UNIVERSITA' DEGLI STUDI MILANO - BICOCCA

Dottorato in EPIDEMIOLOGIA E BIOSTATISTICA

INDICATORI SOCIO-OCCUPAZIONALI,PSICOLOGICI E NUOVI BIOMARCATORI NELLA PREDIZIONE DI EVENTI CARDIOVASCOLARI MAGGIORI

Tutor: Chiar.mo Prof. Marco M FERRARIO

Co-tutor: Chiar.mo Prof. Vincenzo BAGNARDI

Tesi di Dottorato di:

Dr.ssa Lorenza Bertù

Matr. Nr. 528429

Anno Accademico 2014-2015

a Emma e Rachele

"Complicare è facile, semplificare è difficile. Per complicare basta aggiungere, tutto quello che si vuole: colori, forme, azioni, decorazioni, personaggi, ambienti pieni di cose. Tutti sono capaci di complicare. Pochi sono capaci di semplificare. [...]

Per semplificare bisogna togliere e per togliere bisogna sapere cosa togliere. [...]

Togliere invece che aggiungere vuol dire riconoscere l'essenza delle cose e comunicarle nella loro essenzialità."

B. Munari

Ringraziamenti

Il percorso di dottorato mi ha regalato molte opportunità di miglioramento e di conoscenza, sia dal punto di vista professionale che personale. Sono felice delle cose che ho appreso e delle persone che ho avuto la fortuna di conoscere ed apprezzare in questi tre anni: non posso non citare il gruppo di lavoro della Prof.ssa Valsecchi, in particolare Stefania e ovviamente Ilia (sì cara siamo arrivate alla fine!).

Alla termine di un'esperienza di questo tipo i ringraziamenti sono doverosi e, non volendomi dilungare troppo, ne vorrei fare tre ai quali tengo in maniera particolare:

Al Professor Ferrario, che mi ha dato l'opportunità di seguire questo cammino e che, nei miei momenti di crisi, mi ha "acciuffato per i capelli" spronandomi ad andare avanti.

A Enzo, preparato e disponibile ma, soprattutto, amico e vero e proprio *deus ex machina*.

A Giovanni, che col suo modo sempre garbato, mi è stato vicino trovando spesso e volentieri il modo giusto per uscire dai "gineprai" in cui riesco ad infilarmi.

Ad ognuno di voi un sincero grazie!

Lorenza

Sommario

Abbreviazioni

ACV	Accidente CerebroVascolare
AUC	Area Under the Curve
BDI	Beck Depression Inventory
CART	Classification and Regression Tree
CCA	Complete Case Analysis
CHD	Coronary Heart Disease
CI	Confidence Interval
CRp	C-Reactive protein
CVD	Cardiovascular Disease
DL	Decision Latitude
EC	Eventi Coronarici
ERI	Effort Reward Imbalance
HR	Hazard Ratio
JAS	Jenkins Activity Index
JCQ	Job Content Questionnaire
MONICA	MONItoring CArdiovascular diseases
PAMELA	Pressioni Arteriose Monitorate E Loro Associazioni
PJD	Psychological Job Demand
RSF	Random Survival Forest
SDO	Schede di Dimissione Ospedaliera
SEMM	Study of the Employees of the Municipality of Milan

ABSTRACT

Introduzione. L'importanza degli aspetti psicosociali nell'insorgenza delle malattie cardiovascolari è nota ma, ancora non è chiaro, quali aspetti siano maggiormente associati al rischio cardiovascolare e se, tenerne conto, possa migliorare la capacità predittiva dei classici modelli basati su età, sesso, abitudine al fumo, diabete, pressione arteriosa e colesterolo.

Obiettivo. Questa tesi si pone un duplice obiettivo: i) valutare l'associazione tra rischio di eventi cardiovascolari maggiori ed una batteria di indicatori socio-occupazionali e psicologici, integrando, quando possibile, l'informazione soggettiva con quella oggettiva ricavata dall'analisi dei biomarcatori; ii) identificare, attraverso tecniche di statistical learning, gli item maggiormente implicati nella predizione di evento cardiovascolare.

Metodi. Il lavoro di tesi è stato incentrato nei primi due anni all'estensione del follow-up degli eventi fino alla fine dell'anno 2008 (in particolare per la coorte PAMELA), ed alla valutazione dell'associazione tra singoli questionari ed endpoint cardiovascolari (in particolare disturbi del sonno ed evento CVD, e strain lavorativo ed evento CHD). Nel corso dell'ultimo anno si sono approfonditi gli aspetti metodologici legati alle tecniche di Statistical Learning al fine di arrivare alla selezione degli item delle scale psico-sociali maggiormente legati all'insorgenza di eventi cardiovascolari.

Per la selezione degli item si sono considerati gli uomini delle coorti MONICA-Brianza e PAMELA, in età 25-64 anni ed si è applicata la tecnica dei Random Survival Forest. Il contributo aggiuntivo degli item identificati è stata valutata in termini di Area Under the Curve rispetto ad

un modello contenente i principali fattori di rischio CVD (età, pressione arteriosa, colesterolo, fumo e diabete).

Risultati. L'analisi delle scale psico-sociali raccolte nelle coorti MONICA, PAMELA e SEMM ha evidenziato l'effetto di disturbi e durata del sonno sugli eventi CVD (HR = 1.80; 95% CI: 1.07-3.03 disturbi severi vs. nessun disturbo; HR = 1.56; 1.10-2.22 dormire 9 ore o più vs. 7-8 ore). Altra associazione rilevante messa in luce nelle analisi dei questionari è stato quella tra strain lavorativo e rischio di eventi CHD.

L'analisi attraverso i Random Survival Forest ha identificato, negli uomini delle coorti MONICA Brianza e PAMELA, come possibili predittori del rischio cardiovascolare 12 item relativi a: scala del sonno, strain lavorativo, tratti di personalità e depressione. Considerando, in un modello dove sono già presenti età, pressione arteriosa, colesterolo, fumo e diabete, il contributo aggiuntivo a 15 anni di questi item è dell'1.2%.

Conclusioni. I risultati confermano che, tenere conto degli aspetti psicologici, sociali e lavorativi nello studio del rischio di primo evento cardiovascolare è importante. Dal punto di vista dell'associazione i nostri dati hanno consentito di mettere in evidenza il rischio aumentato legato a situazioni di stress lavorativo e di disturbo del sonno, mentre per altre scale non si evidenziata nessuna associazione. L'analisi con i Random Survival Forest ha mostrato come sia possibile identificare un numero ridotto di item in grado di riconoscere soggetti che potrebbero sperimentare con maggior frequenza un evento cardiovascolare.

1. INTRODUZIONE

L'importanza degli aspetti psicosociali nell'insorgenza delle malattie cardiovascolari è nota (Perk J, 2012) tuttavia, non è ancora chiaro quali di questi aspetti sia maggiormente associato al rischio cardiovascolare e se tenerne conto possa migliorare la capacità predittiva dei classici modelli di rischio cardiovascolare basati su età, sesso, abitudine al fumo, diabete, pressione arteriosa, colesterolo.

Le difficoltà che si incontrano nel raccogliere e misurare i fattori psicosociali riguardano sostanzialmente due aspetti: da un lato la difficoltà di fornire una definizione univoca del fattore in sé in quanto influenzato dal contesto culturale e sociale in cui viene analizzato (Neylon A, 2013); dall'altra il metodo di raccolta delle informazioni, nella maggior parte dei casi affidata a questionari auto-compilati, che può portare ad una mancanza di obiettività nella misurazione. Negli ultimi anni accanto alla valutazione dei fattori psicosociali attraverso l'uso di questionari (misura soggettiva del problema), si sono imposti all'attenzione dei ricercatori anche una serie di biomarcatori che fungono da proxy per la valutazione dei fattori psicosociali (ad esempio cortisolo - stress) consentendone una misura oggettiva. La possibilità di disporre congiuntamente di entrambe le tipologie di informazioni risulta interessante al fine di studiare la relazione tra valutazione soggettiva ed oggettiva ed approfondire ulteriormente le conoscenze sulla relazione tra patologie cardiovascolari e fattori psicosociali.

Recentemente si è, inoltre, osservato che l'impiego congiunto di più scale che colgono aspetti diversi della sfera psicosociale offre una miglior comprensione della relazione tra questi aspetti e gli outcome di interesse. Alcuni ricercatori sono arrivati a definire, utilizzando metodi di analisi delle componenti principali, un fattore composito che, introdotto

nell’analisi, ha portato ad un significativo miglioramento nella capacità predittiva di eventi cardiovascolari rispetto all’analisi delle singole scale (Whittaker KS, 2012).

Un’ulteriore criticità in un contesto nel quale la misurazione dei fattori di interesse avviene attraverso la somministrazione di test o questionari costituiti da più item è l’aspetto legato alla presenza di dati mancanti (Rubin, 1976; Schafer & Graham, 2002). Ovviamente all’aumentare del numero di item considerati aumenta la probabilità di avere almeno un dato mancante e di conseguenza il suo impatto sull’analisi.

In generale, la ricerca di variabili associate al rischio di un evento in studi che raccolgono informazioni su molteplici caratteristiche è condotta utilizzando due approcci: (i) modellizzazione statistica, basata sull’assunto che i dati siano generati secondo un determinato modello probabilistico (ii) modellazione algoritmica, che considera sconosciuto il processo di generazione dei dati. In ambito epidemiologico si è sempre preferito il primo strumento (modello di regressione logistica e modello di Cox) che, se da un lato fornisce risultati immediatamente interpretabili in termini probabilistici, dall’altro, basandosi su assunti parametrici spesso troppo semplificativi, rischia di non identificare fattori e interazioni tra fattori importanti. La modellazione algoritmica, classicamente basata sugli alberi decisionali, è stata invece spesso criticata perché ritenuta instabile (piccole perturbazioni nei dati possono portare alla scelta di alberi che selezionano gruppi di variabili molto diverse tra loro in termini di significato). Questo problema è stato affrontato e risolto con l’introduzione delle foreste casuali in cui i singoli alberi sono costruiti estraendo con tecniche di boostrap dei campioni delle osservazioni ed utilizzando ad ogni nodo un campione casuale di variabili predittrici come candidate per la divisione.

In questo scenario si colloca l'idea di questa tesi che, partendo dai dati disponibili su 5 coorti (quattro di popolazione – area Monza Brianza - ed una di lavoratori del settore pubblico – comune di Milano) ha come scopi:

i) Valutare l'associazione tra rischio di eventi cardiovascolari maggiori ed una batteria di indicatori socio-occupazionali e psicologici valutati attraverso la somministrazione di questionari. Integrando, quando disponibile, l'informazione soggettiva con quella oggettiva ricavata dall'analisi dei biomarcatori;

ii) Identificare, tra tutti gli item dei questionari disponibili, quelli maggiormente implicati nella predizione di evento cardiovascolare. Per questo scopo verranno utilizzate tecniche di statistical learning, principalmente Random Forest, sfruttando i punti di forza di questi algoritmi in particolare nella gestione di elevato numero di variabili di natura anche diversa e la flessibilità di questi metodi nel risolvere problemi legati alla presenza di dati mancanti.

2. REVISIONE DELLA LETTERATURA

In questo capitolo si vuole sia, presentare una panoramica della letteratura che ha valutato la relazione tra gli indicatori psicologici e socio-occupazionali ed il rischio di eventi cardiovascolari sia delineare il razionale che porta alla scelta dell'uso di tecniche di statistical learning per la selezione delle variabili (in particolare singoli item di questionari) maggiormente predittive dell'outcome di interesse.

2.1 Indicatori socio-occupazionali e psicologici

Al fine di valutare quanto già noto in letteratura sulla relazione tra variabili psicosociali e rischio cardiovascolare è stata fatta una ricerca bibliografica utilizzando come database bibliografici PUBMED e la piattaforma Web of Knowledge. Sono stati inizialmente inseriti come criteri di ricerca i termini "*cardiovascular disease* AND *psychosocial risk factor*" . La ricerca è stata poi ripetuta utilizzando di volta in volta la dicitura della scala psicosociale considerata. Il materiale selezionato è stato sottoposto ad una revisione manuale i cui risultati sono riportati nel seguito.

Sebbene negli ultimi 20 anni si sia assistito ad una diminuzione della mortalità cardiovascolare, a progressi in campo medico e farmaceutico e ad uno sforzo culturale per la riduzione dei principali fattori di rischio, le patologie cardiovascolari restano la principale causa di morte e invalidità nei paesi sviluppati e si apprestano a diventarla anche nei paesi in via di sviluppo (WHO, 2011). Fra i fattori che contribuiscono a questo scenario vi è anche la sottostima dell'importanza dei fattori psicosociali nello sviluppo, nella progressione e nella prognosi di queste malattie.

Accanto ai fattori di rischio universalmente considerati predittori per queste patologie (dislipidemia, obesità, insulino resistenza, diabete mellito, predisposizione genetica) e ad una serie di parametri comportamentali come l'abitudine al fumo, il consumo di alcool, le abitudini alimentari e l'attività fisica, la cui associazione con gli eventi cardiovascolari è anch'essa nota, esistono un insieme di fattori psicosociali che i ricercatori hanno messo in relazione con il rischio cardiovascolare. Questi elementi hanno la tendenza a clusterizzare in un unico fattore che descrive la tendenza a perdere il controllo nei confronti del contesto in cui si vive e che viene definito "affettività negativa" (Katsarou, 2012). L'affettività negativa include una serie di emozioni e tratti della personalità specifici (i.e. ostilità, rabbia, disperazione), stati psicologici (i.e. depressione, ansia, strain lavorativo) ed altri parametri quali disturbi del sonno, il carico familiare, il supporto sociale ed il livello socio-economico. Queste componenti possono condizionare i comportamenti dei singoli soggetti portandoli ad assumere atteggiamenti che nel tempo si potrebbero rivelare dannosi per la salute (fumo, cattive abitudini alimentari, scarsa attività fisica) ed anche influenzando negativamente l'aderenza ai trattamenti prescritti causando un aumento del rischio di sviluppare malattie cardiovascolari o peggiorandone il decorso clinico (Lichtman JH, 2008; Dimsdale, 2008). La componente psicosociale può agire anche a livello fisiologico portando ad un aumento del battito cardiaco e della pressione arteriosa, alla secrezione di ormoni (catecolamine, ormone della crescita e glucocorticoidi) e ad un generale indebolimento del sistema immunitario (Charmandari, 2005; Cohen, 1992; Glass, 1977).

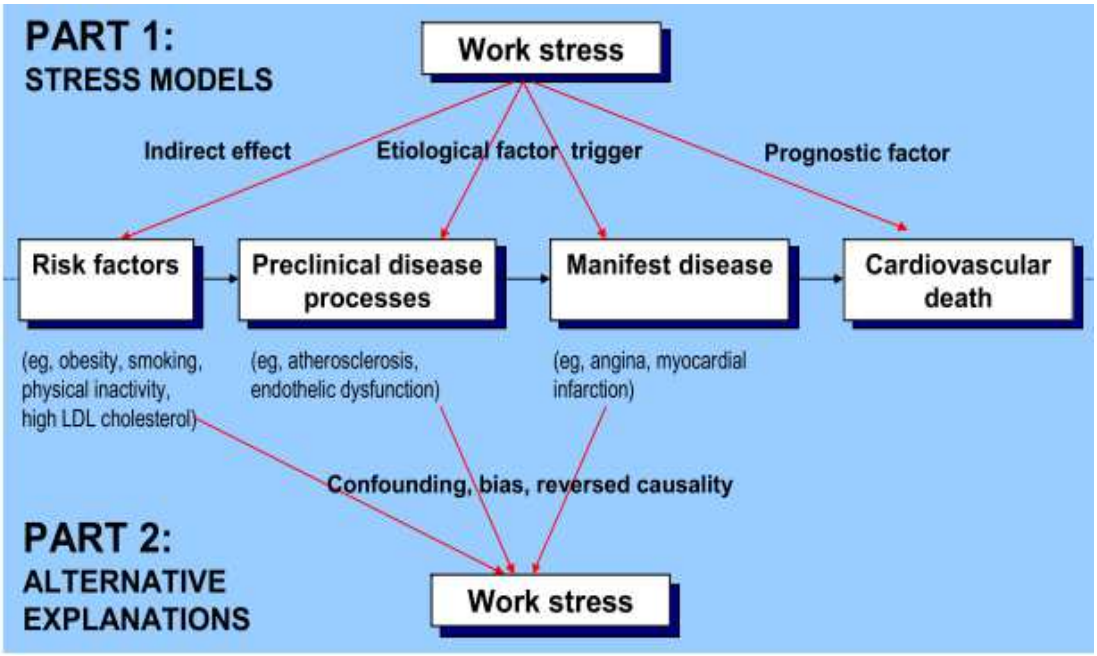
La relazione tra le variabili psicosociali e lo sviluppo e progressione delle malattie cardiovascolari (CVD) è ampiamente studiata: atteggiamenti aggressivi ed ostili sono predittivi di infarto al miocardio e aritmie (Eaker

ED, 2004), uno scarso supporto sociale è stato confermato essere collegato alla morbidità e mortalità per CVD (Barth J, 2010). Si osserva, inoltre, come l'effetto delle variabili psicosociali si mantenga anche dopo aver corretto le stime per i principali fattori di rischio e come questo effetto sia indipendente dallo stato socioeconomico e ubiquitario nelle aree geografiche, nelle classi d'età e tra i sessi (Rosengren A, 2004). Molti studi prospettici hanno mostrato che, in ambo i sessi, soggetti con un basso livello socio-economico (inteso come basso livello educazionale, basso salario o residenti in aree residenziali disagiate) hanno un aumento del rischio di morte per tutte le cause come anche per malattia cardiovascolare pari ad un rischio relativo che oscilla tra 1.3 e 2.0 a seconda degli studi (Stringhini, 2010; Tonne, 2005).

Recenti revisioni sistematiche confermano che, le **persone socialmente isolate**, hanno un maggior rischio di morte prematura per CVD e che, similmente, un basso supporto sociale è collegato ad una diminuzione della sopravvivenza e ad una peggior prognosi tra i soggetti con evidenze cliniche di disturbi cardiovascolari con un rischio relativo associato che varia tra 1.5-3.0 a seconda degli studi (Mookadam & Arthur, 2004; Lett, 2005; Barth J, 2010). In un lavoro pubblicato nell'ottobre 2014 su "Psychosomatic Medecine" gli autori riportano un aumento della mortalità nei soggetti socialmente isolati rispetto ai non isolati (HR=1.50 IC95%: 1.07-2.11 – Kreibig, Whooley, & Gross, 2014). In alcuni lavori si sottolinea come l'isolamento sociale agisca soprattutto nelle donne ed in relazione allo stroke (HR pari a 2.7 IC95%: 1.1-6.7 - Rutledge et al., 2008).

L’ipotesi che lo **stress**, nei suoi molteplici aspetti, sia un fattore di rischio per gli eventi cardiovascolari è esplorata da tempo; primi riferimenti si trovano in pubblicazioni risalenti agli inizi del ‘900 (Cannon, 1915). In anni recenti gli aspetti dello stress che maggiormente sono stati analizzati in relazione alle patologie cardiovascolari sono quelli relativi allo stress lavoro-correlato ed a quello dovuto all’isolamento sociale (Steptoe & Kivimaki, 2013). Nella meta-analisi pubblicata nel 2013 Steptoe e Kivimäki riportano una stima pooled, aggiustata per età e genere, dell’Hazard Ratio di CHD fra i soggetti con job strain rispetto ai soggetti senza strain lavorativo pari a 1.34 (95% CI: 1.18;1.51), mentre l’associazione tra stress lavorativo e stroke ha risultati inconsistenti. I risultati presentati da Steptoe e Kivimäki supportano l’ipotesi che lo stress sul luogo di lavoro sia un fattore di rischio per gli eventi CHD, e tale effetto viene mantenuto anche dopo aver eliminato dall’analisi gli eventi avvenuti entro 5 anni dall’ingresso nelle rispettive coorti, al fine di ridurre l’effetto di *reverse causation*; tuttavia i risultati sono anche compatibili con l’ipotesi di condivisione di alcuni fattori di rischio sia fisiologici che comportamentali e legati all’ambiente come descritto dalla figura 1 tratta dall’articolo di Kivimäki del 2006 (Kivimaki, et al., 2006). In una revisione sistematica pubblicata nel 2012 che raccoglieva articoli pubblicati tra il 1977 ed il marzo 2010 l’associazione tra CVD e stress lavorativo (indipendentemente dal tipo di modello adottato per la valutazione dello stress) è risultata significativa negli uomini, mentre nelle donne non si osservano associazioni rilevanti con l’unica eccezione di uno studio giapponese del 2005 (Backé, Sidler, Latza, Rossnagel, & Schumann, 2012).

Figura 1: Possibili pathways nella relazione tra stress lavorativo e malattie cardiache e spiegazione alternativa per questa associazione (Kivimäki 2006).



Nelle donne , tuttavia, si osserva un aumento del rischio di CHD (Hazard Ratio 2.7-4.0 a seconda degli studi) in relazione a **condizioni conflittuali e stressanti presenti in famiglia** (Eaker, 2007). Il carico familiare come numero di ore dedicato all'accudimento di figli/nipoti è risultato associato ad un aumentato rischio di CHD (54412 donne 46-71 anni – 321 eventi CHD, RR = 1.82 IC95% 1.08-3.05). Nessuna differenza è stata osservata stratificando per donne lavoratrici e non (Lee, Colditz, Berkman, & Kawachi, 2003). Il Family Responsability Scale è uno strumento utilizzato nel Framingham Heart Study come misura del carico familiare, inteso come responsabilità familiare. I risultati di uno studio sui dati del Framingham suggeriscono che non vi sia differenza rispetto al rischio di eventi CHD tra donne lavoratrici e casalinghe, mentre il numero di figli è positivamente associato al rischio di CHD in particolare per le impiegate "ever-married" e con almeno 3 figli (Haynes & Feinleib, 1980)

La relazione **tra sintomi depressivi e/o vital exhaustion** ed eventi cardiovascolari è estensivamente studiata. Molto spesso questi due fattori sono stati trovati essere una conseguenza piuttosto che una causa di un evento CVD, in particolare, molti studi hanno mostrato che la depressione risulta essere un fattore prognostico per la sopravvivenza dopo infarto miocardico. Tuttavia, anche in soggetti liberi da malattia, è stata osservata un'associazione tra depressione e rischio di CHD che risulta particolarmente forte considerando in soggetti con diagnosi clinica di depressione (HR = 2.69 IC95%: 1.63-4.43) ma è significativamente presente anche nei soggetti con sintomi depressivi ("depressive mood" HR = 1.49 IC95%:1.16-1.92 - (Rugulies R. , 2002). Per quanto riguarda la vital exhaustion è stata messa in relazione con un aumentato rischio di stroke HR = 1.13 IC95%: 1.04-1.23 (Schuitemaker, Dinant, GA, Verhelst, & Appels, 2004).). Vi è anche un'ampia letteratura a supporto della relazione tra **stati ansiosi** e rischio cardiovascolare, in particolare un aumento del rischio di incidenza di eventi cardiovascolari associati ad attacchi di panico (HR 1.7 - Smoller, 2007; 4.2 - Chen, 2009). Due recenti meta-analisi confermano inoltre, che l'ansia è un fattore di rischio per l'incidenza di CHD (HR 1.3- Roest, 2010) e per l'incidenza di eventi avversi conseguenti ad un infarto al miocardio (HR 1.5 e 1.7 - Roest, 2010).

L'ostilità, tratto di personalità caratterizzato da sfiducia verso gli altri, **aggressività** e collera e dalla tendenza ad assumere atteggiamenti aggressivi e disadattivi nelle relazioni sociali, da una recente meta-analisi, è stata confermata essere associata con un aumento del rischio cardiovascolare (HR 1.2, Chida & Steptoe, 2009). Tra i pazienti affetti da patologie cardiovascolari, l'incapacità di esprimere la rabbia è risultata

associata ad un aumento del rischio di eventi cardiaci avversi (HR 2.9, Denollet, 2010).

L'insonnia, definita come la difficoltà ad addormentarsi, di rimanere addormentato o l'avere un sonno disturbato, è una condizione molto frequente nei paesi industrializzati. Più del 30% degli adulti riportano sintomi di questo tipo e la frequenza aumenta nelle donne e all'aumentare dell'età. Fra gli over 65 circa il 50% riferisce un problema legato a questo disturbo (Ohayon, 2002).

Molti studi anche in anni recenti hanno studiato l'associazione tra disturbi del sonno in termini sia qualitativi (come si dorme) che quantitativi (per quante ore si dorme) ed il rischio di malattie cardiovascolari (Quan, 2009) con risultati non sempre consistenti.

In una meta-analisi pubblicata nel 2014, su 13 coorti considerate gli autori riportano una stima pooled del rischio relativo di incorrere in un evento cardiovascolare pari a 1.45 (IC95% 1.29-1.62), fra coloro che riportano disturbi legati all'insonnia rispetto a chi non ne riferisce (Sofi, et al., 2014).

Quale sia il meccanismo che lega i disturbi del sonno al rischio di avere un evento cardiovascolare rimane ancora non chiaro; l'insonnia è associata ad un aumento dell'attività del sistema nervoso simpatetico e questo potrebbe essere responsabile di un aumento nella morbidità cardiovascolare; l'insonnia potrebbe anche portare ad un aumento degli ormoni correlati allo stress o, infine, potrebbe essere indice della presenza di altri fattori di rischio legati a CVD o ad altre co-morbidità.

Per quanto riguarda la durata del sonno, in molti articoli si sottolinea come l'andamento del rischio di eventi cardiovascolari in relazione alla durata del sonno sembri essere descritta da una U-shape, suggerendo come

sia periodi di sonno troppo brevi o, al contrario, eccessivamente lunghi abbiano effetti negativi sulla salute (tale andamento viene descritto anche quando si consideri la mortalità in generale, il diabete di tipo 2, l'ipertensione o problemi respiratori e obesità). Nella meta-analisi pubblicata da Cappuccio e colleghi nel 2011 un ridotto numero di ore di sonno (tipicamente meno di 5/6 ore) è risultato significativamente associato al rischio di CHD (RR =1.48 CI95%: 1.22-1.80) e stroke (RR=1.15 CI95%: 1.00-1.31) e solo debolmente e non in maniera significativa con CVD in generale (RR=1.03; CI95% 0.93-1.15); dormire per un numero di ore superiore a 9 risulta essere significativamente associato al rischio di CHD (RR =1.38 CI95%: 1.15-1.66), stroke (RR=1.65 CI95%: 1.45-1.87) ed a CVD in generale (RR=1.41; CI95% 1.19-1.68) (Cappuccio, Cooper, D'Elia, & al, 2011).

Malgrado non vi sia uniformità di giudizio nel ritenere che uno screening degli indicatori psicosociali possa portare ad un miglioramento negli outcomes cardiovascolari, si osserva che esiste una relazione tra questi fattori ed un aumento della morbidità e della mortalità e che, la presenza di questi fattori, è associata ad una minor probabilità di modificare positivamente altri fattori di rischio e ad un generale peggioramento della qualità della vita. A riprova di ciò sia la *European Society of Cardiology* che l'*American Heart Association* supportano l'utilizzo di questionari standard per la rilevazione e la valutazione delle tematiche psicosociali (Perk J, 2012; Lichtman JH, 2008).

2.2 Statistical Learning e applicazioni in ambito clinico/epidemiologico

Lo Statistical Learning è un insieme di strumenti e tecniche utilizzati per modellare e comprendere le relazioni all'interno di banche dati complesse. Con l'avvento dei "Big Data" queste tecniche sono diventate sempre più rilevanti in molti ambiti quali ad esempio genetica, medicina, marketing, finanza, ma non ancora molto diffusa nell'analisi di grandi studi epidemiologici.

In ambito clinico vi sono molti lavori che sfruttano i punti di forza di queste tecniche per studiare la relazione tra variabili o selezionare i fattori di rischio associati alle diverse patologie.

Per la descrizione degli aspetti teorici si rimanda al capitolo 4 dove verranno descritte le caratteristiche di CART (Classification and Regression Tree), Bagging, Random Forest e, la recente estensione di questi ultimi all'ambito dell'analisi della sopravvivenza, il Random Survival Forest (Ishwaran H. , Kogalur, Blackstone, & Lauer, 2008).

Le caratteristiche comuni che rendono queste tecniche di interesse anche in ambito epidemiologico riguardano principalmente i) la loro capacità di identificare, utilizzando opportuni algoritmi, la presenza di eventuali pattern nei dati anche da database di grandi dimensioni; ii) l'assenza della necessità di ipotizzare a priori la relazione tra le variabili (lineare o non lineare) o di effettuare trasformazioni dei dati; iii) la possibilità di identificare più facilmente eventuali interazioni, anche complesse (Hisch, Gorodeski, Blackstone, Ishwaran, & Lauer, 2011).

In un lavoro di recentissima pubblicazione si è esplorata la possibilità di affiancare queste tecniche a quelle statistiche basate sui modelli osservando come l'utilizzo congiunto porti ad aumentare la capacità di comprendere i fenomeni di interesse (Dipnale, et al., 2016).

Quando si considerino le tecniche di Statistical Learning più avanzate (Random Forest, Boosting, Bagging) alcuni lavori evidenziano anche come vi sia un miglioramento nell'accuratezza complessiva del modello (Miao, Cai, Zhang, & Li, 2015).

Nello studio di Van e colleghi del 2011 vengono considerati un insieme di fattori psicologici, sociali e clinici per evidenziare, utilizzando i Random Forest, gli aspetti che maggiormente influiscono il successo della riabilitazione cardiaca di fase II (Van, Gay, Kennedy, Barin, & Leijdekkers, 2011). L'utilizzo dei Random forest seleziona una serie di item che, tuttavia, non portano un miglioramento significativo nella capacità predittiva del modello.

La capacità degli algoritmi di classificazione, tipo Random Forest, di evidenziare le possibili interazioni tra le variabili è stato ben evidenziato nell'ambito della ricerca genetica (Lunetta, Hayward, Segal, & Van Eerdewegh, 2004) mettendo in luce le capacità dei random forest di cogliere l'importanza di variabili (nel caso specifico SNPs) che pur avendo uno scarso effetto sulla popolazione risultano avere un grande effetto in termini di interazione.

Come già segnalato, la capacità degli algoritmi di statistical learning, di analizzare le variabili senza ipotizzare andamenti a priori o senza doverle trasformare, risulta molto utile quando si lavori con variabili (item) provenienti da questionari; in aggiunta, questi metodi, sono flessibili nella gestione dei dati mancanti, ulteriore elemento che spesso caratterizza le informazioni raccolte tramite questionari, soprattutto se auto-compilati dai pazienti. Nel lavoro di Lu e Petkova del 2013 vengono messe a confronto differenti tecniche di statistical learning proprio con lo scopo di arrivare a definire uno strumento di screening utile e veloce per le patologie psichiatriche partendo da un insieme di item provenienti da differenti

questionari (Lu & Petkova, 2013), mostrando come l’utilizzo di Random forest aiuti a limitare l’impatto sull’analisi dei dati mancanti.

3. COORTI E INDICATORI SOCIO OCCUPAZIONALI OGGETTO DI INDAGINE

3.1 Coorti in studio

In questo lavoro verranno analizzati i dati provenienti dalle coorti MONICA e PAMELA dell'area Brianza e dall'indagine svolta sui lavoratori della municipalità milanese (SEMM).

Il Progetto MONICA è un esteso intervento di monitoraggio cardiovascolare che si è svolto dall'inizio degli anni ottanta alla metà degli anni novanta in circa 30 centri distribuiti in Europa, nord America, Asia ed Australia. Nel presente lavoro si analizzeranno i dati relativi alle 3 coorti reclutate in uno dei due centri italiani (Area Brianza) negli anni 1986-1987 (MONICA 1), 1989-1990 (MONICA 2) e 1993-1994 (MONICA 3). Ogni indagine campionaria è stata condotta su 300 soggetti, selezionati casualmente, per ogni classe di età compresa tra i 25 ed i 66 anni per ambo i sessi. In totale si sono reclutati, per ciascuna coorte, 2400 persone.

Lo studio PAMELA (Pressioni Arteriose Monitorate E Loro Associazioni) è uno studio epidemiologico, il cui obiettivo principale è quello di determinare i valori di normalità della pressione arteriosa misurata autonomamente dai pazienti al proprio domicilio. I soggetti sono stati reclutati nel corso del 1990 tra i residenti nella città di Monza con un campionamento stratificato per genere e classi d'età decennali. Un totale di 2054 tra uomini e donne di età compresa tra i 25 ed i 74 anni hanno accettato di partecipare allo studio.

La coorte SEMM è composta da un campione di dipendenti del comune di Milano, di età compresa tra i 20 anni e l'età pensionabile. I soggetti sono stati reclutati nel periodo giugno 1992/aprile 1996. Sono stati arruolati i lavoratori appartenenti ai settori organizzativi del Nido (900 soggetti), delle Scuole Materne (1265 soggetti), degli Impiegati (895), dei Servizi Sociali

(823), dei Commessi (2058) e dei Vigili (1931) per un totale di 7872 soggetti con una netta prevalenza di donne (67%).

I soggetti sono stati seguiti nel tempo, dalla data di reclutamento fino al 31 dicembre 2008, registrandone lo stato in vita e l'occorrenza di eventi coronarici.

Per le informazioni relative alla mortalità ci si è avvalsi della collaborazione dei comuni di residenza dei soggetti e della Aziende Sanitarie Locali di riferimento per raccogliere le informazioni relative allo stato in vita e le eventuali cause di morte. Grazie agli uffici anagrafe si è potuto recuperare le informazioni su ciascun soggetto al 2008: residente in Italia o all'estero, emigrato in altro comune d'Italia, emigrato all'estero, deceduto o cancellato per irreperibilità. Le persone emigrate in altri comuni italiani sono state seguite nel comune di emigrazione, mentre le migrazioni all'estero figurano come uscite dal follow-up alla data dell'emigrazione. Per i soggetti deceduti sono state raccolte le schede ISTAT di morte presso le ASL di riferimento in modo da poter classificare le morti secondo la categoria MONICA (schema 1).

Schema 1: Codifica MONICA per eventi EC ed ACV.

Categoria diagnostica per Evento Coronarico	**Categoria diagnostica per Evento Cardiovascolare**
Infarto miocardico sicuro	Accidente cerebrovascolare definito
Infarto miocardico possibile o morte coronarica	Accidente cerebrovascolare definito associato ad un evento coronario
Assenza di infarto miocardico acuto	Non accidente cerebrovascolare
Dati insufficienti	Dati insufficienti

Gli eventi coronarici non fatali sono stati identificati applicando metodologie di record linkage agli archivi delle ospedalizzazioni in regione Lombardia relativamente agli anni 1989-2008 considerando come criteri di identificazione degli eventi coronarici e cerebrovascolari quelli riportati nello schema 2.

Si sono utilizzate due metodologie di linkage: deterministico, basato sull'esatta corrispondenza delle variabili di appaiamento ed il linkage probabilistico che mira a minimizzare la probabilità di errori di linkage.

Sulla base dei ricoveri così estratti, dopo una prima verifica di tipo manuale, sono state contattate le direzioni sanitarie degli ospedali lombardi richiedendo conferma che il ricovero appartenesse effettivamente ad un soggetto delle coorti considerate e, in caso di risposta affermativa, si è proceduto alla raccolta della cartella relativa. Gli eventi individuati sono stati validati e codificati secondo i criteri MONICA (Bombelli, et al., 2013)

Schema 2: Criteri di identificazione degli eventi coronarici e cerebrovascolari sospetti.

	Evento Coronarico	**Evento cerebrovascolare**
Certificato di morte (eventi fatali)	Causa principale di morte: ICD-IX 410-414, 798, 799; o 250, 428, 440 in presenza di 410-414 in altra causa	Causa principale di morte: ICD-IX 430-432, 434, 436-437; o 250, 401-404, 427, 440 in presenza di 430-432, 434, 436-437 in altra causa
Codice di dimissione ospedaliera	ICD-IX 410-412 e ICD-IX CM 36.0-9 per rivascolarizzazioni cardiache	ICD-IX 430-432, 434, 436; ICD-IX CM 38.01-39.22 o 39.50-39.52 con almeno un 430-438 tra I codici di dimissione, per endoarterectomie della carotide

Per maggiori dettagli su arruolamento delle coorti e misurazione dei fattori di rischio al basale sono consultabili sia nelle monografie del progetto WHO-MONICA che, per lo studio PAMELA, nel lavoro pubblicato nel 1991 da Cesana, Ferrario ed altri (Cesana, DeVito, & Ferrario, 1991).

Dopo l'estensione del follow-up al 31 dicembre 2008 (ultimo aggiornamento concluso nell'ottobre 2014 per lo studio PAMELA) si dispone dei dati su 5 coorti composte complessivamente da 14422 soggetti tra i 25 ed i 75 anni d'età (5809 uomini e 8613 donne). I soggetti sono stati seguiti fino a:

- ✓ data insorgenza primo evento cardiovascolare fatale o non fatale;
- ✓ data del decesso per altra causa
- ✓ compimento degli 80 anni di età
- ✓ 31 dicembre 2008

I soggetti hanno sperimentato nel corso di circa 17 anni di follow-up un totale di 571 eventi cardiovascolari (412 tra gli uomini e 159 nelle donne).

In tabella 3.1-1 si riporta la descrizione del numero di soggetti e di eventi nelle 5 coorti.

3.2 Indicatori disponibili

La disponibilità degli indicatori socio-occupazionali e psicologici nelle coorti in studio è descritta in tabella 3.2-1.

Tabella 3.2-1: Indicatori socio-occupazionali e psicologici. Coorti MONICA e SEMM

SCALA	MONICA 1	MONICA 2	MONICA 3	SEMM
BERKMAN SYME SCALE	X	X	X	
BECK DEPRESSION INVENTORY (BDI)	X			
SOCIAL & GEOGRAPHIC MOBILITY	X	X	X	
SLEEP DISTURBANCIES	X	X	X	
JOB CONTENT QUESTIONNAIRE (JCQ)	X	X	X	X
JENKINS ACTIVITY SURVEY	X	X	X	
MAASTRICHT QUESTIONNAIRE		X	X	
CARATTERISTICHE LAVORATIVE	X	X	X	X
CARICO FAMILIARE	X	X	X	X

3.2.1 Berkman Syme scale

Il Berkman-Syme Social Network Index considera 4 tipologie di relazioni sociali: lo stato civile (sposato vs. altro); la socializzazione (frequenza e numero di contatti con amici e parenti); l'appartenenza a gruppi religiosi e la partecipazione attiva in altre organizzazioni.

La scala prevede la costruzione di 4 indici che consentono di misurare il grado di integrazione sociale considerando sia il numero che l'importanza dei differenti legami :

1. SCORE1: costruito sulla base del numero di amici e parenti con i quali si ritiene di avere un rapporto di confidenza e fiducia;
2. CONTACT SCORE (CS): costruito sulla base dello SCORE1 e del numero di amici e parenti con i quali si hanno contatti mensili;
3. INDEX OF CLOSE CONTACT (ICC): costruito sulla base del CS e dello stato civile;

4. SOCIAL NETWORK INDEX SCORE (SNIS): costruito sulla base di ICC e della partecipazione attiva a gruppi e/o organizzazioni.

Il Social Network Index Score può assumere 4 livelli: basso (1), medio (2), medio-alto (3) e alto (4). Le persone con un basso livello di relazione sociale saranno quindi caratterizzate dal non essere sposate o conviventi, dall'avere pochi amici o parenti affettivamente vicini e dal non partecipare attivamente in nessun gruppo o organizzazione (Eng, 2002, Berkman & Syme, 1979).
Nelle tabelle seguenti è riassunta la costruzione dei 4 indici della scala.

Tabella 3.2.1-1: Berkman Syme scale – Costruzione dello SCORE1

Numero di amici e parenti	Score 1
0-5	1
6-10	2
11-15	3
>15	4

Tabella 3.2.1-2: Berkman Syme scale – Costruzione del CONTACT SCORE

Score 1	Numero amici/parenti frequentati in un mese	Contact Score
1	0-5	1
2	0-5	2
2	≥ 6	3
3	0-5	3
3	≥ 6	4
4	Qualsiasi	4

Tabella 3.2.1-3: Berkman Syme scale – Costruzione del INDEX of CLOSE CONTACT

Contact Score	Stato civile[§]	ICC
1	1	1
2 o 3	1	1
4	1	2
1	2	1
2 o 3	2	2
4	2	3

[§] 1 = Single/Divorziato/Separato/Vedovo; 2 = Sposato/convivente

Tabella 3.2.1-4: Berkman Syme scale – Costruzione del SOCIAL NETWORK INDEX SCORE

ICC	Group score[§]	Social Network Index Score
1	1	1
1	2	2
2	1	2
2	2	3
3	1	3
3	2	4

[§] 1 =Nessuna partecipazione attiva a gruppi/organizzazioni;
2 = Partecipazione attiva ad almeno un gruppo/organizzazione

3.2.2 Social and Geographical mobility

Nel questionario sono raccolte informazioni sulla "residential mobility"; lo scopo è quello di definire se i soggetti abbiano creato un insieme di rapporti stabili e di sostegno nell'ambiente nel quale vivono. Le informazioni consentiranno inoltre, di individuare soggetti che, cresciuti in un area e trasferitisi poi in un'altra, abbiano abitudini e retaggi culturali differenti che possano in qualche modo influenzare l'insorgenza degli eventi di interesse.

3.2.3 Sleep Disturbancies

La scala utilizzata per raccogliere informazioni sui disturbi del sonno è quella presentata di Jenkins e da lui utilizzata nello studio "Air Traffic Controller Health Change Study" (Rose & al, 1978; Jenkins & al, 1988)

La scala è composta da 5 item, i primi 4 riguardano i problemi del sonno e le risposte valutano la frequenza (in giorni) con la quale si è sofferto del disturbo relativo nel mese precedente l'indagine. Le risposte sono strutturate in una scala a 5 livelli: da 0 (per nulla) a 5 (da 22 a 31 giorni al mese). Il quinto item richiede di riportare le ore di sonno al giorno. Per il calcolo dello score vengono sommati i primi 4 item della scala (range 0-

20). Il punteggio viene poi categorizzato come: 0-4 = problema non presente o scarso; 5-9 = moderato; 10-14= elevato; 15-20= critico.

3.2.4 Job Content Questionnaire (JCQ)

Per la valutazione dello stress lavorativo percepito è stato somministrato il Job Content Questionnaire sviluppato da Karasek (Karasek R. , 1979). Il modello proposto vede nell'elevata domanda lavorativa e nella bassa libertà decisionale la causa scatenante di una condizione di "strain" lavorativo, dove con strain si intende l'insieme negativo degli effetti/risposte che le fonti di stress presenti nel contesto lavorativo possono produrre sull'individuo.

Il questionario è composto da 3 scale che valutano: l'impegno lavorativo richiesto (JOB DEMAND), la professionalità e la capacità di programmare ed organizzare il lavoro (DECISION LATITUDE) e il supporto ottenuto da colleghi e superiori (SOCIAL SUPPORT). Gli item che concorrono alla costruzione delle tre scale, misurati su una scala likert a 4 gradi, vengono sommati per costruire i 3 indicatori.

Dicotomizzando al valore mediano le scale JOB DEMAND e DECISION LATITUDE vengono identificate 4 categorie di strain: *high strain*, caratterizzati da alta domanda e basso controllo; *active*, caratterizzati da alta domanda e alto controllo; *passive*, bassa domanda e basso controllo; *low strain*, bassa domanda e alto controllo.

3.2.5 Jenkins Activity Survey(JAS)

Il Jenkins Activity Survey (versione short N) è stato utilizzato per valutare il tratto di personalità di tipo A, caratterizzato da un atteggiamento impaziente, aggressivo, fortemente competitivo finanche ostile. Questo tratto di personalità è stato considerato, tra la fine degli anni '70 e negli

anni ’80, un possibile fattore di rischio per le patologie cardiache . Negli anni successivi non si sono trovate forti evidenze a supporto di quest’ipotesi e l’attenzione, più che sul tratto di personalità nel suo complesso, si è spostata su alcuni aspetti che lo caratterizzano, in particolare l’ostilità.

La scala JAS è composta da 13 item a ciascuna delle possibili risposte, compreso l’eventuale dato mancante, è attribuito un peso che differisce da item ad item. Lo score totale (X) è calcolato come somma dei pesi delle 13 risposte.

Tale score viene inserito all’interno della seguente formula al fine di calcolare il punteggio della scala: Type A Scaled Score= [(X-22.54)/0.62].

In caso di punteggio superiore a 0.5 il soggetto viene identificato come avente un tratto di personalità di tipo A.

3.2.6 Maastricht Questionnaire e Beck Depression Inventory

Per la valutazione della “vital exhaustion” è stato utilizzato il Maastricht Questionnaire sviluppato da Meesters e Appels (Meesters & Appels, 1996). Il questionario è composto da 14 item che valutano un insieme di caratteristiche quali stanchezza, irritabilità e senso di demoralizzazione. Alle risposte viene assegnato un punteggio di 0 o 1 e vengono poi sommate per costruire lo score che viene poi categorizzato come: 0-4 = exhaustion bassa o assente; 5-9 = exhaustion moderata; 10-14 = exhaustion critica. La scala Maastricht contiene al suo interno alcuni item che indagano aspetti legati specificatamente allo stato depressivo, e che sono sovrapponibili ad alcuni item della scala Beck Depression Inventory adottata nella coorta MONICA1 per la valutazione della depressione. Per poter disporre di informazioni sulla depressione per le tre coorti MONICA e per la coorte PAMELA si è quindi proceduto ad uniformare le due scale avvalendosi del supporto di uno psicologo che ha identificato gli item equivalenti nelle due

scale. In tabella 3.2.6-1 si riporta l'armonizzazione delle due scale che ha portato ad identificare 11 item comuni per la valutazione della depressione. Il punteggio attribuito a ciascun item era 0= assenza del sintomo, 1=presenza del sintomo.

Tabella 3.2.6-1: Armonizzazione scale BDI e Maastricht Q. Coorti MONICA e PAMELA

ITEM	Maastricht Q	BDI
DEP1	Ex1. Si sente spesso stanco	BDI17 - Stanchezza
DEP2	Ex2. Fa spesso fatica ad addormentarsi	BDI16 – Sonno
DEP3	Ex3. Si sveglia ripetutamente durante la notte	BDI16 – Sonno
DEP4	Ex4. Si sente spesso completamente fiacco	BDI15 – Perdita di energia
DEP5	Ex5.Ha la sensazione chiara di essere nel fiore degli anni	BDI14 – Aspetto fisico
DEP6	Ex6. Sente di non aver realizzato molto negli ultimi anni	BDI3 – Fallimento
DEP7	Ex7. Crede di essere arrivato ad un punto morto	BDI2 – Pessimismo
DEP8	Ex8. La sua vita sessuale è attiva e soddisfacente come sempre	BDI22 – Vita sessuale
DEP9	Ex9. Le piccole contrarietà la irritano ora più di prima	BDI11 – Irritabilità
DEP10	Ex11. Non le è mai capitato di pensare che sarebbe meglio morire	BDI9 – Pensieri suicidi
DEP11	Ex13. Si sente abbattuto	BDI1 – Tristezza profonda

3.2.7 Carico Familiare

Per il calcolo dello score si sono utilizzate le informazioni sullo stato civile e sul numero di persone che vivono in famiglia costruendo l'indicatore come riportato nella tabella sottostante.

Stato Civile	**N° famigliari**	**CARICO FAMILIARE**
Single	Qualsiasi	1 – Single
Coniugato/convivente	2	2 – Living with partner only
Coniugato/convivente	3/4	3 – Living with partner and 1/2 kids/relatives
Coniugato/convivente	≥ 5	3 – Living with partner + 3 or more kids/relatives
Separato/divorziato/vedovo	≥ 2	4 – Alone + 1 or more kids/relatives

4. STATISTICAL LEARNING: alberi decisionali, bagging, random forest e random survival forest

I metodi algoritmici di Statistical Learning si riferiscono ad un insieme di strumenti utilizzati per costruire modelli che predicano o stimino un outcome sulla base di una o più variabili di input e sono applicabili sia a modelli di regressione che di classificazione. Il metodo più semplice, il singolo albero decisionale, predice l'esito di una certa osservazione utilizzando la media, nel caso della regressione, o la moda, per la classificazione, della regione nella quale tale osservazione ricade; è un metodo semplice e di facile interpretazione ma non molto accurato in termini predittivi, per questo motivo sono stati introdotti approcci che, aggregando singoli alberi decisionali, riescono ad ottenere migliori performance. In particolare in questo capitolo verranno presentati, oltre ai CART, Bagging, Random Forest e Random Survival Forest (James, Witten, Hastie, & Tibshirani, 2014).

4.1 Alberi decisionali

L'idea che sta alla base della costruzione di un albero decisionale è quello di classificare gli individui correttamente, per fare questo l'algoritmo divide lo spazio dei predittori in un numero di regioni che risultino quanto più possibili omogenee in relazione all'outcome di interesse. I soggetti in ciascuna regione saranno identificati dalla loro media (regression tree) o dal loro valore modale (classification tree). Nel caso degli alberi di classificazione, spesso nell'interpretazione dei risultati non si tiene solo conto del valore preminente nella classe ma anche della proporzione di eventi/non eventi presenti nella classe stessa, questo dato

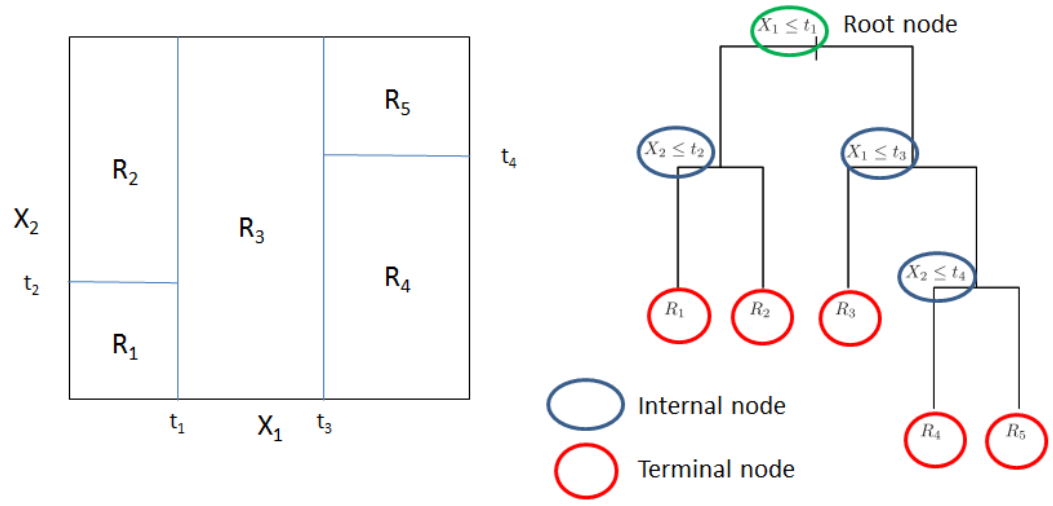
risulta più informativo soprattutto quando si studino outcome con tasso di evento basso.

La flessibilità di questo metodo lo fa talvolta preferire alle analisi più tradizionali, in particolar modo quando gli assunti su cui queste si poggiano vengono meno o non possono essere facilmente verificate.

Un albero di classificazione si presenta come un diagramma ad albero rovesciato che si biforca in corrispondenza di un "nodo". L'algoritmo parte dai dati raggruppati in un unico nodo (root node) ed esegue ad ogni passo una ricerca fra tutte le possibili suddivisioni, ad ogni passo viene scelta la suddivisione migliore, cioè quella che produce rami il più possibile omogenei fra loro. Si distinguono due tipologie di nodo: i nodi interni (internal node), che hanno due discendenti diretti, ed i nodi terminali o foglie (external node) che non subiscono ulteriori bipartizioni. In questo tipo di albero ogni split, basato sul valore di una singola variabile, divide sempre un nodo genitore in due nodi figli. In questo tipo di analisi è possibile far rientrare sia variabili qualitative che quantitative.

La costruzione dell'albero porta alla suddivisione dello spazio, in aree contenenti soggetti simili dal punto di vista dell'outcome di interesse come esemplificato in figura 4.1-1 nel caso di uno spazio costituito da 2 variabili continue X_1 e X_2.

Figura 4.1-1: Esempio di partizione di uno spazio bidimensionale (a sinistra) e relativa rappresentazione tramite albero decisionale (a destra).



L'algoritmo di costruzione degli alberi implementato in R, è del tipo CART, prevede la divisione del dataset di partenza in modo da avere un dataset di training (usato per sviluppare l'albero) ed uno di test, generalmente nella proporzione di 9 a 1 (10cross-validation). Nel caso di alberi di classificazione per valutare la bontà di uno split si possono utilizzare tre indici: *classification error rate*, l'indice di *Gini*, ed il *cross-entropy*.

Il *classification error rate* è definito come la frazione di osservazioni che ricadono in una regione in cui il valore modale non corrisponde.

L'indice di Gini è definito come:

$$G = \sum_{k=1}^{K} \hat{p}_{mk} \, (1 - \hat{p}_{mk})$$

E fornisce una misura della varianza totale tra le k classi dove $\hat{p}_{mk}$ rappresenta la proporzione di osservazioni nella regione m-esima relativa alla classe k-esima. L'indice di Gini, misurando il livello di purezza di un

nodo, ci dice quanto ciascun nodo contiene osservazioni che cadono prevalentemente in una singola classe.

L'indicatore di *cross-entropy* è definito come

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log(\hat{p}_{mk})$$

Ciascuno di questi indicatori può essere utilizzato quando si effettua la procedura di *pruning* dell'albero ("potatura"). Dato che l'algoritmo prevede la crescita di ciascun albero fino al massimo dettaglio possibile, rendendo, quindi, ciascun albero difficilmente generalizzabile in altri campioni, la tecnica di pruning viene applicata, utilizzando il dataset di training, per identificare a quale livello (nodo) l'albero debba essere potato per arrivare a definire il miglior trade-off tra complessità dell'albero e test error rate.

Figura 4.1-2: Algoritmo per la costruzione di un albero di regressione/classificazione (tratta da "An introduction to Statistical Learning" – James, Witten, Hastie e Tibshirani).

Algorithm 8.1 *Building a Regression Tree*

1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α.
3. Use K-fold cross-validation to choose α. That is, divide the training observations into K folds. For each $k = 1, \ldots, K$:
   (a) Repeat Steps 1 and 2 on all but the kth fold of the training data.
   (b) Evaluate the mean squared prediction error on the data in the left-out kth fold, as a function of α.
   Average the results for each value of α, and pick α to minimize the average error.
4. Return the subtree from Step 2 that corresponds to the chosen value of α.

Fra i **punti di forza** di questo metodo c'è sicuramente la facilità con la quale può essere spiegata e la possibilità di riassumere la procedura attraverso un grafico che risulta, generalmente, facilmente interpretabile e comprensibile; questo metodo consente, inoltre, di gestire facilmente predittori differenti, qualitativi o quantitativi senza necessità di effettuare trasformazioni sui dati, creare dummies o di valutarne a priori i possibili andamenti.

Le **limitazioni** di questo metodo risiedono principalmente nel fatto che non tengono conto dell'influenza che la scelta di una particolare divisione ha sulle divisioni successive. In altre parole, la decisione della divisione avviene ad ogni nodo dell'albero, in un preciso momento durante l'esecuzione dell'algoritmo, e non è più riconsiderata in seguito. Dato che tutte le suddivisioni vengono scelte sequenzialmente e ognuna di esse di fatto dipende dalle precedenti, si ha che tutte le divisioni sono dipendenti dal nodo radice dell'albero; una modifica del nodo radice potrebbe portare alla costruzione di un albero completamente differente. Questo fa sì che il singolo albero decisionale sia soggetto ad una grande variabilità, l'introduzione di un nuovo dato o di una nuova variabile porta a variazioni importanti, il risultato è quello di avere un'accuratezza inferiore rispetto ad altri metodi.

Per questo motivo si ha la necessità di introdurre dei metodi che, sfruttando l'idea base del singolo albero, riescano a superare questa criticità.

4.2 Bagging

Come ricordato in precedenza, il singolo albero decisionale non è stabile (grande variabilità), questo vuol dire che se si dividesse il dataset di training in due parti in maniera casuale e si costruisse un albero decisionale su ciascuna delle due parti il risultato sarebbe molto diverso. Il Bagging o Boostrap aggregation è nato proprio per cercare di ridurre questa variabilità.

L'idea di base è che, un metodo efficacie per ridurre la variabilità e, quindi, migliorare l'accuratezza, sia quello di considerare più training set e fare la media dei risultati. Dato che nella realtà non si dispone di più datasets di training questi vengono ottenuti campionando con la tecnica del boostrap dai dati disponibili. Su ciascuno dei datasets viene costruito un albero decisionale che viene fatto crescere fino al massimo livello possibile, ottenendo alberi con alta varianza ma basso bias. Quindi ipotizzando di generare B dataset di training si calcola per ognuno $\hat{f}^b(x)$, valore predetto nel punto x, ottenendo poi il Bagging come media

$$\hat{f}^{*b}{}_{AVG}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x)$$

Il Bagging è stato dimostrato avere un notevole miglioramento in termini di accuratezza combinando centinai o anche migliaia di alberi. L'altro punto di forza del Bagging, come anche dei Random Forest presentati nel paragrafo successivo, è che non necessitano di effettuare una cross-validation; infatti, utilizzando il metodo di campionamento boostrap è possibile calcolare una stima del test error, utilizzando i soggetti che di volta in volta non entrano nel training set. Queste osservazioni vengono definite *out-of bag* (OOB), ed ammontano a circa un terzo del campione

complessivo. Il risultante OOB error, calcolato sulle osservazioni out-of-bag, rappresenta una stima valida per il test error. Ovviamente questo metodo risulta particolarmente utile quando si lavori su datasets molto ampi per i quali la cross-validation sarebbe computazionalmente onerosa.

Per contro nel passare dal singolo albero decisionale al bagging perdiamo in interpretabilità, infatti non è possibile rappresentare la somma dei differenti alberi in un unico grafico. Esistono però degli indici che permettono di calcolare l'importanza delle variabili nel processo. In particolare possiamo, ancora una volta, ricorrere all'indice di impurità di Gini, un valore elevato di questo indice identifica una variabile importante dal punto di vista della classificazione.

4.3 Random Forest e Random Survival Forest

L'algoritmo Random Forest è stato proposta da Breiman all'inizio degli anni 2000 e fornisce un'ulteriore miglioramento rispetto al Bagging, in particolare adotta una tecnica per costruire alberi che risultino maggiormente diversificati rispetto a quelli costruiti nel Bagging.

Si supponga di disporre di un campione di n soggetti di cui è disponibile la classificazione in k classi secondo un dato fattore di interesse; su ognuno dei soggetti si misurano p variabili (categoriali o continue).

1. Si seleziona un campione di n soggetti (campionamento con reinserimento - boostrap) dal campione in esame. Questo set di dati verrà usato nella costruzione dell'albero come *training set*. I casi non selezionati (out-of-bag, oob) che rappresentano circa un terzo del totale, vengono usati per valutare l'error rate dell'albero e per stabilire l'importanza delle p variabili in sede di classificazione;

2. Si sceglie un numero k ≤ p. Ad ogni split k predittori sono selezionati casualmente e solo essi vengono usati per valutare la bipartizione ottimale;
3. Ogni albero viene fatto crescere fino alla sua massima estensione.

L'error rate globale della foresta dipende da due fattori: la correlazione esistente fra gli alberi e la bontà dei singoli alberi. Aumentando la correlazione aumenta l'error rate, mentre aumentando la bontà dei singoli alberi l'error rate diminuisce. Diminuendo il valore di k si riducono correlazione fra gli alberi e bontà dei singoli alberi, mentre aumentandolo si aumentano entrambi. Un valore di k intermedio risulterà quindi ottimale (generalmente nel caso di alberi di classificazione viene fissato un valore k pari alla radice quadrata del numero di variabili disponibili). Anche per nei Random Forest come per nel Bagging l'OOB costituisce una stima non distorta dell'error rate rendendo non necessario la validazione su un campione indipendente.

La procedura di Random Forest fornisce una stima di quali variabili siano importanti per la classificazione, offrendo la possibilità di selezionare solo un sottoinsieme che risulti ottimale dal punto di vista statistico. I due parametri che forniscono una stima dell'importanza di ciascuna variabile sono la "*Variable Importance*" e, ancora una volta, l'indice di "*Impurità di Gini*".

La *Variable Importance* viene costruito considerando per il j-esimo albero C_j il numero di casi classificati correttamente, si calcolano $CP_{ij,}$ numero di casi classificati correttamente dopo aver permutato casualmente i valori della i-esima variabile per i casi oob. Si calcola il valore di importanza della i-esima variabile I_{ij} come differenza $C_j - CP_{ij.}$

Si ripete il procedimento su tutti gli alberi della foresta e si fa la media dei valori ottenendo l'indice I_i che stima l'importanza della i-esima variabile. L'errore standard di tale stima viene valutato come se tali valori fossero indipendenti. Dividendo la stima I_i per il suo errore standard si ha la stima normalizzata dell'importanza della i-esima variabile.

Il Random Survival Forest è l'estensione del metodo dei Random forest nel caso di analisi della sopravvivenza. E' stato introdotto da Ishwaran e colleghi (Ishwaran H. , Kogalur, Blackstone, & Lauer, 2008) ed usa come metodo per la costruzione dei differenti alberi lo stesso pensato da Breiman per il Random Forest, ma utilizzando come criterio per la creazione dei nodi (split) il Log Rank.

4.4 Simulazione

Per esemplificare quanto descritto metodologicamente nei pragrafi precedenti si riporteranno qui i risultati dell'applicazione delle tecniche per la costruzione di alberi decisionali, bagging e random forest su dati simulati.

Si è creato un dataset simulato contenente 5000 osservazioni e 40 variabili. L'outcome è di tipo dicotomo con probabilità di accadimento pari al 15%. Le variabili predittrici sono continue e seguono una distribuzione di tipo normale, per ciascuna variabile è stato simulato anche il corrispettivo beta che associa ciascuna delle X all'outcome Y, secondo una distribuzione normale (media= 0, DS= 0.3).

Tabella 4.4-1: Dati simulati

Variabile	β	Rank
X2	-0.834	1
X22	-0.462	2
X25	0.451	3
X36	-0.409	4
X35	0.404	5
X17	-0.360	6
X8	-0.348	7
X24	-0.299	8
X14	0.290	9
X23	-0.281	10
X5	0.275	11
X26	0.243	12
X28	0.226	13
X29	-0.224	14
X27	0.204	15
X16	-0.199	16
X19	-0.183	17
X12	0.183	18
X30	-0.170	19
X32	-0.167	20
X18	0.163	21
X34	0.159	22
X4	-0.158	23
X7	0.152	24
X10	-0.145	25
X38	-0.104	26
X15	-0.101	27
X39	0.096	28
X1	-0.084	29
X21	-0.0668	30
X33	-0.0654	31
X11	0.065	32
X13	0.063	33
X6	-0.057	34
X40	-0.055	35
X37	-0.032	36
X31	0.029	37
X3	0.025	38
X20	-0.022	39
X9	0.009	40

Si considera un training set composto da 4000 osservazioni ed un test di 1000. Nel grafico 4.4-2 è rappresentato l'albero decisionale nella sua massima estensione costruito sul dataset di training, sulla base dell'andamento dell'error rate nel dataset test viene effettuata la procedura di pruning a 4 nodi.

Grafico 4.4-2: Albero decisionale - dati simulati

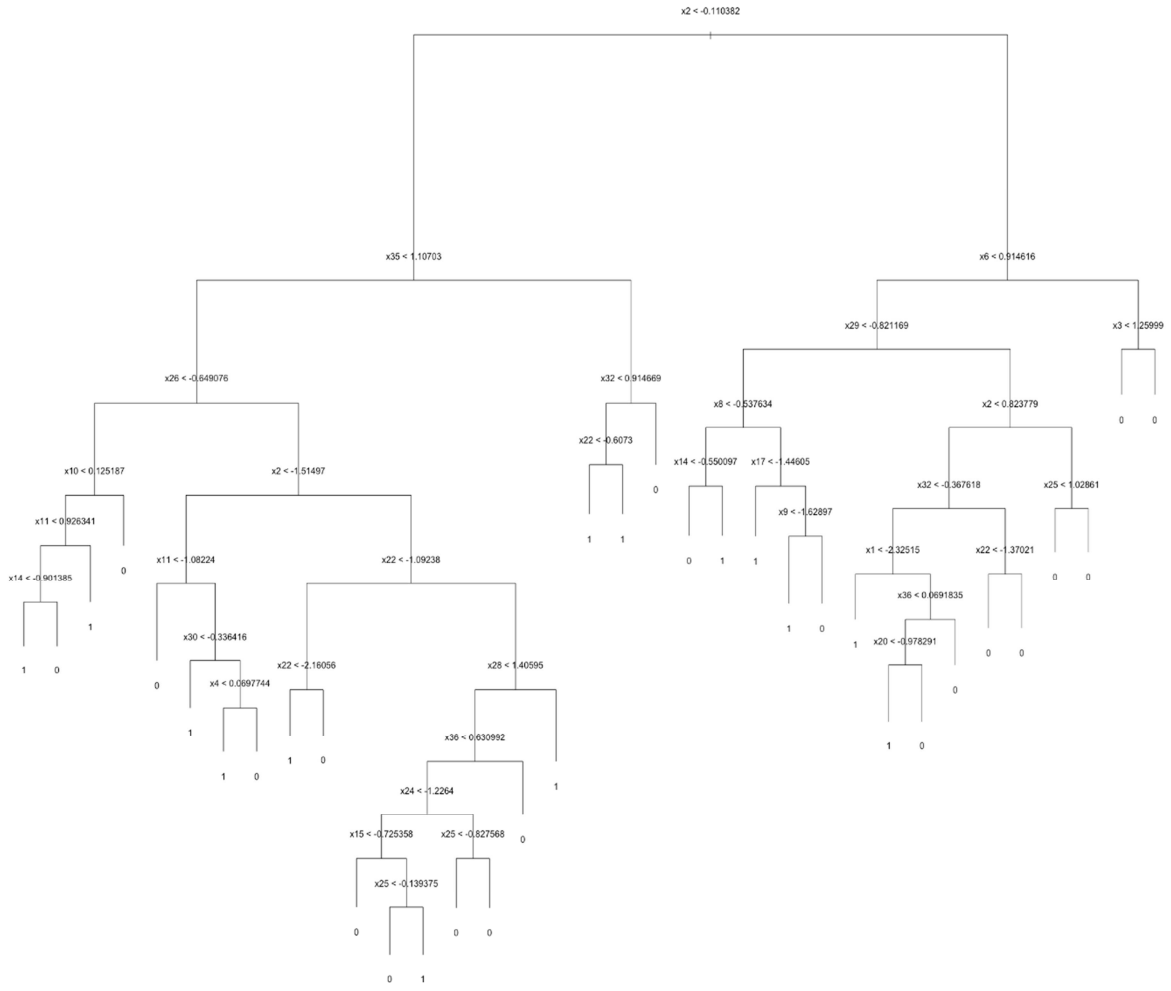
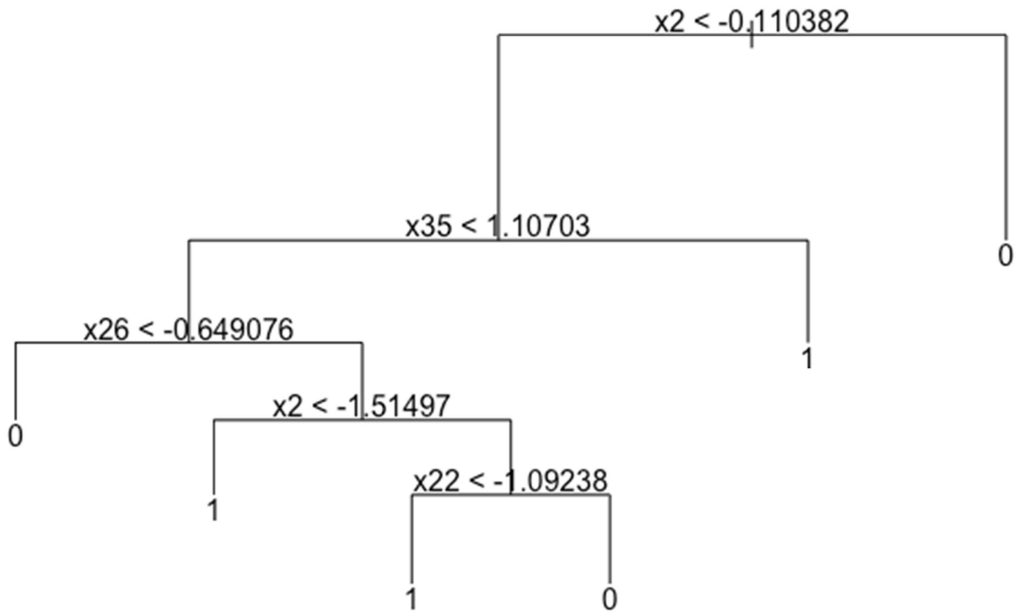


Grafico 4.4-3: Albero decisionale dopo procedura di pruning- dati simulati



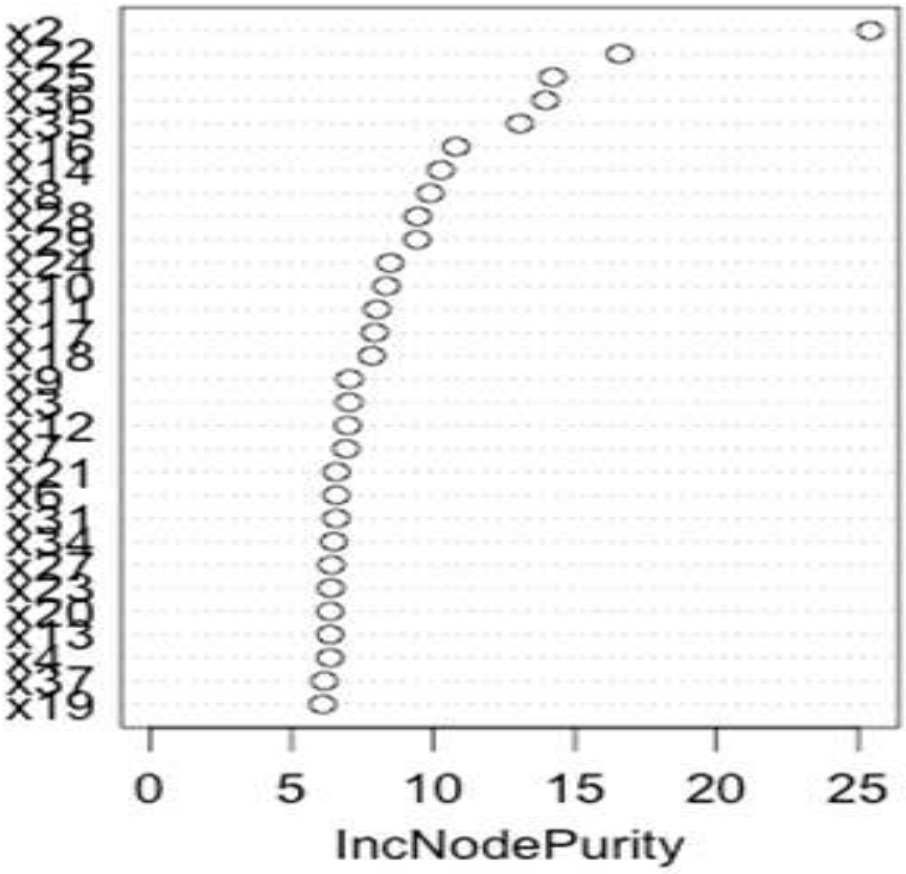
Dopo aver applicato la procedura di pruning le variabili più rilevanti per l'outcome risultano essere X2, X35, X26 e X22 che, dal punto di vista del ranking costruito sulla base dei parametri β, rappresentano rispettivamente la prima, la quinta, la dodicesima e la seconda variabile.

Come detto il singolo albero decisionale è soggetto a ampia variabilità. Ad esempio campionando casualmente altri 4000 soggetti per il dataset di training le variabili che risultano più rilevanti risultano essere X2, X22, X29 e X36.

Per diminuire questa variabilità applichiamo il bagging. Dato che iil bagging, come anche i random forest, generano un numero elevati di alberi per poi calcolarne la media, non è possibile rappresentare la procedura utilizzando il grafico ad albero come fatto per l'algoritmo precedente; per decidere quali siano le variabili maggiormente implicate nella predizione

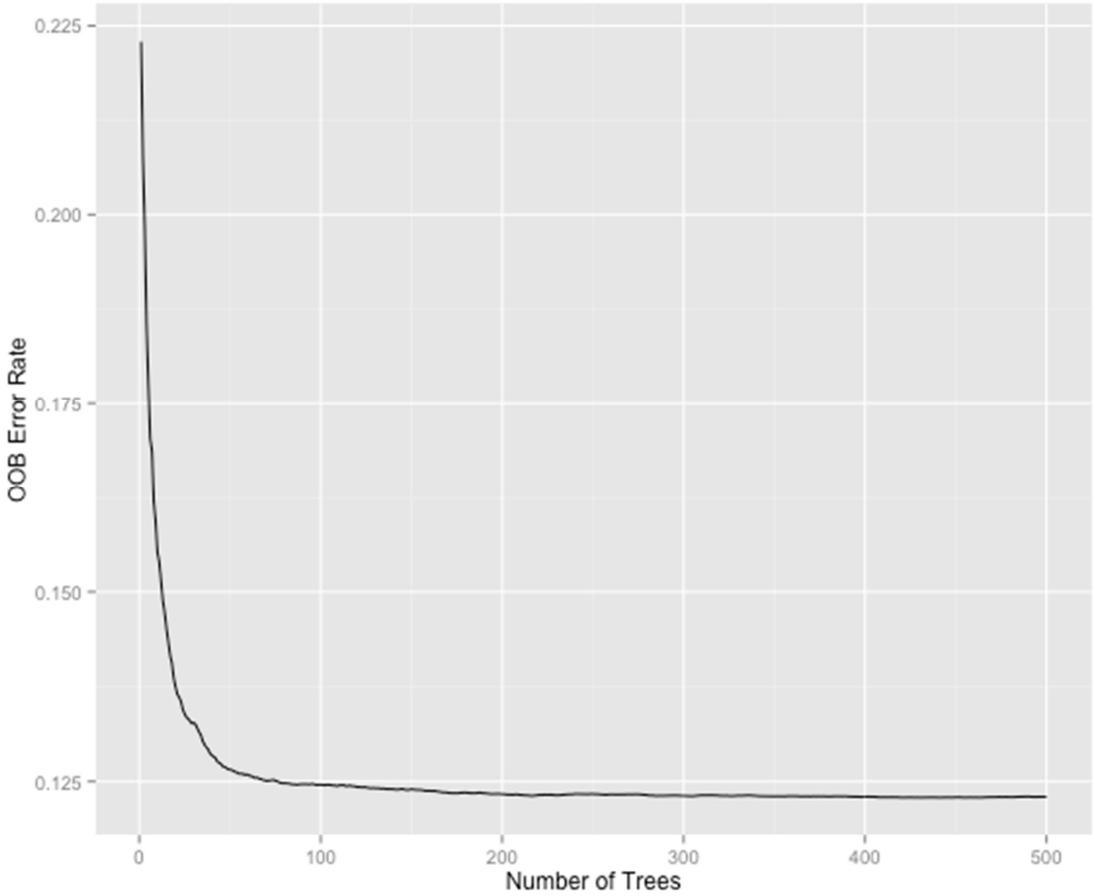
dell’outcome si utilizzano, essendo Y in questa simulazione dicotoma, l’indice di Gini.

Grafico 4.4-4: Bagging, indice di impurità di Gini - dati simulati



Applichiamo ora ai dati simulati l'algoritmo Random Forest, è necessario definire il numero di item che entrano nella definizione di ogni nodo, questo valore per le analisi di classificazione come la radice del numero di variabili totale, nel nostro caso sarà quindi 6. L'altro parametro da definire è il numero di alberi da costruirsi, questo valore viene definito quel valore in corrispondenza del quale l'OOB si stabilizza.

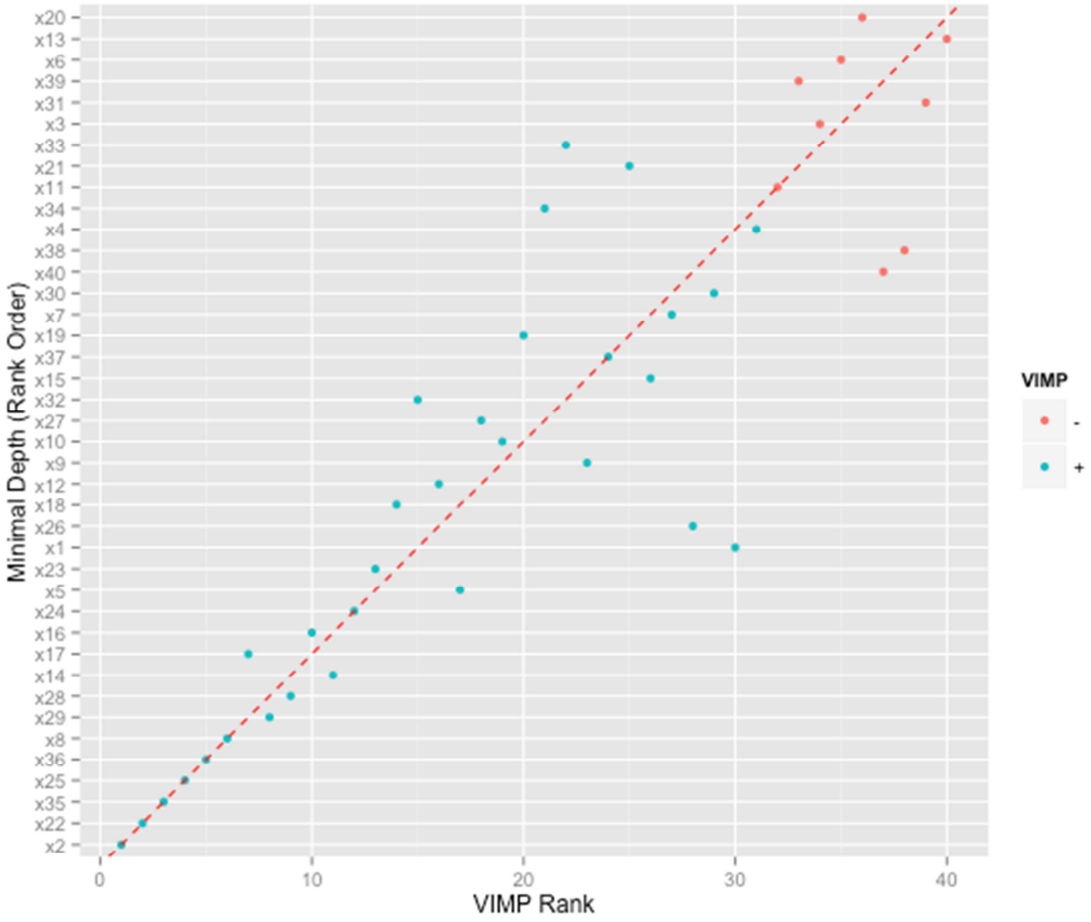
Grafico 4.4-5: Random Forest, andamento del numero di alberi in funzione dell'Out of Bag error - dati simulati



Sulla base del grafico 4.4-5 si generano 200 alberi.

Per definire quali variabili siano maggiormente rilevanti si analizzano due parametri: la variable importance e la minimal depth che possono essere rappresentate contemporaneamente in unico grafico.

Grafico 4.4-6: Random Forest, Variable importance e Minimal Depth - dati simulati



Rispetto all'ordinamento delle variabili atteso sulla base della simulazione, si osserva che considerando le prime 5 posizioni, l'albero decisionale semplice ne riconosce 3 ma solo la prima (X2) nell'ordinamento corretto, il bagging 4 collocandone correttamente le prime 2, il random forest riconosce e colloca correttamente tutte e 5 le variabili.

Nel successivo capitolo, verrà applicata la tecnica dei Random Forest nella sua estensione Random Survival Forest ai dati delle coorti a nostra disposizione.

5. SELEZIONE DEGLI ITEM

Per identificare, tra i quasi 50 item delle scale psicosociali disponibili, quelli di maggior rilevanza al fine di predire un evento cardiovascolare si è applicata la procedura di Random Survival Forest (RSF). Questo metodo ha permesso di effettuare l'analisi dei dati senza: i) definire a priori il tipo di relazione tra item e outcome (lineare o meno), ii) effettuare trasformazioni sugli item per migliorarne l'adattabilità al modello e, non meno importante, iii) consentendo una gestione agevole dei dati mancanti direttamente all'interno dell'algoritmo di classificazione.

L'analisi è stata condotta sui 6238 soggetti appartenenti alla coorti MONICA-Brianza e PAMELA, di età 25-65 anni e liberi da malattia cardiovascolare al momento del reclutamento. Durante un follow-up mediano di 17 anni si sono osservati 438 eventi cardiovascolari. Sono stati esclusi dalla fase di selezione delle variabili i soggetti della coorte SEMM per la quale erano disponibili solo il Job Content Questionnaire e le scale su caratteristiche lavorative e carico familiare.

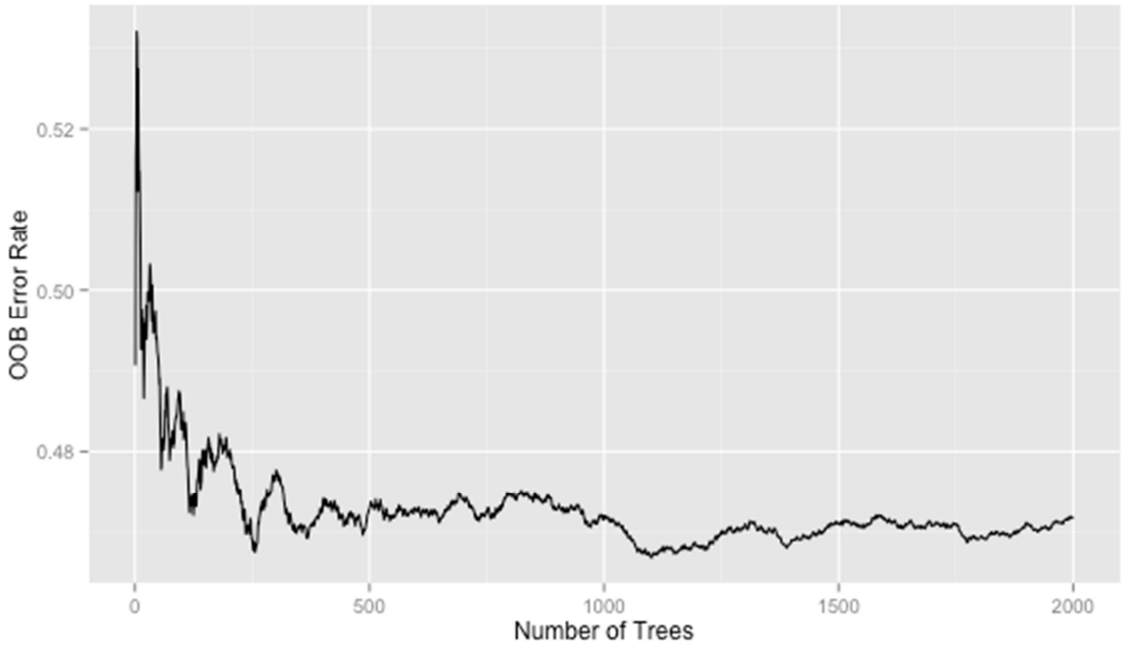
Poiché uno dei principali problemi delle foreste casuali è la difficoltà di cogliere segnali in presenza di dati fortemente sbilanciati (cioè, come nel nostro caso, quando gli eventi sono molti meno dei non eventi) si è scelto di lavorare su un dataset in cui ad ogni evento è stato appaiato un non-evento dello stesso sesso, età (±5 anni) e in osservazione al momento dell'insorgenza dell'evento. Dato il basso numero di eventi nelle donne si è deciso di considerare nell'analisi solo gli uomini.

Il campione consta quindi di 626 soggetti (313 eventi).

La procedura RSF è stata applicata estraendo 1000 campioni boostrap dal datasets originale, il numero di campioni è stato definito dopo analisi dell'andamento dell'Error rate in funzione del numero di alberi (figura 5.1.1). Come da definizione della procedura, da ciascun campione

rimaneva escluso circa un terzo delle osservazioni, queste rappresentano i dati out-of-bag (OOB) usati dalla procedura RSF per effettuare la cross validation.

Figura 5.1.1: Error rate per la procedura RSF. Uomini 35-65 anni

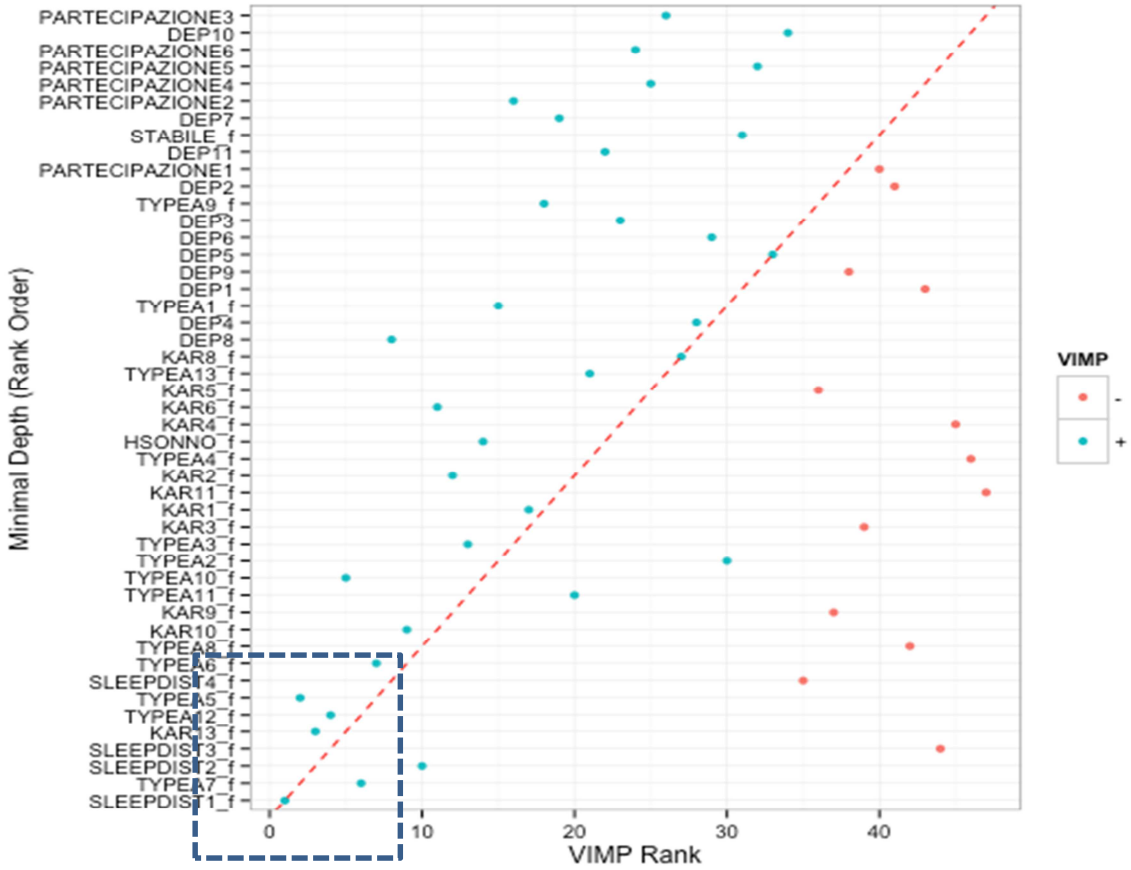


Per ciascun campione è stato costruito un *survival tree* partendo dai 47 item delle scale raccolte. In corrispondenza di ciascun nodo dell'albero un campione casuale di 7 dei 47 item disponibili è stato utilizzato per la ramificazione (7 è equivalente alla radice quadrata del numero di variabili disponibili, come definito da Ishwaran e colleghi nel lavoro del 2008). L'item in grado di identificare i due gruppi con massima differenza in termini di sopravvivenza veniva selezionato.

Al termine della procedura la rilevanza delle variabili è stata valutata utilizzando la **Variable Importance (VIMP)** ed il **Minimal Depth** (figura 5.1.2).

L’analisi è stata effettuata utilizzando il pacchetto “RandomSurvivalForest” implementato nella versione 3.2.0 di R (Ishwaran & Kogalur, 2014).

Figura 5.1.2: Variable Importance e Minimal Depth plot. Uomini 35-65 anni



Per l’individuazione degli item maggiormente indicativi di rischio cardiovascolare si sono considerati quelli che risultavano avere punteggi migliori sia per Minimal Depth che per Variable Importance (quadrato in basso a destra del grafico) ottenendo quindi 8 item: 2 relativi alla scala sui disturbi del sonno (SLEEPDIST1 = difficoltà ad addormentarsi e SLEEPDIST2 = difficoltà a rimanere addormentato), 2 item del Job Content Questionnaire (KAR10 = lavorare duramente e KAR13 = presenza di ordini contrastanti) e 4 item della scala relativa ai tratti di personalità (essere considerati grintosi e competitivi in gioventù -TYPE5, e

attualmente - TYPE6, TYPE7 = avere un livello di attività elevato e TYPE12 = affrontare la vita più seriamente rispetto alle media delle altre persone). Si sono, inoltre, considerati anche 4 item aggiuntivi che risultavano avere alta concordanza rispetto ai due parametri usati per valutare la rilevanza delle variabili (valori molto prossimi alla diagonale del grafico): due ulteriori item del JCQ (KAR1 = il mio lavoro richiede che impari cose nuove; KAR8 = il mio lavoro è sicuro), due item della scala sulla depressione/vital exhaustion (DEP 4 = sentirsi spesso completamente facco; DEP5 = sentirsi vecchio dal punto di vista dell'aspetto fisico).

In una prima fase è stata effettuata un'analisi di associazione tra disturbi cardiovascolari maggiori e le scale i cui item sono stati estratti dalla procedura RSF, in particolare, nel capitolo 6 sono riportati i risultati dell'analisi dell'associazione tra eventi cardiovascolari e disturbi del sonno (*paper in press*, par. 6.1) e tra eventi cardiovascolari e Job Content Questionnaire (*paper under review*, par.6.2). Nel capitolo 7 è stata impostata la costruzione di un indicatore che tenga conto di tutti gli item estratti con l'algoritmo RSF valutandone la rilevanza sulla predizione del rischio CVD.

6. APPLICAZIONI SU SINGOLE SCALE

6.1 Studio dell'associazione tra rischio di evento cardiovascolare maggiore e disturbi del sonno

6.1.1 Campione in studio e scala per la valutazione dei disturbi del sonno

Per lo studio dell'associazione tra eventi cardiovascolari maggiori e disturbi del sonno si sono, inizialmente considerati tutti i soggetti appartenenti alle tre coorti MONICA ed alla coorte PAMELA di età compresa tra 25 e 64 anni. Dato il basso numero di eventi tra le donne e negli uomini sotto i 35 anni questi soggetti non sono entrati nell'analisi; si sono inoltre esclusi gli uomini con storia pregressa di CVD al basale (n=211), con nessun item compilato per il questionario del sonno (n=270) e con informazioni incomplete sui fattori di rischio cardiovascolare (n=34). Il campione è quindi costituito da 2277 uomini che in circa 17 anni di follow-up mediano hanno generato 293 primi eventi CVD (214 coronarici e 96 stroke ischemici).

Per la valutazione dei disturbi del sonno è stato utilizzato il questionario Jenkins, presentato nel paragrafo 3.2.3. Dopo aver calcolato lo score totale come somma dei 4 item si sono costruite tre categorie che identificano livelli crescenti di disturbo: la categoria di riferimento *none/some* (score compreso tra 0 e 9), *moderate* (10-14) e *severe* (15- 20). Il numero di ore dormite mediamente per notte sono state categorizzate come: $\leq$ 6 h, 7-8 h (reference) e $\geq$ 9 h.

6.1.2 Fattori di rischio cardiovascolare misurati al basale

I fattori di rischio cardiovascolare sono stati valutati alla visita basale secondo le procedure standardizzate del progetto WHO-MONICA (http://www.ktl.fi/publications/monica/manual/index.htm).

La pressione arteriosa è stata misurata, da personale opportunamente istruito, su soggetti seduti e a riposo per almeno 10 minuti, utilizzando un sfigmomanometro a mercurio standard; la media delle due misurazioni prese a 5 minuti di distanza l'una dall'altra è stata utilizzata come valore della pressione sistolica. Campioni di sangue venoso sono stati prelevati nei soggetti a digiuno (12h o più). I campioni, refrigerati a -4 ° C sono stati spediti entro 4 ore al Dipartimento di Patologia Clinica dell'Ospedale di Desio dove sono stati misurati colesterolo totale, colesterolo HDL e glicemia con un metodo enzimatico.

Un questionario standardizzato riguardante le abitudine e lo stile di vita dei soggetti è stato somministrato ai partecipanti da intervistatori addestrati, raccogliendo informazioni su abitudine al fumo (utilizzato nell'analisi dopo aver dicotomizzato la variabile in *fumatore attuale vs altro*).

La diagnosi di Diabete Mellito è stata posta in caso di diagnosi di diabete riferita dal soggetto, uso di insulina e/o ipoglicemizzanti orali o glicemia a digiuno superiore a 126 mg / dl.

La presenza di un evento cardiovascolare pregresso è stata valutata chiedendo ai partecipanti informazioni su eventuali ospedalizzazione per infarto miocardico, angina pectoris instabile, rivascolarizzazione cardiaca o ictus.

Il livello di istruzione è stata accertata durante l'intervista, e dicotomizzato come "basso" (al più scuola secondaria di prima grado) e "alta" (liceo o superiore).

L’attività fisica durante il tempo libero (LTPA) è stata valutata mediante il questionario Baecke, e l'indice di LTPA è stato classificato in tre livelli: basso, medio e alto (punteggi 0-2, 3, 4-5, rispettivamente).

La depressione è stata valutata utilizzando la versione a 14-item del Maastricht Vital Exhaustion Questionnaire (MQ), già descritto nel paragrafo 3.2.6.

6.1.3 Analisi statistica

Data la presenza di 411 uomini con almeno 1 item del questionario Jenkins mancante (pari al 18%) si è utilizzata una procedure di Multiple Imputation (SAS MI ANALYZE) per massimizzare il numero di dati disponibili, per un totale di un 6% di item imputati.

Per i maggiori fattori di rischio cardiovascolare è stata calcolata la media o la prevalenza, aggiustata per età, nelle diverse categorie di disturbo e di durata del sonno (tabella 6.1-1) utilizzando un modello di regressione lineare o logistico.

Sono state costruite le curve di Kaplan-Meier ed il log-rank test è stato utilizzato per confrontare la sopravvivenza all’evento tra le categorie dei disturbi e della durata del sonno (grafico 6.1-2). L’associazione tra evento CVD o CHD e disturbi (categoria di riferimento *none/some*) e durata del sonno (categoria di riferimento 7-8h) è stata valutata con un modello di COX considerando l’età sulla scala del tempo. Le stime sono state calcolate considerando sia un modello che prevedeva l’aggiustamento per pressione sistolica, colesterolo totale e HDL, abitudine al fumo, presenza di diabete e classe educazionale; che un secondo modello nel quale oltre alle covariate precedenti sono stati inseriti anche la scala per la valutazione della depressione e l’attività fisica nel tempo libero.

L’effetto congiunto di disturbi e durata del sonno è stato valutato nelle durate di sonno brevi costruendo una variabile così definita:

- ✓ dormire 7-8 h e non avere disturbi (categoria di riferimento);
- ✓ dormire 7-8 h e avere disturbi (moderati o severi)
- ✓ dormire non più di 6 ore e non avere disturbi
- ✓ dormire non più di 6 ore e avere disturbi (moderati o severi)

dato il basso numero di soggetti che riferivano di avere disturbi del sonno pur dormendo a lungo, questa doppia caratterizzazione non è stata ulteriormente indagata.

6.1.4 Risultati

Caratteristiche del campione

Nella tabella 6.1-1 sono descritte le caratteristiche dei 2277 uomini inclusi nell'analisi. La prevalenza di disturbi del sonno moderata e severa sono 8.3% e 3.5%, rispettivamente. La prevalenza di soggetti con durata del sonno breve (meno di 6 ore al giorno) o lunga (9 ore o più) sono 23.3% e 9%, rispettivamente.

Il gruppo di soggetti con disturbi del sonno moderati o severi è tendenzialmente più anziano, meno istruito e con una maggior prevalenza di diabetici e soggetti ipertesi. La prevalenza dei fumatori è più alta tra gli uomini che riferiscono disturbi del sonno severi. Infine, la prevalenza dei sintomi critici di depressione aumenta all’aumentare dei disturbi del sonno. Rispetto agli uomini che dormono 7-8 ore al giorno, gli uomini che dichiarano di dormire 9 ore o più sembrano essere più anziani, meno istruiti, con maggior prevalenza di ipertesi, diabetici, e con sintomi di depressione grave.

La distribuzione congiunta di durata e disturbi del sonno evidenzia, tra coloro che hanno riferito di dormire meno di 6 ore, un 25.0% di soggetti

che presenta disturbi del sonno da moderati a gravi, mentre tra coloro che dormono almeno 9 ore ne soffre solo il 5.4%.

Associazione tra disturbi del sonno e CVD

Nei 17 anni di follow-up mediano (range interquartile: 14.5-19 anni), si sono osservati 293 primi eventi cardiovascolari (214 coronarici e 96 ictus ischemici). Come riportato in Tabella 6.1-3 - Modello 1, un significativo aumento del rischio di primo evento cardiovascolare è stato osservato per i disturbi severi (HR = 1.80; 95% CI: 1.07-3.03) rispetto alla categoria *none/some*, indipendente da fattori di rischio CV e durata del sonno. L'eccesso di rischio è stato in parte ridotto dopo aggiustamento per la LTPA e depressione (Modello 2 - HR = 1.71; 0.99-2.94). Risultati simili sono stati osservati limitando l'analisi ai soli eventi CHD, con un rischio più elevato per i soggetti con disturbi severi (HR = 1.97; 1.09-3.56), ridotto a 1.83 (0.98-3.41) aggiustando per depressione e LTPA.

Un aumento significativo del rischio di malattia cardiovascolare è stato rilevato per gli uomini che dormono 9 o più ore (1.56; 1.10-2.22) rispetto alle 7-8 ore canoniche, i risultati non si modificano dopo aggiustamento per LPTA e depressione. Durate del sonno brevi non hanno mostrato alcun incremento rilevante del rischio: CVD (HR: 1.14, IC 95%: 0.85-1.53) e CHD (HR: 1.14; 0.81-1.61), non modificati dall'aggiustamento per LPTA e punteggi di depressione.

Tabella 6.1-4 visualizza l'effetto congiunto di breve durata e qualità del sonno sull'incidenza di eventi cardiovascolari e di malattia coronarica. Rispetto agli uomini che dormono 7-8 ore, senza disturbi (categoria di riferimento), un aumento del rischio di evento cardiovascolare è stato osservato tra coloro che dichiarano di dormire 6 ore o meno in presenza di

disturbi moderati o severi (HR disturbi = 1.69; 1.08-2.64). Al contrario, la durata breve in assenza di disturbi o durate del sonno “normali” con presenza di disturbi moderati-severi non ha mostrato un aumento significativo del rischio. Quando si considera il rischio di CHD, l'effetto congiunto è ancora più pronunciato (HR = 2.24; 1.39-3.63).

Età all’evento e CVD

Le curve di Kaplan-Meier (Figura 6.1-2) mostrano l'impatto dell'età di insorgenza di eventi cardiovascolari tra le categorie di disturbi del sonno e tra i livelli di durata del sonno. La presenza di disturbi del sonno porta ad una diminuzione della probabilità di sopravvivenza all’evento (log-rank test p-value 0.02) già prima dei 50 anni d’età. Il test overall significativo per le diverse categorie di durata del sonno ($p < 0.01$) è imputabile ad una differenza significativa tra 9 o più ore e 7-8 ore ($p = 0.002$), a partire dall'età di 60 anni. Quest'ultima osservazione indica che l'aumento del rischio tra coloro che dormono per 9 ore o più, sembra essere rilevante solo per gli anziani.

6.1.5 Discussione

In uno studio di coorte prospettico, su una popolazione con bassa incidenza di CVD e CHD, con un lungo periodo di follow-up (mediana 17 anni), abbiamo confermato l'aumento del rischio cardiovascolare tra gli uomini con disturbi del sonno severi, indipendentemente da altri fattori di rischio; tale aumento del rischio è principalmente determinato dagli eventi coronarici. Durate del sonno brevi non erano significativamente associate con un aumentato del rischio di eventi cardiovascolari, mentre considerando congiuntamente durate del sonno brevi e disturbi del sonno si osserva un aumento del rischio di malattie cardiovascolari di quasi il 70%

ed un aumento del rischio più che doppio considerando i soli eventi coronarici. È interessante notare che l'effetto dei disturbi del sonno severi sull'insorgenza di evento cardiovascolare era evidente già all'età di 48 anni. L'aggiustamento per attività fisica nel tempo libero e depressione non hanno modificato le associazioni osservate nel nostro campione.

In confronto con la grande maggioranza degli studi precedenti, i nostri risultati sono caratterizzati dall'avere utilizzato la scala Jenkins (4 item) per la valutazione dei disturbi e la valutazione contemporanea della durata del sonno, che permette di stimare l'associazione con CVD delle due esposizioni simultanee. Inoltre i risultati suggeriscono che l'aumento del rischio cardiovascolare legato a gravi disturbi del sonno può iniziare presto nella vita, già all'età di 50 anni, rimanendo poi stabile nelle età successive.

Alcuni autori suggeriscono che queste associazioni possano essere l'effetto di confondimento non misurato a causa di depressione, basso status socio-economico o inattività fisica, o che ci possa essere un effetto di *reverse causation* data dalla presenza di una malattia cardiovascolare subclinica. I risultati di questo studio forniscono un importante contributo per far luce su questi temi. In primo luogo, abbiamo controllato per tutti i principali fattori e, in particolare, abbiamo trovato un effetto inferiore all'atteso del punteggio della depressione. Inoltre, un'analisi stratificata per durata del follow-up (<10 anni />10 anni) ha rivelato che, mentre il punteggio di depressione è associata con l'endpoint nei primi anni di follow-up ma non nei periodi successivi, le caratteristiche del sonno hanno mostrato la stesse associazioni in entrambi i periodi di follow-up. Questo risultato porterebbe ad escludere la presenza di reverse causation.

I punti di forza di questo studio, sono rappresentati dal disegno longitudinale e dalla lunga durata del follow-up, dalla standardizzazione della valutazione dei fattori di rischio, dalla validazione degli eventi secondo standard codificati e dall'inclusione di una vasta gamma di potenziali variabili confondenti. Inoltre, i partecipanti sono stati selezionati in modo casuale dalla popolazione generale, il che rafforza la generalizzabilità e la validità esterna delle associazioni osservate. Purtroppo a causa del basso numero di eventi nelle donne abbiamo dovuto restringere l'analisi ai soli uomini.

In conclusione abbiamo trovato un aumento del rischio di primo evento CVD e CHD in soggetti con disturbi del sonno severi e in coloro che dormono più delle 7-8 h canoniche, così come in coloro che dormono poco in presenza di disturbi del sonno anche moderati. L'eccesso di rischio era indipendente da tutti i fattori classici di rischio cardiovascolare e da depressione. Il questionario Jenkins sembra essere uno strumento facile e veloce per discriminare livelli di disturbi del sonno in relazione alla valutazione del rischio CVD. Pertanto, può essere fattibile l'adozione di routine in programmi di prevenzione CVD per identificare gli uomini ad aumentato rischio di insorgenza precoce di evento. L'utilità clinica della valutazione dei diversi livelli di disturbi del sonno e il loro contributo supplementare negli algoritmi di stratificazione del rischio di CVD dovrebbe essere valutata in ampi studi prospettici tra cui donne e gli individui tra i diversi gruppi di età.

6.1.6 Approfondimento: analisi biomarcatore C-Reactive protein

Come anticipato nell'introduzione, interesse di questa tesi era anche confrontare le informazioni soggettive provenienti dai questionari con i dati obiettivi ricavati dall'analisi di biomarcatori specifici. In letteratura, spesso

è riportato come possibile spiegazione all'associazione osservata tra disturbi del sonno ed insorgenza di eventi cardiovascolari la presenza di uno stato infiammatorio generale del soggetto. Tra i biomarcatori a nostra disposizione abbiamo considerato la C-Reactive protein (CRp-us), la cui concentrazione elevata è indicativa della presenza di flogosi. L'analisi è stata effettuata sia sui disturbi che sulla durata del sonno anche se in letteratura solo i primi sembrerebbero influenzati dalla presenza di uno stato infiammatorio e quindi da una maggior concentrazione di C-Reactive protein (Liu, et al., 2014).

Metodi. I dati sui biomarcatori sono disponibili solo per le tre coorti MONICA, quindi l'analisi è stata limitata ai soli soggetti reclutati in queste coorti ed in età compresa tra i 35 e 74 anni per uniformarsi alle fasce d'età considerate nell'analisi presentata nei paragrafi precedenti. L'analisi è stata condotta su uomini e donne congiuntamente.

E' stata effettuata una trasformazione logaritmica sul valore della CRp-us per ricondursi ad una distribuzione normale. Dopo verifica dell'omoschedasticità utilizzando il test di Bartlett, un'analisi della varianza è stata applicata per testare l'esistenza di una differenza significativa in termini di concentrazione di CRp-us tra le classi dei disturbi e di durata del sonno. Sono inoltre stati costruiti i contrasti ortogonali per confrontare disturbi assenti vs. disturbi moderati/severi e durata del sonno 7-8h vs durate del sonno minori (meno di 6 ore) o maggiori (9 o più ore).

Risultati. L'analisi è stata effettuata su un totale di 3672 soggetti (1785 uomini e 1885 donne) per un numero di eventi cardiovascolari pari a 337 (242 negli uomini e 95 nelle donne). Il dato sulla CRp-us era disponibile per 3462 di essi, dopo aver eliminato anche i soggetti con

questionario sui disturbi del sonno mancante il campione di analisi si è ulteriormente ridotto a 2473 (1274 uomini e 1199 donne) e 205 eventi CVD (150 negli uomini e 55 nelle donne). In tabella 6.1.6-1 si riportano i valori delle concentrazioni di CRp nelle classi di disturbi e di durata del sonno.

L'analisi della varianza condotta sul totale del campione suggerisce l'esistenza di una differenza nella concentrazione di CRp-us nelle classi di disturbi del sonno, in particolare si osserva una differenza significativa tra disturbi moderati/severi e nessun/alcuni disurbi ($F_{\text{Moderato/severo vs Nessuno/qualche}} = 4.78$ p-value=0.03) mentre non c'è evidenza di differenti concentrazioni di CRp-us tra soggetti con disturbo severo e soggetti con disturbo moderato ($F_{\text{Severo vs Moderato}} = 0.04$ p-value=0.83).

L'analisi rispetto alla durata del sonno ha evidenziato l'esistenza di una differenza significativa nelle concentrazioni di CRp considerando il contrasto tra dormire 7-8 ore e dormirne di meno o di più (F=10.1 p-value=0.002) mentre il confronto tra dormire meno di 6 ore o dormirne 9 o più non è risultato significativo (F=3.12 p-value=0.07).

Concludendo, l'ipotesi che esista una relazione tra disturbi del sonno e concentrazione di CRp sembra supportata dai nostri dati. Inoltre, i risultati suggerirebbero che la stessa relazione possa essere presente anche per la durata del sonno.

6.2 Associazione tra rischio cardiovascolare e stress lavorativo percepito

6.2.1 Campione in studio e questionario adottato per la misura dello stress lavorativo

Per lo studio dell'associazione tra rischio cardiovascolare e stress lavorativo percepito si sono considerati tutti i lavoratori reclutati nelle tre indagini MONICA, nello studio PAMELA e nella coorte SEMM dei lavoratori del Comune di Milano.

Dopo aver escluso i soggetti con dati mancanti per: i) fattori di rischio cardiovascolare, ii) questionario stress lavorativo; iii) informazioni sulla classe occupazionale e ristretto il campione ai soli lavoratori uomini di età compresa tra 25 e 64 anni di età con esclusione dei dirigenti della coorte SEMM a causa della bassa numerosità e dell'esiguo numero di eventi (77 uomini e 9 eventi) il campione complessivo per l'analisi era costituito da 4029 occupati.

Il Job Content Questionnaire (JCQ) è stato utilizzato per valutare il livello di stress (strain) lavorativo. Il questionario è stato somministrato a tutti i lavoratori, utilizzando due diverse versioni. Per le coorti MONICA Brianza e PAMELA è stata utilizzata la versione ridotta del questionario composta da con 5 item per valutare la domanda psicologica lavorativa (PJD) e 6 item per la libertà decisionale (DL). La versione estesa del JCQ è stata, invece, adottata nella coorte SEMM. Le due versioni sono state uniformate secondo i protocolli standard utilizzando pesi opportuni arrivando ad ottenere punteggi di DL e PJD comparabili (Karasek, Choi, Ostergren, Ferrario, & de Smet, 2007). Secondo la procedura standard sono state derivate le tradizionali quattro categorie JCQ basate sul valore delle mediane di DL e PJD nel campione e si sono definiti ad "Alto Strain" i soggetti che presentavano valori della PJD superiore alla mediana e valori

di DL pari o inferiore alla mediana. Le altre tra classi di strain (Attivi, Passivi e Basso strain) sono state utilizzate come unica categoria di riferimento.

La classe occupazionale è stata definita sulla base delle variabili: occupazione attuale, mansione e settore lavorativo e classificandole in base alle classi Erikson-Goldthorpe-Portocarero (EGP). Le classi EGP originali sono state, per motivi legate alla numerosità raggruppate in due macro categorie: "Managers and Employers" comprendente amministratori, dirigenti e lavoratori autonomi; "White and Blue Collars" comprendente impiegati e operai specializzati e non.

6.2.2 Fattori di rischio cardiovascolare misurati al basale

Per la descrizione delle procedure adottate per la misurazione dei fattori di rischio all'ingresso nella coorte si rimanda al paragrafo 6.1.2.

6.2.3 Endpoint dello studio e procedure di follow-up

Tutti i soggetti sono stati seguiti dall'ingresso nelle rispettive coorti fino al verificarsi di una delle seguenti condizioni: i) primo evento CHD, ii) emigrazione, iii) decesso, iv) 80-esimo anno di età o v) 31 dicembre 2008.

Il follow-up è stato completato per il 97,6% dei soggetti. Lo stato in vita è stato attivamente indagato per tutti i soggetti, compresi quelli che emigrano in altre città italiane, ed i certificati di morte sono stati ottenuti dalle ASL di competenza. Gli eventi fatali sospetti sono stati identificati sulla base dei seguenti codici ICD-IX: 410-414 per eventi CHD. I sospetti eventi non fatali sono stati identificati sulla base dei codici di dimissione ospedaliera: 410- 411 per eventi coronarici acuti, e codice intervento 36.0-9 per rivascolarizzazione coronarica. Tutti gli eventi fatali e non fatali sono

stati validati secondo i criteri diagnostici MONICA. L'endpoint dello studio è il verificarsi del primo evento coronarico acuto (infarto miocardico, sindrome coronarica acuta) o rivascolarizzazione coronarica.

6.2.4 Analisi statistica

Abbiamo calcolato la media (prevalenza) dei principali fattori di rischio CHD per classi di EGP e categorie di strain, il modello ANOVA e il test chi-quadrato sono stati adottati per testare, rispettivamente, le differenze tra i gruppi per le variabili quantitative e qualitative.

Per valutare la validità di costrutto dei questionari raccolti è stata effettuata un'analisi fattoriale esplorativa applicando ai fattori una rotazione varimax per massimizzare l'interpretabilità dei risultati. La consistenza interna dei costrutti specifici è stata valutata calcolando il coefficiente α di Cronbach e il coefficiente di correlazione di Pearson è stato adottato per valutare le correlazioni tra costrutti. Come già descritto nel paragrafo 6.2.1 le categorie di strain sono state calcolate sulla base dei valori mediani del campione complessivo di DL (= 37) e PJD (= 30).

Un modello di COX con l'età sulla scala del tempo è stato adottato per studiare le associazioni tra il rischio di eventi CHD e stress lavorativo, utilizzando come categoria di riferimento la macro-classe costituita da passivo, attivo e basso strain. I modelli sono stati aggiustati per i principali fattori di rischio cardiovascolari (pressione arteriosa, colesterolo totale e LDL, diabete, abitudine al fumo).

Sono state condotte analisi stratificate per i sottogruppi del campione, vale a dire le coorti di popolazione (MONICA Brianza e Pamela) e la coorte di lavoratori (SEMM); e per classi EGP aggregate.

Le analisi sono state eseguite utilizzando il Software per analisi statistiche SAS (versione 9.3, SAS Institute Inc, Cary, NC). I grafici sono stati realizzati utilizzando il software R (http://www.R-project.org/).

6.2.5 Risultati

I 4029 uomini lavoratori tra i 24 ed o 65 anni, liberi da malattia coronarica al basale, hanno generato in circa 15 anni di follow-up mediano 163 eventi coronarici maggiori, di cui 21 fatali. La tabella 6.2-1 mostra le distribuzioni delle variabili socio-demografiche e delle covariate utilizzate nell'analisi per il totale del campione e per tipo di coorte e classe EGP. I "White &blue collars" mostrano sia per la coorte MONICA-PAMELA che per la coorte SEMM età inferiori rispetto ai manager ed ai lavoratori autonomi e, solo nel primo campione, anche un livello educazionale inferiore. Come atteso, per costruzione, dal modello JCQ il 25% del totale del campione ricade nella classe "*high job strain*"; dall'analisi per coorte-classe EGP la più alta prevalenza di HS è stata osservata tra i "White&blue collars" della coorte SEMM; *attivi* e *low strain* sono preminenti nel gruppo di "Managers&Employers" mentre i passivi prevalgono nei "White&blue collars" indipendentemente dal tipo di coorte.

Con l'eccezione della percentuale dei fumatori (alta nell'intero campione e in tutte le coorti), tutti gli altri fattori di rischio cardiovascolare si distribuiscono diversamente per tipo di coorte e classe EGP (tabella 6.2-1A), ma non tra categorie JCQ (tabella 6.2-1B). I manager e datori di lavoro delle coorti di popolazione hanno dimostrato e più alta prevalenza di diabete portato e superiori valori medi di pressione arteriosa sistolica e colesterolo totale. Bianco e blu gruppo della coorte SEMM evidenziato mezzi più bassi di colesterolo HDL.

La tabella 6.2-2 mostra i risultati dell'analisi fattoriale. Considerando per gli autovalori accettabile un valore superiore a 0.30 (come suggerito dalla letteratura) emerge come 1 item della DL (ripetitività delle operazioni) e due di PJD (lavorare velocemente e lavorare molto duramente), non contribuiscano alla caratterizzazione dei due costrutti, Factor 1 (DL) e Factor 2 (PJD). Il JCQ "ridotto" esclude, quindi questi tre item e definisce DL-r e PJD-r come somma di quelli restanti. Sulla base della nuova definizione sono state costruite le categorie di strain adottando l'approccio precedentemente descritto, utilizzando le nuove mediane calcolate sull'intero campione: 15 per DL-r e 7 per PJD-r.

I coefficienti α di Cronbach risultano superiori utilizzando la versione ridotta, in particolare, incremento maggiore si osserva per la DL in “Manager & Employers” (DL = 0.56, DL-r =0.70). Considerando la totalità del campione il coefficiente di correlazione di Pearson tra DL e PJD era -0,046, supportando l’ipotesi di ortogonalità dei due costrutti.

In tabella 6.2-3 si riportano i risultati dell'analisi di associazione tra stress lavorativo ed incidenza di eventi CHD, per l'intero campione e per classi EGP (considerando congiuntamente “White&blue collar” delle coorti MONICA/PAMELA e della coorte SEMM). L’analisi è stata effettuata sull'intero periodo di follow-up ed anche escludendo i primi tre anni al fine di limitare il possibile effetto di reverse causation. Considerando la costruzione delle classi di strain con i costrutti standard non emerge nessuna associazione tra *High Strain* ed eventi coronarici, sia per il totale del periodo che dopo esclusione dei primi tre anni di follow-up. Adottando la definizione ridotta dei costrutti, si è evidenziato un eccesso di rischio per gli *High Strain* dell’intero campione, del 36% rispetto ai non-alta tensione (95% CI: 0.96-1.93), e in “White&Blue collar” HR =1.79 (95% CI: 1.21-

2.67). Questi risultati sono confermati escludendo gli eventi nei primi tre anni di follow-up.

Per esplorare la coerenza interna dei nostri risultati, la tabella 6.2-4 mostra gli HR per le quattro classi di strain separate, dove il gruppo di riferimento è quello dei *low strain*. Anche in questo caso considerando le definizioni standard per DL e PJD, nessuna associazione statisticamente significativa è stata osservata. Utilizzando DL-r e PJD-r, si osservano HR significativamente aumentati per *high strain* e *active* nel gruppo di "White&Blue collar" HR di 2.84 (95% CI: 1.51-5.84) e HR 2.44 (95% CI: 1.14-5.10) rispettivamente.

6.2.6 Discussione

L'analisi sul campione di uomini lavoratori residenti nella provincia di Monza e Brianza ha evidenziato un tasso complessivo di incidenza di CHD pari a 2.68 per 1000 persone-anno. L'analisi stratificata dell'associazione tra livello di strain e rischio di evento CHD ha mostrato che, utilizzando DL e PJD secondo l'approccio standard non si evidenziava un aumento del rischio per i soggetti ad alto strain, e questo sia per il totale del campione che nell'analisi stratificata. L'adozione di un metodo alternativo, che ha determinato una definizione ridotta dei costrutti DL e PDJ sulla base dei risultati dell'analisi di validità e di consistenza interna ha portato ad evidenziare un eccesso di rischio del 36% nei soggetti ad altro strain rispetto agli altri tre gruppi, questo legato soprattutto al gruppo dei "White & Blue collars" dove HR è risultato pari a 1.79 (1.21-2.67). "Managers & Employers" non hanno mostrato alcuna associazione tra JS ed incidenza di CHD, indipendentemente dal tipo di definizione dei costrutti adottata. L'esclusione degli eventi avvenuti nei primi tre anni di

follow-up non ha modificato sostanzialmente i risultati, suggerendo un effetto minimo di *reverse causation* nei nostri dati.

I risultati di questo lavoro suggeriscono che si dovrebbe porre attenzione quando si analizzano gli effetti a lungo termine degli aspetti comportamentali e psicologiche. In particolare il ricorrere a questionari validati e ampiamente riconosciuti a livello internazionale per la valutazione di queste variabili, sempre consigliato, non deve prescindere da un'analisi di validità di costrutto e di consistenza interna sui dati raccolti, questo perché questo tipo di variabili e la loro percezione è strettamente legata al contesto culturale della popolazione indagata. Nel nostro campione i risultati dell'analisi fattoriale indicano che alcuni degli item originali non sono stati ben compresi dai partecipanti e quindi non identificano correttamente i costrutti JCQ attesi; ad esempio l'item "non ripetere le cose più e più volte" può avere un significato diverso a seconda dei profili professionali, degli ambienti di lavoro e dei paesi; e non necessariamente descrive un lavoro monotono. Anche i due item PJD "lavoro molto veloce" e "lavoro molto duro", possono essere rilevanti in contesti prevalentemente industriali e manifatturieri, ma può avere minor rilevanza nelle società postindustriali e lavori d'ufficio. Invece, gli elementi che descrivono le condizioni di lavoro creativo, la pressione e la presenza di richieste lavorative contrastanti possono probabilmente caratterizzare meglio ed in maniera ubiquitaria le criticità nei luoghi di lavoro.

7. INDICATORE COMPOSITO

Gli item estratti con la tecnica dei Random Survival Forest sono stati dicotomizzati e/o raggruppati definendo i costrutti psicosociali come descritto in tabella 7.1. Le variabili così definite sono state utilizzate per studiare l'Area Under the Curve (AUC) a 10, 15 e 20 anni confrontando il miglioramento tra modello con età, pressione arteriosa, colesterolo, fumo e diabete e il modello con l'aggiunta, dapprima delle singole variabili una alla volta e poi con il modello che considera contemporaneamente tutte le variabili psico-sociali di interesse.

Tabella 7.1: Costrutti psicosociali definiti sulla base degli item estratti con RSF. Uomini 25-65 anni, coorti MONICA e PAMELA.

ITEM		COSTRUTTI
SLEEPDIST1	Nel mese scorso ho avuto difficoltà ad addormentarmi	SLEEP
SLEEPDIST2	Nel mese scorso ho avuto difficoltà a rimanere addormentato	
KAR1	Il mio lavoro richiede che impari cose nuove	DL
KAR8	Il mio lavoro è sicuro	STABILITY
KAR10	Il mio lavoro è duro	PJD
KAR13	Mi vengono dati ordini contrastanti	
DEP4	Mi sento spesso completamente fiacco	DEPRESSIONE
DEP5	Mi sento vecchio dal punto di vista dell'aspetto fisico	
TYPEA5	Quando ero giovane venivo considerato grintoso e competitivo	TYPEA
TYPEA6	Oggi sono considerato grintoso e competitivo	
TYPEA7	Il mio generale livello di attività è considerato eccessivo/necessita di rallentare	
TYPEA12	Rispetto alle altre persone affronto la vita molto più seriamente	

I costrutti sono stati definiti come:

SLEEP: almeno uno dei due disturbi presente per più di 8 giorni nel mese precedente l'indagine;

DL: la variabile è stata dicotomizzata considerando le risposte "completamente d'accordo/d'accordo" come presenza di HIGH DL e "in disaccordo/completamente in disaccordo" come LOW DL.

STABILITY: la variabile è stata dicotomizzata considerando le risposte "completamente d'accordo/d'accordo" come presenza di HIGH STABILITY e "in disaccordo/completamente in disaccordo" come LOW STABILITY.

PJD: la variabile è stata dicotomizzata al valore mediano della somma dei due item.

DEPRESSIONE: variabile dicotoma cha ha valore SI= risponde si ad entrambe le risposte, NO altrimenti.

TYPEA: variabile dicotoma cha ha valore SI= se presenta tutte e quattro le caratteristiche rilevate dagli item, NO altrimenti.

In tabella 7.2 si riportano, per ciascuno dei costrutti, i valori deli Odds Ratio a 15 anni ed i relativi intervalli di confidenza al 95% relativi a 3 modelli logistici differenti: nel model 1 si considerano i costrutti singolarmente in un modello aggiustato per età, nel modello 2 tutti i costrutti vengono inseriti insieme nel modello aggiustato per età, mentre, nel modello 3 si inseriscono nel modello 2 anche i fattori di rischio standard (pressione arteriosa, colesterolo totale e HDL, abitudine al fumo e diabete).

Nessuno degli OR risulta significativamente associato al rischio di evento CVD negli uomini e 15 anni, ad eccezione della stabilità lavorativa, peraltro nel verso contrario all'atteso (chi non ha un lavoro sicuro presenta un OR inferiore pari a 0.42 IC95%:0.20-0.87 rispetto a chi ha un lavoro sicuro). Si può tuttavia notare un OR per chi riferisce di avere bassa libertà decisionale sul lavoro pari a 1.49 (IC95%:0.88-2.50) e per chi riporta alto

carico psicologico lavorativo pari a 1.12 (IC95%:0.76-1.76). Fra coloro che riportano di soffrire di disturbi del sonno l'OR è pari a 1.14 (IC95%:0.72-1.81). Si può, inoltre, osservare come non vi siano differenze nella stima degli OR anche aggiustando per gli altri costrutti psico-sociali o per i fattori di rischio.

Aggiungendo al modello con i principali fattori di rischio gli item psico-sociali selezionati attraverso la tecnica del Random Survival Forest porta ad un incremento dell'1.2% dell'AUC (da 0.7913 a 0.8034), tale incremento non risulta, tuttavia, significativo (tabella 7.3 e grafico curva ROC 7.4).

Nella tabella 7.5 si mettono a confronto i contributi aggiuntivi di ciascuna delle variabili partendo da un FULL (contenente età, pressione arteriosa, colesterolo totale e HDL, abitudine al fumo, diabete e fattori psicosociali) e togliendo di volta in volta una delle variabili. Si osserva che, l'eliminazione dei fattori di rischio psico-sociale produce il decremento maggiore in termini di AUC (1.2%) seguito dal colesterolo totale (0.8%), dalla pressione arteriosa (0.5%) e dal fumo (0.4%).

In tabella 7.6 si riportano gli andamenti degli AUC a 10, 15 e 20 confrontando il modello BASE con il modello al quale si aggiungono gli item psicosociali selezionati.

8. DISCUSSIONE E CONCLUSIONI

Il presente lavoro rappresenta un primo esempio di applicazione congiunta di tecniche di statistical learning, per la selezione delle variabili e di tecniche di modellizzazione statistica per la stima dell'associazione / predizione del rischio di evento, con particolare riferimento a variabili discrete relative a informazioni raccolte attraverso questionari.

L'associazione tra rischio di eventi cardiovascolari maggiori e alcune caratteristiche legate alla sfera psicologica, sociale e lavorativa è stata supportata dalle analisi svolte. In particolare si è vista la rilevanza che possono avere variabili come la qualità e la durata del sonno e la presenza di stress lavorativo. I risultati trovati giustificherebbero l'utilizzo di questionari ad-hoc che raccolgano informazioni su questi aspetti da somministrare ai pazienti sia in fase di prevenzione primaria che secondaria (questo secondo aspetto già recepito dalle linee guida sia in Europa che negli Stati Uniti).

Ovviamente la difficoltà di codifica delle informazioni relative a queste variabili e la criticità legata alla presenza di dati mancanti, tanto più presenti quanto più numerose sono le domande dei questionari, rende necessario effettuare una riduzione delle variabili in gioco portando alla definizione di un questionario core limitato.

L'analisi condotta sui dati delle coorti MONICA Brianza e PAMELA utilizzando le tecniche di Statistical Learning ed in particolare i Random Survival Forest ha consentito, partendo da circa 50 item relativi a 8 questionari che raccoglievano informazioni su vari aspetti psico-sociali e lavorativi, di ridurre il numero di item di circa un quarto (passando da 47 a 12 item). L'applicazione di queste tecniche ormai ampiamente utilizzate in molti ambiti (economia, genetica ecc.) consente di effettuare una riduzione

delle variabili di interesse in maniera agevole, senza necessità di ipotizzare a priori la relazione esistente tra parametri ed endpoint.

I Random Forest presentano come ulteriore punto di forza quello di non necessitare di una cross-validation in quanto la procedura stessa (dato il metodo di selezione del campione che entra di volta in volta in analisi) stima l'errore Out-of-Bag.

L'analisi di un evento poco frequente, come nel caso dei nostri dati, rappresenta sicuramente un limite a questo tipo di approccio, che è stato tuttavia risolto, nel caso specifico, appaiando ciascun evento con un non evento per sesso e fascia d'età.

Un'altra limitazione alle analisi è legata al fatto che non è stato possibile effettuare l'analisi sulle donne, dato l'ancor più esiguo numero di eventi osservato.

In conclusione, l'utilizzo dei Random Survival Forest ha portato ad evidenziare come rilevanti nell'insorgenza di primo evento cardiovascolare item relativi ai disturbi del sonno, a aspetti legati allo stress lavorativo, alla depressione ed al tratto caratteriale aggressivo (TIPOA). L'analisi dei questionari legati a questi aspetti, in particolare sonno e stress hanno evidenziato un'associazione con eventi CVD. In termini di predizione, considerando i 12 item selezionati si è ottenuto un miglioramento di circa l'1% in termini di AUC rispetto al modello predittivo contenente i principali fattori di rischio (età, pressione arteriosa, colesterolo totale e HDL, fumo e diabete), tale miglioramento non risulta tuttavia significativo.

Sulla base dei risultati, è ragionevole ritenere che, la somministrazione di un questionario inerente aspetti psicologici e lavorativi possa essere uno strumento utile per identificare soggetti che potrebbero sperimentare con maggior frequenza un evento cardiovascolare.

9. ICONOGRAFIA

Tabella 3.1-1: Coorti MONICA, PAMELA e SEMM: numerosità, durata follow-up e numero eventi incidenti. Uomini e donne 25-75 anni, liberi da malattia cardiovascolare al basale (Follow-up al 31 dicembre 2008).

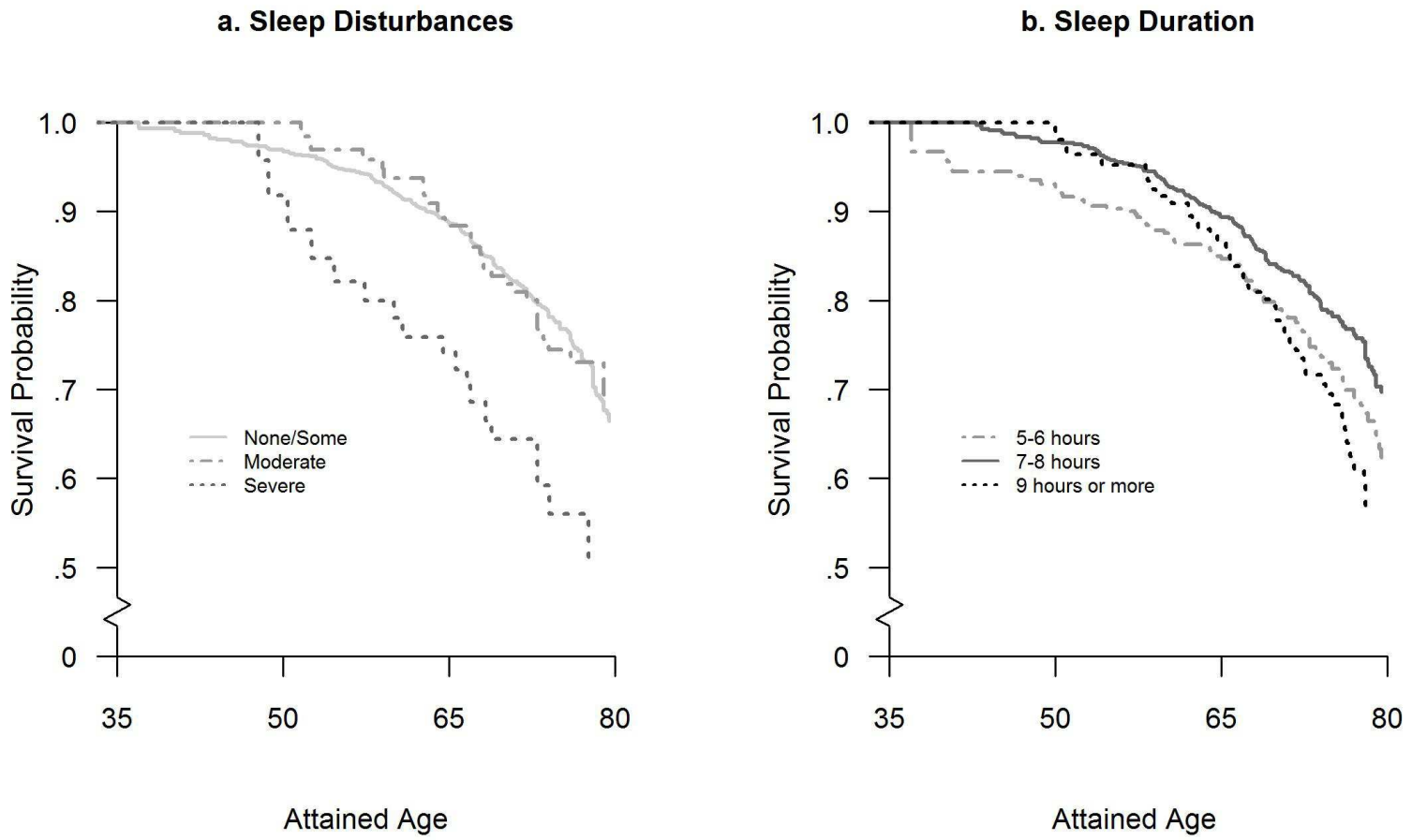
	MONICA 1		MONICA 2		PAMELA		MONICA3		SEMM	
	Uomini	**Donne**	**Uomini**	**Donne**	**Uomini**	**Donne**	**Uomini**	**Donne**	**Uomini**	**Donne**
Numerosità (n)	775	817	776	781	968	984	721	777	2569	5254
Età – anni (Media e DS)	44.7 (11.1)	45.5 (11.0)	45.5 (11.4)	45.9 (11.0)	50.1 (13.7)	49.7 (13.5)	46.6 (11.6)	45.3 (11.3)	39.8 (9.0)	37.3 (8.2)
Follow-up – anni (Mediana)	22.1	22.1	19.1	19.1	17.1	17.3	14.7	14.7	13.5	14.6
Evento coronarico (n)	71	21	58	15	72	18	41	15	77	32
Evento ischemico (n)	19	10	30	10	35	15	16	7	15	22
Evento cardiovascolare (n)	86	30	82	25	99	31	55	20	90	53

Tabella 6.1-1. Age-adjusted mean (SD) and prevalence* of major CVD risk factors at baseline, in the overall sample and by categories of sleep disturbances and sleep duration, and joint distribution of sleep duration by categories of sleep disturbances. Men, 35-74 years old and free of CVD at baseline.

		Sleep disturbances				Sleep duration			
	Entire sample	None/Some	Moderate	Severe	p-value°	≤6 hrs	7-8 hrs	9+ hrs	p-value°
Number of subjects (%)	2277	2010 (88.3)	187 (8.2)	80 (3.5)		528 (23.2)	1545 (67.9)	204 (9.0)	
Age, years (SD)	50.9 (9.5)	50.4	54.4	51.3	<0.0001	51.0	50.4	54.8	<0.0001
High school diploma or higher, %	31.8	32.9	29.2	12.1	0.0004	29.4	35.1	13.5	<0.0001
Systolic BP, mmHg	134.7 (19.4)	134.6	134.6	138.9	0.1	134.7	134.5	136.7	0.2
Hypertensive^ subjects, %	48.2	47.2	53.5	62.2	0.005	46.1	47.5	58.6	0.003
Total cholesterol, mg/dL (SD)	223.0 (42.0)	223.4	221.4	218.4	0.4	222.0	222.9	227.1	0.6
HDL cholesterol, mg/dL (SD)	50.6 (13.0)	50.5	52.2	49.8	0.4	50.0	50.7	51.7	0.5
Body Mass Index, kg/m^2 (SD)	26.1 (3.5)	26.1	26.1	26.0	0.8	26.0	26.1	26.4	0.08
Diabetes, %	6.2	6.1	7.3	6.9	0.01	6.2	5.7	9.7	<0.0001
Current smokers, %	36.1	36.8	27.0	41.8	0.001	36.5	35.3	40.8	0.2
Low LTPA, %	26.5	25.3	29.1	41.3	0.01	28.6	25.7	23.4	0.4
Critical depression, %	6.8	4.3	20.4	37.6	<0.0001	11.8	4.9	8.0	<0.0001
Sleep Duration, n (row %)									
≤6 hrs		396 (75.0)	73 (13.8)	59 (11.2)					
7-8 hrs		1422 (92.0)	105 (6.8)	18 (1.2)	<0.0001				
9+ hrs		193 (94.6)	8 (3.9)	3 (1.5)					

* Age-adjusted mean and prevalence at the sample mean age of 51 years. BP = blood pressure; HDL = high density lipoprotein; LTPA= Leisure Time Physical Activity. ^ hypertensive subjects if systolic BP => 140 mm. Hg or diastolic BP => 90 mm. Hg or under high BP drug treatment. ° Analysis of covariance, F-test for continuous variable and chi-square test for dichotomous variables.

Grafico 6.1-2. Kaplan Meier curves according to Sleep disturbances (a) and Sleep duration (b). CVD events – Men 35-74 years old and CVD free at baseline



Log-rank test chi-square and p-values:

Overall:	chi-square (2df)=8.0, p-value=0.02
Moderate vs None/Some:	chi-square (1df)=0.1, p-value=0.8
Severe vs None/Some:	chi-square (1df)=8.0, p-value=0.005
Severe vs Moderate:	chi-square (1df)=4.4, p-value=0.04

Log-rank test chi-square and p-values:

Overall:	chi-square(2df)=9.6, p-value=0.008
5-6h vs7-8h:	chi-square(1df)=1.1, p-value=0.3
9+h vs7-8h:	chi-square(1df)=9.5, p-value=0.002
9+h vs5-6h:	chi-square(1df)=3.8, p-value=0.05

Tabella 6.1-3. Hazard ratios (95% confidence intervals) of first CVD and CHD events for sleep disturbances and sleep duration, adjusted for major CVD risk factors, educational classes, depression and leisure time physical activity. Men 35-74 years old and CVD free at baseline.

	CVD event					CHD event				
	Ev/N	Model1		Model2		Ev/N	Model1		Model2	
		HR	95%CI	HR	95%CI		HR	95%CI	HR	95%CI
Sleep disturbances										
None/Some	248/2010	1	Ref.	1	Ref.	178/2010	1	Ref.	1	Ref.
Moderate	28/187	1.19	0.79-1.79	1.14	0.74-1.77	22/187	1.42	0.89-2.27	1.35	0.83-2.21
Severe	17/80	**1.80**	**1.07-3.03**	1.71	0.99-2.94	13/80	**1.97**	**1.09-3.56**	1.83	0.98-3.41
p-value*			0.07		0.2			**0.04**		0.1
Sleep duration										
6 hrs or less	72/528	1.14	0.85-1.53	1.14	0.84-1.53	54/528	1.14	0.81-1.61	1.14	0.80-1.61
7-8 hrs	177/1545	1	Ref.	1	Ref.	132/1545	1	Ref.	1	Ref.
9 hrs or more	44/204	**1.56**	**1.10-2.22**	**1.55**	**1.08-2.21**	27/204	1.36	0.87-2.12	1.32	0.85-2.07
p-value§			**0.04**		0.05			0.4		0.4

Ev= number of events ; N = number of subjects
Model1: Cox model with attained age during follow-up on the time scale, adjusted for systolic BP, total cholesterol, HDL cholesterol, diabetes, smoking habits, educational level, plus mutual adjustment for sleep disturbances and sleep duration
Model2: Cox model with attained age during follow-up on the time scale, adjusted for variable in Model 1 plus LTPA and depression (both categorized in three levels). * trend chi-square test (1 dF); § heterogeneity Wald chi-square test (2 dF)

Tabella 6.1-4. Joint effect of sleep disturbances and duration: hazard ratios (95% confidence intervals) adjusted for major CVD risk factors, educational class, depression and leisure time physical activity. Men 35-74 years old and CVD free at baseline.

	CVD event					CHD event				
	Ev/N	Model 1		Model 2		Ev/N	Model 1		Model 2	
		HR	95%CI	HR	95%CI		HR	95%CI	HR	95%CI
Sleep duration no disturbances										
6 hrs or less	49/396	1.12	0.80-1.56	1.12	0.79-1.57	33/396	0.99	0.66-1.51	0.99	0.66-1.51
7-8 hrs	157/1422	1	Ref.	1	Ref.	118/1422	1	Ref.	1	Ref.
Sleep duration, disturbances										
6 hrs or less	23/132	**1.69**	**1.08-2.64**	**1.68**	**1.03-2.74**	21/132	**2.24**	**1.39-3.63**	**2.21**	**1.31-3.74**
7-8 hrs	20/123	1.27	0.77-2.07	1.28	0.76-2.13	14/123	1.27	0.70-2.28	1.27	0.69-2.34
p-value*			0.1		0.2			**0.01**		**0.03**

Ev= number of events; N= number of subject. No disturbances: non/some sleep problems. Disturbances: moderate or severe sleep problems
Model1: Cox model with attained age during follow-up on the time scale, adjusted for systolic BP, total cholesterol, HDL cholesterol, diabetes, smoking habits, educational level, plus mutual adjustment for sleep disturbances and sleep duration
Model2: Cox model with attained age during follow-up on the time scale, adjusted for variable in Model 1 plus LTPA and depression (both categorized in three levels). *: heterogeneity Wald chi-square test (3 dF)

Tabella 6.1.6-1: Concentrazione di CRp-us (mg/L) nelle classi di distrubo e durata del sonno. Donne e Uomini 35-74 anni coorti MONICA.

	N	Mean	SD
Sleep disturbances			
Nessuno/Qualche	2093	2.9	5.2
Moderato	224	3.4	4.9
Severo	156	3.9	8.2
Sleep Duration			
5-6 h	558	3.3	6.7
7-8 h	1607	2.8	4.2
9+	308	4.4	9.5

Tabella 6.2-1A. Distribution of socio-demographic characteristics and major CVD risk factors at baseline, in the entire sample and by cohort-type and EGP-aggregate classes. Men 25-64 years old and currently employed at baseline.

		MONICA Brianza -PAMELA		SEMM	
	Entire sample	Managers& Employers	White/Blue collars	White/Blue collars	p-value
Subjects CHD-free at baseline, n	4029	745	1198	2086	
Age, years	40.7 (9.2)	43.3 (10.4)	40.7 (9.3)	39.7 (8.5)	<0.0001*
High School diploma or higher, %	39.0	44.3	30.1	42.2	<0.0001^
High Job Strain, %	25.0	10.5	21.0	32.4	
Active, %	19.1	36.2	14.2	15.8	<0.0001^
Passive, %	32.1	17.3	38.5	33.8	
Low Job Strain, %	23.8	36.0	26.3	18.1	
Systolic Blood Pressure, mm Hg	127.2 (16.2)	128.7 (17.5)	127.2 (16.2)	126.7 (15.7)	0,01*
Total Cholesterol, mg/dl	211.1 (41.3)	218.4 (43.2)	212.3 (42.3)	207.8 (39.7)	<0.0001*
HDL-cholesterol,mg/dl	49.7 (12.9)	50.4 (12.8)	50.5 (12.4)	49.1 (13.1)	0,005*
Current cigarette smokers,%	39.4	36.1	39.2	40.6	0,1^
Diabetes mellitus,%	2.7	4.8	3.3	1.5	<0.0001^
Median follow-up, years	14.6	17.5	17.7	13.5	
CHD first fatal or non-fatal events, n	163	55	52	56	
CHD fatal events, n	21	10	7	4	

Unless, otherwise indicated, the numbers reported in the table are means and standard deviations (SD)
*ANOVA F-test and ^ Chi-square test.

Table 6.2-1B. Distribution of socio-demographic characteristics and major CVD risk factors at baseline, by job strain categories. Men 25-64 years old and currently employed at baseline

	HIGH STRAIN	ACTIVE	PASSIVE	LOW STRAIN	p-value
Subjects CHD-free at baseline, n	1005	770	1294	960	
Age, years	39.2 (8.7)	40.6 (9.7)	41.3 (9.2)	41.5 (9.3)	<0.0001*
High School diploma or higher, %	38.7	49.2	30.9	41.9	<0.0001^
Systolic Blood Pressure, mm Hg	126.1 (15.6)	126.9 (15.2)	127.9 (16.9)	127.7 (16.8)	0,04*
Total Cholesterol, mg/dl	208.5 (40.0)	212.3 (42.0)	211.5 (42.2)	212.3 (41.0)	0,13*
HDL-cholesterol,mg/dl	49.8 (12.9)	49.7 (13.2)	49.3 (12.7)	50.2 (12.8)	0,49*
Current cigarette smokers, %	41.3	35.7	39.3	40.3	0,1^
Diabetes mellitus, %	2.1	2.5	2.7	3.3	0,38^
Median follow-up, years	13.9	14.7	14.6	14.9	
CHD first fatal or non-fatal events, n	30	32	65	36	
CHD fatal events, n	6	2	7	6	

Unless, otherwise indicated, the numbers reported in the table are means and standard deviations (SD)
*ANOVA F-test and ^ Chi-square test.

Table 6.2-2. Factor pattern matrix after Varimax rotation according to the standard JCQ scales and definition of the Restricted JCQ scales

	MONICA Brianza -PAMELA				SEMM		Restricted item JCQ scales	
	Managers&Employers		White/Blue collars		White/Blue collars			
JCQ Items - scales	FACTOR1	FACTOR2	FACTOR1	FACTOR2	FACTOR1	FACTOR2	DLr	PJDr
Learn new things - SK	0.4714	0.0037	0.5862	-0.021	0.7250	0.0064	x	
High level of skill - SK	0.6521	-0.0783	0.6986	-0.008	0.7570	0.0874	x	
Be creative - SK	0.6340	-0.0547	0.7084	-0.11199	0.5428	-0.03206	x	
Not repeat things over and over - SK	-0.1004	0.2396	0.0608	0.0696	0.2104	-0.18787		
I decide how much work I have to do - DA	0.5054	-0.1675	0.5078	-0.30962	0.5712	-0.11794	x	
Freedom to decide what do at job - DA	0.5069	-0.1422	0.5369	-0.31541	0.1879	-0.19076	x	
Working very fast - PJD	0.2895	0.0610	0.2294	0.1104	-0.01815	0.0653		
Working very hard - PJD	0.2925	0.0618	0.1857	0.1582	0.1235	0.3751		
Excessive amount of work required - PJD	0.1077	0.4987	-0.07602	0.5605	-0.06576	0.6442		x
Not enough time to get the job done - PJD	0.1082	0.8178	0.000	0.6458	-0.06736	0.5922		x
Conflicting demands - PJD	-0.075	0.3402	-0.134	0.5124	-0.00562	0.3605		x

SK = skill discretion, DA = decision authority; PJD = psychological job demand; DL = decision latitude; JCQ= Job Content Questionnaire

Table 6.2-3. Multivariate-adjusted Hazard Ratios (HR) and 95% confidence intervals (95%CI) of first CHD event, for hish jb strain (HJS) versus non-high job strain (noHJS), as reference caterogy. Job strain categories defined by Psychological Job Demand (PJD) and Decision Latititude (DL), using the standard JCQ items (above) and the restricted JCQ items (below). Men 25-64 years old and currently employed at baseline.

		Entire sample					Managers&Employers					White&Blue collars				
		N	#CHD	HR	95%CI		N	#CHD	HR	95%CI		N	#CHD	HR	95%CI	
		All events in the entire follow-up period														
PJD and DL based on standard items &	no HS	3024	133		REF		667	49		REF		2357	84		REF	
	HS	1005	30	0.86	0.58	1.28	78	6	0.77	0.32	1.86	927	24	0.93	0.59	1.47
		Events occurred after the first three years of follow-up only														
	no HS	2981	113		REF		650	40		REF		2258	73		REF	
	HS	998	27	0.92	0.61	1.41	77	5	0.72	0.28	1.90	899	22	1.00	0.62	1.62
		All events in the entire follow-up period														
PJD and DL based on restricted items $	no HS	2992	120		REF		667	51		REF		2325	69		REF	
	HS	1037	43	1.36	0.96	1.93	78	4	0.61	0.22	1.70	959	39	**1.79**	**1.21**	**2.67**
		Events occurred after the first three years of follow-up only														
	no HS	2957	103		REF		652	42		REF		2305	61		REF	
	HS	1022	37	1.41	0.96	2.10	75	3	0.54	0.17	1.76	947	34	**1.81**	**1.18**	**2.77**

MONICA Brianza, PAMELA and SEMM cohorts. Men 25-64 years old, current employed and with all DL and PJD items availables.
HR estimated from Cox regression models with age as the time scale, adjusted for systolic blood pressure, total cholesterol, HDL cholesterol, diabetes and current smokers & 5 items for PJD and 6 items for DL.

$ restricted based on the results of contruct validity analysis: 3 items for PJD and 5 items for DL.

Table 6.2-4. Multivariate-adjusted Hazard Ratios (HR) and 95% confidence intervals (95%CI) of first CHD event, according to job strain category (Low strain as reference). Men 25-64 years old and currently employed at baseline.

JCQ categories	Entire sample					Managers&Employers					All White&Blue collars				
	N	#CHD	HR*	95%CI		N	#CHD	HR*	95%CI		N	#CHD	HR*	95%CI	
Standard PJD and DL definitions															
HIGH STRAIN	1005	30	1.07	0.66	1.74	78	6	0.84	0.33	2.14	927	24	1.42	0.76	2.65
ACTIVE	770	32	1.32	0.82	2.14	270	14	0.82	0.41	1.66	500	18	1.89	0.97	3.68
PASSIVE	1294	65	1.39	0.92	2.09	129	16	1.68	0.85	3.33	1165	49	1.72	0.99	2.99
LOW STRAIN	960	36		REF		268	19		REF		652	17		REF	
Restricted-item PJD and DL definitions															
HIGH STRAIN	1037	43	1.54	0.98	2.41	78	4	0.61	0.21	1.80	959	39	**2.84**	**1.51**	**5.84**
ACTIVE	585	24	1.29	0.76	2.16	173	9	0.69	0.31	1.50	412	15	**2.44**	**1.14**	**5.10**
PASSIVE	1445	61	1.16	0.77	1.76	181	20	1.26	0.68	2.34	1264	41	1.67	0.90	3.13
LOW STRAIN	962	35		REF		313	22		REF		649	13		REF	

* HR from Cox regression models with age on the time scale, adjusted for systolic blood pressure, total cholesterol, HDL cholesterol, diabetes and current smokers.
CHD occurred in the entire follow-up period.

Tabella 7.2: Studio dell'associazione tra item psicosociali e evento CVD a 15 anni. Modello di regressione logistico. Uomini 25-65 anni, coorti MONICA e PAMELA.

Costrutti		Model 1			Model 2			Model 3		
		OR	IC95%	p-value§	OR	IC95%	p-value§	OR	IC95%	p-value§
SLEEP*	high vs low	1.14	0.72-1.81	0.58	1.13	0.70-1.81	0.61	1.22	0.75-1.97	0.49
DL	low vs high	1.49	0.88-2.50	0.14	1.53	0.90-2.60	0.12	1.46	0.85-2.51	0.17
PJD	high vs low	1.12	0.76-1.76	0.58	1.14	0.77-1.69	0.53	1.15	0.77-1.72	0.49
STABILITY	low vs high	0.42	0.20-0.87	0.02	0.39	0.19-0.83	0.01	0.39	0.18-0.83	0.01
DEPRESSIONE	si vs no	1.02	0.69-1.49	0.90	1.02	0.69-1.50	0.93	1.03	0.69-1.53	0.88
TYPEA	si vs no	0.81	0.49-1.34	0.25	0.79	0.48-1.32	0.37	0.79	0.47-1.34	0.39

Model1 - age adjusted
Model 2 - age and psychosocial items
Model 3- age, psychosocial items and other risck factors (SBP, Colesterolo, fumo e diabete)

SLEEP* : high = aver sofferto di disturbi del sonno almeno 8 giorni nel mese precedente l'indagine
§ Wald Chi-square p-value

Tabella 7.3: Area sotto la curva per il rischio di evento cardiovascolare a 15. Uomini 25-65 anni, coorti MONICA e PAMELA.

Modello	AUC	DIFF	95% IC LOWER	95% IC UPPER
BASE	0.7913			
BASE + 2 ITEM DISTURBI SONNO	0.7919	0.001	-0.004	0.006
BASE+ JCQ (PJD DL SECURITY)	0.8004	0.009	-0.003	0.021
BASE+ 2 ITEM DEPRESSIONE	0.7926	0.001	-0.002	0.005
BASE+ 4 ITEM TYPEA	0.7943	0.003	-0.003	0.009
BASE+TUTTI GLI ITEM SELEZIONATI CON RSF	0.8034	0.012	-0.003	0.027

BASE=AGE SBP CHOLDL HDLDL FUMOBI DIAB

Figura 7.4: Curva ROC per rischio di evento cardiovascolare a 15 e relativi AUC. Uomini 25-65 anni, coorti MONICA e PAMELA

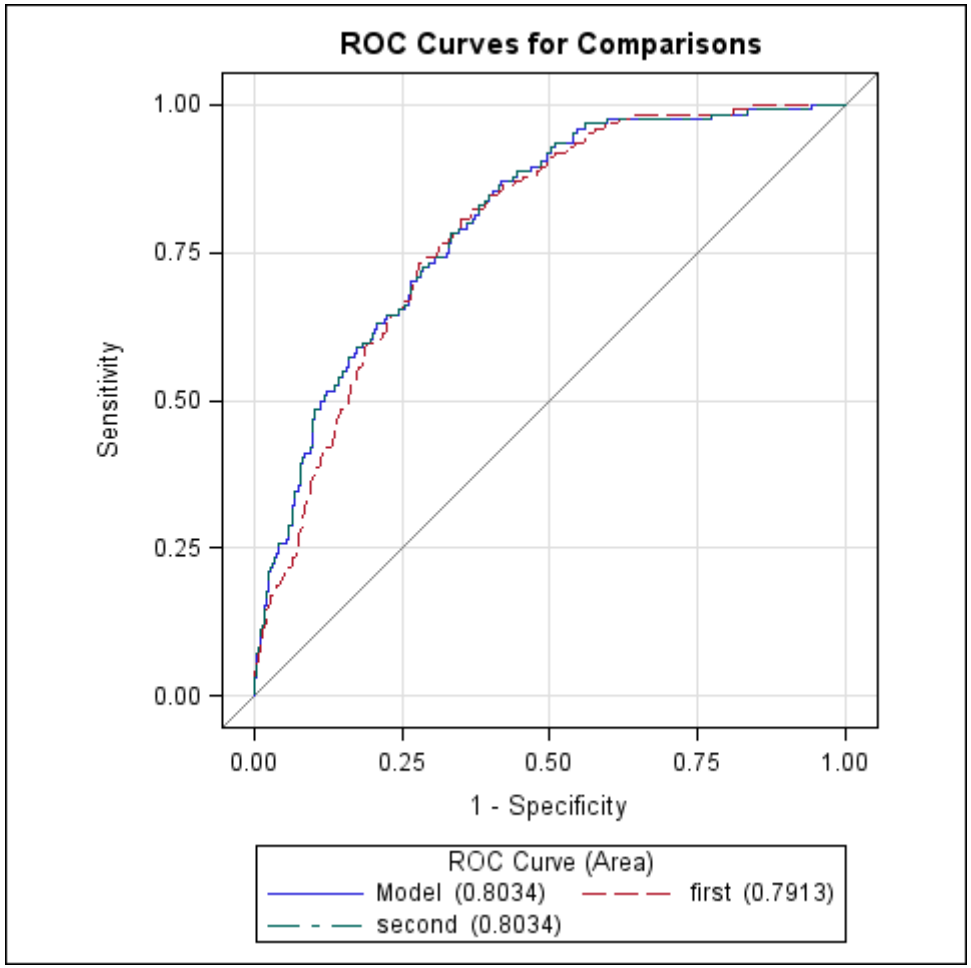


Tabella 7.5: Area sotto la curva per il rischio di evento cardiovascolare a 15. Uomini 25-65 anni, coorti MONICA e PAMELA.

	AUC	DIFF	95% IC LOWER	95% IC UPPER
FULL MODEL	0.8034			
FULL MODEL - SBP	0.7984	-0.005	-0.013	0.003
FULL MODEL - CHOLDL	0.7951	-0.008	-0.018	0.002
FULL MODEL - HDL	0.8002	-0.003	-0.011	0.004
FULL MODEL - FUMO	0.7990	-0.004	-0.012	0.003
FULL MODEL - DIABETE	0.8015	-0.002	-0.007	0.003
FULL MODEL - PSYCHOSOCIAL	0.7913	-0.012	-0.030	0.003

FULL MODEL=AGE SBP CHOLDL HDLDL FUMO DIABETE PSYCHOSOCIAL ITEM

Tabella 7.6: Confronto tra Aree sotto la curva per il rischio di evento cardiovascolare a 10, 15 e 20 anni. Uomini 25-65 anni, coorti MONICA e PAMELA.

	AUC		
	10 ANNI	**15 ANNI**	**20 ANNI**
BASE	0.822	0.791	0.788
BASE+TUTTI GLI ITEM SELEZIONATI CON RSF	0.828	0.803	0.795
Diff	0.006	0.012	0.007
p-value	0.46	0.11	0.20

BASE=AGE SBP CHOLDL HDLDL FUMOBI DIAB

10. BIBLIOGRAFIA

Backé, E. M., Sidler, A., Latza, U., Rossnagel, K., & Schumann, B. (2012). The role of psychosocial stress at work for the development of cardiovascular disease: a systematic review. *International Archives of Occupational and Environmental Health*, 85:67-79.

Baecke, J., & al, e. (1982). A short questionnaire for the measurement of habitual physical activity in peidemiological studies. *The American Journal of Clinical Nutrition*, 936-942.

Barth J, e. a. (2010). Lack of social support in the etiology and the prognosis of coronary heart disease: a systematic review and meta-analysis. *Psychosomatic medecine*, 229-238.

Barth, J. e. (2004). Depression as a risk factor for mortality in patients with coronary heart disease: a meta-analysis. *Psychosom Med*, 802-813.

Beck, A., Ward, C., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of general psychiatry*, 561-571.

Berkman, L., & Syme, L. (1979). Social networks, host resistance, and mortality: a nine-year follow-up study of alameda county residetns. *American Journal of Epidemiology*, 186-204.

Bombelli, M., Toso, E., Peronio, M., Fodri, D., Volpe, M., Brambilla, G., . . . Mancia, G. (2013). The Pamela Study: Main Findings and Perspectives . *Curr Hypertens Rep*, 15:238-243.

Cannon, W. (1915). *Bodily Changes in Pain, Hunger, Fear and Rage.* London: Routledge and Kegan Paul.

Cappuccio, F., Cooper, D., D'Elia, L., & al, e. (2011). Sleep duration predicts cardiovascular outcomes: a sistematic review and meta-analysis of prospective studies. *Eur Heart J*, 32: 1484-1492.

Cesana, G., DeVito, G., & Ferrario, M. e. (1991). Ambulatory blood pressure normality: The PAMELA Study. *Journal of Hypertension Suppl*, 17-23.

Charmandari, E. e. (2005). Endocrinology of the stress response. *Annu Rev Physiol*, 259-284.

Chen, Y. e. (2009). Increased risk of acute myocardial infarction for patients with panic disorder: a nationwide population-based study. *Psychosom Med*, 798-804.

Chida, Y., & Steptoe, A. (2009). The association of anger and hostility with future coronary heart disease: a meta-analytic review of prospective evidence. *J Am Coll Cardiol*, 936-946.

Cohen, S. (1992). *The meaning and measurement of social support.* New York: Veiel HOF & Baumann U.

Denollet, J. e. (2010). A general propensity to psychological distress affects cardiovascular outcomes: evidence from research on the type D (distressed) personality profile. *Circ Cardiovasc Qual Outcomes*, 546-557.

Denollet, J. e. (2010). Anger, suppressed anger, and risk of adverse events in patients with coronary artery disease. *Am J Cardiol*, 1555-1560.

Dimsdale, J. (2008). Psychological stress and cardiovascular disease. *J Am Coll Cardiol*, 1237-1246.

Dipnale, J., Pasco, J., Berk, M., Willimas, L., Dodd, S., Jacka, F., & D, M. (2016). Fusing dtata mining, machine learning and traditional statistics to detect biomarkers associated with depression. *Plos One*.

Eaker ED, e. a. (2004). Anger and hostility predict the development of atrial fibrillation in men in the Framingham Offspring Study. *Circulation*, 1267-1271.

Eaker, E. e. (2007). Marital status, marital strain and risk of coronary heart disease or total mortality: the Framingham Offspring Study. *Psychosom Med*, 509-513.

Eller, N. e. (2009). Work-related psychosocial factors and the development of Ischemic heart disease: a systematic review. *Cardiol Rev*, 83-97.

Eng, P. e. (2002). Socila ties and change in social ties in relation to subsequent total and cause-specific mortality and coronary heart disease incidence in men. *American Journal of Epidemiology*, 700-709.

Glass, D. (1977). Stress, behavior patterns, and coronary disease. *Am Science*, 177-187.

Haynes, S., & Feinleib, M. (1980). Women, Work and coronary heart disease: prospective findings from the Framingham Heart Study. *American Journal of Public Health*, 133-141.

Hisch, E., Gorodeski, E. Z., Blackstone, E. H., Ishwaran, H., & Lauer, M. S. (2011). Identifying Important Risk Factors for Suvival in patient with systolic heart failure using Random Survival Forest. *Circulation: cardiovascular quality and outcomes*, 39-45.

Ishwaran, H., & Kogalur, U. (2014). Random Forest for Survival, Regression and Classification (RF-SRC), R package version 1.6 . *URL://http://CRAN.R-project.org/package=randomForestSRC*.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random Survival Forest. *The Annals of Applied Statistics*, 841-860.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning - with application in R.* New York: Springer.

Jenkins, D., & al, e. (1988). A scale for estimation of sleep problems in clinical research. *Journal of Clinical Epidemiology*, 313-321.

Karasek, R. (1979). Job Demand, Job Decision Latitude, and Mental Strain: Implication for job redesign. *Administrative Science Quarterly*, 285-308.

Karasek, R., Choi, B., Ostergren, P., Ferrario, M., & de Smet, P. (2007). Testing two methods to create comparable scale scores between the job content questionnaire (JCQ) and JCQ_like questionnaires in the European JACE study. *International Journal Behavioral Medecin.*, 189-201.

Katsarou, A. e. (2012). Perceived stress and vascular disease: where are we now? *Angiology*, 529-534.

Kivimaki, M., Marianna, V., Elovainio, M., Kouvonen, A., Vaananen, A., & Vahtera, J. (2006). Work stress in the etiology of coronary heart disease - a meta-analysis. *Scandinavian Journal of Work, Environmental and Health*, 431-442.

Kreibig, S. D., Whooley, M. A., & Gross, J. J. (2014). Social Integration and mortality in patients with coronary heart disease: findings from the Heart and Soul Study. *Psychosomatic Medecine*, 659-668.

Lee, S., Colditz, G., Berkman, L., & Kawachi, I. (2003). Caregiving to children and grandchildren and risk of coronary heart disease in women. *American Journal of Public Health*, 1939-1944.

Lett, H. e. (2005). Social support and coronary heart disease: epidemiologic evidence and implications for treatment. *Psychosom Med*, 869-878.

Lichtman JH, e. a. (2008). Depression and Coronary Heart Disease. *Circulation*, 1768-1775.

Liu, R., Liu, X., Phyllis, Z., Hou, L., Sheng, S., Wei, Y., & Du, J. (2014). Association between Sleep Quality and C-Reactive Protein: results from National Helath and Nutrition Examination Survey 2005-2008. *Plos one*.

Lu, F., & Petkova, E. (2013). a comparative study of variable selection methods in the context of developing psychiatric screening instruments. *Statistics in Medecine*, 401-421.

Lunetta, K. L., Hayward, B. L., Segal, J., & Van Eerdewegh, P. (2004). Screening large-scale association study data: exploiting interactions using ranodm forests. *BMC Genetics*.

Meesters, C., & Appels, A. (1996). An interview to measure vital exhaustion. Development and comparison with the Maastricht Questionnaire. *Psychology and Health*, 557-571.

Miao, F., Cai, Y.-P., Zhang, Y.-T., & Li, C.-Y. (2015). Is Random Survival Forest and alternative to Cox Proportional Model on predicting cardiovascular disease? *6TH European conference of the international federation for medical and biological engineering* (p. 740-745). Springer - Verlag Berlin.

Mookadam, F., & Arthur, H. (2004). Social support and its relationship to morbidity and mortality after acute myocardial infarction: systematic overview. *Arch Int Med*, 1514-1518.

Neylon A, e. a. (2013). A global perspective on psychosocial risk factors for cardiovascular disease. *Progress in Cardiovascular Diseases*, 574-581.

Nicholson, A. e. (2006). Depression as an aetiologic and prognostic factor in coronary heart disease: a meta-analysis of 6362 events among 146538 participants in 54 observational studies. *Eur Heart J*, 2763-2774.

Ohayon, M. (2002). Epidemiology of insomnia: what we know and what we still need to learn. *Sleep Med Rev*, 6:97-111.

Perk J, e. a. (2012). European Guidelines on cardiovascular disease prevention in clinical practice (Version 2012); The Fifth Join Task Force of the European Society of Cardiology and other societies on Cardiovascular Disease PRevention in Cinical Practice. *E. Heart Journal*, 1365-1701.

Practice, F. J. (2012). European Guidelines on cardiovascular disease prevention in clinical practice. *European Heart Journal*, 1365-1701.

Quan, S. (2009). Sleep disturbancies and theri relationship to cardiovascular disease. *Am J Lifestyle Med*, 3: 55-59.

Roest, A. e. (2010). Anxiety and risk of incident coronary heart disease: a meta-analysis. *J Am Coll Cardiol*, 38-46.

Roest, A. e. (2010). Prognostic association of anxiety post myocardial infarction with mortality and new cardiac events a meta-analysis. *Psychosom Med*, 563-569.

Rose, R., & al, e. (1978). Air Traffic Controller Health Change Study: a prospective investigation of physical, psychological and work-related changes. *Psychosomatic Medecine*, 142-165.

Rosengren A, e. a. (2004). Association of psychological risk factors with risk of acute myocardial infarction in 11119 cases and 13648 controls from 52 countries (the INTERHEART study): case-control study. *Lancet*, 953-962.

Rubin, D. (1976). Inference and missing data. *Biometrika*.

Rugulies, R. (2002). Depression as a predictor for coronary heart disease a review and meta-analysis. *Am J Prev Med*, 51-61.

Rugulies, R. (2002). Depression as predictor for coronary heart disease. A review and meta-analysis. *American Journal Preventive Medecine*, 51-61.

Rutledge, T., Linke, S., Olson, M., Francis, J., Johonson, D., Bittner, V., . . . Mer. (2008). Social Networks and incidence stroke among women with suspected myocardial ischemia. *Psychosomatic Medecine*, 282-287.

Schafer, J. (1997). *Analysis of incomplete multivariate data.* Chapman and Hall.

Schafer, J., & Graham, J. (2002). Missing data: our view of the state of the art. *Psychological methods*, 147-177.

Schafer, J., & Graham, J. (2002). Missing data: our view of the State of the Art. *Psychological Methods*.

Schafer, J., & Olsen, M. (1998). Multiple imputation for multivariate missing data problems. A data analyst's perspective. *Multivariate Behavioral Research*, 545-571.

Schuitemaker, G., Dinant, G., GA, V. D., Verhelst, A., & Appels, A. (2004). Vital exhaustion as a risk indicator for first stroke. *Psychosomatic*, 114-118.

Siegrist, J. (1996). Adverse health effects of high-effort/ low-reward conditions. *Journal of Occupational Health and Psychology*, 1:27-41.

Smoller, J. e. (2007). Panic attacks and risk of incident cardiovascular events among postmenopausal women in the Women's Health Initiative Observational Study. *Arch Gen Psychiatry*, 1153-1160.

Sofi, F., Cesari, F., Casini, A., Macchi, C., Abbate, R., & Gensini, G. F. (2014). Insomnia and risk of cardiovascular disease: a meta-analysis. *European Journal of Preventive Cardiology*, 57-64.

Steptoe, A., & Kivimaki, M. (2013). Stress and Crdiovascular Disease: an update on current knowledge. *Annual Review of Public Health*, 34:337-354.

Stringhini, S. e. (2010). Association of socioeconomic position with health behavior and mortality. *JAMA*, 1159-1166.

Sulis, I., & Porcu, M. (2008). Assessing the effectiveness of a stochastic regression imputation method for ordered categorical data. *Working paper - Centro Ricerche Economiche .*

Tonne, C. e. (2005). Long-Term survival after acute myocardial infarction is lower in more depreived neighborhoods. *Circulation*, 3063-3070.

Van, A., Gay, V., Kennedy, P., Barin, E., & Leijdekkers, P. (2011). Understanding risk factors in cardiac rehabilitation patients with random forest and decision trees. *Proceedings of the 9-th Australasian data mining conference* (p. 11-22). Ballarat - Australia: Australian Computer Society.

Whittaker KS, e. a. (2012). Combining psychosocial data to improve prediction of cardiovascular disease risk factors and events: The NHLBI- Spnsored Women's Ischemia Syndrome Evaluation (WISE) study. *Psychosomatic medecin*, 263-270.

WHO. (2011). *Global Atlas on Cardiovascular Disease Prevention and Control.* Geneva: WHO.

WHO-MONICA, P. (s.d.). *WWW-publications from the WHO-MONICa Project. MONICA manual.* http://www.ktl.fi/publications/minica/index.html.

Wulsin, L., & Singal, B. (2003). Do depressive symptoms increase the risk for the onset of coronary disease? A systematic quantitavie review. *Psychosom Med*, 201-210.