *window_anni*). Questo dato ci viene anche confermato dalle tabelle di probabilità (CPT), nelle quali, si può notare che, quando *edss_max* ha un valore alto (maggiore di 3.468) la probabilità che in una finestra temporale che corrisponde a più di 15 anni dall'esordio, l'*edss_mean* passi da un valore medio al tempo t ad un valore alto al tempo $t + 1$ (secondo gli intervalli di quantizzazione riportati in Tabella 4.5) è: circa del 71.3% data la residenza in una città di grandi dimensioni, del 46.1% data la residenza in una cittadina, mentre è del 40% data la residenza in un'area rurale. In modo

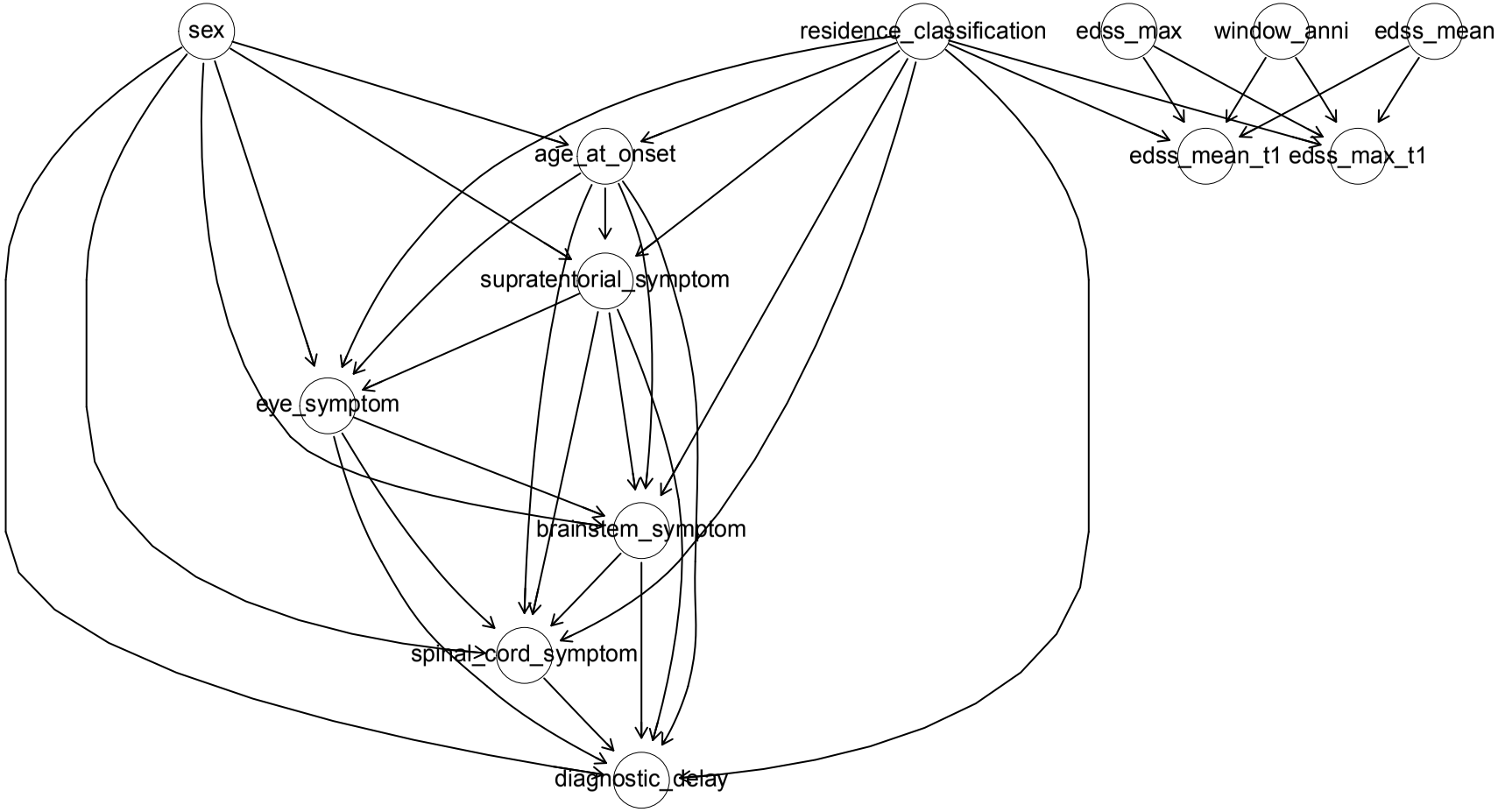

Figura 5.1: Plot prima DBN

simile si nota che, quando *edss_max* ha valori medi la probabilità che entro sette anni dall'esordio si passi da un *edss_mean* di livello basso ad un *edss_mean* di livello medio (secondo i valori di quantizzazione) sia del 40.3% se si vive in una grande città e solo del 14.5% se si vive in un'area rurale.

Come si può osservare dal plot della rete bayesiana dinamica inoltre, l'età di esordio della malattia sembrerebbe influenzare la presenza di sintomi nei quattro siti anatomici considerati. Sebbene non ci sia un ampissimo riscontro in letteratura, Cossburn et al evidenziano che secondo un'analisi di regressione logistica fatta gli Odds Ratio (ovvero il rapporto tra la frequenza con la quale un evento si verifica in un gruppo di pazienti e la frequenza con la quale lo stesso evento si verifica in un gruppo di pazienti di controllo) implicano ci sia lo 0.3% di aumento di rischio per quei sintomi associati all'aumento di età, e una riduzione di almeno lo 0.9% del rischio di neurite ottica per ogni anno di aumento dell'età di esordio [34]. Lo studio suggerisce che l'età sia, almeno in parte, legata all'insorgenza della malattia, avendo osservato che la neurite ottica si presenta più comunemente quando c'è un esordio giovanile della malattia, con un calo dell'occorrenza dopo i 30 anni [34]. Anche in questo caso le CPT concordano con quanto ipotizzato, dato che a prescindere dal sesso e dalla residenza e in assenza di sintomi sopratentoriali, si evidenziano in generale delle probabilità più alte di sviluppare sintomi visivi all'esordio se questo avviene prima dei 25 anni piuttosto che dopo i 33. Sembra quindi possibile che l'età di esordio non sia una variabile casuale, ma piuttosto che le sue proprietà abbiano un impatto sulle caratteristiche dell'esordio definendo presentazioni specifiche per età [34]. Quello che invece ci si sarebbe aspettati, poiché si è trovato anche riscontro in letteratura, ma che non emerge, invece, nella rete è un'influenza dell'età di esordio con i pazienti più anziani, rispetto a quelli di età più giovane, che hanno una tendenza maggiore ad una progressione nei valori dell'EDSS a breve termine [34]. Considerazioni analoghe riguardano il sesso, visto che secondo la letteratura gli uomini sono più inclini ad avere una rapida progressione nei valori EDSS con degli esiti peggiori, avendo quindi un ruolo di influenza nella disabilità della malattia [33]; questa relazione tuttavia non emerge dalla rete ottenuta sui dati.

5.2 Seconda DBN

Figura 5.2 riporta la rete ottenuta a partire dalla seconda versione dei dati, dando come rete iniziale la prima. Si può notare come, rispetto alla precedente rete, *edss_max* si isola e risulta non avere relazioni con altre variabili, e viene persa un'evidenza importante come la relazione tra *residence_classification* e le due variabili EDSS al tempo $t + 1$.

Ciò che si può osservare però, è il ruolo da "nodo centrale" di *BS_T1g*, che risulta essere genitore o figlio di quasi ogni altro nodo della rete. Una possibile spiegazione riguarda il fatto

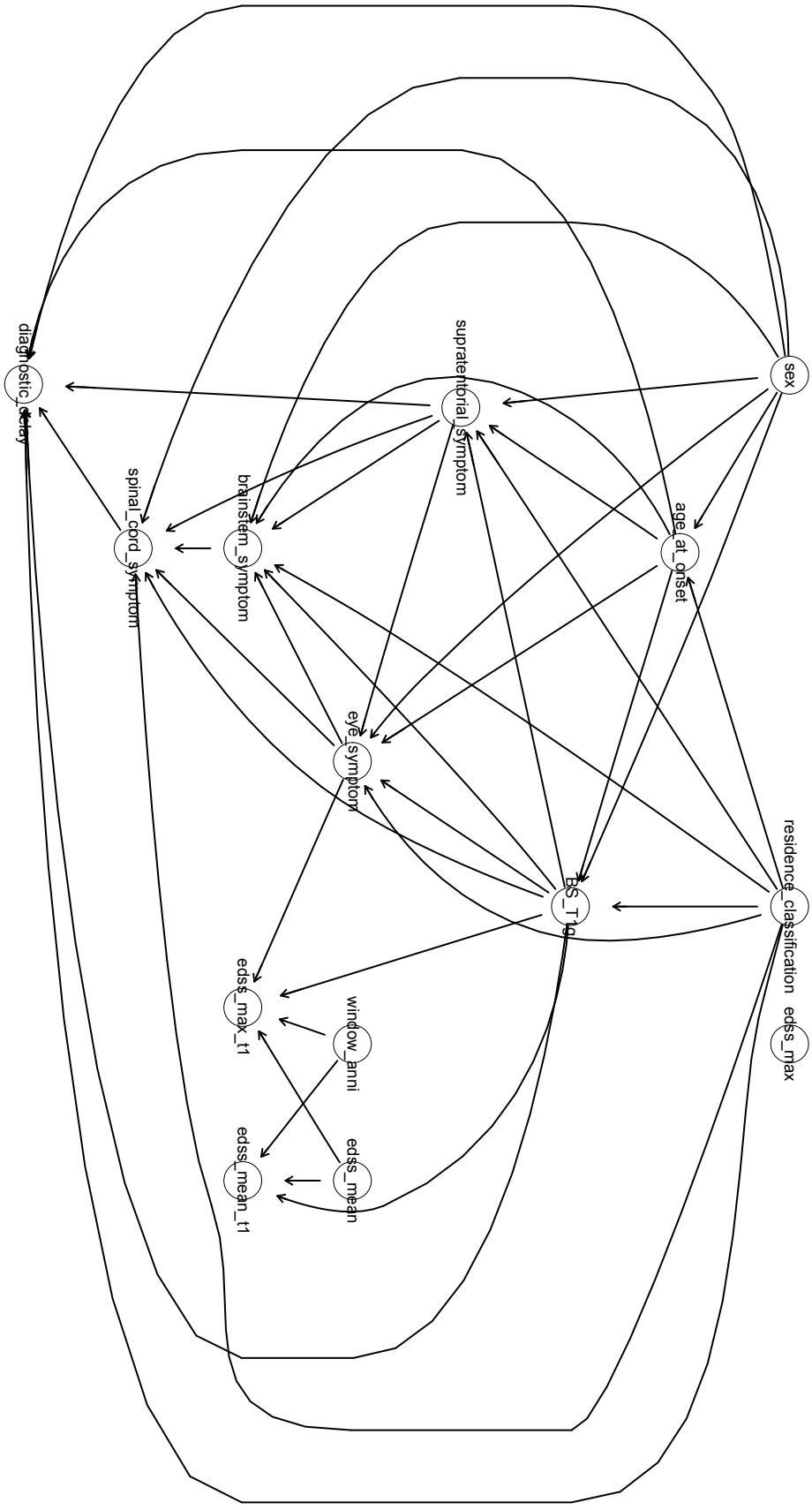

Figura 5.2: Plot seconda DBN

che, come riportato in precedenza, se sono presenti una o più lesioni cerebrali individuate tramite MRI si avrà una probabilità maggiore di sviluppare i sintomi e di conseguenza di avere un peggioramento dell'EDSS, come suggerisce anche la relazione tra la presenza di lesioni T1 con gadolinio con i valori di EDSS medio e massimo. Nel dettaglio, guardando le tabelle di probabilità condizionata, se è presente almeno una lesione al tronco encefalico, c'è generalmente una probabilità maggiore di registrare sintomi riguardanti il tronco encefalico, considerando i primi sette anni dall'esordio e l'assenza di sintomi visivi e sopratentoriali, a prescindere dal sesso e dalla residenza. Inoltre, dalla letteratura si evince come nei pazienti con malattia recidivante-remittente, l'EDSS correla con il numero di lesioni con potenziamento del gadolinio, con un valore del coefficiente di correlazione di 0,52. Un esempio dettagliato della rilevanza clinica del potenziamento con gadolinio è il nervo ottico, in cui la presenza del potenziamento è molto predittiva di segni e sintomi clinici [35].

5.3 Terza DBN

In Figura 5.3 è raffigurata la rete ottenuta a partire dalla terza versione dei dati, fornendo come rete iniziale la seconda. Come nella precedente rete *edss_max* è isolato, con le relazioni tra le variabili statiche che rimangono molto simili. Notiamo che *BS_T1g* è genitore di tutte le variabili al tempo $t + 1$ e specificatamente di *symptomatic_treatment* e *immunoactive_medications*. Poiché come abbiamo detto la presenza di lesioni a volte si manifesta nella comparsa di sintomi, possiamo ipotizzare che questi abbiano comportato la prescrizione di una terapia. A supporto di questa ipotesi, analizzando le CPT corrispondenti osserviamo che: se non si stanno assumendo farmaci immunoattivi al tempo t, la probabilità che al tempo $t + 1$, in una finestra temporale maggiore di 15 anni dall'esordio, sia prescritto un farmaco immunoattivo è del 43% se era presente almeno una lesione nell'intorno dell'esordio considerato in fase di preprocessing, mentre solo del 2% in assenza di lesioni.

La rete evidenzia anche il legame tra la prescrizione di trattamenti sintomatici e i valori dell'EDSS. Viene quindi rilevata l'importanza e l'efficacia dei trattamenti sintomatici, che viene riscontrata anche in uno studio pubblicato nel 2020, nello specifico sulla fampridina (che era l'unico farmaco approvato per il trattamento della difficoltà di deambulazione nei pazienti con SM), in cui dagli studi clinici, circa un terzo dei soggetti sottoposti al trattamento ha ottenuto una risposta clinicamente significativa, ed il totale sembra crescere addirittura fino al 75% nel mondo reale [36].

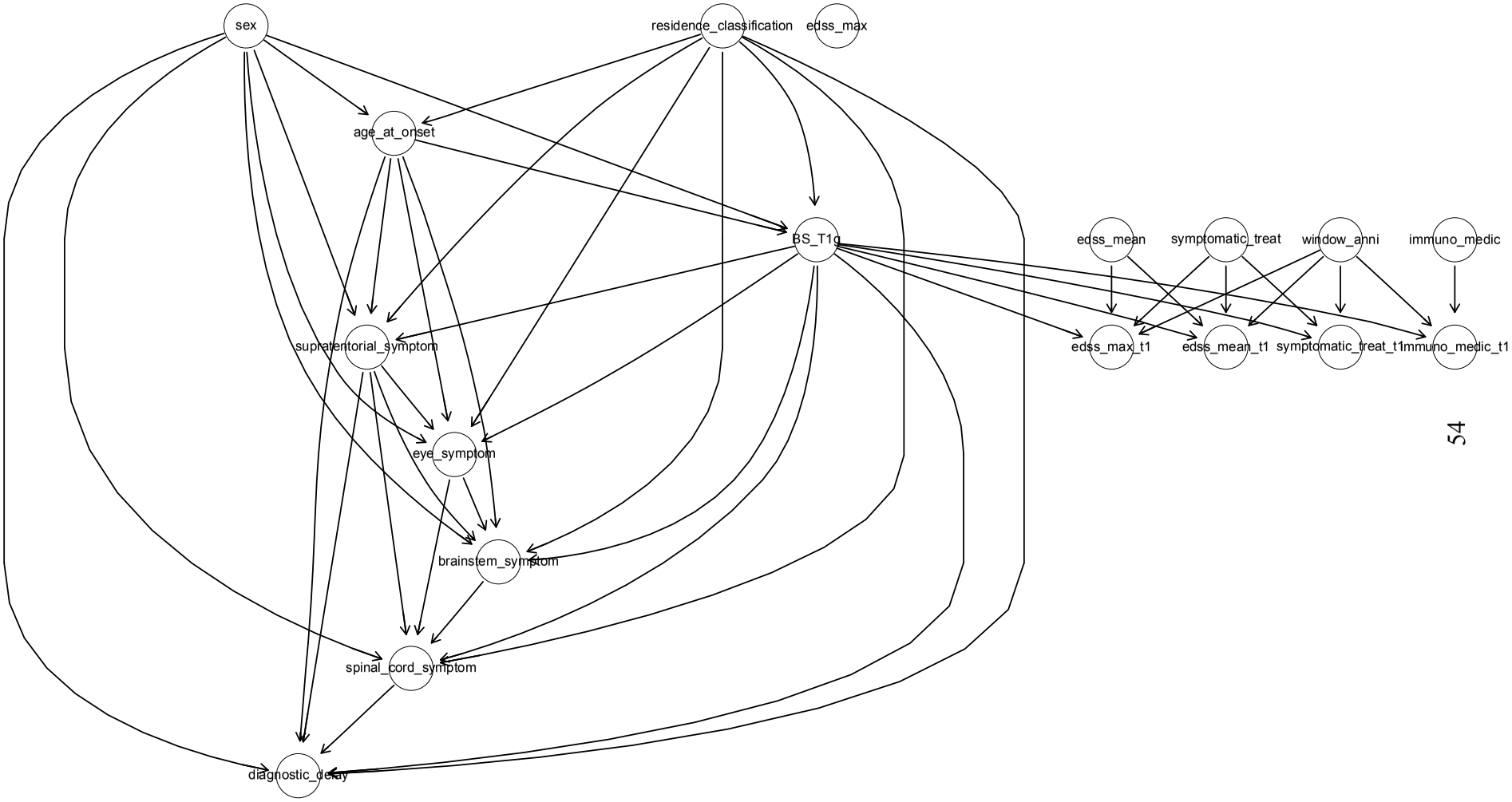

Figura 5.3: Plot terza DBN

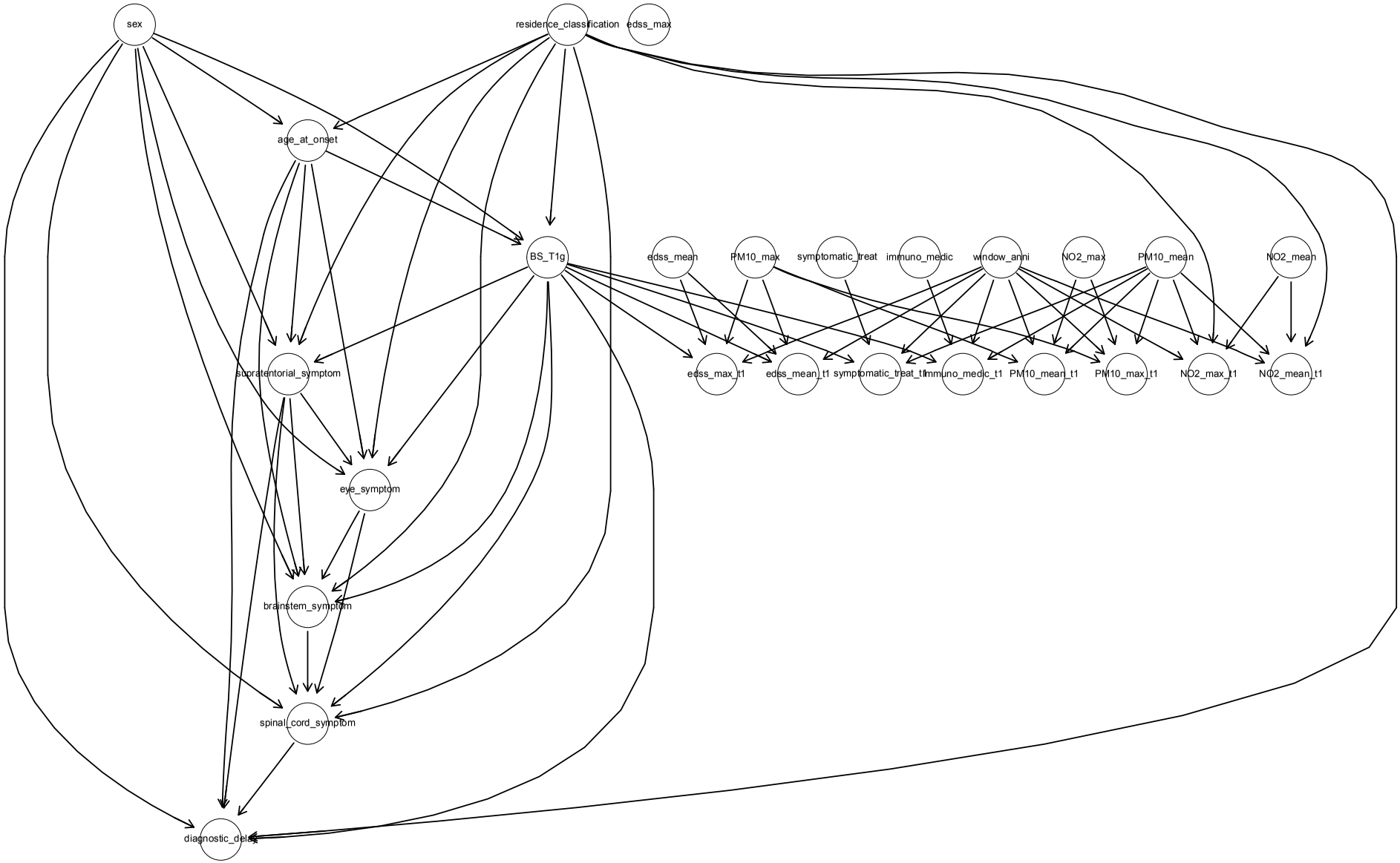

Figura 5.4: Plot quarta DBN

5.4 Quarta DBN

In Figura 5.4 si può osservare la quarta rete dello studio ottenuta a partire dalla quarta versione dei dati, dando come rete iniziale la terza. Questa DBN, come la terza e la seconda, presenta: relazioni molto simili tra le variabili statiche, *edss_max* isolato rispetto agli altri nodi ed in aggiunta *BS_T1g* sempre genitore di tutte le variabili al tempo $t + 1$. A seguito delle evidenze trovate nella prima rete in cui la residenza influenzava il valore dell'EDSS, si è deciso di introdurre le 4 variabili ambientali con l'intento di avvalorare quanto già trovato e capire nel dettaglio come questi due inquinanti possano influenzare le altre variabili.

Innanzitutto è interessante osservare come la residenza influenzi i livelli di NO2, infatti chi vive in città ha probabilità del 99% di avere valori elevati di NO2 medio al tempo $t + 1$ dato che aveva valori elevati sia di PM10 medio che di NO2 medio entrambi al tempo t, considerando una finestra temporale entro sette anni dall'esordio, mentre chi vive in un'area rurale solo del 33%. Si possono osservare valori simili se al tempo $t + 1$ consideriamo NO2 massimo e non medio. Si nota anche che, come ci si aspetterebbe, il valore dei livelli dei due inquinanti siano molto correlati tra loro.

Come riportato precedentemente nel Capitolo 1 i valori di PM10 contribuiscono al peggioramento della malattia, infatti lo si può notare dal fatto che *PM10_max* sia genitore di *edss_max_t1* e *edss_mean_t1*.

Inoltre, l'arco tra *PM10_mean* e i trattamenti sintomatici potrebbe suggerire una relazione tra l'aumento di inquinamento e la prescrizione dei farmaci.

Capitolo 6

Conclusioni

In questo lavoro di tesi è stato analizzato un database contenente informazioni demografiche e cliniche per un insieme di pazienti affetti da sclerosi multipla. Essendo la prognosi di questa malattia molto variabile, con un'elevata eterogeneità fenotipica, lo scopo della tesi è stato quello di modellizzare la progressione della SM al fine di ricavare informazioni sulla natura delle relazioni che legano le variabili presenti nel database e la prognosi osservata. In particolare, tramite l'utilizzo di reti bayesiane dinamiche, ci si è posto l'obiettivo di dare una preliminare rappresentazione di quali variabili influenzino maggiormente lo stato generale dei pazienti, espresso dal punteggio della scala EDSS. Dopo un accurato preprocessing, si sono aggregati, quantizzati ed elaborati i dati di 1524 pazienti, al fine di costruire quattro reti bayesiane via via più complesse in termini di variabili incluse.

Si è partiti da una prima rete, il cui learning è stato rafforzato tramite una procedura di cross-validazione per definire la rete iniziale, dove si sono considerati solo alcuni dati statici dei pazienti e i valori EDSS. Si sono poi costruite altre tre reti aggiungendo via via altre variabili, come: la presenza di eventuali lesioni nel tronco encefalico rilevate tramite risonanza magnetica (con pesatura T1 con gadolinio), la prescrizione di trattamenti sintomatici ed immunoattivi ed infine i dati sull'inquinamento ambientale dell'aria riguardo i valori di PM10 ed NO2.

Osservando i grafi ottenuti possiamo trarre delle conclusioni, che almeno in parte, soddisfano gli obiettivi che ci si era posti per l'analisi. Infatti, notiamo l'influenza dell'età di esordio della malattia sulla presenza dei sintomi all'esordio nei quattro siti anatomici considerati: in particolare, come si riscontra anche nelle CPT della prima rete, si evidenziano in generale delle probabilità più alte di sviluppare sintomi visivi all'esordio se questo avviene prima dei 25 anni piuttosto che dopo i 33, a prescindere dal sesso e dalla residenza e in assenza di sintomi sopratentoriali; sembrerebbe quindi possibile che l'età d'esordio della malattia non sia una variabile casuale, ma che invece influenzi le caratteristiche dell'esordio.

Rilevante sembra essere anche come, nella seconda rete, se è presente almeno una lesione al

tronco encefalico, c'è generalmente una probabilità maggiore di registrare sintomi riguardanti il tronco encefalico stesso, considerando i primi sette anni dall'esordio e l'assenza di sintomi visivi e sopratentoriali, a prescindere dal sesso e dalla residenza. La presenza di lesioni nel tronco encefalico all'esordio sembrerebbe comportare di conseguenza un peggioramento nei valori EDSS.

Per quanto riguarda i trattamenti terapeutici inclusi nella terza DBN, osserviamo che, se consideriamo una finestra temporale maggiore di 15 anni dall'esordio e non si stanno assumendo farmaci immunoattivi al tempo t, la probabilità che al tempo $t + 1$ sia prescritto un farmaco immunoattivo è del 43% qualora fosse presente almeno una lesione nell'intorno dell'esordio, mentre solo del 2% in assenza di lesioni; si evidenza così come la presenza di lesioni comporti una probabilità maggiore nella prescrizione di un farmaco immunoattivo nel tempo.

L'evidenza più notevole di tutto il lavoro di tesi risulta essere il legame tra la residenza del paziente e i valori dell'EDSS: tale risultato emerge già dalla prima rete, in cui si osserva che, quando *edss_max* ha un valore alto (maggiore di 3.468) la probabilità che, in una finestra temporale che corrisponde a più di 15 anni dall'esordio, l'*edss_mean* passi da un valore medio al tempo t ad un valore alto al tempo $t + 1$ è circa del 71.3% data la residenza in una città di grandi dimensioni, contro il 46.1% se il paziente vive in una cittadina, e del 40% data la residenza in un'area rurale. Queste osservazioni, che trovano riscontro con quanto presente in letteratura in merito alla relazione tra inquinamento ambientale e prognosi, hanno suggerito di affinare l'analisi considerando l'informazione specifica disponibile di due inquinanti ambientali, PM10 e NO2. La quarta rete, che comprende in aggiunta queste due variabili, ha confermato che chi vive in aree più densamente popolate ha maggiori probabilità di avere i livelli dei due inquinanti (che sono correlati tra loro) più elevati e che questi abbiano un effetto sulla prognosi della malattia, con *PM10_max* che risulta essere genitore di *edss_max_t1* e *edss_mean_t1*.

Nonostante il buon numero di relazioni rilevanti trovate tra le variabili e diverse conferme sulla direzione di tali relazioni tramite l'analisi delle CPT, si osserva come, aggiungendo via via più variabili (alcune non valorizzate per circa la metà dei pazienti, come quelle ambientali), emergano alcune differenze significative tra le varie reti. Tra queste, si osserva: l'isolamento di *edss_max* e la perdita della relazione tra la residenza e i valori EDSS nelle reti 2, 3 e 4, oppure la mancanza di una relazione tra l'età d'esordio e il sesso con i valori EDSS. Inoltre, si osserva come solo la variabile *PM10_max* risulti genitore dello score EDSS al tempo $t + 1$, probabilmente a causa dell'alta correlazione di questa variabile con le altre relative ai livelli di inquinamento ambientale.

6.1 Sviluppi futuri

Considerata la multifattorialità della sclerosi multipla sia per quanto riguarda l'esordio che la prognosi, la caratterizzazione dei pazienti da un punto di vista clinico, ambientale e genetico ampio può contribuire a fare luce su questa malattia. Come sviluppo futuro di questo lavoro di tesi, si potrebbe pensare quindi ad introdurre o valorizzare meglio variabili che permettano di definire più nel dettaglio lo stato di salute dei pazienti nel tempo.

Focalizzandosi nello specifico sull'utilizzo delle reti bayesiane dinamiche per la modellizzazione di patologie come la sclerosi multipla, inoltre, si potrebbe estendere il lavoro presentato in questa tesi indagandone il loro uso per scopi predittivi. Utilizzando i dati storici dei pazienti affetti da SM di una popolazione di riferimento e le informazioni sui fattori di rischio disponibili in letteratura, si può pensare, infatti, di utilizzare le DBN per prevedere i futuri risvolti clinici dei pazienti, come ad esempio la progressione della disabilità tramite una stima del punteggio EDSS, oppure la risposta alle cure dopo i trattamenti farmacologici. Una previsione accurata e personalizzata dello sviluppo della malattia renderebbe possibile ottimizzarne le strategie di gestione e di individualizzazione dei trattamenti, contribuendo a perfezionare una medicina mirata con interventi specifici e tempestivi. Tutto ciò aiuterebbe a migliorare l'esito clinico dei pazienti affetti da malattie croniche come la sclerosi multipla, aiutando e supportando i clinici a prendere decisioni informate e personalizzate.

Bibliografia

[1] EpiCentro, *Sclerosi multipla epidemiologia*. indirizzo: https://www.epicentro.iss.it/sclerosi-multipla/epidemiologia.

[2] C. Cordis, *BRinging Artificial INTelligencE home for a better cAre of amyotrophic lateral sclerosis and multiple SclERosis*, dic. 2020. indirizzo: https://cordis.europa.eu/project/id/101017598.

[3] *La diffusione della sclerosi multipla a livello globale*. indirizzo: https://www.multiplesklerose.ch/it/attualita/dettaglio/la-diffusione-della-sclerosi-multipla-a-livello-globale/.

[4] R. Dobson e G. Giovannoni, «Multiple sclerosis – a review,» *European Journal of Neurology*, vol. 26, n. 1, pp. 27–40, nov. 2018. doi: 10.1111/ene.13819. indirizzo: https://doi.org/10.1111/ene.13819.

[5] A. Compston e A. Coles, «Multiple sclerosis,» *The Lancet*, vol. 372, n. 9648, pp. 1502–1517, ott. 2008. doi: 10.1016/s0140-6736(08)61620-7. indirizzo: https://doi.org/10.1016/s0140-6736(08)61620-7.

[6] S. Abbaszadeh, M. Tabary, A. Aryannejad et al., «Air pollution and multiple sclerosis: a comprehensive review,» *Neurological Sciences*, vol. 42, n. 10, pp. 4063–4072, ago. 2021. doi: 10.1007/s10072-021-05508-4. indirizzo: https://link.springer.com/article/10.1007/s10072-021-05508-4.

[7] R. Bergamaschi, A. Cortese, A. Pichiecchio et al., «Air pollution is associated to the multiple sclerosis inflammatory activity as measured by brain MRI,» *Multiple Sclerosis Journal*, vol. 24, n. 12, pp. 1578–1584, ago. 2017. doi: 10.1177/1352458517726866. indirizzo: https://journals.sagepub.com/doi/10.1177/1352458517726866.

[8] M. Elgabsi, L. Novack, S. Yarza, M. Elgabsi, A. Shtein e G. Ifergane, «An impact of air pollution on moderate to severe relapses among multiple sclerosis patients,» *Multiple Sclerosis and Related Disorders (Print)*, vol. 53, p. 103 043, ago. 2021. doi: 10.1016/j.msard.2021.103043. indirizzo: https://www.sciencedirect.com/science/article/pii/S2211034821003102.

[9] *Robbins e Cotran - Le basi patologiche delle malattie*, 9ª ed. Edra, 2017, vol. 2.

[10] D. Maimone, «Terapia immunoricostituente nella sclerosi multipla: razionale e potenzialità,» 2019. indirizzo: https://www.smilejournal.it/media/smile2S-2019/38-43.pdf.

[11] *Sclerosi multipla: quante forme esistono? | AISM | Associazione Italiana Sclerosi Multipla*. indirizzo: https://www.aism.it/sclerosi_multipla_forme.

[12] I. Cerillo, «I nuovi criteri diagnostici per la sclerosi multipla (McDonald 2017) e il loro impatto nel real world setting,» 2019. indirizzo: https://www.smilejournal.it/media/SMILE-supp3_2019.pdf.

[13] J. F. Kurtzke, «Rating neurologic impairment in multiple sclerosis,» *Neurology*, vol. 33, n. 11, p. 1444, nov. 1983. doi: 10.1212/wnl.33.11.1444. indirizzo: https://pubmed.ncbi.nlm.nih.gov/6685237/.

[14] *Expanded Disability Status Scale (EDSS)*. indirizzo: https://mstrust.org.uk/a-z/expanded-disability-status-scale-edss#edss.

[15] S. Sen, «NEUROSTATUS AND EDSS CALCULATION WITH CASES,» *Noropsikiyatri Arsivi*, gen. 2018. doi: 10.29399/npa.23412. indirizzo: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6278622/.

[16] contributori di Wikipedia, *Digrafo aciclico*, apr. 2023. indirizzo: https://it.wikipedia.org/wiki/Digrafo_aciclico.

[17] G. Zweig e S. Russell, «Probabilistic modeling with Bayesian networks for automatic speech recognition,» in *Proc. 5th International Conference on Spoken Language Processing (ICSLP 1998)*, 1998, paper 0858. doi: 10.21437/ICSLP.1998-198.

[18] L. Nakhleh, *Probabilistic Modeling: Bayesian Networks*, 2015. indirizzo: https://www.cs.rice.edu/~nakhleh/COMP571/Slides-Spring2015/ProbabilisticModels.pdf.

[19] *Causal and statistical dependence*. indirizzo: https://probmods.org/chapters/dependence.html.

[20] K.-L. Du e M. N. S. Swamy, «Probabilistic and Bayesian networks,» in *Springer eBooks*. dic. 2013, pp. 563–619. doi: 10.1007/978-1-4471-5571-3_19. indirizzo: https://link.springer.com/chapter/10.1007/978-1-4471-5571-3_19.

[21] W. contributors, *Markov blanket*, gen. 2024. indirizzo: https://en.wikipedia.org/wiki/Markov_blanket.

[22] A. Darwiche, «Recursive conditioning,» *Artificial Intelligence*, vol. 126, n. 1–2, pp. 5–41, feb. 2001. doi: 10.1016/s0004-3702(00)00069-2. indirizzo: https://www.sciencedirect.com/science/article/pii/S0004370200000692.

[23] F. Ruggeri, R. Kenett e F. Faltin, *Encyclopedia of Statistics in Quality and Reliability*. ott. 2007, vol. 2.

[24] M. Piot, F. Bertrand, S. Guihard, J.-B. Clavier e M. Maumy, «Bayesian Network structure learning algorithm for highly missing and non imputable data: Application to breast cancer radiotherapy data,» *Artificial Intelligence in Medicine*, vol. 147, p. 102 743, gen. 2024. doi: 10.1016/j.artmed.2023.102743. indirizzo: https://www.sciencedirect.com/science/article/pii/S0933365723002579?via%3Dihub.

[25] A. Franzin, F. Sambo e B. Di Camillo, «bnstruct: an R package for Bayesian Network structure learning in the presence of missing data,» *Bioinformatics*, vol. 33, n. 8, pp. 1250–1252, dic. 2016. doi: 10.1093/bioinformatics/btw807. indirizzo: https://pubmed.ncbi.nlm.nih.gov/28003263/.

[26] T. Silander e P. Myllymaki, *A simple approach for finding the globally optimal Bayesian network structure*, giu. 2012. indirizzo: https://arxiv.org/abs/1206.6875.

[27] I. Tsamardinos, L. E. Brown e C. F. Aliferis, «The max-min hill-climbing Bayesian network structure learning algorithm,» *Machine Learning*, vol. 65, n. 1, pp. 31–78, mar. 2006. doi: 10.1007/s10994-006-6889-7. indirizzo: https://link.springer.com/article/10.1007/s10994-006-6889-7.

[28] N. Friedman, D. Geiger e M. Goldszmidt, «Bayesian network classifiers,» *Machine Learning*, vol. 29, n. 2/3, pp. 131–163, gen. 1997. doi: 10.1023/a:1007465528199. indirizzo: https://doi.org/10.1023/a:1007465528199.

[29] L. M. de Campos, «A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests,» *Journal of Machine Learning Research*, vol. 7, n. 77, pp. 2149–2187, 2006. indirizzo: http://jmlr.org/papers/v7/decampos06a.html.

[30] P. Dagum, A. Galper e E. J. Horvitz, *Dynamic Network Models for Forecasting*, 2013. arXiv: 1303.5396 [cs.AI].

[31] R. E. Neapolitan, «Learning Bayesian networks,» *Prentice Hall*, ago. 2007. doi: 10.1145/1327942.1327961. indirizzo: https://doi.org/10.1145/1327942.1327961.

[32] Apr. 2024. indirizzo: https://www.who.int/.

[33] M. C. Ysrraelit e J. Correale, «Impact of sex hormones on immune function and multiple sclerosis development,» *Immunology*, vol. 156, n. 1, pp. 9–22, ott. 2018. doi: 10.1111/imm.13004. indirizzo: https://onlinelibrary.wiley.com/doi/10.1111/imm.13004.

[34] M. D. Cossburn, G. Ingram, C. L. Hirst, Y. Ben-Shlomo, T. Pickersgill e N. Robertson, «Age at onset as a determinant of presenting phenotype and initial relapse recovery in multiple sclerosis,» *Multiple Sclerosis Journal*, vol. 18, n. 1, pp. 45–54, ago. 2011. doi: 10.1177/1352458511417479. indirizzo: https://journals.sagepub.com/doi/10.1177/1352458511417479.

[35] F. Barkhof, «MRI in multiple sclerosis: correlation with expanded disability status scale (EDSS),» *Multiple Sclerosis Journal*, vol. 5, n. 4, pp. 283–286, ago. 1999. doi: 10.1177/135245859900500415. indirizzo: https://journals.sagepub.com/doi/abs/10.1177/135245859900500415.

[36] E. C. Arpín, «Efficacy and safety of fampridine for walking disability in multiple sclerosis,» *Neurodegenerative Disease Management (Print)*, vol. 10, n. 5, pp. 277–287, ott. 2020. doi: 10.2217/nmt-2020-0024. indirizzo: https://www.futuremedicine.com/doi/full/10.2217/nmt-2020-0024.