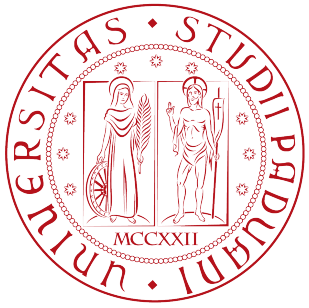
Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in Scienze Statistiche



Tesi di Laurea

Approcci di data mining in farmacovigilanza applicati ai dati di segnalazioni spontanee

Relatrice: Prof.ssa Giovanna Boccuzzo
Dipartimento di Scienze Statistiche

Correlatore: Dott. Pietro Belloni
Dipartimento di Scienze Statistiche

Laureando: Matteo Cortivo
Matricola N. 1242357

Anno Accademico 2021/2022

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

"La farmacovigilanza è la scienza e le attività designate ad identificare, valutare, capire e prevenire gli eventi avversi o qualsiasi altro problema collegato ad un medicinale o ad un vaccino" è la definizione di farmacovigilanza data dall'Organizzazione Mondiale della Sanità (OMS). Per fare questo, tutti gli enti incaricati della farmacovigilanza sfruttano, come strumento primario per reperire informazioni circa gli eventi avversi causati da un farmaco, i *dataset* di segnalazioni spontanee, che però hanno delle complicazioni intrinseche, come si vedrà nella sezione 1.3.

Questi *dataset*, come per esempio quello gestito e mantenuto dalla *Food and Drug Administration* (*FDA*), l'agenzia del farmaco degli Stati Uniti d'America, sono *database* con un elevato numero di segnalazioni spontanee effettuate da medici, farmacisti e specialisti del settore, ma anche da persone comuni. Questo crea dei problemi in fase di analisi statistica del dati e, più nello specifico, in fase di applicazione di metodi di *data mining* e metodi *data driven*. Il problema principale è la mancanza dei cosiddetti "controlli", ovvero soggetti a cui è stato somministrato il farmaco, ma che non hanno avuto nessun tipo di effetto avverso. Questa mancanza provoca dei problemi nel calcolo delle usuali statistiche utilizzate in questi casi, ovvero i metodi di disproporzionalità, per valutare se un farmaco risulti essere associato o meno con un effetto avverso, per fare in modo che vengano poi effettuati ulteriori test per valutare un'eventuale associazione causale, impossibile da stabilire utilizzando solo metodi statistici.

L'obiettivo del lavoro svolto e presentato di seguito è stato quello di identificare dei metodi di *data mining* che possano facilitare il più possibile il meccanismo di identificazione dei segnali, ovvero coppie farmaco - evento avverso che risultino essere statisticamente associate. Fissato un evento avverso, si vogliono identificare tutti i farmaci (o i principi attivi di un

farmaco) che possano portare allo sviluppo di eventi avversi nei soggetti e, in un secondo momento, passare queste informazioni a medici ed eventualmente ad aziende farmaceutiche per portare avanti altri studi per capire se questi farmaci siano realmente collegati con l'evento avverso indicato.

Da notare come in fase di sviluppo di un farmaco tutte le aziende sono tenute a svolgere dei *trials* clinici ma, come si vedrà in seguito (sezione 1.1.4), data la natura dei soggetti inclusi nei *trials*, non è scontato che alcuni eventi avversi non si verifichino o si verifichino in numero molto inferiore a quanto riportato nei *database* di farmacovigilanza, e dunque seguendo le indicazioni fornite dall'analisi di questi *database* le aziende andranno ad effettuare ulteriori test per stabilire eventuali nessi causali e in caso di esistenza di un nesso, si andrà a valutare il ritiro del farmaco dal mercato o l'imposizione di alcune restrizioni per l'uso.

La tesi si sviluppa nel seguente modo: nel capitolo 1 verranno introdotte le fasi di sviluppo e approvazione di un farmaco e verranno presentati i *database* di farmacovigilanza con le loro limitazioni. Nel capitolo 2 vengono descritti i metodi di disproporzionalità più utilizzati in farmacovigilanza al giorno d'oggi. Nel capitolo 3 verranno presentati i *database* che saranno utilizzati per l'applicazione dei metodi presentati nel capitolo 4. Prima dell'applicazione a questi dati, presentata nel capitolo 7, viene effettuato uno studio di simulazione presentato nei capitoli 5 e 6. Nel capitolo 5 è presentato uno studio di simulazione per valutare la capacità del metodo LASSO di identificare le coppie farmaco - evento avverso. Successicamente nel capitolo 6 verrà indagata la capacità dei metodi di identificare oltre alle singole coppie anche le possibili interazioni tra farmaci.

Capitolo 1

La farmacovigilanza

1.1 Le fasi di sperimentazione di un farmaco

Quando le aziende farmaceutiche intraprendono il processo di sviluppo di un farmaco, questo deve sottostare a delle procedure molto complesse, che possono durare anni, e che sono orientate a minimizzare i rischi alle persone che ne usufruiranno. Per questo motivo le agenzie del farmaco, comprese la statunitense *Food and Drug Administration* (*FDA*) e l'europea *European Medicines Agency* (*EMA*), prima di dare il via libera alla commercializzazione di un nuovo farmaco richiedono che siano eseguiti test molto scrupolosi. Gli esiti di tali test vengono presi in esame da un comitato delle agenzie per valutarne i risultati e prendere l'ultima decisione sulla possibile commercializzazione.

I farmaci che vengono sviluppati non hanno l'obbligo di essere immessi sul mercato, e molti di essi non arrivano neanche alle fasi più avanzate della sperimentazione poiché vengono riscontrati problemi in fasi preliminari dello sviluppo (fasi 1 e 2 della sperimentazione). Spesso può capitare anche che, come nel caso del *Retrovir* (*zidovudine*, anche noto come AZT), venga lanciata la sperimentazione per curare una tipologia di malattia (in questo caso il cancro), senza che il farmaco venga approvato, ed essere riproposto in seguito per un'altra tipologia di malattia (AIDS), e in questo caso venir approvato (FDA, 2021).

L'approvazione di un farmaco viene solitamente divisa in quattro distinte fasi, che verranno discusse nelle successive sezioni (viene portato l'esempio delle fasi di approvazione seguite dall'FDA): ognuna di queste fasi ha obiettivi specifici da raggiungere.

1.1.1 Fase 1 - Ricerca preclinica

La prima fase della sperimentazione di un farmaco viene definita **preclinica**. In questa fase viene scoperto un nuovo composto molecolare che può avere maggiori effetti (a volte sono inaspettati, ma non sempre) rispetto a trattamenti già esistenti, o vengono create nuove tecnologie che consentono metodi innovativi per fare in modo che un farmaco raggiunga più rapidamente o in maniera più efficiente la parte del corpo interessata.

Solo dopo che questi composti o queste tecnologie sono stati identificati inizia il vero e proprio sviluppo del farmaco. Inizialmente non vengono subito testati su esseri umani, ma su animali, anche di varie specie, in modo tale da poter raccogliere informazioni sulla sua tossicità, sicurezza ed efficacia.

Successivamente, gli sponsor del farmaco (che possono essere aziende farmaceutiche, istituti di ricerca o altre organizzazioni che si assumono la responsabilità dello sviluppo di un farmaco) redigono il cosiddetto *Investigational New Drug* (*IND*) (DPAC, 2021). L'*IND* deve includere i risultati dei test iniziali svolti su animali, la composizione e produzione del farmaco e lo sviluppo di un piano di sperimentazione di questo farmaco su soggetti umani.

La fase di ricerca preclinica può essere fatta in due modi differenti: *in vitro* o *in vivo*. La ricerca *in vitro* fa riferimento ad esperimenti effettuati in ambiente controllato al di fuori di organismi viventi, mentre nella ricerca *in vivo*, gli esperimenti sono effettuati in organismi viventi. Entrambe le tipologie di ricerca devono comunque seguire pratiche di laboratorio regolate, conosciute come *good laboratory practice* (*GLP*), che delineano le regole da seguire per ricercatori, i requisiti necessari per le strutture e le attrezzature, etc (DPAC, 2021).

1.1.2 Fase 2 - Ricerca clinica

Dopo che la fase di ricerca preclinica ha avuto parere positivo dalla commissione dell'agenzia del farmaco designata, ha inizio la cosiddetta fase **clinica**, ovvero la fase in cui il farmaco viene somministrato a soggetti umani per stabilire la dose massima per evitare tossicità, quali sono gli eventi avversi che possono verificarsi con maggior frequenza, stabilire la gravità di questi eventi avversi, e capire l'efficacia del farmaco.

In questa fase prendono luogo i *trials* clinici che devono rispondere a specifiche richieste e devono seguire un protocollo predeterminato.

Ognuno dei *trials* clinici consta poi di quattro sottofasi:

- **Step 1**: è la prima fase dei *trials*, che normalmente include tra le 20 e le 100 persone, usualmente volontari sani (FDA, 2021); durante questa fase viene enfatizzata la sicurezza. L'obiettivo è quello di determinare quali sono gli eventi avversi più frequenti, e spesso viene indagato come il farmaco viene metabolizzato ed espulso. Inoltre viene determinato il dosaggio del farmaco per garantire la sicurezza dei pazienti che lo assumeranno. Circa il 70% dei farmaci che arrivano alla prima fase passano poi a quella successiva.

- **Step 2**: vengono incluse in questa fase dello studio centinaia di persone che presentano la patologia potenzialmente curabile dal farmaco o una condizione tale per cui il farmaco può essere utile. L'obiettivo è ottenere dei dati iniziali per capire se il farmaco abbia effetti sulla patologia o la condizione dei soggetti interessati. In studi controllati, i pazienti che assumono il nuovo farmaco vengono comparati con soggetti di simili caratteristiche (età, sesso, patologie, etc) che però ricevono un altro trattamento, che può essere un placebo (ovvero una sostanza priva di principi attivi) o la terapia usuale in quelle condizioni. La sicurezza continua ad essere sotto controllo ed eventi avversi a breve termine vengono studiati. Questa fase solitamente dura da pochi mesi fino anche a due anni; solo il 30% circa dei farmaci che arrivano a questa fase passano poi a quella successiva.

- **Step 3**: in questa fase vengono coinvolte migliaia di persone volontarie, con lo scopo di raccogliere maggiori informazioni circa la sicurezza e l'efficacia del farmaco somministrato a differenti popolazioni, utilizzando dosaggi diversi, anche in combinazione con altri farmaci. Rimane comunque principale anche in questa fase il controllo degli eventi avversi, dal momento che la lunghezza di questo step permette di tenere sotto controllo eventi avversi a lungo termine. Questa fase della sperimentazione può arrivare a durare fino a quattro anni. Circa il 25-30% dei farmaci testati in questa fase arrivano all'ultima fase dei *trials* clinici.

- **Step 4**: vengono inclusi anche in questa fase migliaia di volontari (fino a 3000) che hanno la patologia o la condizione di interesse. Gli studi effettuati in questa fase vengono effettuati una volta che il farmaco è

già stato immesso sul mercato e svolgono un ruolo di monitoraggio della sicurezza e dell'efficienza. A questo punto, se un farmaco passa anche questa fase entra in gioco l'agenzia del farmaco compentente che dovrà effettuare una revisione di tutto lo studio per determinare se questo possa o meno essere mantenuto sul mercato.

Gli studi di fase 4 vengono effettuati una volta che il farmaco o il dispositivo è stato approvato dalla FDA durante il monitoraggio della sicurezza post-commercializzazione.

1.1.3 Fase 3 - Revisione

Una volta che un farmaco ha passato con successo tutte le sottofasi della ricerca clinica viene effettuata una richiesta all'agenzia del farmaco competente per far si che ci sia una revisione di tutti i dati raccolti da un *panel* di esperti composto da medici, chimici, statistici, microbiologi, farmacologi ed altri. L'obiettivo del *panel* non è solamente quello di giudicare se il farmaco sia sicuro per i soggetti che lo assumono, ma anche di valutare se questo porta ad un miglioramento delle condizioni di salute nei pazienti affetti dalla patologia o condizione per cui il farmaco è stato creato. Se il farmaco non crea pericoli nei soggetti, ma non aiuta neanche a curare gli stessi soggetti, questo farmaco non verrà approvato. Va considerato che nessun farmaco è assolutamente sicuro, poiché tutti i farmaci hanno degli eventi avversi più o meno gravi: l'obiettivo del *panel* di esperti è quindi quello di valutare se i benefici portati da questo farmaco superano i rischi dello stesso, per poter dare poi il via libera alla commercializzazione.

In alcune situazioni si ricorre all'approvazione accelerata: questa procedura viene accettata in situazioni di malattie molto serie e pericolose per la vita per cui mancano dei trattamenti efficaci (un esempio del giorno d'oggi è l'approvazione dei vaccini per curare il COVID-19).

In una normale richiesta di approvazione, definita *New Drug Application Review* (*NDA review*) dall'agenzia del farmaco statunitense, vengono richiesti i risultati dei *trials* clinici effettuati, informazioni circa l'etichetta del farmaco (viene verificato che siano fornite le appropriate informazioni per salvaguardare la salute dei pazienti ed informare adeguatamente i medici sul suo utilizzo), informazioni sulla sicurezza del farmaco e i risultati nel caso di abuso del

farmaco. Alla fine della revisione, se il gruppo di esperti dà parere positivo alla commercializzazione, il farmaco entra sul mercato.

In situazioni di estrema necessità, agenzie come la FDA hanno però sviluppato delle procedure per accelerare i processi di approvazione dei farmaci:

- ***Fast track***: questo processo è pensato per accelerare lo sviluppo e la revisione dei farmaci che sono orientati a trattare gravi condizioni o che "soddisfino un bisogno medico insodisfatto";
- ***Breakthrough therapy*** (terapia rivoluzionaria): questo processo permette di accelerare l'approvazione di farmaci che si trovano ad essere più efficaci per una certa condizione o patologia rispetto a quelli già presenti sul mercato;
- ***Accelerated approval*** (approvazione accelerata): questo processso è per quei farmaci che soddisfano un bisogno medico insoddisfatto e hanno prove di potenziali benefici clinici anche se non sono ancora dimostrati dai dati;
- ***Priority review*** (revisione prioritaria): questo processo significa che l'agenzia del farmaco ha l'obiettivo di prendere una decisione sul farmaco entro sei mesi.

1.1.4 Fase 4 - *Post-marketing*

poiché non è mai possibile riuscire ad identificare tutti gli eventi avversi che possono insorgere con la somministrazione di un farmaco, e in alcuni casi ne possono insorgere di molto seri, la fase di *post-marketinging* gioca un ruolo molto importante per la sicurezza dei farmaci. Nel caso di reazioni molto serie è anche possibile rivalutare il farmaco per eventualmente ritirarne la vendita. Questa fase viene gestita non solo dalle agenzie del farmaco, ma anche dagli stessi sponsor del farmaco che devono fornire regolarmente degli aggiornamenti sulla sicurezza del farmaco.

L'obiettivo della fase quattro di sviluppo di un farmaco non mira solamente a reperire informazioni circa i nuovi effetti avversi (più o meno gravi) che si possono verificare quando il farmaco viene assunto da centinaia di migliaia o milioni di persone, ma è anche quello di valutarne l'efficacia nel mondo reale, che risulta differente da quella ottenuta durante le fasi precedenti poiché in

quel caso i soggetti erano selezionati tramite attenti criteri di inclusione ed esclusione.

Per questa fase sono stati sviluppati degli appositi *database* di segnalazioni spontanee. Il loro obiettivo è quello di raccogliere i cosiddetti dati di *real world*, poiché, come detto in precedenza, non è possibile conoscere tutti gli eventi avversi che si possono sviluppare con la somministrazione di un farmaco. Questi *database* vengono analizzati dalle agenzie del farmaco per valutare se esista un collegamento tra l'evento avverso sviluppato e segnalato (definiti **SUSARs**, *serious and unexpected suspected adverse reactions*, sospette reazioni avverse serie e non note) e il farmaco assunto (Suvarna, 2010).

1.2 La farmacovigilanza

"Medicinali e vaccini hanno trasformato la prevenzione e il trattamento delle malattie. Purtroppo però i prodotti medicinali possono far si che si sviluppino eventi avversi che spesso sono inaspettati e/o non desiderati. La farmacovigilanza è la scienza e le attività designate ad identificare, valutare, capire e prevenire gli eventi avversi o qualsiasi altro problema collegato al medicinale o al vaccino" (WHO, 2021).

Come visto in precedenza, tutti i farmaci sono sottoposti a rigorosi test e revisioni prima del loro inserimento sul mercato. Ma poiché i *trials* effettuati somministrano il farmaco ad un numero relativamente ristretto di persone, molti eventi avversi possono rimanere estranei. Infatti, alcuni effetti avversi possono insorgere solo in determinate persone con caratteristiche diverse da quelle selezionate nei *trails*, o che magari hanno comorbidità, per cui magari assumono anche altri tipi di farmaco che potrebbero creare in interazione eventi avversi sconosciuti e in alcuni casi anche seri.

Le stesse aziende farmaceutiche hanno obbligo di monitorare e raccogliere continuamente dati sui loro farmaci in commercio e condurre studi di farmacovigilanza. Le aziende non sono però le uniche che sono adibite al controllo degli eventi avversi: infatti anche le agenzie del farmaco gestiscono sistemi per raccogliere dati su di essi. L'FDA gestisce e mantiene il *database FDA Adverse Event Reporting System* (**FAERS**) nel quale medici, professionisti sanitari, case farmaceutiche e anche singoli soggetti che assumono un far-

maco possono segnalare l'evento avverso. L'EMA gestisce **EudraVigilance**, l'equivalente del FAERS.

1.3 I *database* di farmacovigilanza

I sistemi di segnalazione spontanea sono diventati il punto di partenza per il controllo dei farmaci dopo la loro immissione sul mercato, come anche testimoniato dal grafico in figura 1.1, che evidenzia un costante aumento delle segnalazioni negli ultimi anni. Nonostante la loro grande utilità nell'identificazione di relazioni farmaco - evento avverso, questi *dataset* presentano anche delle grosse criticità che verranno spiegate nelle prossime sezioni (Stephenson e Hauben, 2007).

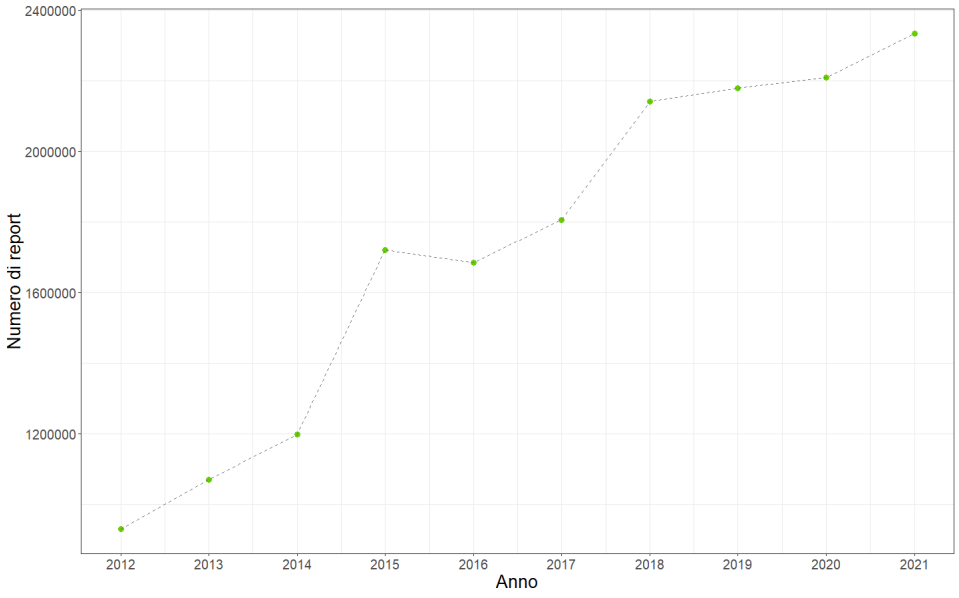


Figura 1.1: Evoluzione del numero di segnalazioni annuali di eventi avversi nel *database* FAERS nel periodo 2013 - 2021 (secondo trimestre).

1.3.1 Spontaneità delle segnalazioni

I sistemi di segnalazione spontanea sono afflitti dalla sottostima dei sospetti eventi avversi così come dalla sovrastima degli eventi che non sono sospetti di essere eventi avversi. Capita infatti che i *database* includano eventi avversi che sono stati oggetto di specifici studi da parte di alcuni medici, risposte sollecitate a pazienti a cui è stato sottoposto un questionario ed

eventi avversi che derivano da specifici *database* di malattie o specifici farmaci (Stephenson e Hauben, 2007).
Per questo motivo la completezza, la qualità e l'accuratezza di queste informazioni possono variare molto in base alla persona o all'azienda che ha effettuato la segnalazione e possono anche dipendere dall'evento avverso segnalato o da quanto recente è il farmaco.

Non avendo numeratori e denominatori certi, i *database* con questa struttura rappresentano dei campioni di convenienza senza però avere una chiara struttura di probabilità e per questo motivo non si può fare affidamento solamente sui risultati ottenuti da questi *database*, neanche per il calcolo di tassi di incidenza, in quanto non abbiamo a disposizione il numero di coloro che assumono il farmaco.

Infatti, poiché ci sono fattori che influenzano i tassi e poiché è difficile da ottenere il numero totale di pazienti trattati (la popolazione di riferimento) e per quanto tempo questi hanno ricevuto il trattamento, risulta poco adatta la comparazione dei tassi dei vari farmaci. Questo è uno dei problemi più importanti dei *database* di *Spontaneous Reporting System* (**SRS**), ovvero la mancanza di un denominatore, che ha portato allo sviluppo dei metodi di disproporzionalità (capitolo 2). Questo problema è legato al fatto che in *database* di questo tipo non vengono riportati i casi di assunzione di un farmaco che non hanno avuto eventi avversi, e quindi manca un'indicazione di quanti soggetti abbiano effettivamente assunto il farmaco (mancano i controlli all'interno del *dataset*) ed inoltre anche per coloro che hanno sviluppato eventi avversi non sono sempre presenti tutte le segnalazioni.

Inoltre, un ulteriore problema è da ricercarsi nel cosiddetto *selection bias*: infatti le differenti frequenze di rappresentazione dei farmaci potrebbero essere dovute anche ad altre motivazioni, come per esempio la prescrizione selettiva dei farmaci o al fatto che i medici potrebbero essere propensi a segnalare solamente determinati eventi avversi perché considerati più importanti rispetto ad altri. Questo crea quindi una distorsione che può inficiare i risultati finali di qualsiasi metodo utilizzato per l'identificazione di segnali.

Nonostante queste limitazioni, gli SRS, sistemi di segnalazione spontanea, rimangono un importante strumento per monitorare la sicurezza dei farmaci sul mercato e rimangono, al momento, l'unico mezzo per individuare eventi avversi rari e seri nei primi anni dopo la commercializzazione dei farmaci, perché, come visto nella sezione 1.1, durante la sperimentazione di un farmaco

non è scontato che alcuni eventi avversi vengano rilevati tra i soggetti: questo perché il campione di persone che partecipano alla fase di sperimentazione di un farmaco è selezionato con strette regole di inclusione ed esclusione, che nel mondo reale non sono presenti, e questo potrebbe far sì che sottogruppi di persone, con caratteristiche simili, sviluppino delle reazioni avverse sconosciute e molto serie.

Gli **SRS** risultano anche molto utili nell'identificazione non solo degli eventi avversi rari e inaspettati, ma anche di eventi avversi più comuni che nei *trials* clinici si sono però manifestati in popolazioni con determinate caratteristiche e non in altre, in quanto sottorappresentate a causa dei criteri di inclusione/esclusione. Infatti, una volta che il farmaco viene commercializzato, questo viene venduto a chiunque ne abbia bisogno, ad esclusione di pochi casi con determinate patologie o situazioni cliniche (ad esempio donne incinte), e dunque ci sono molte più possibilità che gruppi di soggetti con caratteristiche non incluse nei *trials* vengano a contatto con questo farmaco, andando ad aumentare di molto il numero di reazioni avverse più comuni che si erano viste durante i test.

1.3.2 *Weber effect*

É ben noto che il tempo sul mercato di un farmaco ha un effetto sulla segnalazione degli eventi avversi associati a quel particolare farmaco. Questo fenomeno inizialmente descritto da Weber, e per questo noto come *Weber effect*, fu notato in uno studio sui farmaci antiinfiammatori non steroidei (noti anche come FANS) commercializzati nel Regno Unito durante gli anni 70 e i primi anni 80 (Weber, 1984). Più recentemente questo fenomeno è stato ripreso e replicato su dati del *database* FAERS (Hartnell e Wilson, 2004).

L'effetto Weber è definito come un effetto epidemiologico, il quale definisce che il numero di segnalazioni di eventi avversi per un farmaco cresce fino approssimativamente a metà/fine del secondo anno di commercializzazione, periodo in cui raggiunge il picco di segnalazioni, e poi inizia una descrescita nonostante le prescrizioni di quel farmaco aumentino (Weber, 1984).

Di seguito vengono mostrati dei grafici che riportano l'andamento del numero di segnalazioni durante gli anni di alcuni farmaci la cui data di entrata sul mercato è il 2013. I dati sulle segnalazioni sono stati reperiti dal *database* FAERS, sfruttando il pacchetto R `faers.db` (Lanera, Belloni

e Guidone, 2021). I farmaci considerati con alcune specifiche di essi sono riportati nella tabella 1.1

Marchio (nome generico)	Casa produttrice	Data di approvazione	Indicazione	Numero totale di report
Adempas (riociguat)	Bayer HealthCare	8 ottobre 2013	Ipertensione polmonare tromboembolica cronica	288229
Gilotrif (afatinib)	Boehringer Ingelheim	12 luglio 2013	Cancro del polmone non a piccole cellule	52286
Opsumit (macitentan)	Actelion	18 ottobre 2013	Ipertensione arteriosa polmonare	17962
Xofigo (radium Ra 223 dichloride)	Bayer HealthCare	15 maggio 2013	Cancro alla prostata resistente alla castrazione	24479

Tabella 1.1: Informazioni generali sui farmaci considerati per valutare il Weber effect

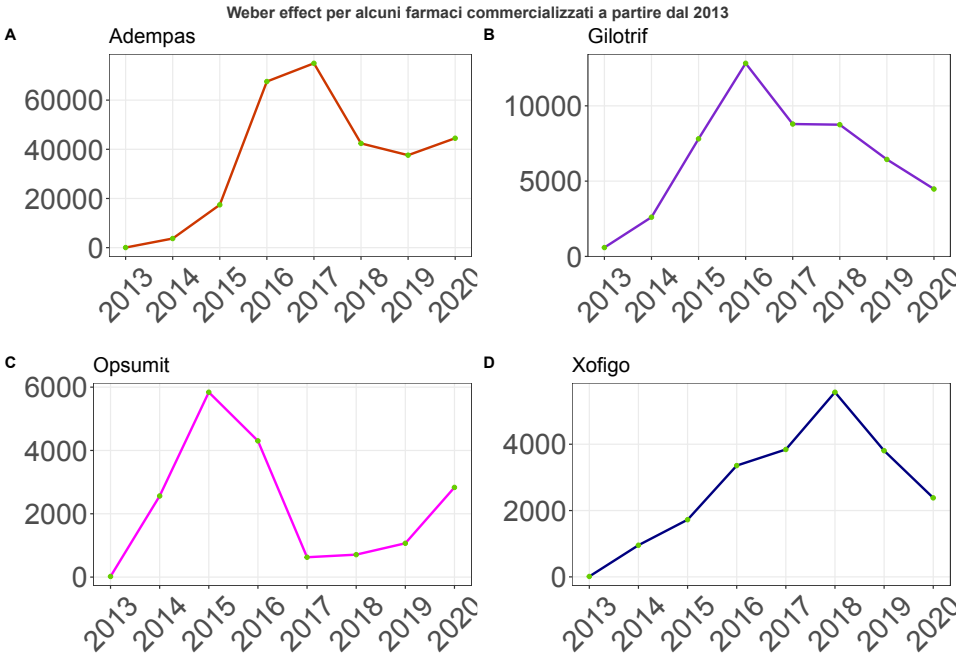


Figura 1.2: Numero di eventi avverse segnalati nel FAERS per i farmaci riportati in tabella 1.1 dall'anno di commercializzazione (2013) al 2020 (ultimo anno con dati completi del FAERS)

Dai grafici in figura 1.2 possiamo notare come non sia sempre evidente il *Weber effect*. Partendo dal grafico A in figura 1.2, riferito al farmaco *Adempas*, possiamo notare come un effettivo calo nelle segnalazioni degli eventi avversi ci sia stato, ma un po' più diluito nel tempo anche considerando la data di commercializzazione del farmaco (inizio ottobre 2013): possiamo notare come

un calo ci sia stato dopo il 2017, quindi dopo tre anni circa dall'entrata sul mercato del farmaco, inoltre possiamo vedere che nell'ultimo periodo il trend è in crescita.

Il secondo farmaco, *Gilotrif*, rappresentato nel grafico B in figura 1.2 ha un andamento molto simile a quello del farmaco discusso in precedenza, se non che il suo trend continua ad essere decrescente; il picco delle segnalazioni viene raggiunto al quarto anno di commercializzazione del farmaco.
Nel grafico C, possiamo trovare l'andamento delle segnalazioni di eventi avversi per il farmaco *Opsumit*, che mostra in questo caso il *Weber effect* così com'è definito: infatti notiamo che il picco viene raggiunto nel 2015 quindi dopo due anni dalla commercializzazione del farmaco. Come per il primo farmaco analizzato possiamo notare un trend crescente. L'ultimo farmaco analizzato, mostrato nel grafico D, evidenzia un importante picco che si verifica 5 anni dopo la commercializzazione, e un trend calante negli ultimi anni.

Non avendo informazioni circa le campagne pubblcitarie fatte attorno a questi farmaci è difficile trarre delle conclusioni, però sembra evidente che alcuni farmaci abbiano risentito maggiormente di questo effetto con andamenti molto diversi a livello di segnalazione di eventi avversi.

1.3.3 *Publicity bias*

Strettamente legato al *Weber effect* è un altro tipo di distorsione che può essere comune in *database* di questo tipo, ovvero il *publicity bias*. Questa distorsione è dovuta a vari eventi occorsi negli anni, da quando il farmaco viene approvato, che potrebbero influenzare le segnalazioni di eventi avversi. Nel seguito si riporta un esempio citato da Hartnell e Wilson, (2004), che illustra le complesse dinamiche che si intrecciano fra segnalazioni e comunicazione pubblica. L'FDA, dopo un aumento di segnalazioni di problemi valvolari causati da due farmaci prescritti per la perdita di peso, ha annunciato il ritiro di questi dal mercato nel settembre del 1997. É molto probabile che questa situazione abbia sensibilizzato i medici così come il pubblico in generale sugli effetti avversi.

Inoltre, in concomitanza con questo fatto, l'FDA rilasciò delle linee guida per televisioni e radio, chiedendo che per le pubblicità riguardanti farmaci non venissero più solamente scritti (per le televisioni) gli eventi avversi, ma anche riferiti a voce. Facendo questo tutte le persone che ascoltavano furono

rese consapevoli degli eventi avversi collegati ad un determinato farmaco: il risultato fu un aumento delle segnalazioni di reazioni avverse.

Esistono altri fattori che possono influenzare le segnalazioni associate agli eventi avversi presentate spontaneamente. In effetti, Wallenstein e Fife (2001) hanno ipotizzato che l'aumento iniziale delle segnalazioni, caratteristico del *Weber effect*, sia dovuto all'aumento dell'esposizione del paziente, e il calo dei rapporti sia dovuto alla familiarità del medico con il farmaco e alla perdita di interesse nella segnalazione di eventi avversi associati allo stesso. Weber ha ipotizzato che, poiché i medici sono esposti a molti nuovi farmaci ogni anno, 2 anni potrebbe essere il tempo minimo in cui i medici possono mantenere l'interesse per un particolare farmaco. É anche possibile che eventi altamente pubblicizzati e correlati ad eventi avversi, insieme ad eventi come l'introduzione di nuovi farmaci generici o l'approvazione di nuove indicazioni per i farmaci, siano responsabili dei picchi osservati nelle segnalazioni dei farmaci studiati. poiché i farmaci generici solitamente hanno un costo più basso, questo fa si che i farmaci principali subiscano un calo nelle prescrizioni e dunque nelle vendite facendo si che le segnalazioni di reazioni avverse per quei determinati farmaci calino drasticamente.

Un esempio di *publicity bias*

I problemi associati ai vari FANS che erano disponibili contemporaneamente a quelli analizzati da Weber potrebbero aver influenzato le segnalazioni. Alcuni di questi FANS, nel marzo 1983, furono ritirati dal mercato a seguito di decessi che furono segnalati come reazioni avverse agli stessi farmaci. Questa sospensione seguì ad un telegiornale che citava come le morti fossero associate al farmaco.

Questi eventi associati ai FANS furono molto pubblicizzati. I problemi riscontrati con questi farmaci mentre erano sul mercato e la pubblicità associata al loro ritiro, hanno probabilmente fatto sì che la consapevolezza degli eventi avversi associati ai FANS commercializzati in quel periodo aumentatasse e questo potrebbe aver influenzato le segnalazioni per molti farmaci di questa classe.

Nel periodo tra fine 1998 e inizio 1999, furono introdotti sul mercato farmaceutico dei farmaci considerati superiori ai FANS. L'ingresso di questi nuovi farmaci sul mercato ha fatto si che gli eventi avversi gastrointestinali legati ai tradizionali FANS venissero all'attenzione degli operatori sanitari

in maniera diversa da quanto accadeva prima: perciò questi nuovi farmaci hanno fatto sì che le segnalazioni per alcuni FANS già sul mercato venissero influenzate.

Capitolo 2

I metodi di disproporzionalità

Attualmente la procedura che genera i segnali di associazione fra farmaci ed eventi avversi segue i seguenti passi: su base trimestrale le liste di potenziali coppie farmaco - evento avverso sono generate, sfruttando le segnalazioni contenute nei vari *database* gestiti dai vari centri di farmacovigilanza, attraverso tecniche come il *Proportional Reporting Ratio* (**PRR**) e il *Reporting Odds Ratio* (**ROR**). A questo punto un panel di esperti riceve i segnali generati e li valuta sulla base delle loro conoscenze. Da queste valutazioni viene redatta una lista finale di segnali che viene inviata ai centri di farmacovigilanza nazionali, e dunque spetta a quest'ultimi prendere le decisioni finali. Ci sono però delle ovvie limitazioni a questo metodo: gli esperti sono in grado di considerare un contenuto numero di segnalazioni, inoltre la loro conclusione viene tratta dalle loro conoscenze a priori, che creano una distorsione verso l'identificazione di segnali che sono già sospetti o sono stati evidenziati per altre ragioni (Meyboom et al., 1997).

Negli anni sono stati sviluppati una serie di strumenti noti come **analisi di disproporzionalità** o metodi di disproporzionalità (di cui fanno parti anche i due citati precedentemente), atti a scremare le associazioni più significative da sottoporre al parere degli esperti. Possiamo distinguere due famiglie di questi metodi: metodi frequentisti e metodi bayesiani.

2.1 I metodi frequentisti

I due metodi frequentisti più utilizzati e citati in letteratura, ovvero il **PRR** (Evans, Waller e Davis, 2001) e il **ROR** (Rothman, Lanes e Sacks,

2004), sfruttano delle tabelle di contingenza costruite in maniera differente rispetto alle usuali; un esempio, in cui consideriamo solamente tre eventi avversi, è riportato in Tabella 2.1:

Evento avverso	Farmaco di interesse	Tutti gli altri farmaci nel database	Totale
Evento A	a	b	a + b
Evento B	c	d	c + d
Evento C	e	f	e + f
Eventi totali	a + c + e	b + d + f	n

Tabella 2.1: Tabella di esempio per il calcolo del PRR e del ROR

Partendo da questa tabella è possibile ottenere i due valori che vengono calcolati come:

$$PRR = \frac{a}{b} / \frac{a+c+e}{b+d+f} = \frac{\text{tasso eventi avversi A}}{\text{tasso eventi avversi generali}} \tag{2.1}$$

$$ROR = \frac{a/b}{c/d} \tag{2.2}$$

Si possono notare subito delle differenze nel calcolo delle due statistiche: infatti il PRR confronta l'occorrenza di uno specifico evento avverso rispetto a tutti gli eventi avversi dovuti al farmaco, con la stessa occorrenza calcolata con riferimento a tutti gli altri farmaci. Il ROR è invece l'*odds ratio* (**OR**) che confronta un farmaco (quello di interesse) con tutti gli altri farmaci presenti nel *dataset* rispetto a due eventi avversi selezionati. Guardando la Tabella 2.1 e le formule per il calcolo delle due statistiche possiamo notare come il PRR, in questo caso generico poiché non sono stati specificati eventi avversi specifici e nemmeno farmaci, è calcolato per valutare se un farmaco di interesse è collegato all'evento avverso di interesse A, mentre il ROR valuta se il farmaco può avere un effetto maggiore sull'evento A rispetto all'evento C, confrontando il tutto con tutti gli altri farmaci segnalati come causa dell'evento avverso A o B.

Per entrambe le quantità è possibile ottenere un intervallo di confidenza, che permette di individuare quali segnali sono significativi: un segnale viene considerato significativo quando l'estremo inferiore dell'intervallo è superiore ad 1; quando una delle quantità è circa pari a 1, con limite inferiore del-

l'intervallo minore di 1, questo ci indica che i farmaci sono sostanzialmente uguali in termini di eventi avversi. Il PRR e il ROR hanno delle distribuzioni asimmetriche, poiché hanno un limite inferiore pari a zero, ma non hanno un limite superiore. Tuttavia, il logaritmo di PRR e ROR hanno distribuzione approssimativamente Normale quando a, b, c e d sono sufficientemente grandi. Perciò, l'intervallo per il PRR può essere calcolato come

$$e^{\log(PRR)\pm z_\alpha \cdot SE} = \left(\frac{PRR}{e^{z_\alpha \cdot SE}}, \quad PRR \cdot e^{z_\alpha \cdot SE} \right) \tag{2.3}$$

dove z_α è il quantile di livello α della distribuzione Normale standard e SE viene definito come $SE = \sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{a+b} - \frac{1}{c+d}}$ (M. H. Pham, 2018).
L'intervallo di confidenza del ROR è invece definito come

$$e^{\log(ROR)\pm z_\alpha \cdot SE} = \left(\frac{ROR}{e^{z_\alpha \cdot SE}}, \quad ROR \cdot e^{z_\alpha \cdot SE} \right) \tag{2.4}$$

dove z_α è il quantile di livello α della distribuzione Normale standard e SE viene definito come $SE = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$ (M. H. Pham, 2018). Il calcolo di questi intervalli è subordinato all'assunzione di normalità.

2.2 I metodi bayesiani

Tra i metodi per l'analisi di disproporzionalità sono presenti anche due metodi bayesiani largamente studiati in letteratura: questi due metodi sono il *Multi-item Gamma Poisson Shrinkage* (**MGPS**) (DuMouchel, 1999, DuMouchel e Pregibon, 2001, Fram, Almenoff e DuMouchel, 2003) e il *Bayesian Confidence Propagation Neural Network* (**BCPNN**) (Bate et al., 1998, Norén et al., 2006).

2.2.1 MGPS - *Multi-item Gamma Poisson Shrinkage*

In *dataset* come il FAERS, il nostro interesse è incentrato sulla ricerca di coppie farmaco - evento avverso che abbiano una frequenza abbastanza grande da attirare la nostra attenzione. In situazioni come queste, metodi che si basano su metodi empirici bayesiani risultano essere molto raccomandati.

Un metodo largamente utilizzato è stato proposto da DuMouchel ed è chiamato **MGPS**. Questo metodo utilizza due misure derivate internamente definite frequenza di base (*baseline frequency*) E_{ij} e N_{ij} (frequenze osservate),

ovvero il numero di report in cui sono riportati insieme sia il farmaco i che l'evento avverso j. Usando questa frequenza di base come denominatore, possiamo individuare e definire come statistica di interesse il *relative report rate* $RR_{ij} = N_{ij}/E_{ij}$, che indica se la combinazione farmaco i ed evento avverso j è stata riportata più frequentemente rispetto a quanto ci si aspetterebbe se farmaco ed evento avverso non fossero associati.

La quantità RR_{ij} è però inaffidabile per due differenti cause:

- a causa della varianza di campionamento quando la *baseline* e le frequenze osservate sono piccole. Ad esempio nel caso in cui $N = 1$ e $E = 0.001$, $RR = 1000$, ma l'interpretazione è molto diversa rispetto al caso $N = 100$ e $E = 0.1$, nonostante il RR abbia lo stesso valore;

- a causa del *reporting bias*, una distorsione che si verifica quando la diffusione di risultati scientifici viene influenzata dalla natura e dalla direzione dei risultati stessi.

Nello specifico il *reporting bias* può svilupparsi in diversi modi: infatti risultati statisticamente significativi e "positivi", che indicano un miglior funzionamento di un trattamento o di una tecnica, è più facile che vengano pubblicati. Questa cosa fa si che si formi una distorsione nei risultati che si trovano negli articoli poiché ci si focalizza sulla pubblicazione e non sul risultato scientifico. Va ricordato che anche risultati non significativi possono portare ad un contributo e non solamente quelli significativi. Questo concetto è legato al problema della varianza di campionamento, in quanto se il RR non viene riportato insieme ai valori delle frequenze osservate e quello della *baseline*, può essere male interpretato e utilizzato per trarre conclusioni errate.

La metodologia empirica bayesiana sviluppata da DuMouchel va a migliorare l'effetto della varianza di campionamento sull'interpretazione del RR, ma non è in grado di annullare o almeno ridurre l'effetto portato dal *reporting bias*.

Il metodo **MGPS** per valutare i conteggi N_{ij} consiste di due step:

1. definizione della frequenza di base E_{ij};

2. definizione di una misura comparativa tra N_{ij} e E_{ij}.

Definizione della frequenza di base

Definiamo alcune quantità che torneranno poi utili anche in seguito:

$$N_{i\cdot} = \sum_j N_{ij} \tag{2.5a}$$

$$N_{\cdot j} = \sum_i N_{ij} \tag{2.5b}$$

$$N_{\cdot\cdot} = \sum_i \sum_j N_{ij} \tag{2.5c}$$

$$E_{ij} = N_{i\cdot} N_{\cdot j} / N_{\cdot\cdot}. \tag{2.5d}$$

La quantità E_{ij}, come detto in precedenza, indica la frequenza di base e possiamo anche interpretarla come l'ipotesi nulla per la frequenza di una determinata cella ($D = i$, $EA = j$), mentre la quantità N_{ij} indica il numero di report dell'evento avverso j provocati dal farmaco i. Questa è quindi la frequenza attesa quando le variabili D e EA sono tra di loro indipendenti. L'obiettivo, però, non è quello di effettuare un test per l'indipendenza, poiché viene assunto che D e EA siano associate, ma sarà quello di identificare una misura per comparare le M frequenze delle celle (con $M = I \cdot J$, dove I è il numero di farmaci, mentre J è il numero di eventi avversi): la quantità N_{ij} diventa interessante solo quando risulta essere grande se confrontata con E_{ij}.

Definizione della misura comparativa

Una volta definita la formula per la frequenza di base come visto nella sottosezione precedente, il secondo step consiste nell'identificare una funzione di N ed E e successivamente classificare tutte le celle in accordo con questa funzione. DuMouchel ha individuato, sperimentato e confrontato tre metodi differenti, ovvero il *relative report rate*, la significatività statistica e l'*empirical Bayes*:

- il *relative report rate* è il criterio più semplice dei tre proposti da DuMouchel ed è basato sul rapporto $RR_{ij} = N_{ij}/E_{ij}$. Il grande vantaggio di questa misura è la sua facile interpretabilità, ma per contro abbiamo una elevata varianza di campionamento quando la frequenza di base e la frequenza osservata sono molto basse;

- il problema introdotto dalla varianza di campionamento nel *relative report rate* può essere oltrepassato utilizzando un test statistico per valutare l'ipotesi nulla che $E[N_{ij}] = E_{ij}$. Definiamo quindi la misura **LogP** come $\log P_{ij} = -\log_{10}(Pr[X \geq N_{ij}])$, dove $X \sim Poisson(E_{ij})$. Questa non è comunque l'unica misura che possiamo ottenere poiché considerando altri test statistici possiamo arrivare ad altre misure che potrebbero essere di interesse (un esempio è quella che possiamo ottenere da un χ^2). Il concetto che sta dietro a queste misure non è che l'ipotesi nulla venga presa a riferimento, ma solamente il test statistico o il grado di significatività possano essere una misura utile per classificare il grado di associazione di celle differenti;

- l'ultimo approccio sviluppato da DuMouchel è quello dell'*empirical Bayes* che ha cercato di raccogliere i migliori aspetti dei due metodi precedentemente presentati, in particolare l'interpretabilità del *relative report rate*, ma cerca anche di aggiustare per la varianza di campionamento.

Empirical Bayes

Assumiamo che la frequenza N_{ij} sia distribuita come una Poisson con media sconosciuta μ_{ij} e concentriamo il nostro interesse sui rapporti $\lambda_{ij} = \mu_{ij}/E_{ij}$. DuMouchel ha proposto di trattare ognuno dei λ invece che come costanti non legate tra loro, come delle estrazioni da una comune distribuzione a priori. Questa distribuzione a priori viene assunta essere una mistura tra due Gamma. La distribuzione Gamma ha media α/β e varianza α/β^2, con funzione di densità definita come

$$g(\lambda; \alpha, \beta) = \beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}/\Gamma(\alpha). \tag{2.6}$$

La densità a priori che assumiamo per λ è definita come

$$\pi(\lambda; \alpha_1, \beta_1, \alpha_2, \beta_2, P) = Pg(\lambda; \alpha_1, \beta_1) + (1-P)g(\lambda; \alpha_2, \beta_2). \tag{2.7}$$

Di conseguenza la media a priori di λ assumendo il modello mistura è $P\alpha_1/\beta_1 + (1-P)\alpha_2/\beta_2$ e la varianza a priori è $P(1-P)(\alpha_1/\beta_1 - \alpha_2/\beta_2)^2 + P\alpha_1/\beta_1^2 + (1-P)\alpha_2/\beta_2^2$. In realtà la scelta della distribuzione per λ non è fondamentale,

volendo si può anche utilizzare una distribuzione stimata utilizzando i dati osservati.

La famiglia delle distribuzioni Gamma è spesso utilizzata per modellare popolazioni derivanti da distribuzioni di Poisson in virtù della relazione di coniugazione presente tra le distribuzioni Poisson e Gamma (Johnson e Kotz, 1969, Bryck e Raudenbush, 1992, O'Hagan, 1994).
I calcoli vengono semplificati sotto due aspetti:

- le distribuzioni marginali di ogni N sono una mistura di distribuzioni binomiali negative;
- la distribuzione a posteriori di ogni λ è una mistura di due distribuzioni Gamma con parametri modificati.

Assumendo quindi θ ed E come noti, otteniamo che la distribuzione di N è

$$Pr(N = n))Pf(n; \alpha_1, \beta_1, E) + (1 - P)f(n; \alpha_2, \beta_2, E), \tag{2.8}$$

$$f(n; \alpha, \beta, E) = (1 + \beta/E)^{-n}(1 + E/\beta)^{-\alpha}(\alpha + n)/\Gamma(\alpha)n!. \tag{2.9}$$

Definiamo Q_n come la probabilità a posteriori che λ derivi dalla prima componente della mistura, dato $N = n$. Dalla legge di Bayes, otteniamo la formula di Q_n

$$Q_n = Pf(n; \alpha_1, \beta_1, E)/[Pf(n; \alpha_1, \beta_1, E) + (1 - P)f(n; \alpha_2, \beta_2, E)]. \tag{2.10}$$

La distribuzione a posteriori di λ, dopo aver osservato $N = n$, può essere rappresentata come

$$\lambda|N = n \sim \pi(\lambda; \alpha_1 + n, \beta_1 + E, \alpha_2 + n, \beta_2 + E, Q_n) \tag{2.11}$$

dove $\pi()$ è dato dalla formula 2.7. Usando le ben note proprietà delle distribuzioni Gamma, otteniamo il valore atteso a posteriori di λ e $\log(\lambda)$ come

$$E[\lambda|N = n] = Q_n(\alpha_1 + n)/(\beta_1 + E) + (1 - Q_n)(\alpha_2 + n)/(\beta_2 + E) \tag{2.12}$$

e

$$E[\log(\lambda)|N = n] = Q_n[\Psi(\alpha_1+n)-\log(\beta_1+E)]+(1-Q_n)[\Psi(\alpha_2+n)-\log(\beta_2+E)] \tag{2.13}$$

dove $\Psi(x)$ è la funzione digamma, la derivata di $\log(\Gamma(x))$. La misura *empirical Bayes* usata per classificare le frequenze delle celle da DuMouchel è chiamata $EB\log 2$ e viene definita come

$$EB\log 2_{ij} = E[\log_2(\lambda_{ij})|N_{ij}] = E[\log(\lambda)|N = N_{ij}]/\log(2) \tag{2.14}$$

dove il secondo valore atteso dell'equazione sopra espressa viene definito dall'equazione 2.13. La quantità $EB\log 2$ non è altro che una versione bayesiana dell'informazione statistica $\log_2(RR)$.

Quando E o N/E non sono grandi, l'effetto di usare $EB\log 2$ è quello di restringimento di $\log_2(RR)$ verso valori più bassi, che è esattamente l'effetto desiderato quando la varianza di campionamento rende incerto il vero grado di associazione tra $A = i$ e $B = j$.
Per ottenere una quantità sulla stessa scala del RR, basta semplicemente fare l'esponenziale della quantità $EB\log 2$ per ottenere

$$EBGM_{ij} = 2^{EB\log 2_{ij}} \tag{2.15}$$

che è la media geometrica della distribuzione a posteriori empirica bayesiana del "vero" RR.

2.2.2 BCPNN - *Bayesian Confidence Propagation Neural Network*

Il *data mining* è stato largamente utilizzato in ambiti di farmacovigilanza per il controllo delle reazioni avverse provocate dai farmaci immessi sul mercato. Nello specifico questi metodi vengono utilizzati per l'identificazione di segnali, che in accordo con l'Organizzazione Mondiale della Sanità (**OMS**) vengono definiti come "informazioni riportate su una possibile relazione causale tra un evento avverso e un farmaco, relazione sconosciuta o precedentemente documentata in modo incompleto. Solitamente per generare un segnale è necessaria più di una segnalazione, a seconda della gravità dell'evento e dalla qualità dell'informazione" (Delamothe, 1992).

L'avanzamento dell'informatica in combinazione con una ben consolidata teoria bayesiana ha permesso lo svilupparsi di metodi di *data mining* basati su reti neurali bayesiane come quelle introdotte da Bate et al., 1998. Questo metodo ci aiuta a minimizzare le limitazioni di quelli che erano i sistemi

correnti di individuazione dei segnali poiché le combinazioni farmaco - evento avverso venivano considerate come non distorte. Questi metodi andranno quindi a facilitare i panel di esperti che non dovranno più valutare migliaia di combinazioni, ma saranno chiamati a dare un loro giudizio su un numero molto più ridotto di combinazioni che verrà evidenziato da metodi come il BCPNN.

Nello specifico, questo metodo si basa su reti neurali, un tipo di modello inzialmente pensato per cercare di replicare la biologia dei neuroni e realizzare quindi una ideale rete basata su questi elementi. In termini computazionali questi neuroni sono semplici, ma quando vengono usati in combinazione tra di loro permettono di eseguire compiti anche molto complessi. I neuroni di queste reti sono idealmente legati tra loro da degli archi e ad ognuno dei nodi viene attribuito un peso in fase di stima della rete. La rete specifica che viene usata in ambito di farmacovigilanza viene definita *Bayesian Confidence Propagation Neural Network* (BCPNN): è una rete cosiddetta *feed-forward* (ovvero un tipo di rete neurale dove le connesioni tra unità non formano cicli), dove la stima è effettuata utilizzando i principi della legge di Bayes (Bate et al., 1998).

In questo ambito viene utilizzata una rete ad uno strato, ma può essere estesa anche ad una rete multistrato per esempio nel caso in cui si voglia anche andare ad investigare delle possibili reazioni avverse derivanti da una combinazione di farmaci. In letteratura però, questa alternativa non viene mai implementata e dunque non è possibile utilizzarla a livello pratico per l'identificazione di relazioni tra più farmaci rispetto ad uno stesso evento avverso.

Uno dei maggiori vantaggi di questo metodo è da ricercarsi nella facile interpretabilità delle quantità che questo metodo fornisce come i pesi, che vengono interpretati come quantità probabilistiche e che vengono utilizzate per quantificare la dipendenza farmaco - evento avverso.

Nel *FAERS* le segnalazioni sono riportate in modo tale che ogni riga del *database* contenga la reazione avversa sviluppata dal soggetto e il farmaco che ha provocato tale reazione o che sembra abbia provocato la reazione, insieme ad altre informazioni sul soggetto. Dunque è possibile arrivare ad ottenere una quantità c_{ij} che indica quante volte una combinazione farmaco i - evento avverso j si è verificata all'interno del *database*. Questo valore c_{ij} non deve trarre in inganno perché non rappresenta una misura di forza dell'associazione

tra farmaco ed evento avverso, poiché è solamente un numero assoluto. Quello che si vuole cercare è un valore di c_{ij}, ma che sia più elevato di un valore atteso considerando c_i, numero di report in cui è presente il farmaco i, e c_j, numero di report in cui è presente l'evento avverso j.

Possiamo quindi introdurre in queste relazioni dei concetti di statistica bayesiana: in particolare possiamo considerare come **probabilità a priori** la probabilità che un certo evento avverso sia presente in un report. Se questo specifico record ha al suo interno un certo farmaco, allora la probabilità dell'evento avverso potrebbe cambiare e questa viene definita **probabilità a posteriori**. Se la probabilità a posteriori risulta essere più elevata di quella a priori, allora la presenza del farmaco nella segnalazione ha incrementato le possibilità che quell'evento avverso si sviluppasse e la coppia tra farmaco ed evento avverso è presente più spesso di quanto ci si possa aspettare.
La legge di Bayes si esprime come

$$P(A|D) = \frac{P(A,D)}{P(D)} \tag{2.16}$$

e in quest'ambito possiamo riscriverla come

$$P(A|D) = P(A) \cdot \frac{P(A,D)}{P(A) \cdot P(D)} \tag{2.17}$$

dove con $P(A|D)$ esprimiamo la probabilità a posteriori, ovvero la probabilità che una specifica reazione avversa A sia presente in una segnalazione insieme al farmaco D; $P(A)$ è la probabilità a priori, ovvero la probabilità che l'evento avverso A sia presente in una segnalazione; $P(D)$ è la probabilità a priori, ovvero la probabilità che il farmaco D sia presente in una segnalazione; $P(A,D)$ è la probabilità congiunta, ovvero la probabilità che il farmaco D e l'evento avverso A siano entrambi presenti nella stessa segnalazione.

L'informazione comune, come viene definita in *information theory* (Pearl, 1988), misura la quantità di informazione che abbiamo relativamente ad una variabile (X) quando abbiamo informazioni sullo stato di un altra variabile (Y), e dunque misura la forza dell'associazione tra le due variabili X e Y, e viene definita come

$$I(X,Y) = \sum_x \sum_y P(x,y) \cdot \log \frac{P(x,y)}{P(x) \cdot P(y)} \tag{2.18}$$

dove x rappresenta uno specifico stato della variabile X e y rappresenta uno specifico stato della variabile Y. Si nota che nella formula è presente una quantità logaritmica: questo implica che nel caso di eventi indipendenti l'informazione di quest'ultimi sia additiva. Bate et al. hanno quindi definito la quantità *Information Component* (IC) che indica la forza dell'associazione tra uno specifico stato in ciascuna delle due variabili ed è la forma logaritmica del fattore di simmetria (definito come $P(A,D)/P(A) \cdot P(D)$) legato alle probabilità a priori e posteriori citate in precedenza:

$$IC = \log_2 \frac{P(x,y)}{P(x)P(y)}. \tag{2.19}$$

Considerando questa quantità possiamo identificare quattro combinazioni di stati delle variabili differenti, ma si è interessati solo alla combinazione che considera la presenza di uno specifico farmaco e uno specifico evento avverso all'interno della stessa segnalazione.

Quando il valore dell'IC sarà positivo per una combinazione farmaco - evento avverso, questo significherà che la coppia è più fortemente associata di quello che ci si aspetta confrontato con quanto raccolto dal *database* (c_i e c_j); al contrario più il valore risulta essere vicino a zero più quanto si evince dal *database* indica indipendenza tra la coppia farmaco - evento avverso.

La quantità IC, in una BCPNN, viene stimata come il peso tra un neurone nello strato degli eventi avversi e un neurone nello strato dei farmaci.
Per questa quantità è anche possibile fornire un intervallo di confidenza per ottenere una misura di certezza del valore IC.

poiché non siamo a conoscenza delle vere probabilità $p(A)$, $p(D)$ o $p(A,D)$, viene assunta, per convenienza, una distribuzione Beta per ciascuna di queste probabilità (Heckerman, 1997). Tramite questa assunzione siamo quindi in grado di calcolare i valori attesi e le varianze delle distribuzioni di ogni variabile poiché queste quantità per una distribuzione Beta sono note.
Queste distribuzioni diventano sempre più strette man mano che si ottengono maggiori informazioni, poiché la varianza decresce. Perciò, man mano che i conteggi aumentano di valore, le precedenti distribuzioni a posteriori diventeranno le nuove distribuzioni a priori e un nuovo insieme di distribuzioni a posteriori verrà calcolato.

Per il calcolo della varianza, Bate et al. decidono di sfruttare l'approssimazione normale poiché permette il calcolo della varianza dell'IC ($V(IC)$)

partendo dalle varianze di $p(A)$, $p(D)$ e $p(A, D)$. Usando questo metodo possiamo calcolare $V(IC)$ come:

$$V(IC) \approx \left(\frac{1}{\log 2}\right)^2 \left[\frac{C - c_{ij} + \gamma - \gamma_{11}}{(c_{ij} + \gamma_{11})(1 + C + \gamma)} + \frac{C - c_i + \alpha - \alpha_1}{(c_i + \alpha_1)(1 + C + \alpha)} + \frac{C - c_j + \alpha - \alpha_1}{(c_j + \alpha_1)(1 + C + \alpha)}\right] \tag{2.20}$$

dove α_1 e α_0 sono fattori delle distribuzioni Beta di $p(A)$ e $p(D)$, e γ_{11} e γ sono i corrispondenti fattori della distribuzione di probabilità congiunta $p(A, D)$. Entrambe le coppie di fattori rispecchiano la nostra fiducia nelle probabilità fornite dalle priori Beta. Un'assunzione a priori viene fatta sull'uguaglianza delle distribuzioni a priori $p(A)$ e $p(D)$; in una distribuzione Beta questo corrisponde alle costanti α_1 e α_0 definite come $\alpha_1 = \alpha_0 = 1$. γ_{11} e γ definiscono la distribuzione congiunta. Viene fissato quindi $\gamma_{11} = 1$ e si definisce

$$\gamma = \frac{\gamma_{11}}{p(A)p(D)}. \tag{2.21}$$

Da questa definizione possiamo vedere come l'IC tenda a zero quando c_{ij} e C tendono a zero, perché viene assunta indipendenza nella relazione tra farmaco e reazione avversa quando non abbiamo nessuna segnalazione presente nel *database* del farmaco o dell'evento avverso.

2.3 Fasi del lavoro di analisi

Nel capitolo 4 verranno presentati dei metodi basati sulla regressione logistica con penalizzazione LASSO da utilizzare per identificare segnali di farmacovigilanza in *dataset* di segnalazioni spontanee che verranno presentati nel capitolo 3. Verrà dapprima effettuato uno studio di simulazione, costruendo degli insiemi di dati che possano permettere il test dei metodi presentati in sezione 4. Verranno seguite due strade parallele: una con un *dataset* in cui l'obiettivo sarà quello di identificare coppie farmaco - evento avverso, e una seconda con un *dataset* in cui l'obiettivo sarà quello di testare la capacità dei metodi di identificare associazioni tra due farmaci che interagiscono tra di loro rispetto ad un evento avverso. Successivamente a questa fase di simulazione verrà applicato il metodo LASSO BIC ai dati reali: in particolare verranno selezionati dai dati FAERS relativi al terzo e quarto trimestre del 2019, un

sottoinsieme di dati che contenga al suo interno un numero di eventi avversi ridotto e un numero di farmaci non elevato per motivi computazionali. Per l'identificazione invece delle coppie di farmaci in interazione verrà sfruttato il *dataset TWOSIDES*, presentato nella sezione 3.3. Anche in questo secondo caso, il numero di eventi avversi e di farmaci, per motivi computazionali, sarà ridotto. L'idea è che se i metodi sono in grado di identificare le corrette associazioni in entrambi i casi, con un numero ristretto di osservazioni, con un numero più elevato il risultato sia lo stesso se non addirittura più preciso. Il procedimento che verrà svolto di seguito viene presentato anche graficamente nel diagramma di figura 2.1.

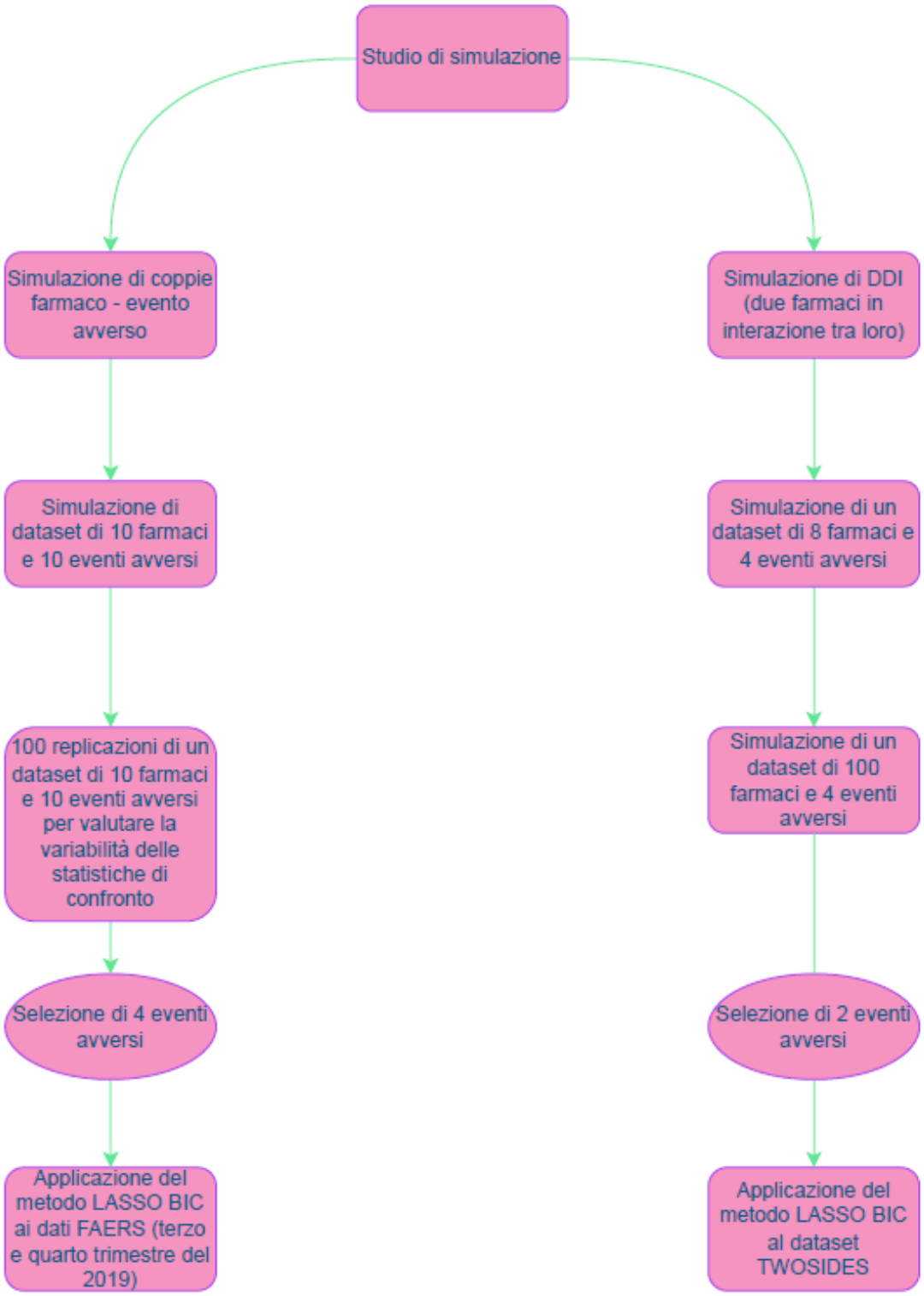


Figura 2.1: Procedimento delle analisi svolte nel seguito.

Capitolo 3

I dati

3.1 FDA *Adverse Event Reporting System*

Per cercare di individuare e sviluppare questi metodi di *data mining* si è deciso di utilizzare i dati pubblici della *FDA*, raccolti nel cosidetto *FDA Adverse Event Reporting System* (*FAERS*) negli anni in cui esso risulta disponibile online. Questo vuol dire che i dati disponibili vanno dal quarto trimestre dell'anno 2012 fino ai più recenti, ovvero il terzo trimestre dell'anno 2021, per un totale di 17.269.569 segnalazioni spontanee.

I dati di questo database sono suddivisi in diverse tabelle, nello specifico sei tabelle che vengono suddivise in base alle informazioni che contengono:

- la tabella *drug*: contiene le informazioni relative ai farmaci. Nello specifico contiene informazioni sul principio attivo del farmaco, il nome del farmaco, informazioni varie sulle dosi somministrate al paziente, informazioni su come è stato somministrato il farmaco e una coppia di due variabili che forma una chiave univoca del *database*, utile per poter unire i vari *dataset* tra di loro;

- la tabella *reaction*: contiene le informazioni relative agli effetti avversi sviluppati da un paziente a seguito della somministrazione del farmaco: in particolare possiamo trovare indicazioni circa il tipo di reazione avversa che si è sviluppata. Tutte le reazioni avverse che sono presenti nel *dataset* sono termini contenuti nel *Medical Dictionary for Regulatory Activities* (**MedDRA**, Brown, 2007), che permette dunque di unificare tutte le terminologie di eventi avversi;

- la tabella *demographic*: contiene informazioni demografiche del soggetto che ha sviluppato gli eventi avversi;
- la tabella *therapy*: contiene le informazioni relative alle date di inizio e fine della terapia prescritta per il farmaco che ha creato evento avverso;
- la tabella *indication*: contiene i termini MedDRA per le indicazioni d'uso (diagnosi) per i farmaci segnalati;
- la tabella *outcome*: contiene informazioni circa lo stato del soggetto dopo il verificarsi dell'evento avverso: dunque ci sono informazioni circa una eventuale ospedalizzazione, morte, disabilità a seguito della reazione avversa;
- la tabella *Report_ Sources*: contiene informazioni su chi ha segnalato la reazione avversa del farmaco: possono essere gli stessi utilizzatori, le aziende farmaceutiche, dei professionisti del settore medico (medici stessi o infermieri), o provenire da letteratura medica o studi (come per esempio gli stessi *trials* clinici).

3.2 Il *database OFFSIDES*

Il *FAERS* presenta delle problematiche a livello di codifica dei farmaci e dei loro principi attivi. Infatti, è impossibile ottenere l'intera lista di farmaci dal *FAERS* perché i farmaci o principi attivi per un singolo caso sono tutti concatenati in una stringa senza un separatore univoco: può infatti capitare di avere un principio attivo che si presenta come "(1-743)-(1638-2332)-BLOOD-COAGULATION FACTOR VIII (SYNTHETIC HUMAN) FUSION PROTEIN WITH IMMUNOGLOBULIN G1 (SYNTHETIC HUMAN FC DOMAIN FRAGMENT), (1444-6'),(1447-9')-BIS(DISULFIDE) WITH IMMUNOGLOBULIN G1 (SYNTHETIC HUMAN FC DOMAIN FRAGMENT)", dove non è presente un separatore chiaro e dunque riuscire ad ottenere una lista utilizzabile risulta impossibile (M. Pham, Cheng e Ramachandran, 2019).

Si è quindi deciso di utilizzare come *database* di partenza, un *database* sviluppato e presentato da un gruppo di studio americano, che sfrutta dati provenienti dall'*AERS*, antecedente al FAERS (nel periodo compreso tra il primo trimestre del 2004 e il primo trimestre del 2009), dal *database*

SIDER e dal *Canada MedEffect* (considerando le segnalazioni fino all'anno 2009), un *database* molto simile all'AERS (Tatonetti et al., 2012). Questo database chiamato *OFFSIDES* contiene al suo interno informazioni derivanti dai *database* sopracitati: la particolarità di questi dati, presentati da Tatonetti et al., è il fatto di essere dati aggregati, infatti ogni riga non rappresenta una segnalazione, ma bensì riporta tutte le informazioni utili per calcolare misure di disproporzionalità, per ciascuna coppia farmaco - evento avversa presente nei *database*. Per poter utilizzare quindi questo *dataset* si deve procedere ad una disgregazione dei dati per ottenere un *dataset* che ha raggiunto 55.902.234 segnalazioni. Inoltre, una particolarità di questo *database* è la possibilità di sfruttarlo come *silver standard*, una sorta di *gold standard* non validato, ma calcolando alcune misure di disproporzionalità frequentiste è dunque possibile utilizzarlo per confrontare i risultati provenienti da altri metodi in fase di sviluppo.

3.3 Il *database TWOSIDES*

Lo stesso gruppo di ricerca ha creato insieme al *database OFFSIDES* precedentemente presentato (sezione 3.2), un'ulteriore insieme di dati denominato *TWOSIDES*. La struttura dei due è pressoche la stessa, con le quantità a, b, c e d classiche di una tabella di contingenza (2.1) e il PRR con relativo errore. C'è però una differenza tra i due: infatti mentre il *database OFFSIDES* contiene le informazioni relative all'associazione tra un singolo farmaco ed un avento avverso, il *TWOSIDES* contiene le informazioni relative alle interazioni tra farmaci (*Drug-Drug Interaction*, **DDI**). Risulta quindi utile per validare metodi statistici per identificare se coppie di farmaci sono associate a un evento avverso nel soggetti. A differenza delle reazioni avverse causate da un singolo farmaco dove, se pur in numero ridotto, possono essere identificati durante i *trials* clinici, gli eventi avversi causati da più farmaci non sono identificabili durante i *trials* perché i soggetti che assumono già altri farmaci vengono spesso esclusi dagli studi per evitare problemi gravi.

Le problematiche evidenziate in precedenza hanno fatto si che venissero sviluppati nel tempo dei metodi definiti di disproporzionalità, che sono stati introdotti nella sezione 2. Successivamente, con lo sviluppo della tecnologia e dunque con lo sviluppo di computer sempre più potenti a livello computazionale, sono stati sviluppati, implementati e testati degli ulteriori metodi

di *data mining* e di *machine learning* che sfruttano le numerose quantità di dati per creare algoritmi per ottenere segnali di farmacovigilanza. Nello specifico, nel seguito, verrà presentato un metodo che sfrutta la penalizzazione LASSO modificata per utilizzare, come criterio per la selezione delle variabili, il *Bayesian Information Criterion* (BIC), proposto da un gruppo di studio francese e testato sul *databse* di farmacovigilanza nazionale francese (Courtois, Tubert-Bitter e Ahmed, 2021).

Capitolo 4

Metodi

I metodi statistici attualmente utilizzati in farmacovigilanza sono quelli appartenenti alla famiglia dei metodi di disproporzionalità frequentisti e bayesiani descritti nei capitolo 2. Con l'avanzare della tecnologia però nuovi metodi si stanno inserendo in questo ambito per rendere sempre più efficiente ed efficace il processo di identificazione delle associazioni fra farmaci ed eventi avversi cercando di ridurre al minimo l'identificazione di coppie non associate. Le alternative che più vengono esplorate in questi anni sono quelle degli algoritmi di *machine learning* e metodi di *data mining*. Tra questi metodi, uno che ha una struttura adatta a dati di questo tipo, come vedremo in seguito, è la regressione logistica con penalizzazione LASSO presentato in Tibshirani (1996) e ripreso poi con modifiche da altri autori. In particolare, quello che verrà presentato e utilizzato in seguito è una modifica che sfrutta il BIC (*Bayesian Information Criterion*) per la selezione delle variabili, utilizzato per la prima volta da Courtois, Tubert-Bitter e Ahmed (2021) in ambito di farmacovigilanza. Il principale motivo per cui non viene utilizzata una più semplice regressione logistica, comunque implementata nel seguito come confronto nelle prime simulazioni, è il fatto che le penalizzazioni LASSO introducono nella stima dei parametri di penalizzazione che permettono una miglior scelta e stima dei coefficienti delle variabili (in questo caso dei farmaci). Questo procedimento dovrebbe quindi permettere di ottenere una selezione delle variabili migliore rispetto a quella ottenibile con la tecnica *stepwise*. Il LASSO inoltre risulta essere una soluzione computazionalmente più efficiente per stimare una regressione, logistica in questo contesto, in dati ad elevata dimensionalità.

Di seguito verrà presentato il metodo LASSO così come è stato introdotto inizialmente da Tibshirani, 1996. Successivamente verrà introdotto il metodo innovativo basato sul BIC proposto da Courtois, Tubert-Bitter e Ahmed, 2021 e una versione pensata appositamente per l'identificazione delle interazioni tra variabili proposta da Lim e Hastie (2015).

4.1 LASSO

4.1.1 Definizione

Supponiamo di avere N coppie (x_i, y_i) dove $x_i = (x_{i1}, \cdots, x_{ip})^T$ sono le variabili esplicative per la i-esima unità statistica e y_i è la variabile risposta relativa alla medesima unità statistica i. Come in un normale *set up* in ambito di regressione, anche nel LASSO viene assunto che le osservazioni siano tra di loro indipendenti, o almeno che le y_i siano condizionatamente indipendenti date le x_{ij} $\forall i, j$. Assumiamo che le x_{ij} siano standardizzate e quindi che $\sum_i x_{ij}/N = 0$, $\sum_i x_{ij}^2/N = 1$. Sia $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)^T$, la stima LASSO $(\hat{\alpha}, \hat{\beta})$ è definita da

$$(\hat{\alpha},\ \hat{\beta}) = \text{argmin} \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \quad \text{soggetto a } \sum_j |\beta_j| \leq s. \quad (4.1)$$

In questo contesto $s \geq 0$ viene definito parametro di regolazione. Ora, per tutti i valori di s, la soluzione per α è $\hat{\alpha} = \bar{y}$. Si può assumere senza perdita di generalità che $\bar{y} = 0$ e dunque omettere α.

Il parametro $s \geq 0$ controlla la quantità di *shrinkage* che viene applicata alle stime. Siano $\hat{\beta}_j^0$ le stime complete ai minimi quadrati e sia $s_0 = \sum |\hat{\beta}_j^0|$. Valori di $s < s_0$ causeranno uno *shrinkage* delle soluzioni verso zero, mentre alcuni coefficienti saranno esattamente 0.

4.1.2 Geometria del LASSO

Il criterio $\sum_{i=1}^{N} (y_i - \sum_j \beta_j x_{ij})^2$ equivale alla funzione quadratica $(\beta - \hat{\beta}^0)^T X^T X (\beta - \hat{\beta}^0)$ più una costante. I contorni ellittici di questa funzione sono mostrati dalle curve in Figura 4.1. Le curve sono centrate nella stima OLS (*Ordinary Least Squares*). La regione costante è un rombo inscritto in un cerchio di centro 0 e raggio 1. La soluzione LASSO è il primo posto

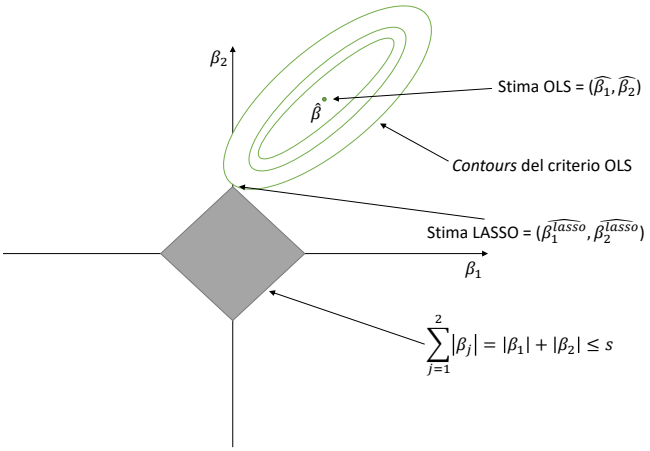


Figura 4.1: *Contours* di RSS (*Residual Sum of Squares*, somma dei quadrati dei residui) e funzione costante per il LASSO. L'area romboidale grigia è la regione costante definita da $|\beta_1| + |\beta_2| \leq s$.

in cui i contorni toccato il rombo e questo avverà talvolta in un angolo, corrispondente a un coefficiente pari a 0.

4.1.3 Stima del parametro di regolazione

Il parametro di regolazione del LASSO può essere stimato in modi differenti, metodo che può venire scelto anche in base al contesto in cui si sta utilizzando il LASSO. Tra i metodi conosciuti per la stima di s si può trovare la *cross validation*, la *generalized cross validation* e la stima analitica non distorta del rischio. Questi metodi sono i primi conosciuti per la stima del parametro in quanto vengono già citati e presentati da Tibshirani nel 1996. Come vedremo in seguito soprattutto nel contesto di farmacovigilanza un ulteriore metodo per scegliere il valore ottimale di s è dato dall'utilizzo del criterio BIC.

4.2 LASSO BIC

Come abbiamo già visto nelle sezioni precedenti, in ambito di farmacovigilanza esistono già molti metodi che analizzano i grandi *dataset* di segnalazioni spontanee che si hanno a disposizione con l'obiettivo di identificare coppie

farmaco - evento avverso, che poi necessiteranno di investigazione medica per delineare la causalità o meno. I metodi di disproporzionalità non sono molto accurati per interazioni tra farmaci e sono soggetti all'effetto di mascheramento. Per questo motivo, negli anni più recenti si è sperimentato l'utilizzo di modelli di regressione logistica con penalizzazione LASSO per cercare di aggirare le limitazioni dei metodi di disproporzionalità. Un metodo che sfrutta questa penalizzazione, è stato introdotto da Courtois, Tubert-Bitter e Ahmed e viene presentato di seguito.

Si considera come unità statistica il singolo report presente nel *dataset* a disposizione: questo report sarà composto dalla variabile risposta, rappresentata dalla presenza o assenza di uno specifico evento avverso, e dalle variabili esplicative, rappresentate dalle variabili indicatrici di presenza o assenza di uno specifico farmaco. Possiamo dunque scrivere il *dataset* in forma matriciale come

$$\boldsymbol{AE_{ij}} = \boldsymbol{d_{ijz}} \quad \text{con } \boldsymbol{d_{ijz}} = [\boldsymbol{d_{ij1}}, \cdots, \boldsymbol{d_{ijD}}] \tag{4.2}$$

$$\begin{bmatrix} AE_{1j} \\ AE_{2j} \\ ... \\ AE_{ij} \\ ... \\ AE_{Nj} \end{bmatrix} = \begin{bmatrix} d_{1j1} & ... & d_{1jz} & ... & d_{1jD} \\ d_{2j1} & ... & d_{2jz} & ... & d_{2jD} \\ ... & ... & ... & ... & ... \\ d_{ij1} & ... & d_{ijz} & ... & d_{ijD} \\ ... & ... & ... & ... & ... \\ d_{Nj1} & ... & d_{Njz} & ... & d_{NjD} \end{bmatrix}, \text{ per } \forall j = 1, \cdots, M \tag{4.3}$$

con $j = 1, \cdots, M$, dove M è il numero di eventi avversi presenti nel *dataset*, $i = 1, \cdots, N$, con N numero di segnalazioni spontanee nel dataset e $z = 1, \cdots, D$, con D numero di farmaci nel *dataset*. La matrice relativa alle covariate, rappresentata nell'equazione 4.3, è grande, binaria ed estremamente sparsa, ed inoltre c'è anche un forte sbilanciamento tra presenza e assenza di un dato evento avverso.

La penalizzazione LASSO si presta molto all'ambito delle regressioni a grandi dimensioni per la sua efficienza computazionale. La parsimonia indotta dalla norma L_1 (definita come la somma dei valori vettoriali assoluti) è una condizione molto importante per algoritmi applicabili alla farmacovigilanza. Tuttavia, mentre la convalida incrociata è normalmente utizzata per migliore la capacità predittiva di un modello, è meno semplice scegliere il miglior parametro di regolarizzazione che controlla la sparsità del modello nel contesto

di selezione delle variabili. É stato dimostrato da Courtois, Tubert-Bitter e Ahmed che non c'è un parametro di regolazione che permetta al LASSO di godere delle proprietà di predittore ideale presentate da Fan e Li (2001). Questo significa che la selezione delle variabili fornita dal LASSO potrebbe essere non corretta. Per questo motivo tecniche di sottocampionamento sono state proposte per diminuire l'importanza della scelta del parametro di regolazione; in particolare Ahmed, Pariente e Tubert-Bitter (2018) hanno proposto un algoritmo di sottocampionamento chiamato CISL, specificatamente implementato per tenere conto del grande squilibrio nei dati di segnalazione spontanea.

Il LASSO adattivo è un'alternativa implementata per cercare di migliorare la selezione delle variabili propria del LASSO. Consiste nell'uso di pesi adattivi (AW) per penalizzare diversamente le covariate rispetto alla penalizzazione L_1. Inizialmente, gli AW (*Adaptive Weigth*) sono ottenuti dalle stime di massima verosimiglianza.

La versione proposta da Courtois, Tubert-Bitter e Ahmed, che verrà utilizzata in seguito, è basata su un nuovo metodo automatico di identificazione dei segnali basato sul LASSO adattivo che punta a migliorare la selezione delle variabili implementata dal LASSO tramite delle penalizzazioni adattive specifiche per ciascuna covariata. Questa nuova strategia include anche l'uso del criterio BIC. Sono stati proposti due diversi tipi di AW:

1. una regressione LASSO dove il parametro di regolazione viene scelto utilizzando il BIC;

2. una versione che sfrutta l'algoritmo CISL.

Questi AW ottenuti mediante questi due metodi sono poi integrati in una regressione logistica con penalizzazione LASSO usando il BIC per la selezione del parametro di regolazione.

4.2.1 Il LASSO logistico

Sia N il numero di segnalazioni spontanee presenti nel *dataset* e P il numero di farmaci presenti. Indichiamo con $\boldsymbol{X}$ la matrice binaria di dimensione $N \times P$ relativa all'esposizione ai farmaci e $\boldsymbol{y}$ il vettore di dimensione N di risposte binarie relativo alla presenza o assenza di uno specifico evento avverso di interesse (avremo quindi tanti vettori e tante matrici quanti sono

gli eventi avversi di nostro interesse). Per $i \in \{1, \cdots, N\}$, il corrispondente modello logistico sarà

$$\text{logit}(Pr(y_i = 1|x_i)) = \beta_0 + \sum_{p=1}^{P} \beta_p \cdot x_{ip}, \tag{4.4}$$

dove β_0 è l'intercetta, e $\boldsymbol{\beta}$ è un vettore di dimensione P dei coefficienti di regressione associati ai farmaci.
Si ottengono dunque le stime del LASSO logistico penalizzato come

$$(\hat{\beta_0}, \hat{\boldsymbol{\beta}})_\lambda = \text{argmax}_{\beta_0, \boldsymbol{\beta}}\{l((\beta_0, \boldsymbol{\beta}), y, X) - \text{pen}(\lambda)\}, \tag{4.5}$$

dove l è la log-verosimiglianza del modello 4.4, λ è il parametro di regolazione e $pen(\lambda)$ è definito come

$$\text{pen}(\lambda) = \lambda|\boldsymbol{\beta}|_1 = \lambda \sum_{p=1}^{P} |\beta_p|. \tag{4.6}$$

Indichiamo con $\hat{\boldsymbol{\beta}_\lambda}$ il vettore di dimensione P nella stima LASSO $(\hat{\beta_0}, \hat{\boldsymbol{\beta}})_\lambda$. Grazie all'inserimento della norma L_1 nella penalizzazione espressa in 4.6, alcuni coefficienti $\hat{\beta}_\lambda$ vengono fissati pari a 0, in questo modo le covariate ad essi associate vengono escluse dal modello comportando dunque una selezione delle variabili. Perciò, il parametro λ di regolazione è fortemente legato al numero di coefficienti non pari a 0. Nell'ambito di farmacovigilanza, ciò che si ricerca sono covariate con valori del coefficiente $\hat{\beta}_\lambda$ positivi.

Scelta del parametro di penalizzazione

Le strategie considerate da Courtois, Tubert-Bitter e Ahmed per la selezione del parametro di penalizzazione sono tre:

- *cross-validation*: ogni ciclo di *cross-validation* implica il partizionamento del *dataset* in n_f sottoinsiemi, chiamati *folds*. Nello specifico $n_f - 1$ sottoinsiemi vengono sfruttati per l'addestramento dell'algoritmo (il modello viene stimato basandosi su questi dati), e il rimanente sottoinsieme viene invece utilizzato per validare il modello ottenuto tramite una matrice di confusione o delle misure atte a valutare la *performance* previsiva del modello. La procedura viene ripetuta f volte, in modo tale che tutti i sottoinsiemi del *dataset* iniziale svolgano il

ruolo di *dataset* di verifica. Questo avviene in ambito di previsione. Nel contesto di una regressione LASSO, invece, la *cross-validation* viene eseguita per ciascun valore di λ testato. Il λ che viene scelto risulterà essere quello che offre le migliori risposte in termini di *performance* previsive per il modello;

- BIC: un'alternativa alla sopracitata *cross-validation* è il criterio di selezione BIC (*Bayesian Information Criterion*). Per ogni valore λ testato viene calcolato il valore del BIC come

$$BIC_\lambda = -2l_\lambda + \mathrm{df}(\lambda) + \log(N), \tag{4.7}$$

  dove l_λ è la log-verosimiglianza di un modello di regressione logistico, dove vengono incluse le covariate con coefficiente $\hat{\beta}_\lambda$ diverso da zero e $df(\lambda) = |\hat{\beta}_\lambda \neq 0|$. La conseguenza dell'uso di questo metodo porta ad ottenere il sottoinsieme di covariate che corrisponde al modello che minimizza il valore del criterio BIC, piuttosto che selezionare uno specifico valore del parametro λ;

- permutazioni: l'ultimo tipo di approccio presentato da Courtois, Tubert-Bitter e Ahmed è quello basato sulle permutazioni. In particolare, definito π come ogni permutazione di $\{1, \cdots, N\}$, prendiamo $y_{\pi_l} = (y_{\pi_{(1)}}, \cdots, y_{\pi_{(N)}})$, definita come una versione permutata della variabile risposta y con $1 \leq l \leq K$. Una regressione LASSO viene quindi stimata per ciascuna di queste permutazioni ottenute utilizzando come variabile risposta y_{π_l} e come matrice delle covariate $\boldsymbol{X}$. Si ottiene perciò $\lambda_{max}(y_{\pi_l})$, cioè il valore più piccolo del parametro di penalizzazione tale che nessuna covariata è selezionata nella regressione LASSO utilizzando la variabile risposta y_{π_l}.

LASSO e sottocampionamento

Per cercare di aggirare il problema della selezione del parametro di regolazione nella regressione LASSO, Meinshausen e Bühlmann (2010) hanno proposto l'algoritmo *stability selection*. Questo consiste nel perturbare i dati, sottocampionandoli molto volte, eseguendo una regressione LASSO su questi sottoinsiemi selezionati casualmente senza reinserimento, e scegliendo le covariate che compaiono nella maggior parte dei risultati inducendo quindi una

selezione tramite LASSO. Una variazione a questo algoritmo è stata proposta da Ahmed, Pariente e Tubert-Bitter per tenere conto di forti sbilanciamenti nelle variabili relative agli eventi avversi di *database* di farmacovigilanza. Nell'algoritmo, chiamato *Class-Imbalanced Subsampling LASSO* (CISL), si ha che i sottoinsiemi sono estratti seguendo un campionamento non equiprobabile con reinserimento in modo tale da ottenere una miglior rappresentazione dei soggetti che hanno sviluppato l'evento di interesse. Una regressione LASSO viene poi eseguita su tutti questi sottoinsiemi con lo scopo di ottenere la quantità di seguito definita:

$$\hat{\pi}_p^b = \frac{1}{B} \sum_{\eta=1}^{E} \mathbb{1}[\hat{\beta}_p^{\eta,b} > 0], \tag{4.8}$$

dove E è il massimo numero di covariate selezionate da tutte le regressioni LASSO, $\eta \in \{1, \cdots, E\}$ è il numero di covariate selezionate e $\hat{\beta}_p^{\eta,b}$ è il coefficiente di regressione stimato dal LASSO logistico per il farmaco p, per il sottoinsieme $b \in \{1, \cdots, B\}$ per un modello che include η covariate. Perciò, per ogni farmaco avremo una distribuzione empirica di $\hat{\pi}_p^b$ ottenuta dai B sottoinsiemi. Un farmaco viene dunque selezionato quando un quantile (normalmente il decimo percentile, q_{10}) della distribuzione di $\hat{\pi}_p^b$ è diverso da zero.

4.2.2 LASSO adattivo

Come definito da Fan e Li (2001), una procedura ottimale dovrebbe avere le seguenti proprietà: identificare il sottoinsieme corretto di veri predittori e produrre stime non distorte. Nel loro lavoro questi autori hanno dimostrato che il LASSO non gode di queste proprietà poiché ci sono alcune situazioni in cui la selezione delle variabili apportata dal LASSO può essere incosistente. Inoltre, si ha che con lo stesso valore di penalizzazione per tutte le covariate, il LASSO tende a sovrapenalizzare quelle più importanti e può produrre stime distorte. Per ovviare a questo problema, Zou ha proposto il LASSO adattivo dove gli AW sono utilizzati per penalizzare le covariate in modo differente rispetto alla penalizzazione L_1, definita come

$$\text{pen}(\lambda) = \lambda \sum_{p=1}^{P} \omega_p |\beta_p|. \tag{4.9}$$

La penalizzazione applicata alla p-esima covariata è definita da $\lambda_p = \lambda \cdot \omega_p$. Più è alto il valore del peso ω_p maggiormente sarà penalizzata la variabile p e minori solo le possibilità che la variabile sia inclusa nel modello.

Per costruire gli AW, Zou (2006) ha proposto di usare inizialmente uno stimatore consistente di $\boldsymbol{\beta^*}$, il vettore di dimensione P dei coefficienti di regressione. Per questo fine, consideriamo $\boldsymbol{\hat{\beta}_p^{mle}}$, la stima di massima verosimiglianza per la p-esima covariata e definiamo il peso associato $\omega_p = \frac{1}{|\hat{\beta}_p^{mle}|^\gamma}$, con $\gamma > 0$. Tuttavia, nel contesto dimensionale proprio dei *dataset* di farmacovigilanza, non è banale una stima consistente per la costruzione degli AW, poiché il calcolo della massima verosimiglianza non è fattibile.

Nel caso lineare, Bühlmann e Geer hanno proposto di usare le stime dei coefficienti di regressione penalizzati tramite una regressione LASSO per determinare gli AW considerando $\gamma = 1$. In entrambi i casi, LASSO e LASSO adattivo, il parametro di penalizzazione λ viene selezionato tramite *cross-validation*. Questa procedura in due step, che include un primo passo utilizzando la *cross-validation* insieme al LASSO, è stata proposta nel caso logistico e viene definita **LASSO iterativo** (Huang, Ma e Zhang, 2008b). Definiamo $\boldsymbol{\hat{\beta}^{lcv}}$ il vettore di dimensione P dei coefficienti di regressione LASSO ottenuti tramite una regressione LASSO sfruttando la *cross-validation*; gli AW associati con i p farmaci nel passo di LASSO adattivo sono definiti come

$$\omega_p^{lcv} = \begin{cases} \frac{1}{|\hat{\beta}_p^{lcv}|} & \text{se } \hat{\beta}_p^{lcv} \neq 0 \\ \infty & \text{se } \hat{\beta}_p^{lcv} = 0. \end{cases} \tag{4.10}$$

Perciò, una covariata che non viene selezionata con il LASSO implementato con la *cross-validation* nel primo step è automaticamente esclusa nello step del LASSO adattivo.

Nel caso lineare, Huang, Ma e Zhang dimostrano che sotto determinate condizioni, l'uso dei coefficienti di regressione univariati per determinare gli AW presenta delle buone proprietà. Definiamo quindi $\hat{\beta}_p^{univ}$ i coefficienti univariati associati al p-esimo farmaco; definiamo quindi gli AW associati al p-esimo farmaco nello step del LASSO adattivo come

$$\omega_p^{univ} = \frac{1}{|\hat{\beta}_p^{univ}|}. \tag{4.11}$$

4.2.3 Estensione del LASSO adattivo alla farmacovigilanza

Poiché l'obiettivo della farmacovigilanza è quello di selezionare il vero sottoinsieme dei farmaci associati ad un evento avverso, per questo scopo viene quindi utilizzato il BIC, come definito in precedenza, per identificare il sottoinsieme di farmaci finale nello step del LASSO adattivo, piuttosto dell'alternativa della *cross-validation* discussa in precedenza. Courtois, Tubert-Bitter e Ahmed propongono due alternative di AW rispetto a quelli precedentemente citati.

La prima versione consiste nell'usare il BIC nel primo step. Definiamo quindi $\hat{\beta}^{lbic}$, il vettore di dimensione P delle stime dei coefficienti di regressione non penalizzati, ottenuti nel primo step; definiamo inoltre l'AW associato al p-esimo farmaco nello step del LASSO adattivo come

$$\omega_p^{lb} = \begin{cases} \frac{1}{|\hat{\beta}_p^{lbic}|} & \text{se } \hat{\beta}_p^{lbic} \neq 0 \\ \infty & \text{se } \hat{\beta}_p^{lbic} = 0. \end{cases} \tag{4.12}$$

Dunque una covariata che non viene selezionata dal LASSO BIC nel primo step, viene automaticamente esclusa anche dal secondo step di LASSO adattivo.

La seconda proposta, invece, permette di ottenere gli AW sfruttando l'approccio basato sull'algoritmo CISL. Prima viene implementato l'algoritmo CISL considerando un vincolo non nullo nel calcolo della quantità 4.8 invece del vincolo positivo originale.

$$\hat{\tau}_p^b = \frac{1}{E} \sum_{\eta=1}^{E} \mathbb{1}[\hat{\beta}_p^{\eta,b} \neq 0]. \tag{4.13}$$

Questa quantità misura la proporzione in cui una variabile è stata selezionata in E primi modelli forniti dal percorso di regolarizzazione LASSO.
Definiamo quindi gli AW per le p covariate in accordo con il vettore di dimensione B $\hat{\boldsymbol{\tau}}_{\boldsymbol{p}}$ come

$$\omega_p^{cisl} = \begin{cases} \frac{1}{B} & \text{se } \forall b \in \{1, \cdots, B\}\ \hat{\tau}_p^b > 0 \\ \infty & \text{se } \forall b \in \{1, \cdots, B\}\ \hat{\tau}_p^b = 0 \\ 1 - \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}[\hat{\tau}_p^b > 0] & \text{altrimenti.} \end{cases} \tag{4.14}$$

Perciò, più $\hat{\boldsymbol{\tau}}_p$ è non nullo nei sottocampioni B, più gli AW associati sono piccoli.

Nelle analisi che verranno presentate in seguito verranno implementati tutti i metodi presentati in precedenza, in particolare una versione LASSO con scelta del parametro di regolazione effettuata tramite *cross-validation*, con numero di *folds* pari a 10 e un versione del LASSO con scelta del parametro di regolazione tramite BIC. Il primo metodo verrà utilizzato come confronto per il secondo metodo LASSO, in quanto il primo è già stato largamente utilizzato. Lo scopo sarà quello di capire se una selezione del parametro di regolazione tramite criterio BIC possa portare dei vantaggi nell'identificazione di coppie farmaco - evento avverso associate riducendo il numero di variabili selezionate.

4.3 *Hierarchical Group-LASSO*

Nel sezione 6.2 come metodo di confronto per il LASSO BIC introdotto in precedenza, sono stati utilizzati due metodi differenti: il primo basato su un LASSO che sfrutta la *cross-validation* ed il secondo basato sul *hierarchical group-LASSO* introdotto da Lim e Hastie (2015).

Questo secondo metodo è stato proposto appositamente per cogliere interazioni tra variabili, poiché capita che una variabile risposta non venga spiegata a dovere dalle relative variabili esplciative e si necessita quindi di uno o più parametri di interazione per spiegarla al meglio.

4.3.1 Notazione

Nel seguito verrà indicata con $\mathbf{Y}$ la variabile risposta, casuale, relativa ad un evento avverso, $\mathbf{F}$ indicherà le variabili categoriali relative ai farmaci e $L = 2$ è il numero di livelli per ciascuna variabile categoriale (in questo caso sarà costante pari a 2 poiché le variabili relative ai farmaci indicheranno presenza o assenza del farmaco e saranno quindi dicotomiche). A differenza dei casi precedenti, in questo la matrice del disegno $\mathbf{X}$ avrà dimensione maggiore al numero dei singoli farmaci presenti, in quanto dovrà includere anche le interazioni tra i diversi farmaci. Avremo quindi che dato il vettore $\mathbf{Y}$ di dimensione $N\ x\ 1$, e la matrice $\mathbf{F}$ di dimensione $n\ x\ P$, si avrà che la

matrice $\mathbf{X}$ è definita come

$$\begin{bmatrix} x_{11} & ... & x_{1j} & ... & x_{1P} & x_{11} : x_{1ij} & ... \\ x_{21} & ... & x_{2j} & ... & x_{2P} & x_{21} : x_{1j} & ... \\ ... & ... & ... & ... & ... & ... & ... \\ x_{N1} & ... & x_{Nj} & ... & x_{NP} & x_{N1} : x_{Nj} & ... \end{bmatrix} . \tag{4.15}$$

. Sia $\mathbb{E}(Y|F_1 = i, F_2 = j) = \mu_{ij}$, la media condizionata di Y dato che la variabile F_1 assume il valore i e la variabile F_2 assume il valore j (entrambe le variabili sono categoriale dunque i e j sono due livelli della variabile: presenza o assenza del farmaco). Si possono quindi individuare cinque possibili casi:

1. $\mu_{ij} = \mu$, nessun effetto principale o di interazione;
2. $\mu_{ij} = \mu + \theta_1^i$, solo l'effetto principale del farmaco 1;
3. $\mu_{ij} = \mu + \theta_2^i$, solo l'effetto principale del farmaco 2;
4. $\mu_{ij} = \mu + \theta_1^i + \theta_2^j$, entrambi gli effetti principali dei due farmaci;
5. $\mu_{ij} = \mu + \theta_1^i + \theta_2^j + \theta_{1:2}^{ij}$, sia gli effetti principali dei due farmaci che l'interazione tra di essi.

Si può notare come tutti tranne il primo caso sono sovraparametrizzati e una soluzione comune è quella di imporre dei vincoli sulla somma degli effetti principali e di interazione

$$\sum_{i=1}^{L_1} \theta_1^i = 0, \quad \sum_{j=1}^{L_2} \theta_2^j = 0, \text{ con } L_1 = L_2 = 2, \tag{4.16}$$

e

$$\sum_{i=1}^{L_1} \theta_{1:2}^{ij} = 0 \text{ per } j \text{ fissato }, \quad \sum_{j=1}^{L_2} \theta_{1:2}^{ij} = 0, \text{ per } i \text{ fissato, con } L_1 = L_2 = 2 \tag{4.17}$$

Lim e Hastie (2015).

Si dice che un modello di interazione rispetta a una gerarchia forte se un'interazione può essere presente solo se entrambi i suoi effetti principali sono presenti. Si parla invece di gerarchia debole quando almeno uno dei due effetti principali viene mantenuto all'interno del modello insieme all'effetto di interazione.

Si può quindi esprimere il modello per una variabile risposta binaria Y come

$$logit(\mathbb{P}(Y = 1|X)) = \mu + \sum_{i=1}^{p} X_i\theta_i + \sum_{i<j} X_{i:j}\theta_{i:j}. \tag{4.18}$$

Per la stima di questo modello viene minimizzata un'appropriata funzione di perdita $\mathcal{L}$, soggetta ad alcuni vincoli. Se il numero di variabili non è molto grande, è necessario imporre anche i vincoli di identificabilità 4.16 e 4.17 per i coefficienti θ. Questi vincoli si arriva a rispettarli aggiungendo appropriate penalizzazioni alla funzione di perdita $\mathcal{L}$. Di seguito viene riportata la funzione di perdita (*logistic loss*, log-verosimiglianza di una Bernoulli negativa):

$$\begin{aligned}\mathcal{L}(\mathbf{Y};\mu,\theta) = -\Bigg[&\mathbf{Y}^T\left(\mu\cdot\mathbf{1} + \sum_{i=1}^{p}\mathbf{X}_i\theta_i + \sum_{i<j}\mathbf{X}_{i:j}\theta_{i:j}\right)\\ &-\mathbf{1}^T\log\left(\mathbf{1}+\exp\left(\mu\cdot\mathbf{1} + \sum_{i=1}^{p}\mathbf{X}_i\theta_i + \sum_{i<j}\mathbf{X}_{i:j}\theta_{i:j}\right)\right)\end{aligned} \tag{4.19}$$

dove log e exp sono intesi componente per componente.

Capitolo 5

Studio di simulazione

Come primo passo per testare il metodo LASSO BIC vengono simulati degli insiemi di dati di dimensioni relativamente ridotte, per motivi computazionali, ma che mantengono sostanzialmente la scala di quelli presenti nel *database* FAERS. Per creare questi *dataset* di simulazione si è deciso di sfruttare il pacchetto `R` chiamato `SRSim`, implementato e presentato da Dijkstra et al., 2020, permette di ottenere grandi *dataset* di segnalazioni spontanee simulati. Questo risulta essere molto utile perché permette di avere dei dati puliti, dunque senza problematiche legati ai nomi dei farmaci che possono essere sbagliati o scritti talvolta in maniera diversa l'uno dall'altro anche se sono gli stessi, ed inoltre sorpassano uno dei problemi più importanti legati alla farmacovigilanza, ovvero la mancanza di un *gold standard* per poter validare i nuovi metodi che vengono sviluppati.

Nello specifico, per valutare i metodi di nostro interesse, dunque i modelli di regressione con penalizzazione LASSO, abbiamo creato un *dataset* di segnalazioni avverse composto da 10000 segnalazioni, con un numero di farmaci pari a 10 e un numero di reazioni avverse anch'esso pari a 10. Le coppie farmaco - evento avverso tra di loro collegate su un totale di 100 sono 10. Ciò che si ottiene è quindi che ciascun farmaco è associato realmente ad un solo evento avverso. In questa funzione inoltre c'è anche la possibilità di avere un determinato numero di farmaci che vengono definiti *bystander*, che sostanzialmente sono dei farmaci che non sono realmente associati con l'evento avverso, ma che sulla base delle frequenze osservate possono far pensare che invece lo siano: questi sono sostanzialmente dei confondenti, che sono responsabili di un effetto simile all'interazione in fase di analisi e svolgono quindi

un ruolo di farmaci coprescritti che però non sono responsabili dell'evento avverso (Ibrahim et al., 2021). Un modo efficiente per depurare dall'effetto di questi confondenti è la regressione logistica, nel nostro caso viene leggermente modificata utilizzando una regressione logistica con penalizzazione LASSO.
Nel caso specifico di quanto fatto, si ha che la probabilità condizionata che i *bystander* siano pari a uno quando il farmaco è effettivamente la causa dell'evento avverso viene fissata pari a 0.9 in fase di simulazione, valore pari al γ espresso nell'equazione 5.1. Ovvero dato D_i, il farmaco di nostro interesse, e B_j il farmaco *bystander*, possiamo specificare la probabilità sopra citata come:

$$P(D_i = 1|B_j = 1) = \gamma. \tag{5.1}$$

Nel caso in cui invece, non si abbia l'effetto del farmaco *bystander*, la probabilità che il farmaco D_i sia collegato ad un evento avverso è dato da

$$P(D_i = 1|B_j = 0) = \pi_i. \tag{5.2}$$

In questo caso, si ha che la probabilità cambia per ciascun farmaco i, e in particolare viene ipotizzato essere un valore derivante da una distribuzione Beta, una scelta molto comune in ambito di farmacovigilanza. Nel nostro caso i parametri di questa distribuzione Beta, ovvero α e β, nella simulazione vengono assunti rispettivamente pari a 1 e 20 per fare in modo che i farmaci vengano elencati in maniera non molto frequente (Dijkstra et al., 2020). Il numero di *bystanders* è assunto pari a 5.
La relazione tra il farmaco D_i e l'evento avverso AE_j viene espresso in termini di un modello logistico:

$$logit(P[AE_j|D_i = d_i]) = \beta_j + log(OR_{ij}) \cdot d_i \tag{5.3}$$

dove $logit(d) = log[d/(1-d)]$, β_j è l'intercetta e OR_{ij} è l'*odds ratio* tra il farmaco i-esimo e il j-esimo evento avverso. Nel caso in cui il farmaco sia causa dell'evento avverso, $OR_{ij} \in [1, \infty)$ proviene da una distribuzione normale troncata con media 1.5, 3 o 5. Quando il farmaco non causa l'evento avverso, $OR_{ij} = 1$. L'intercetta del modello rappresentato in equazione 5.3 è scelta in modo tale che la probabilità che l'evento avverso appaia sulla segnalazione sia piccola. In fase di simulazione viene scelto il valore del parametro OR_{ij} in modo tale da poter controllare le coppie che risulteranno associate, quelle

con la presenza di *bystanders* e quelle non associate (Dijkstra et al., 2020). Utilizzando quindi questi parametri per la simulazione delle segnalazioni spontanee è stato ottenuto un *dataset* utilizzato per calcolare le usuali misure di dispoporzionalità, utilizzate correntemente in farmacovigilanza ed inoltre sono stati implementati due metodi che sfruttano il LASSO, uno proposto nel pacchetto `R pvm` e uno proposto nel pacchetto `R adapt4pv`. Il primo metodo LASSO utilizza la *cross validation* con un numero di sottoinsiemi (*fold*) pari a 10 per la selezione del parametro λ. Il secondo è un metodo proposto da Courtois, Tubert-Bitter e Ahmed, 2021, che sfrutta per la selezione dei farmaci il *Bayesian Information Criterion* (BIC). Per questi metodi sono poi state ottenute le matrici di confusione per poter valutare le *performance* di questi due metodi. In particolare sono stati ottenuti come statistiche di interesse la precisione ($\frac{A}{A+B}$), la sensibilità ($\frac{A}{A+C}$), la specificità ($\frac{D}{B+D}$) e l'*F1 score* (ottenuto come la media armonica tra precisione e sensibilità o *recall*).

Questi valori assumuno il loro usuale significato, ma va fatta una precisazione per quanto riguarda la specificità e la sensibilità: questi due valori sono spesso utilizzati in ambito medico per valutare la capacità di un test di discriminare tra un malato/positivo o sano/negativo. Nello specifico la sensibilità è la probabilità che un malato/positivo risulti positivo anche al test che viene effettuato. La specificità invece è la probabilità che un sano/negativo risulti negativo anche al test effettuato. In questo contesto i malati/positivi sono i farmaci o coppie di farmaci associati ad un evento avverso, mentre i sani/negativi sono i farmaci o coppie di farmaci non associati ad un evento avverso, mentre il test è il metodo LASSO che viene utilizzato per identificare le coppie farmaco - evento avverso. In particolare, la classe di riferimento per il calcolo delle matrici di confusione in tutto il lavoro presentato di seguito è il valore che indica la non associazione tra farmaci ed eventi avversi. Per cui il valore che in questo contesto risulta più interessante da valutare è la specificità, questo perchè l'interesse principale in farmacovigilanza è l'identificazione di segnali. Anche la sensibilità svolge comunque un ruolo importante nella valutazione di questi metodi poichè escludere coppie farmaco - evento avverso non associate riduce il numero di segnali che vengono riportati agli esperti per la successiva valutazione.

I risultati ottenuti sono riportati in tabella 5.1 e 5.2.

Indice	Evento avverso	LASSO CV	LASSO BIC
Precisione	1	1.00	1.00
	2	0.44	1.00
	3	0.00	1.00
	4	1.00	1.00
	5	1.00	1.00
	6	1.00	1.00
	7	1.00	1.00
	8	1.00	0.89
	9	1.00	1.00
	10	1.00	1.00
Indice	Evento avverso	LASSO CV	LASSO BIC
Sensibilità	1	1.00	0.77
	2	1.00	0.86
	3	0.00	1.00
	4	1.00	0.84
	5	1.00	0.67
	6	0.77	0.77
	7	0.90	0.93
	8	0.76	1.00
	9	0.94	0.94
	10	0.91	0.71

Tabella 5.1: Valori di precisione e sensibilità per i due modelli LASSO stimati.

Indice	Evento avverso	LASSO CV	LASSO BIC
Specificità	1	1.00	1.00
	2	0.90	1.00
	3	0.68	0.00
	4	1.00	1.00
	5	1.00	1.00
	6	1.00	1.00
	7	1.00	1.00
	8	1.00	0.00
	9	1.00	1.00
	10	1.00	1.00
Indice	Evento avverso	LASSO CV	LASSO BIC
F1 score	1	1.00	0.87
	2	0.61	0.93
	3	NaN	NaN
	4	1.00	0.91
	5	1.00	0.81
	6	0.87	0.87
	7	0.95	0.96
	8	0.87	0.94
	9	0.97	0.97
	10	0.95	0.83

Tabella 5.2: Valori di *F1 score* e specificità per i due modelli LASSO stimati.

Come possiamo vedere dalle tabelle il metodo LASSO che sfrutta per la selezione delle variabili il BIC risulta avere gli indici molto simili al LASSO che utilizza la *cross-validation*. Questi risultati sono però stati ottenuti solamente da un *dataset* simulato. Una cosa che si può notare è l'uguaglianza di molti valori per i due metodi: questa cosa potrebbe essere spiegata dalla bassa numerosità di farmaci presi in considerazione. Di seguito viene proposta la medesima analisi effettuata però su 100 simulazioni di *dataset* di segnalazioni spontanee in modo tale da avere anche una misura di variazione degli indici proposti in precedenza e mostrata in figura 5.1.

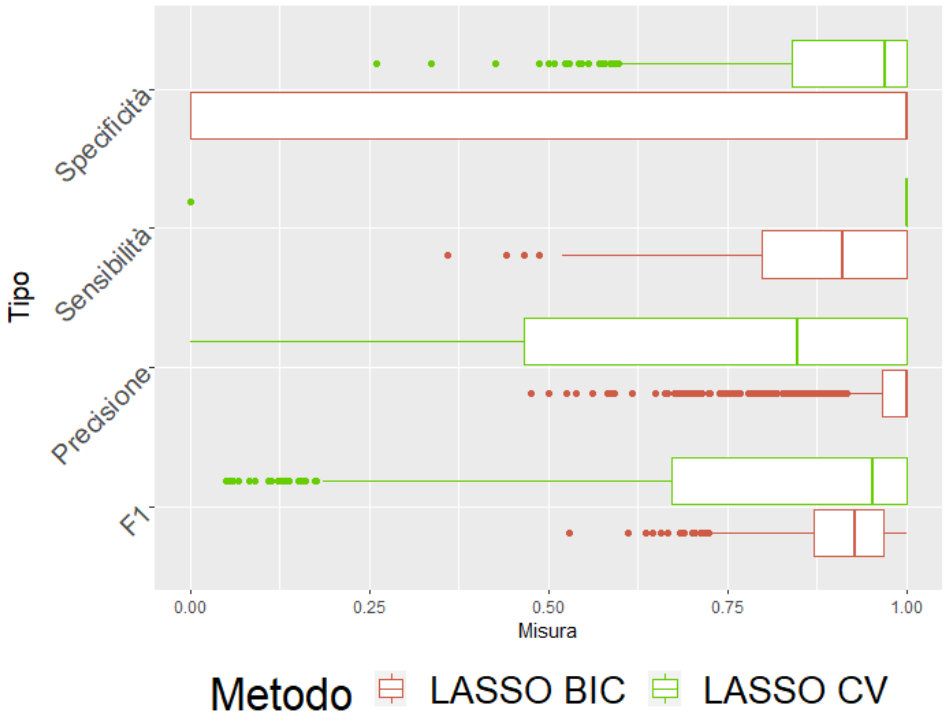


Figura 5.1: Boxplot di precisione, *recall*, *F1 score*, sensibilità e specificità per i due metodi LASSO implementati.

Dalla figura 5.1 si ha che il metodo LASSO che sfrutta il BIC risulti essere un buon metodo confrontato con l'altro metodo proposto: possiamo notare che, ad esclusione dei soli valori di specificità, dunque che una coppia farmaco - evento avverso risulti non correlata quando in realtà lo è, il LASSO BIC risulti essere quello che fornisce i valori migliori e con una minor variabilità nelle 100 simulazioni effettuate. Nello specifico della specificità, si può notare dal grafico come il metodo LASSO BIC assegni valori pari o a 1 o a 0: si nota inoltre che il valore mediano è pari a 1. Notiamo infatti come il metodo LASSO che sfrutta la *cross validation* dia un gran numero di valori anomali. In particolare notiamo che il valore della precisione e del *F1 score* del metodo LASSO BIC sono quelli meno variabili e con una valore mediano che risulta comunque essere buono e di poco inferiore a quello del LASSO che sfrutta la *cross-validation*.

Come ulteriore metodo sono stati stimati dieci modelli di regressione logistica, a cui è stata applicata una selezione delle variabili *stepwise* per mantenere nel modello solo le variabili relative a farmaci significativi rispetto

ad uno specifico evento avverso. I valori delle statistiche utilizzate anche in precedenza vengono riportati in tabella 5.3. Come si può notare i valori in alcune circostanze sembrano indicare una buona identificazione dei farmaci da parte della regressione logistica. Se si va però a vedere il numero di segnali identifcati e quali segnali vengono identificati (riportati in tabella 5.4) si può notare come questo la semplicità di questo metodo non sia adatta a questo tipo di dati.

Evento avverso	Precisione	Specificità	Sensibilità	*F1 score*
1	1.00	1.00	0.70	0.82
2	1.00	1.00	0.55	0.71
3	1.00	0.00	0.79	0.88
4	1.00	1.00	0.51	0.68
5	0.56	0.00	0.44	0.49
6	1.00	1.00	0.71	0.83
7	1.00	1.00	0.65	0.79
8	1.00	1.00	0.85	0.92
9	1.00	1.00	0.68	0.81
10	0.48	0.00	0.30	0.37

Tabella 5.3: Valori di *F1 score* e specificità per i due modelli LASSO stimati.

Modello	# di segnali attesi	# di segnali identificati	# di segnali correttamente identificati
Modello per EA 1	1	3	1
Modello per EA 2	1	5	1
Modello per EA 3	1	2	0
Modello per EA 4	1	5	1
Modello per EA 5	1	7	0
Modello per EA 6	1	3	1
Modello per EA 7	1	4	1
Modello per EA 8	1	3	1
Modello per EA 9	1	4	1
Modello per EA 10	1	7	1

Tabella 5.4: Numero di segnali attesi, identificati e correttamente identificati per i quattro modelli stimati.

In conclusione possiamo dire che dei due metodi testati quello che sembra avere una miglior *performance*, vautando tutti e quattro gli indici presi in considerazione, è il modello LASSO che sfrutta il BIC per la selezione delle variabili.

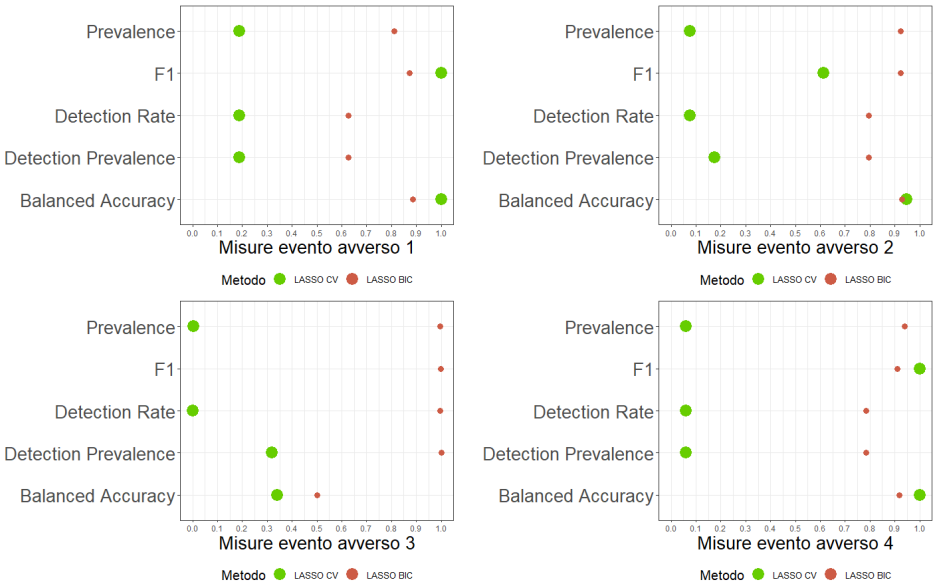


Figura 5.2: Confronto di prevalenza, *F1 score*, *detection rate*, *detection prevalence* e accuratezza bilanciata per i metodi LASSO CV e LASSO BIC.

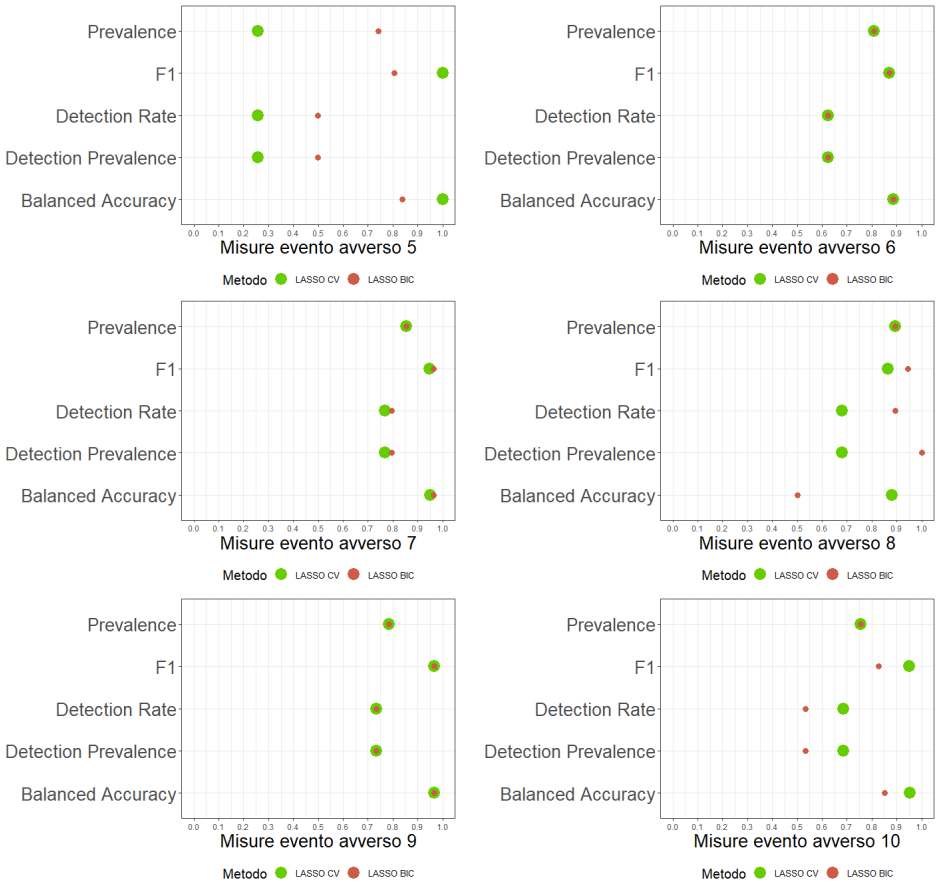


Figura 5.3: Confronto di prevalenza, *F1 score*, *detection rate*, *detection prevalence* e accuratezza bilanciata per i metodi LASSO CV e LASSO BIC.

Un ulteriore confronto tra i metodi LASSO BIC e CV si ha nei grafici riportati nelle figure 5.2 e 5.3 in cui possiamo vedere i valori di cinque misure fornite dalle tabelle di contingenza ottenute dai modelli citati in precedenza. Nello specifico si confrontano la prevalenza (ottenuta come $\frac{A+C}{A+B+C+D}$), l'*F1 score*, il *detection rate* (ottenuta come $\frac{A}{A+B+C+D}$), la *detection prevalence* (ottenuta come $\frac{A+B}{A+B+C+D}$) e la accuratezza bilanciata (ottenuta come $\frac{A/(A+C)+D/(B+D)}{2}$, ovvero sensibilità più specificità diviso 2). Considerando tutti gli eventi avversi presenti nel *dataset*, non si notano grandi differenze tra i due metodi ad esclusione di alcuni casi in cui il LASSO BIC sembra adattarsi meglio.

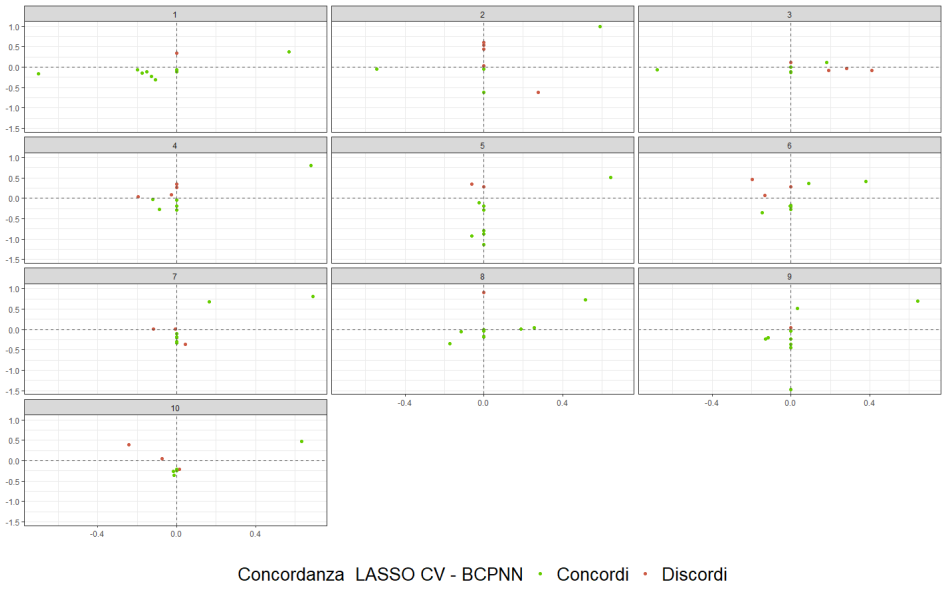


Figura 5.4: Concordanza tra i risultati di associazione farmaco - evento avverso utilizzando i metodi LASSO CV e BCPNN. Sull'asse delle ascisse abbiamo i valori relativi al parametro stimato del modello LASSO CV per i vari farmaci, mentre sull'asse delle ordinate i valori relativi all'indice BCPNN.

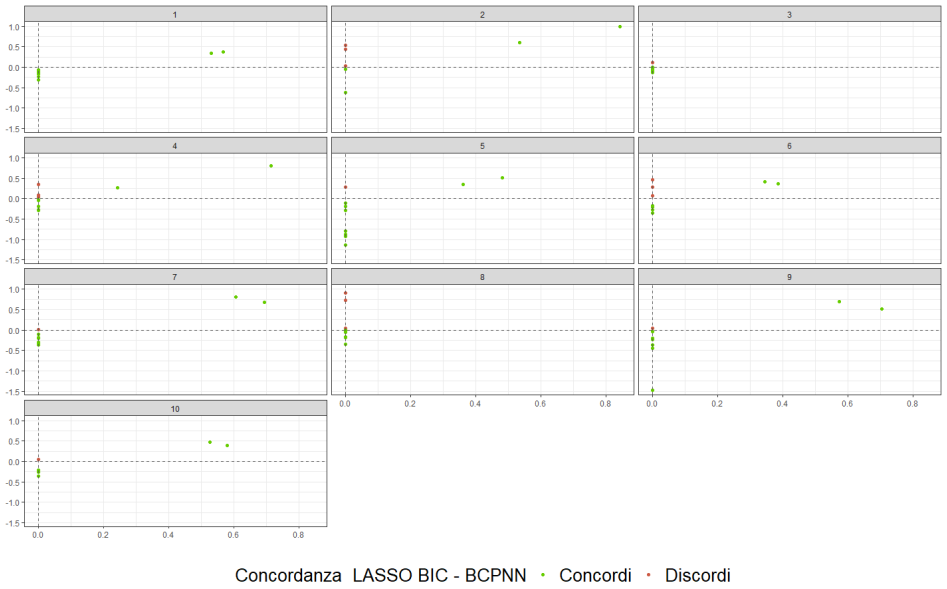


Figura 5.5: Concordanza tra i risultati di associazione farmaco - evento avverso utilizzando i metodi LASSO BIC e BCPNN. Sull'asse delle ascisse abbiamo i valori relativi al parametro stimato del modello LASSO BIC per i vari farmaci, mentre sull'asse delle ordinate i valori relativi all'indice BCPNN.

Per valutare poi se i metodi implementati fossero concordi o meno nelle conclusioni, si può guardare ai grafici delle figure 5.4, 5.5 e 5.6.
In particolare, nel grafico di figura 5.4 possiamo vedere il confronto tra uno dei metodi di disproporzionalità, il BCPNN, e un metodo di *data mining* come il LASSO che sfrutta anche la *cross validation*. Possiamo notare come in questo caso, quando una coppia farmaco - evento avverso risulti essere associata, i due metodi sono tra di loro concordi, mentre in alcuni casi il metodo LASSO CV tende ad individuare relazioni farmaco - evento avverso che il metodo BCPNN non individua. In generale i due metodi non sembrano essere in totale accordo come testimoniato dai punti rossi presenti nel grafico.
In figura 5.5, possiamo valutare l'accordo tra il metodo BCPNN e il LASSO BIC: in particolare possiamo notare come i due metodi siano sempre concordi nell'identificare un farmaco come associato ad un evento avverso quando l'associazione è molto evidente. Ci sono però alcuni casi di discordanza, in cui il LASSO BIC individua dei farmaci come associati, mentre il BCPNN no. Per quanto riguarda i farmaci non associati invece i due metodi risultano sempre essere concordi tra di loro. Rispetto al caso precedente la concordanza tra i due metodi sembra essere molto più forte.

Per quanto riguarda il confronto tra due misure di disproporzionalità, una frequentista e una bayesiana (rispettivamente ROR e BCPNN), possiamo notare dai grafici di figura 5.6, come i due metodi siano sempre concordi, tranne in un solo caso (evento avverso 8).

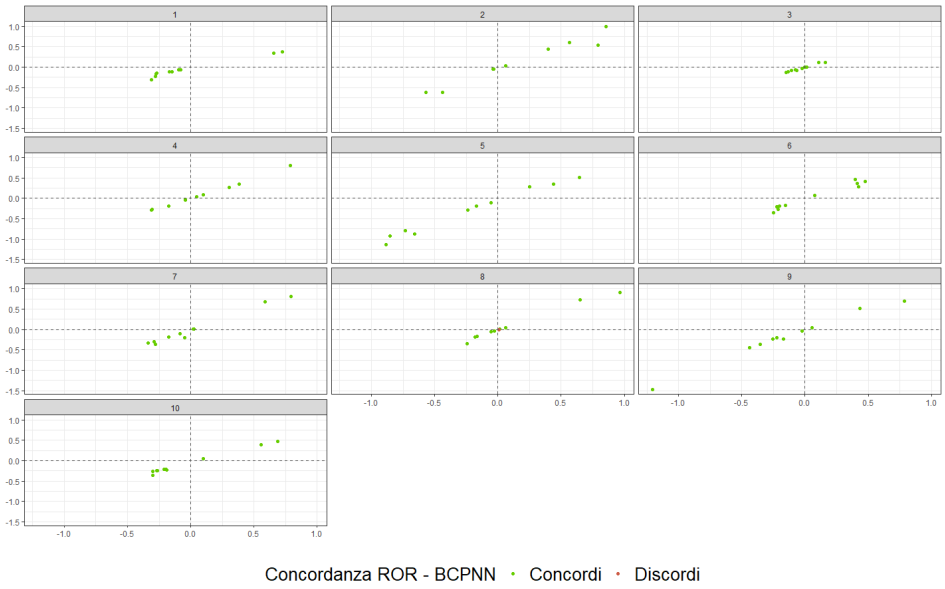


Figura 5.6: Concordanza tra i risultati di associazione farmaco - evento avverso utilizzando i metodi ROR e BCPNN. Sull'asse delle ascisse abbiamo i valori relativi all'indice ROR, mentre sull'asse delle ordinate i valori relativi all'indice BCPNN.

Capitolo 6

Studio di simulazione con interazioni

Nel mondo reale può capitare che un evento avverso si verifichi per l'interazione di più farmaci prescritti insieme per curare una stessa tipologia di malattia o anche per differenti cause. Per questo motivo risulta molto utile e interessante studiare come si comportano questi metodi nell'identificazione di eventi avversi in queste situazioni. Come prima cosa, come nel caso precedente dei singoli farmaci, sono stati valutati i metodi in un contesto di simulazione per capire la loro capacità di predire le associazioni tra più farmaci con un singolo evento avverso.

6.1 Scenario con numero ristretto di farmaci

É stato creato un dataset di segnalazioni spontanee con la presenza di interazione tra più farmaci sfruttando l'algoritmo proposto da Dijkstra et al. (2020) e modificandone l'output in modo tale da ottenere l'interazione tra due farmaci. In particolare, il dataset ottenuto contiene un totale di 901689 segnalazioni spontanee simulate, con 8 farmaci e 4 eventi avversi. Si ha dunque che degli 8 farmaci totali presenti nel dataset, 5 di questi sono associati in varie combinazioni con gli eventi avversi, mentre solo 3 farmaci non hanno nessuna associazione con eventi avversi, come si nota dal grafico presentato in figura 6.1. In particolare si ha che l'evento avverso 1 ha un'associazione con i farmaci 1 e 5 e con l'interazione tra di essi; l'evento avverso 2 ha un'associazione con il farmaco 2 e 5 e con l'interazione di essi;

l'evento avverso 3 è associato al farmaco 3 e 5 e all'interazione dei due; infine l'evento avverso 4 è associato con i farmaci 4 e 5 e l'interazione di essi. Si ha quindi che i farmaci 6, 7 e 8 non hanno un'associazione verificata con nessuno degli eventi avversi presenti nel *dataset*.

Sul *dataset* così ottenuto sono stati stimati quattro modelli LASSO BIC, così come erano stati utilizzati anche per la simulazione senza interazione. L'unica differenza con il procedimento precedente si trova nella matrice del disegno che è stata utilizzata: in questo caso abbiamo infatti che sono presenti anche le interazioni delle covariate relative ai farmaci coinvolti. Oltre al modello LASSO BIC, anche in questo caso è stato stimato un modello con penalizzazione LASSO che sfrutta la *cross-validation*, e come nel caso precedente, il numero di *fold* è pari a 10.

I risultati ottenuti dai due modelli sono presentati nel grafico di figura 6.2.

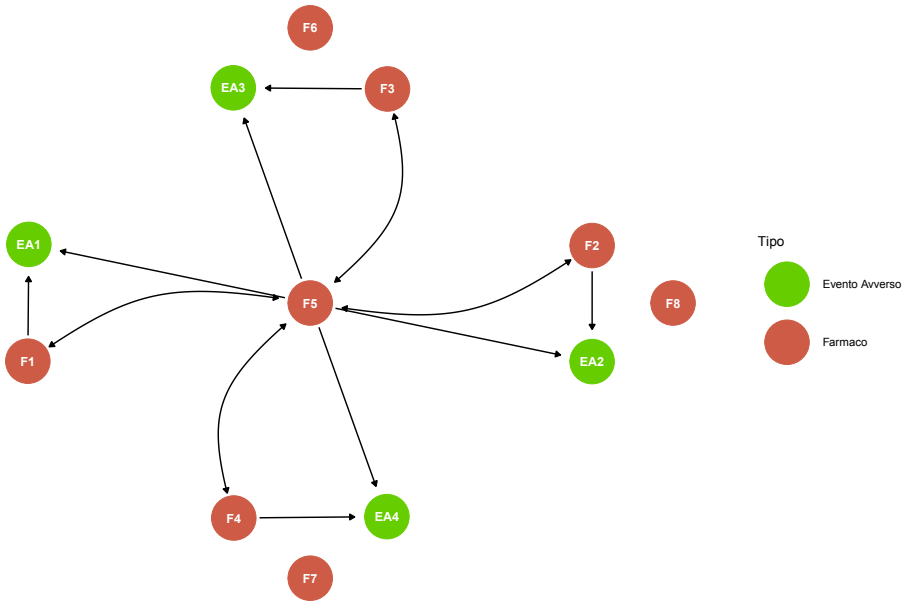


Figura 6.1: Rappresentazione delle interazioni tra farmaci e collegamenti con il relativo evento avverso all'interno del *dataset*.

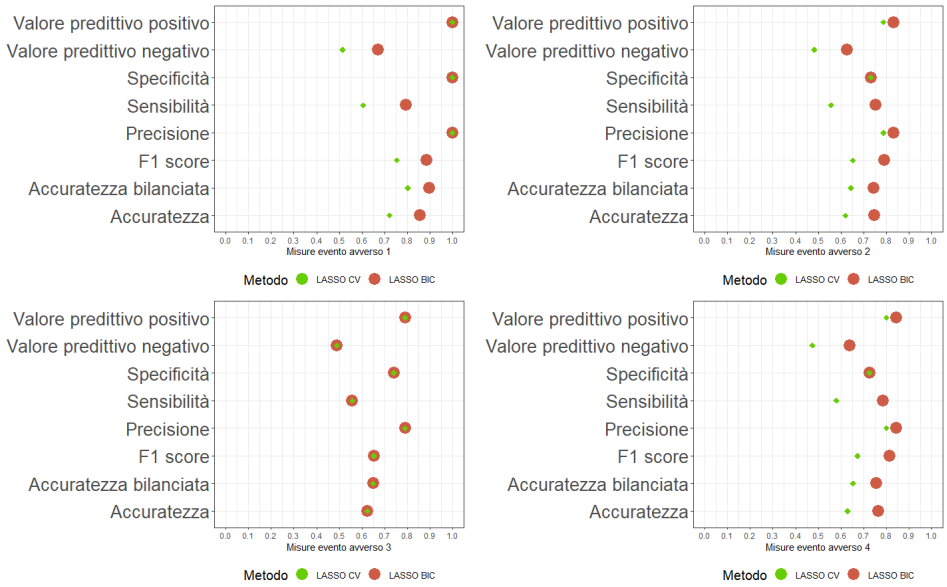


Figura 6.2: Confronto di valore predittivo positivo e negativo, specificità, sensibilità, precisione, *F1 score*, accuratezza bilanciata e accuratezza per i metodi LASSO CV e LASSO BIC.

Dal grafico di figura 6.2 è evidente come, in tutti e quattro i modelli stimati (per i quattro eventi aversi considerati), il metodo LASSO BIC abbia risultati sempre migliori del LASSO CV, tranne nel caso del modello stimato per il terzo evento avverso, in cui i risultati sono uguali. Andando a valutare nello specifico ciascun modello, quello che risulta avere i risultati migliori è quello stimato per il primo evento avverso.

Modello	# di segnali attesi	# di segnali identificati	# di segnali correttamente identificati
Modello per EA 1 LASSO BIC	3	4	3
Modello per EA 2 LASSO BIC	3	5	2
Modello per EA 3 LASSO BIC	3	4	2
Modello per EA 4 LASSO BIC	3	3	2
Modello per EA 1 LASSO CV	3	5	3
Modello per EA 2 LASSO CV	3	4	2
Modello per EA 3 LASSO CV	3	4	2
Modello per EA 4 LASSO CV	3	4	2

Tabella 6.1: Numero di segnali attesi, identificati e correttamente identificati per i quattro modelli stimati.

Nella tabella 6.1 possiamo vedere il numero di segnali identificati per i quattro eventi avversi dai due metodi implementati per i dati simulati con interazione. Quanto emerge da questa tabella è che tutti gli otto modelli

ottenuti sono in linea l'uno con l'altro a livello di risultati e segnali selezionati. Nello specifico si ha che:

- modello per EA 1 LASSO BIC: vengono individuati 4 segnali complessivamente di cui tre sono quelli attesi (i due effetti singoli e l'interazione tra di essi) e un segnale tra evento avverso 1 e farmaco 6 che viene identificato erroneamente;

- modello per EA 2 LASSO BIC: in questo caso gli eventi individuati sono 5 e il metodo risulta essere meno preciso del caso precedente in quanto quelli veritieri identificati sono solamente 2 (l'effetto singolo di un farmaco e l'interazione tra i due farmaci). Viene inoltre colta un'associazione non vera tra evento avverso 2 e il farmaco 6, ed inoltre sono colte due associazioni di interazioni protettive per cui si ha che la combinazione tra farmaco 1 e 5 e tra 3 e 5 porta ad una protezione contro l'evento avverso 2;

- modello per EA 3 LASSO BIC: in questo caso viene colta la relazione tra farmaco 3 e l'evento avverso 3 e l'interazione tra il farmaco 3 e 5. Vengono inoltre individuati erroneamente due segnali che coinvolgono i farmaci 6 e 7;

- modello per EA 4 LASSO BIC: il numero di segnali che si ottengono in questo caso è pari a 3 di cui due di questi sono veri (quello tra l'effetto singolo del farmaco 4 e l'interazione tra farmaco 4 e 5). Il segnale errato che viene identificato è quello per il farmaco 8;

- modello per EA 1 LASSO CV: anche per il modello stimato utilizzando il LASSO CV otteniamo l'identificazione corretta per i due effetti singoli dei farmaci e anche per l'interazione tra di essi. Vengono però identificati anche altri due segnali non veri tra i farmaci 6 e 7 con l'evento avverso di interersse;

- modello per EA 2 LASSO CV: l'effetto principale di uno dei due farmaci viene identificato e anche l'interazione viene identificata. Vengono poi individuati i segnali relativi ai farmaci 6 e 7, anche se il coefficiente LASSO stimato per il farmaco 7 risulta essere prossimo allo zero;

- modello per EA 3 LASSO CV: come per il caso precedente vengono identificati gli effetti di un farmaco e dell'interazione, e vengono anche individuati gli effetti singoli dei farmaci 6 e 7;
- modello per EA 4 LASSO CV: dei quattro segnali identificati in questo caso si ha che due sono corretti (effetto principale di un farmaco ed effetto di interazione) mentre due no (gli effetti principali dei farmaci 6 e 8).

In conclusione, possiamo dire che le interazioni tra farmaci vengono identificate da entrambi i modelli in tutti i casi esaminati che porta a dire che il metodo LASSO BIC sia una valida alternativa per la ricerca di segnali di interazione all'interno di *dataset* di segnalazioni spontaee nel contesto di farmacovigilanza.

6.2 Scenario con numero elevato di farmaci

Per approfondire ulteriormente le capacità del metodo LASSO BIC di cogliere associazioni tra farmaci ed eventi avversi, è stato ottenuto un ulteriore *dataset* in cui il numero di farmaci presenti è stato fissato pari a 100, mantenendo sempre pari a 4 il numero di eventi avversi. In questo caso, il numero di farmaci collegati ad ogni singolo evento avverso è pari a tre, e sono presenti le interazioni tra un farmaco e gli altri due:

- **evento avverso 1**: associato ai farmaci 1, 51 e 71 e all'interazione tra il farmaco 1 e gli altri due;
- **evento avverso 2**: associato ai farmaci 2, 56 e 64 e all'interazione tra il farmaco 2 e gli altri due;
- **evento avverso 3**: associato ai farmaci 3, 14 e 77 e all'interazione tra il farmaco 1 e gli altri due;
- **evento avverso 4**: associato ai farmaci 4, 12 e 70 e all'interazione tra il farmaco 2 e gli altri due.

Essendo aumentato notevolmente il numero di farmaci rispetto al caso precedente, e con esso anche la dimensione della matrice del disegno necessaria per stimare il modello, in questo caso si è deciso di implementare il solo metodo LASSO BIC.

Nel grafico di figura 6.3 vengono rappresentati alcuni indici per valutare la capacità del modello di identificare non solo l'effetto del farmaco singolarmente, ma anche quello dell'interazione tra due di essi. Gli stessi valori sono poi riportati anche nelle tabelle 6.2.

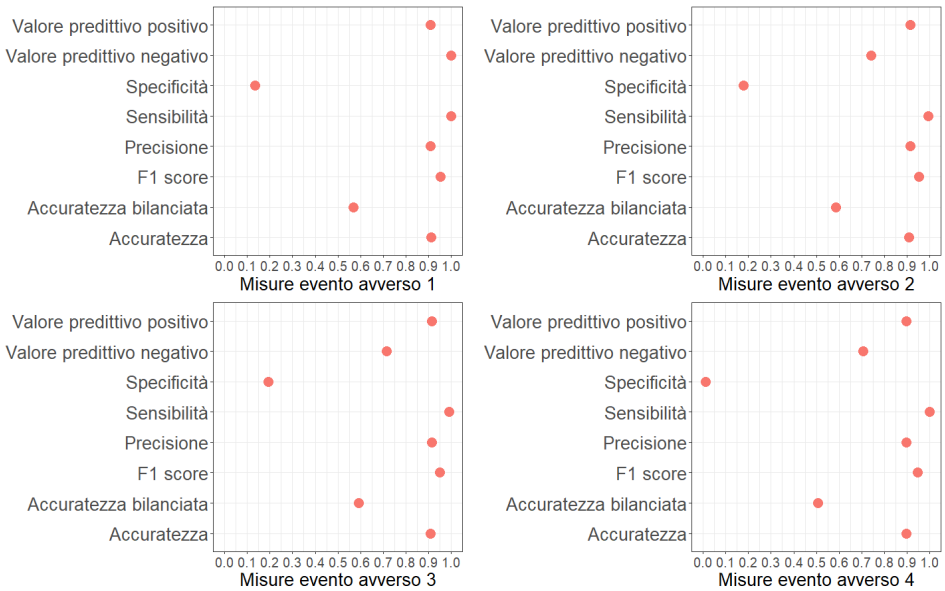


Figura 6.3: Valore predittivo positivo e negativo, specificità, sensibilità, precisione, *F1 score*, accuratezza bilanciata e accuratezza per il metodo LASSO BIC.

Modello	Accuratezza	LI CI accuratezza	LS CI accuratezza
Modello per EA 1	0.9108	0.9102	0.9115
Modello per EA 2	0.9089	0.9082	0.9095
Modello per EA 3	0.9088	0.9081	0.9094
Modello per EA 4	0.8976	0.8969	0.8983

Modello	Sensibilità	Specificità	Valore predittivo positivo
Modello per EA 1	1.0000	0.1359	0.9095
Modello per EA 2	0.9928	0.1795	0.9131
Modello per EA 3	0.9911	0.1933	0.9143
Modello per EA 4	0.9994	0.0129	0.8979

Modello	Valore predittivo negativo	Precisione	*Recall*
Modello per EA 1	1.0000	0.9095	1.0000
Modello per EA 2	0.7416	0.9131	0.9928
Modello per EA 3	0.7153	0.9143	0.9911
Modello per EA 4	0.7076	0.8979	0.9994

Modello	*F1 score*	Accuratezza bilanciata
Modello per EA 1	0.9526	0.5680
Modello per EA 2	0.9513	0.5862
Modello per EA 3	0.9512	0.5922
Modello per EA 4	0.9459	0.5062

Tabella 6.2: Misure per valutare la capacità del modello di identificare i farmaci e le interazioni associate ad un evento avverso.

Come si può notare dai valori presenti sul grafico di figura 6.3, i valori delle statistiche considerate sono tutti molto buoni ad esclusione dei soli valori della specificità. Nello specifico si può notare come i valori di accuratezza (con relativo intervallo di confidenza), prima tabella 6.2, ottenuta come $\frac{A+D}{A+B+C+D}$ (dove A, B, C e D fanno riferimento alle celle di una tabella $2x2$), siano per tutti e quattro i modelli stimati superiori al valore di 0.90 indicando quindi una buona capacità del modello di identificare i veri positivi e i veri negativi, che nel contesto di farmacovigilanza, sono rispettivamente i farmaci realmente associati con un evento avverso e i farmaci non associati. Possiamo notare dei valori costanti ed elevati anche nei valori della precisione (ottenuta come $\frac{A}{A+B}$). Andando a valutare l'accuratezza bilanciata, va invece considerato il forte impatto che ha la specificità, molto bassa in questo caso, nel calcolo del valore; complessivamente i valori si affermano sopra lo 0.50 per tutti e

quattro i casi. Tutti questi valori vanno pesati con i valori di tabella 6.3: infatti i valori sopra riportati valutano la capacità del modello di classicifare il singolo report del *dataset*, ma l'interesse in farmacovigilanza riguarda le singole coppie farmaco - evento avverso.

Soffermandosi sui valori della specificità e della sensibilità, si può notare come i modelli stimati in tutti e quattro i casi siano molto precisi nell'individuare le coppie che non sono realmente associate: questo può aiutare nel contesto di farmacovigilanza poiché avere una buona precisione nell'escludere a priori i farmaci non associati facilita il passo successivo ovvero quello che implica ulteriori studi sui farmaci identificati dal modello. Al contrario il modello pecca sull'identificazione delle coppie realmente associate, non cogliendole tutte alla perfezione, anche se per i risultati ottenuti va considerato l'elevato numero di segnalazioni presenti nel *dataset*.

Per questo motivo nella tabella 6.3 sono riportati il numero di segnali che ci si attendeva venissero identificati, il numero di segnali identificati e il numero di quelli correttamente identificati.

Modello	# di segnali attesi	# di segnali identificati	# di segnali correttamente identificati
Modello per EA 1	5	4	4
Modello per EA 2	5	4	3
Modello per EA 3	5	6	3
Modello per EA 4	5	6	3

Tabella 6.3: Numero di segnali attesi, identificati e correttamente identificati per i quattro modelli stimati.

Si può notare come in nessuno dei quattro casi il metodo LASSO BIC identifichi correttamente tutti i segnali: in particolare il segnale che non viene identificato in tutti e quattro i casi è quello degli effetti singoli dei farmaci in interazione. Questo significa che il modello è in grado di identificare correttamente l'interazione tra due farmaci, e il fatto che non vengano identificati gli effetti singoli dei farmaci non è particolarmente problematico, dato che non è scontato che un farmaco provochi un evento avverso anche quando assunto da solo se lo provoca quando assunto in concomitanza con un altro.

Dunque, in precedenza si è visto che questo metodo risulta essere valido in un contesto in cui si hanno pochi farmaci, ma possiamo estendere le stesse considerazioni fatte in precedenza anche a questo scenario di simulazione: si è infatti visto che anche con un numero dieci volte superiore di farmaci e con più interazioni da dover identificare, il metodo LASSO BIC risulta comunque

essere valido e un buon metodo per l'identificazione di questa tipologia di segnali.

Come metodo di confronto con il LASSO BIC, è stato utilizzato in questo caso il GROUPED LASSO, un metodo LASSO ideato appositamente per cogliere interazioni tra variabili sia continue che categoriali che combinazioni di esse. Di seguito in figura 6.4 vengono riportati i valori della funzione di perdita in riferimento al valore assegnato al parametro λ: in totale per ciascuno dei quattro modelli sono stati implementati modelli con 15 valori diversi del parametro λ.

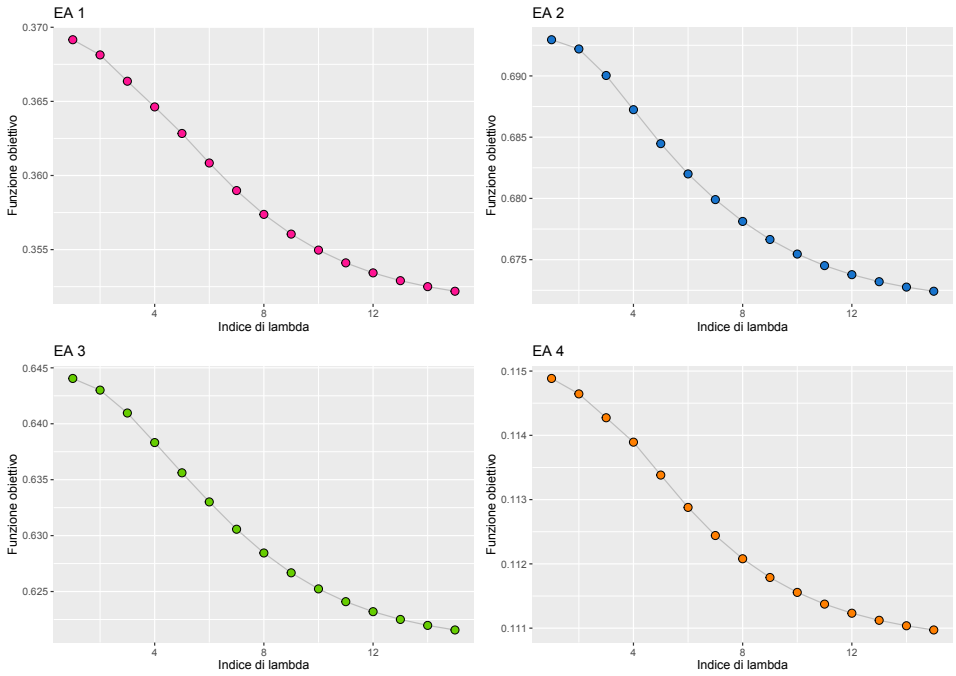


Figura 6.4: Valori di λ rispetto al valore della funzione obiettivo per i quattro modelli stimati.

Sulla base di questo grafico, e dunque del valore della funzione obiettivo, in combinazione con un criterio di parsimonia, sono stati scelte i farmaci e le interazioni tra farmaci per effettuare un confronto con il LASSO BIC precedente. Quanto è emerso da questo confronto è che il GROUPED LASSO in nessuno dei quattro casi mantiene gli effetti principali dei farmaci all'interno della regressione, e solo in alcuni casi identificando le corrette interazioni tra farmaci. Tutto questo viene comunque fatto non selezionando un basso numero di farmaci, ma al contrario l'insieme di farmaci selezionati è molto

alto.

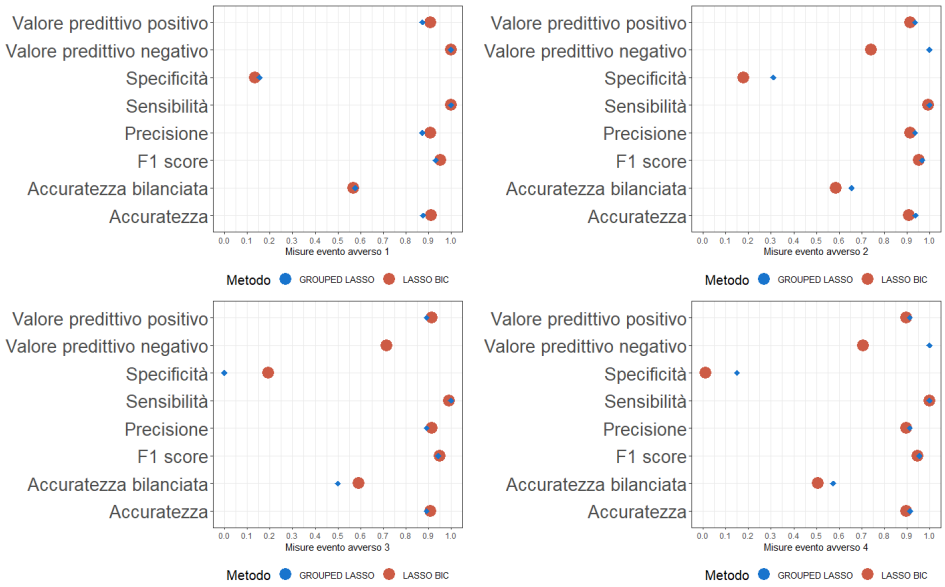


Figura 6.5: Confronto di valore predittivo positivo e negativo, specificità, sensibilità, precisione, *F1 score*, accuratezza bilanciata e accuratezza per i metodi GROUPED LASSO e LASSO BIC.

Dai grafici di figura 6.5 possiamo notare come in alcuni casi i valori delle statistiche scelte per il GROUPED LASSO siano anche superiori a quelli del LASSO BIC: non bisogna però farsi trarre in inganno da questa cosa poiché i valori del LASSO BIC nei quattro casi risultano essere mediamente alte per tutte le statistiche in gioco a differenza del GROUPED LASSO. Inoltre andando a valutare il numero di segnali identificati dai due metodi si potrà notare come il LASSO BIC sia molto più parsimonioso nella scelta di un farmaco o di un'interazione mentre il GROUPED LASSO tende a selezionare molte più variabili e quindi complessivamente risulta essere meno utile, in quanto in farmacovigilanza quello che si cerca è un metodo che restringa le coppie farmaco - evento avverso che gli esperti dovranno poi andare ad ingagare. In tabella 6.4 possiamo vedere il numero di segnali trovati, identificati correttamente e di quelli attesi.

Modello	# di segnali attesi	# di segnali identificati	# di segnali correttamente identificati
Modello per EA 1	5	5	1
Modello per EA 2	5	33	2
Modello per EA 3	5	32	0
Modello per EA 4	5	30	1

Tabella 6.4: Numero di segnali attesi, identificati e correttamente identificati per i quattro modelli stimati.

Da notare come il modello per il terzo evento avverso non sia stato in grado di identificare nessuno dei segnali che erano attesi. I modelli stimati per il primo, secondo e quarto evento avverso, invece sono stati in grado di identificare rispettivamente 1, 2 e 1 segnali, e in entrambi i casi sono delle interazioni tra farmaci e in generale, come detto in precedenza nessun metodo ha identificato segnali relativi ai farmaci presi singolarmente. Dunque i risultati ottenuti a livello di statistiche e mostrati precedentemente in figura 6.5 vanno pesati con queste informazioni, in quanto rendono questo specifico metodo poco utilizzabile in contesto di farmacovigilanza.

Capitolo 7

Applicazioni a dati reali

Dopo aver verificato tramite lo studio di simulazione che il metodo LASSO BIC può competere con i metodi di disproporzionalità attualmente utilizzati in farmacovigilanza, nel seguito verrà applicato ai dati FAERS, in particolare sarà applicato ai dati relativi agli ultimi due trimestri dell'anno 2019, per la verifica delle potenzialità del metodo su dati con la presenza di sole coppie farmaco - evento avverso, mentre per la verifica del metodo rispetto all'identificazione di DDI (*Drug - Drug - Interaction*) verrà utilizzato il *dataset TWOSIDES*, presentato nella sezione 3.3.

Vista la mancanza di un *gold standard* per dati di questo tipo come confronto del metodo LASSO BIC, per capire se questo metodo possa competere con quelli già utilizzati per l'identificazione di segnali di farmacovigilanza, si utilizzerà il metodo di disproporzionalità bayesiano BCPNN.

Essendo che nel FAERS ci sono decine di migliaia di coppie farmaco - evento avverso (e, analogamente, in *TWOSIDES* ci sono migliaia di DDI), è stato selezionato un numero ristretto di eventi avversi, mantenendo però invariato il numero originale di farmaci. Nello specifico, gli eventi avversi mantenuti per la stima dei modelli LASSO BIC per l'identificazione delle coppie farmaco - evento avverso sono quattro in totale: **malattia renale cronica**, **danno renale acuto**, **danno renale** e **bronchite**, per un totale di 158883 segnalazioni spontanee presenti nel *dataset*. Gli eventi avversi scelti invece per stimare il modello con le interazioni tra farmaci sono **anemia** e **infarto miocardico**, per un totale di 23574 segnalazioni spontanee.

Nelle tabelle 7.1 e 7.2, vengono riportati il numero di segnalazioni spontanee presenti nei tre *dataset* per ciascuno degli eventi avversi selzionati.

Malattia renale cronica	Danno renale acuto	Danno renale	Bronchite
67501	52393	25398	13591

Tabella 7.1: Frequenza assoluta degli eventi avversi selezionati per la seconda analisi sui dati del quarto trimestre del 2019.

Anemia	Infarto miocardico
10125	13449

Tabella 7.2: Frequenza assoluta degli eventi avversi selezionati per l'analisi sui dati *TWOSIDES* ristretti.

7.1 Risultati

Per capire le potenzialità del metodo LASSO BIC è stato utilizzato come metodo di confronto il BCPNN: i risultati di questa comparazione vengono riportati in figura 7.1. Questi valori sono calcolati assegnando a tutte le segnalazioni presenti nel *dataset* che contengono la coppia farmaco - evento avverso sotto analisi un valore logico positivo, in modo da poter ottenere le matrici di confusione e successivamente i valori sopra presentati.

In particolare, si ha che per il primo evento avverso, "malattia renale cronica", questo è legato a 12 segnali, di cui 11 sono identificati anche dal BCPNN. Il secondo evento avverso, "danno renale acuto", è legato a 35 segnali, di cui 30 sono identificati anche dal BCPNN. Il terzo evento avverso, "danno renale", è legato a 15 segnali, di cui 13 sono identificati anche dal BCPNN. L'ultimo evento avverso, "bronchite", è legato a 46 segnali, di cui 43 sono identificati anche dal BCPNN.

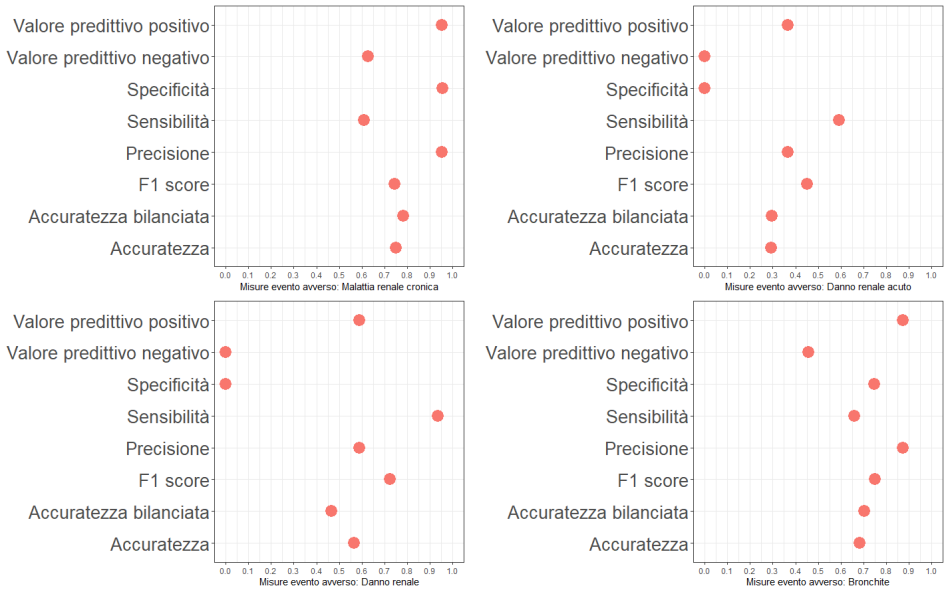


Figura 7.1: Valore predittivo positivo e negativo, specificità, sensibilità, precisione, *F1 score*, accuratezza bilanciata e accuratezza per il metodo LASSO BIC rispetto al metodo BCPNN applicato ai dati FAERS del quarto trimestre 2019, secondo caso.

Per quanto riguarda la parte di applicazione del metodo LASSO BIC al *dataset TWOSIDES*, per l'identificazione delle interazioni l'insieme di dati ottenuti ha una dimensione di 23574 segnalazioni. Su questo *dataset* sono stati stimati i due modelli per i due eventi avversi citati in precedenza. Quello che si è ottenuto conferma la capacità del metodo di individuare le interazioni tra farmaci rispetto ad un evento avverso, però a differenza di quanto visto nelle simulazioni non è in grado di identificarle con la stessa precisione.

Infatti, per l'evento avverso "anemia" il numero di coppie di farmaci ritenute associate all'evento avverso (associazione stabilita tramite BCPNN) identificate dal LASSO è pari a 57 (su un totale di 96 segnali trovati dalla stima), ma quelle realmente associate sono pari a 59. Il metodo dunque permette di identicare delle coppie, ma non tutte quelle che venivano identificate con il BCPNN. Ovviamente non avendo un *gold standard* che indichi le coppie associate ed effettivamente collegate ad un evento avverso risulta impossibile stabilire la capacità di questo metodo di identificare i segnali.

Nel caso del secondo evento avverso, ovvero l'"infarto miocardico", il numero di coppie di farmaci identificate è pari a 4 (su un totale di 4 segnali trovati dalla stima) contro 21 risultate associate secondo il BCPNN. Anche in questo caso vale quanto detto in precedenza.

Capitolo 8

Conclusioni

Quanto è emerso dagli studi effettuati su questi nuovi metodi sviluppati appositamente per la farmacovigilanza hanno evidenziato sicuramente dei vantaggi che in futuro potrebbero sostituire i classici metodi di disproporzionalità, anche virtù della crescente mole di dati che ogni anno viene raccolta dalle agenzie del farmaco.

Sono stati implementati diversi metodi, alcuni già conosciuti e attualmente utilizzati in farmacovigilanza, come i metodi di disproporzionalità (sia frequentisti che bayesiani), che metodi innovativi come la regressione LASSO con selezione delle variabili effettuata tramite criterio BIC, e anche un metodo più comune come la regressione logistica. I metodi di disproporzionalità sono stati utilizzati prevalentemente come confronto per i metodi LASSO, poiché come visto non esistono *gold standard* in questo contesto e si è quindi utilizzata l'informazione ricavata da questi metodi per validare quelli nuovi.

Nei tre scenari di simulazione si è visto come il metodo LASSO BIC sia una valida alternativa ai metodi di disproporzionalità identificando in quasi tutti i casi le associazioni farmaco - evento avverso, anche se in contesto ridotto rispetto a quello che si troverebbe nell'analisi di *database* di farmacovigilanza come il FAERS. Si è notato inoltre una buona capacità anche nell'identificazione di interazioni tra farmaci. In quasi tutti i casi analizzati, sia in un contesto di un basso numero di farmaci, che in un contesto con un numero ben più elevato, le coppie sono sempre state identificate a dovere, anche se in alcuni casi gli effetti singoli dei farmaci non sono stati selezionati come significativi. Questo comunque in farmacovigilanza risulta essere un problema minore poiché può capitare che i due farmaci non siano associati

ad un evento avverso, ma in combinazione tra di loro sviluppino l'evento.

I risultati negli scenari simulati che sembravano poter indicare questo metodo come una buona alternativa ai metodi correntemente utilizzati, hanno trovato un parziale accordo nell'applicazione a dati reali. Quanto si è visto applicando il metodo LASSO BIC ai dati FAERS relativi agli ultimi due trimestri del 2019 è che in quasi tutti i casi il nuovo metodo identifica gli stessi segnali identificati anche dal BCPNN. Il metodo LASSO BIC identifica normalmente più segnali di quelli identificati dal BCPNN, ma non è possibile affermare se questo sia positivo o meno non potendo confrontare questi risultati le reali associazioni tra farmaci ed eventi avversi. Per motivi computazionali, l'applicazione del metodo è stata svolta su un numero ridotto di segnalazioni spontanee relative a soli quattro eventi avversi. Va notato comunque che non tutti i *database* di farmacovigilanza mondiali dispongono di moli di dati elevate come il FAERS, che può essere un lato positivo di questi dati unitamente alla loro accessibilità (i dati FAERS sono *open source*), ma allo stesso tempo è una limitazione. Possibili studi futuri possono incentrarsi sull'applicazioni di questi metodi ai *database* mantenuti dalle agenzie del farmaco nazionali che contano un numero di dati inferiore rispetto a questi.

Quanto visto per l'applicazione con lo scopo di identificazione delle coppie farmaco - evento avverso può essere esteso anche all'applicazione al *database TWOSIDES* per l'identificazione di interazioni tra farmaci. Anche in questo caso, il numero di osservazioni per motivi computazionali è stato ridotto. A differenza del caso precedente, nell'applicazione a questi dati sono stati selezionati un elevato numero di segnali (per l'evento avverso "anemia"), e tra questi, 57 sono stati evidenziati come significativi anche con il metodo BCPNN. Nel secondo caso il numero di segnali si è ridotto a 4, ma tutti quelli ritenuti significativi dal LASSO BIC sono stati evidenziati anche dal BCPNN. Complessivamente però questo secondo caso risulta più impreciso poiché il BCPNN ha evidenziato 21 segnali totali.

Concludendo, si può affermare che il metodo LASSO BIC potrà, in futuro, essere utilizzato come strumento per l'identificazione di segnali di farmacovigilanza da confermare poi tramite studi clinici o da pareri di esperti. Una buona soluzione potrà essere quella di utilizzare parallelamente metodi come il BCPNN o altri metodi di disproporzionalità bayesiani, che sembrano essere i migliori, per l'identificazione dei segnali. Queste conclusioni dovranno comunque essere confermate con ulteriori applicazioni su altri *database* gestiti

da altre agenzie.

Bibliografia

Ahmed, I., A. Pariente e P. Tubert-Bitter (2018). «Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions». In: *Statistical Methods in Medical Research* 27.3. PMID: 27114328, pp. 785–797. DOI: 10.1177/0962280216643116. eprint: https://doi.org/10.1177/0962280216643116. URL: https://doi.org/10.1177/0962280216643116.

Bate, A. et al. (1998). «A Bayesian neural network method for adverse drug reaction signal generation». In: *European Journal of Clinical Pharmacology* 54.4, pp. 315–321. ISSN: 00316970. DOI: 10.1007/s002280050466.

Brown, E. (2007). «Medical Dictionary for Regulatory Activities (MedDRA®)». In: *Pharmacovigilance: Second Edition* 20.September 1998, pp. 168–183.

Bryck, A. e S. Raudenbush (1992). «Hierarchical Linear Models». In: *Sage Publications*.

Bühlmann, P. e S. van de Geer (mag. 2011). «Lasso for linear models». In: DOI: 10.1007/978-3-642-20192-9_2.

Courtois, É., P. Tubert-Bitter e I. Ahmed (2021). «New adaptive lasso approaches for variable selection in automated pharmacovigilance signal detection». In:

Delamothe, T. (1992). «Reporting adverse drug reactions». In: *British Medical Journal*, pp. 304–465.

Dijkstra, L.J. et al. (2020). «Adverse Drug Reaction or Innocent Bystander? A Systematic Comparison of Statistical Discovery Methods for Spontaneous Reporting Systems». In: *Pharmacoepidemiology and Drug Safety*. URL: https://DOI:10.1002/PDS.4970.

DPAC (2021). *How the FDA Drug Approval Process Works - DPAC*. URL: https://diabetespac.org/fda-drug-approval-process/ (visitato il 28/09/2021).

DuMouchel, W. (1999). «Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system». In: *American Statistician* 53.3, pp. 177–190. ISSN: 15372731. DOI: 10.1080/00031305.1999.10474456.

DuMouchel, W. e D. Pregibon (2001). «Empirical Bayes screening for multi-item associations». In: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 67–76. DOI: 10.1145/502512.502526.

Evans, S. J.W., P. C. Waller e S. Davis (2001). «Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports». In: *Pharmacoepidemiology and Drug Safety* 10.6, pp. 483–486. ISSN: 10538569. DOI: 10.1002/pds.677.

Fan, J. e R. Li (2001). «Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties». In: *Journal of the American Statistical Association* 96.456, pp. 1348–1360. DOI: 10.1198/016214501753382273. eprint: https://doi.org/10.1198/016214501753382273. URL: https://doi.org/10.1198/016214501753382273.

FDA (2021). *FDA's Drug Review Process ensuring drugs are safe and effective | FDA*. URL: https://www.fda.gov/drugs/information-consumers-and-patients-drugs/fdas-drug-review-process-ensuring-drugs-are-safe-and-effective (visitato il 28/09/2021).

Fram, D. M., J. S. Almenoff e W. DuMouchel (2003). «Empirical Bayesian data mining for discovering patterns in post-marketing drug safety». In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 359–368. DOI: 10.1145/956750.956792.

Hartnell, N. R. e J. P. Wilson (2004). *Replication of the Weber effect using postmarketing adverse event reports voluntarily submitted to the United States Food and Drug Administration.*

Heckerman, D. (1997). «Bayesian Networks for Data Mining». In: 119, pp. 79–119.

Huang, J., S. Ma e C. Zhang (2008a). «ADAPTIVE LASSO FOR SPARSE HIGH-DIMENSIONAL REGRESSION MODELS». In: *Statistica Sinica* 18, pp. 1603–1618.

— (2008b). «The Iterated Lasso for High-Dimensional Logistic Regression». In:

Ibrahim, H. et al. (2021). «Signal Detection in Pharmacovigilance: A Review of Informatics-driven Approaches for the Discovery of Drug-Drug Interaction Signals in Different Data Sources». In: *Artificial Intelligence in the Life Sciences* 1, p. 100005. ISSN: 2667-3185. DOI: https://doi.org/10.1016/j.ailsci.2021.100005. URL: https://www.sciencedirect.com/science/article/pii/S2667318521000052.

Johnson, N. e S. Kotz (1969). «Discrete distribution». In: *Houghton Mifflin*.

Lanera, C., P. Belloni e N. Guidone (2021). *faers.db: FAERS database in R*. R package version 0.0.0.9001. URL: https://github.com/UBESP-DCTV/faers.db.

Lim, Michael e Trevor Hastie (2015). «Learning Interactions via Hierarchical Group-Lasso Regularization». In: *Journal of Computational and Graphical Statistics* 24.3, pp. 627–654. ISSN: 15372715. DOI: 10.1080/10618600.2014.938812.

Meinshausen, N. e P. Bühlmann (2010). «Stability selection». In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4, pp. 417–473. DOI: https://doi.org/10.1111/j.1467-9868.2010.00740.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2010.00740.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00740.x.

Meyboom, R. H. et al. (1997). «Principles of Signal Detection in Pharmacovigilance». In: *Drug Safety* 16.6, pp. 355–365.

Norén, G. N. et al. (2006). «Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events». In: *Statistics in Medicine* 25.21, pp. 3740–3757. ISSN: 02776715. DOI: 10.1002/sim.2473.

O'Hagan, A. (1994). «Kendall's Advanced Thoery of Statistics». In: *Bayesian inference* 2B.

«THE MORGAN KAUFMANN SERIES IN REPRESENTATION AND REASONING» (1988). In: *Probabilistic Reasoning in Intelligent Systems*. A cura di Judea Pearl. San Francisco (CA): Morgan Kaufmann, p. i. ISBN: 978-0-08-051489-5. DOI: https://doi.org/10.1016/B978-0-08-051489-5.50001-1. URL: https://www.sciencedirect.com/science/article/pii/B9780080514895500011.

Pham, M., F. Cheng e K. Ramachandran (2019). «A Comparison Study of Algorithms to Detect Drug–Adverse Event Associations: Frequentist,

Bayesian, and Machine-Learning Approaches». In: *Drug Safety* 42.6, pp. 743–750. ISSN: 1179-1942. DOI: 10.1007/s40264-018-00792-0. URL: https://doi.org/10.1007/s40264-018-00792-0.

Pham, M. H. (2018). «Signal detection of adverse drug reaction using the adverse event reporting system: literature review and novel methods». In: March.

Rothman, K. J., S. Lanes e S. T. Sacks (2004). «The reporting odds ratio and its advantages over the proportional reporting ratio». In: *Pharmacoepidemiology and Drug Safety* 13.8, pp. 519–523. ISSN: 10538569. DOI: 10.1002/pds.1001.

Stephenson, W. P. e M. Hauben (2007). «Data mining for signals in spontaneous reporting databases: proceed with caution». In: *Pharmacoepidemiology and drug safety* 16.4, pp. 359–365. ISSN: 1053-8569.

Suvarna, V. (2010). «Phase IV of Drug Development.» In: *Perspectives in clinical research* 1.2, pp. 57–60. ISSN: 2229-5488. URL: http://www.ncbi.nlm.nih.gov/pubmed/21829783%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3148611.

Tatonetti, N. P. et al. (mar. 2012). «Data-Driven Prediction of Drug Effects and Interactions». In: *Science Translational Medicine* 4.125, 125ra31 LP –125ra31. DOI: 10.1126/scitranslmed.3003377. URL: http://stm.sciencemag.org/content/4/125/125ra31.abstract.

Tibshirani, R. (1996). «Regression Shrinkage and Selection via the Lasso». In: *Journal of the Royal Statistical Society (Series B)* 58, pp. 267–288.

Wallenstein, E. J. e D. Fife (2001). «Temporal patterns of NSAID spontaneous adverse event reports: the Weber effect revisited». In: *Drug safety* 24.3, pp. 233–237. ISSN: 0114-5916. DOI: 10.2165/00002018-200124030-00006. URL: https://doi.org/10.2165/00002018-200124030-00006.

Weber, J. C. P. (1984). «Epidemiology of adverse reactions to nonsteroidal anti-inflammatory drugs.» In: *Side-effects of anti-inflammatory drugs, advances in inflammation research.*, pp. 1–7.

WHO (2021). *Regulation and Prequalification*. URL: https://www.who.int/teams/regulation-prequalification/regulation-and-safety/pharmacovigilance (visitato il 28/09/2021).

Zou, H. (2006). «The Adaptive Lasso and Its Oracle Properties». In: *Journal of the American Statistical Association* 101.476, pp. 1418–1429. DOI:

10.1198/016214506000000735. eprint: https://doi.org/10.1198/016214506000000735. URL: https://doi.org/10.1198/016214506000000735.

Ringraziamenti

Dopo questi cinque anni e mezzo sono arrivato alla conclusione di un percorso che definire fantastico sarebbe riduttivo. Molte persone hanno fatto si che questo periodo della mia vita sia stato uno dei migliori.

In primis, vorrei ringraziare la Prof.ssa Boccuzzo e il Dott. Belloni che mi hanno seguito con grande disponibilità, attenzione e pazienza durante questi mesi di stesura dell'elaborato, che mi ha dato modo di scoprire un ulteriore ambito di applicazione della statistica a me sconosciuto.

In secondo luogo, un sentito ringraziamento va ad Andrea, Danny e Davide con cui ho condiviso l'interezza di questo percorso condividendo quasi ogni giorno della nostra vita accademica in questi anni e con cui abbiamo creato un bellissima amicizia che ci ha portato a momenti di divertimento e svago anche al di fuori dell'Università.
Inoltre è d'obbligo un pensiero anche a tutti i miei amici, da quelli di vecchia data a quelli più recenti, che in questo periodo mi hanno supportato in questo percorso spronandomi a non mollare.
Una citazione particolare va fatta al G8 e a "i Ragazzi" che tanto hanno fatto per non farmi arrivare a questo giorno.

Infine, un immenso grazie va a tutta la mia famiglia che ha dovuto sopportarmi in questi anni, facendo molti sacrifici che spero siano stati ripagati a pieno.