Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche

IL RUOLO DELL'IPERTENSIONE NELLA RELAZIONE TRA ACIDO URICO E OUTCOME CARDIOVASCOLARI: UNO STUDIO BASATO SU MODELLI AD EQUAZIONI STRUTTURALI E CAUSAL DISCOVERY

Relatore Prof.ssa Giovanna Boccuzzo
Dipartimento di Scienze Statistiche

Correlatore Prof. ssa Valerie Tikhonoff
Dipartimento di Medicina

Laureanda: Claudia Franceschini
Matricola N. 2062596

Anno Accademico 2023/2024

SOMMARIO

L'incremento dell'incidenza degli eventi cardiovascolari rappresenta una delle principali sfide per la sanità pubblica a livello globale. Tra i vari fattori di rischio associati a queste patologie, elevati livelli di acido urico (uricemia) e ipertensione arteriosa emergono come determinanti cruciali. Questo studio si propone di esplorare la relazione causale tra uricemia, ipertensione arteriosa ed eventi cardiovascolari, testando parallelamente diverse misurazioni dell'ipertensione arteriosa per determinarne la rappresentatività nel contesto in esame.

Da un punto di vista metodologico, l'uso dei modelli ad equazioni strutturali (SEM) consente di ragionare circa l'appropriatezza di diverse specificazioni causali ipotizzate a priori, attraverso misure di bontà di adattamento e l'analisi degli effetti stimati. Tuttavia, a causa della natura trasversale dei dati, i modelli SEM presentano il limite di non poter asserire relazioni causali.

Per affrontare questa limitazione, sono stati impiegati i modelli grafici nel contesto della *causal discovery*. L'algoritmo di *causal discovery* di PC (Peter-Clark) è stato impiegato per stimare una rete causale direttamente dai dati osservazionali, fornendo un'evidenza empirica indipendente a integrazione delle ipotesi causali formulate a priori.

I risultati ottenuti tramite i modelli SEM rivelano delle differenze di adattamento tra i *path* causali stimati e le diverse misurazioni dell'ipertensione arteriosa, evidenziandone in ogni modellazione il ruolo centrale per l'insorgenza di eventi cardiovascolari in relazione ad elevati livelli di uricemia. Parallelamente, la *causal discovery* ha offerto ulteriori evidenze per le dipendenze indagate e per la scelta di misura dell'ipertensione, dimostrando il valore aggiunto dell'integrazione di un approccio ipotetico-deduttivo con tecniche di apprendimento non supervisionato e inferenza causale nell'analisi di dati complessi. Per concludere, si sottolinea che i risultati ottenuti non mirano a fornire conclusioni definitive sulle relazioni indagate, ma offrono ulteriori evidenze a supporto di un'analisi integrata, enfatizzando la necessità di successivi studi per la relazione causale tra uricemia, ipertensione arteriosa ed eventi cardiovascolari.

INDICE

1. INTRODUZIONE AL PROBLEMA CLINICO E OBIETTIVO

In questo primo capitolo viene fornita una panoramica sull'entità degli eventi cardiovascolari in Italia, in Europa e nel mondo, a sottolineare come questi rappresentino una questione cruciale per la sanità pubblica e come la ricerca e la pratica clinica contribuiscano alla gestione della problematica.

Vengono poi presentate le principali nozioni mediche necessarie alla comprensione del problema clinico in esame; in particolare, si pone specifica attenzione su uricemia, ipertensione arteriosa, e tutte quelle condizioni cliniche dimostratesi associate all'insorgenza di eventi cardiovascolari, come diabete mellito ed elevati livelli sierici di creatinina. Successivamente, si illustrano fattori di rischio e meccanismi patogenetici che coinvolgono ciascuna variabile di interesse nel rapporto con gli eventi cardiovascolari. Si prosegue con un breve approfondimento riguardante la storia e lo sviluppo dello studio della relazione tra uricemia, ipertensione arteriosa ed eventi cardiovascolari, fino ad arrivare alla presentazione dello studio URRAH e all'obiettivo del presente lavoro di ricerca.

1.1 L'ENTITÀ DEGLI EVENTI CARDIOVASCOLARI IN ITALIA, IN EUROPA E NEL MONDO

1.1.1. LA MORTALITÀ PER EVENTI CARDIOVASCOLARI

Gli eventi cardiovascolari identificano una vasta gamma di patologie che interessano per lo più il cuore tra cui l'infarto del miocardio e l'insufficienza cardiaca ma anche altri organi quali il cervello con l'ictus ischemico e le arterie periferiche con l'arteriopatia periferica obliterante (Benjamin et al., 2019; Mozaffarian et al., 2015).

In Italia, gli eventi cardiovascolari costituiscono la principale causa di mortalità, morbosità ed invalidità. Secondo i dati forniti dall'Istituto Nazionale di Statistica (ISTAT) ed elaborati dall'Istituto Superiore di Sanità (ISS), nel 2021 il 30,8% dei decessi registrati è risultato attribuibile a malattie di tipo cardiovascolare, con una prevalenza più accentuata tra le persone anziane (Ministero della Salute, 2024).

In Europa, il quadro epidemiologico degli eventi cardiovascolari è altrettanto rilevante. La Società Europea di Cardiologia (ESC) ha dichiarato nel rapporto del 2022 come questi siano responsabili di oltre 4 milioni di decessi l'anno, rappresentando circa il 45% di tutti i decessi. La preminenza di queste malattie come causa di mortalità è evidente in quasi tutti i paesi europei, sebbene con qualche differenziazione tra nazioni in termini di incidenza e prevalenza (ESC, 2022).
Anche a livello globale, gli eventi cardiovascolari si confermano una delle principali minacce alla salute pubblica. L'Organizzazione Mondiale della Sanità (OMS) stima che ogni anno 17,9 milioni di persone vadano incontro a morte per eventi cardiovascolari, costituendo circa il 31% di tutte le morti mondiali (OMS, 2021).

1.1.2 IMPATTO SULLA SANITÀ PUBBLICA

Dai dati riportati poco sopra, è evidente come gli eventi cardiovascolari rappresentino una questione di estremo interesse per la sanità pubblica, non solo per l'alto tasso di mortalità che li caratterizza, ma anche per l'enorme impatto economico e sociale ad essi associato: il soggetto che sopravvive ad una forma acuta di evento cardiovascolare infatti, diventa molto spesso un malato cronico, costretto a confrontarsi con notevoli ripercussioni in termini di qualità della vita e costi economici nonché sociali. Le malattie cardiovascolari inoltre rappresentano uno dei maggiori determinanti delle malattie legate all'invecchiamento, producendo disabilità fisica e disturbi della capacità cognitiva. Ancora una volta dunque, la sfida che queste patologie presentano non è solo di natura clinica ma anche economica e sociale, imponendo un onere significativo sui sistemi sanitari ed in primis sulle famiglie dei soggetti coinvolti (OMS, 2021).

1.1.3 FATTORI DI RISCHIO COMPORTAMENTALI ED INDICATORI CLINICI

I principali fattori di rischio comportamentali per le malattie cardiovascolari sono rappresentati da una dieta non salutare, l'inattività fisica, l'uso di tabacco e l'abuso di alcol. Gli effetti di questi fattori di rischio possono manifestarsi negli individui attraverso valori elevati di pressione arteriosa, glicemia, lipidi nel sangue, presenza di obesità etc. Gli indicatori clinici appena citati possono essere misurati nelle strutture di assistenza ospedaliera primaria e rappresentano quindi un

importantissimo mezzo attraverso cui è possibile identificare soggetti a rischio o che presentino una malattia cardiovascolare non conclamata, adottando di conseguenza strategie di prevenzione e gestione della problematica. Attraverso la somministrazione di terapia farmacologica e la riabilitazione cardiovascolare inoltre, è possibile intervenire nel miglioramento degli esiti in pazienti con malattie cardiovascolari diagnosticate, agendo dunque sulla prevenzione terziaria (OMS, 2021).

Dalle considerazioni appena fatte, risultano evidenti le motivazioni a supporto di un grande interesse sul tema delle malattie cardiovascolari, a cui si aggiunge poi ulteriore razionale di carattere di ricerca clinica, oggetto di questa tesi, presentato nei paragrafi 1.3, 1.4, 1.5.

1.2 DEFINIZIONI E CONCETTI FONDAMENTALI

1.2.1 URICEMIA

Per uricemia si fa riferimento alla concentrazione sierica di acido urico, prodotto finale del metabolismo delle purine. Le purine sono basi azotate presenti negli acidi nucleici e possono essere di origine endogena oppure esogena. La fonte principale è costituita dalla sintesi delle purine endogene (adenina e guanina del DNA) mentre l'origine esogena, ovvero l'introito alimentare, da un contributo minore ma non trascurabile all'apporto di purine. La concentrazione di acido urico nell'organismo è il risultato del rapporto tra la quantità prodotta e quella escreta. Per quanto riguarda la produzione, essa dipende dall'introito alimentare di purine, dalla velocità dei processi biosintetici, dalla degradazione e dal recupero delle purine. Per quanto riguarda invece l'eliminazione dell'acido urico, essa avviene per il 70% per via renale e per il 30% per via intestinale.

Elevati livelli di uricemia possono essere associati ad una serie di condizioni patologiche, tra cui ipertensione arteriosa, diabete mellito e malattie cardiovascolari. Inoltre, un'alta concentrazione sierica di acido urico può contribuire alla formazione di cristalli di urato monosodico nelle articolazioni che sono all'origine della patologia gottosa, ma anche promuovere l'infiammazione e la

disfunzione endoteliale contribuendo così alla patogenesi delle malattie cardiovascolari (Feig et al., 2008a; Sharaf El Din et al., 2017).
Il dosaggio sierico dell'uricemia è comunemente utilizzato in ambito clinico per diagnosticare e monitorare la gotta e valutare il rischio di eventi cardiovascolari. Tuttavia, è importante anche considerare che i livelli di uricemia possono variare in base a fattori demografici come l'età, il sesso, l'etnia, e la presenza di comorbidità (Saadat et al., 2018). È necessario dunque mantenere sempre una visione d'insieme della problematica da affrontare, considerando l'effetto di eventuali variabili confondenti.

1.2.2 IPERTENSIONE ARTERIOSA

L'ipertensione arteriosa è una patologia caratterizzata da elevati valori di pressione arteriosa con conseguente danno a livello della pareti dei vasi arteriosi (Whelton et al., 2018b; Williams et al., 2018). Nel Capitolo 2 *I Dati* verrà presentato un approfondimento su questo concetto, inserendolo in un contesto di misurazione per l'analisi statistica e la ricerca clinico-medica.
L'ipertensione è ad oggi riconosciuta come uno dei principali fattori di rischio per le malattie cardiovascolari, inclusi l'ictus ischemico, l'infarto miocardico, l'insufficienza cardiaca, e le malattie renali (Whelton et al., 2018b; Williams et al., 2018). Inoltre, come specificato poco sopra, numerosi studi ne hanno dimostrato l'associazione con elevati livelli di acido urico (Sánchez-Lozada et al., 2020). La sua patogenesi coinvolge una serie di meccanismi tra cui le modificazioni del sistema renina-angiotensina-aldosterone, e la disfunzione endoteliale (Kario, 2018; Williams et al., 2018). L'ipertensione arteriosa può essere classificata come ipertensione essenziale (primaria) o ipertensione secondaria, ovvero secondaria ad una condizione nota (renale, endocrina o vascolare) (Williams et al., 2018).

1.2.3 DIABETE MELLITO

Il diabete mellito è una patologia metabolica caratterizzata da livelli elevati sierici di glucosio, i quali possono essere attribuiti ad una produzione insufficiente di insulina, alla resistenza all'insulina o ad entrambi (ISS, 2024).

I criteri diagnostici più recenti per l'identificazione del diabete mellito si basano sulle misurazioni della glicemia a digiuno, della glicemia postprandiale o dell'emoglobina glicata (HbA1c). Secondo le linee guida internazionali dell'*American Diabetes Association* del 2020, il diabete mellito può essere diagnosticato se si verifica almeno uno dei seguenti criteri:

- Glicemia a digiuno ≥ 126 mg/dl (7.0 mmol/l);
- Glicemia casuale ≥ 200 mg/dl (11.1 mmol/l);
- Glicemia postprandiale ≥ 200 mg/dl (11.1 mmol/l) durante un test di tolleranza al glucosio;
- Emoglobina glicata (HbA1c) ≥ 6.5% (48 mmol/mol).

Questi criteri possono variare leggermente in base alle linee guida regionali e alle pratiche cliniche specifiche.
Le complicanze cardiovascolari associate al diabete mellito sono ben documentate e includono aterosclerosi accelerata, disfunzione endoteliale e ipertensione arteriosa; tutte queste condizioni contribuiscono all'aumento del rischio di sperimentare eventi cardiovascolari (American Diabetes Association, 2020).

1.2.4 CREATININA

La creatinina è un prodotto di scarto del metabolismo muscolare, ottenuto in modo costante dal muscolo scheletrico e dalla creatina fosfato durante il normale *turnover* cellulare. Viene liberata nel sangue e filtrata dai reni, con una *clearance* renale quasi costante, il che la rende un indicatore affidabile della funzionalità renale (Levey et al., 2006; Rule et al., 2004). I livelli ematici di creatinina infatti possono essere utilizzati per stimare il tasso di filtrazione glomerulare (eGFR), che costituisce un importante indicatore della funzionalità renale e della sua capacità filtrante: livelli di eGFR fuori dai valori clinici di riferimento sono a favore di una ridotta *clearance* renale e una compromissione della funzione renale (Levey et al., 2006; Stevens et al., 2007).
L'analisi della creatinina sierica è ampiamente utilizzata nella pratica clinica anche per monitorare l'efficacia di possibili trattamenti in pazienti con malattie renali e altre condizioni mediche; la sua concentrazione nelle urine per esempio,

rappresenta un indicatore per la possibile presenza di malattie muscolari (Patel et al., 2013).

1.2.5 EVENTI CARDIOVASCOLARI

Gli eventi cardiovascolari comprendono una vasta gamma di patologie e condizioni che coinvolgono il cuore e i vasi sanguigni, tra cui l'infarto del miocardio, l'ictus, l'insufficienza cardiaca, l'aritmia cardiaca e altre forme di malattia vascolare. Questi eventi, come visto nel *paragrafo 1.1*, sono annoverati tra le principali cause di morbilità e mortalità a livello mondiale e sono associati ad una serie di fattori di rischio come l'ipertensione arteriosa, il diabete mellito, la dislipidemia, il tabagismo e l'obesità (Benjamin et al., 2019; Mozaffarian et al., 2015).

L'infarto del miocardico si verifica quando vi è o interruzione o riduzione del flusso ematico al tessuto cardiaco con conseguente insufficiente apporto di ossigeno e necrosi del tessuto stesso (Thygesen et al., 2018). Parimente, l'ictus ischemico, o attacco cerebrovascolare, si verifica in conseguenza del mancato o ridotto flusso ematico al cervello che porta ad un danno del tessuto nervoso centrale con potenziale deficit neurologici più o meno invalidanti. (Powers et al., 2019). L'insufficienza cardiaca si verifica quando a seguito di un danno al tessuto cardiaco, vi è una compromissione nell'efficienza della capacità contrattile del tessuto stesso con conseguente affaticamento, dispnea ed edema (Ponikowski et al., 2016).

Gli eventi cardiovascolari sono il risultato di una complessa interazione tra fattori genetici, ambientali e comportamentali, e sono influenzati da svariati meccanismi patogenetici, tra cui l'aterosclerosi, l'ipertensione arteriosa e la disfunzione endoteliale (Benjamin et al., 2019; Libby et al., 2019). Il monitoraggio e il controllo dei fattori di rischio cardiovascolari sono fondamentali per la prevenzione e la gestione efficace di tutti quei pazienti considerati a rischio di sperimentare gli eventi (Arnett et al., 2019; Grundy et al., 2019).

1.3 FATTORI DI RISCHIO E MECCANISMI PATOGENETICI PER LO SVILUPPO DI EVENTI CARDIOVASCOLARI

1.3.1 FATTORI DI RISCHIO

Livelli elevati di uricemia ed ipertensione arteriosa sono noti fattori di rischio per l'insorgenza di malattie cardiovascolari, e la relazione tra questi tre elementi è un argomento di estremo interesse ed attualità nell'ambito della ricerca clinica.

Alti livelli di acido urico possono influenzare negativamente la funzione endoteliale, la quale ha un ruolo importante nella regolazione della pressione arteriosa. Un endotelio danneggiato può favorire la formazione di placche nelle arterie (aterosclerosi) e aumentare il rischio di mortalità per qualsiasi causa, mortalità per eventi cardiovascolari, ma anche il rischio di sperimentare singoli eventi cardiovascolari come l'infarto del miocardio, l'ictus e lo scompenso cardiaco (*American Diabetes Association*, 2020; Bos et al., 2006; Fang & Alderman, 2000; Feig et al., 2008a; Kario, 2018; Niskanen et al., 2004).

L'ipertensione arteriosa invece, aumentando il carico di lavoro del cuore e danneggiando le pareti delle arterie, è associata all'aumento del rischio di aterosclerosi e altri eventi di tipo cardiaco. (Kario, 2018; Williams et al., 2018).

1.3.2 MECCANISMI PATOGENETICI

I processi biologici e fisiologici che coinvolgono ipertensione arteriosa e uricemia contribuendo allo sviluppo di eventi cardiovascolari sono studiati con grande coinvolgimento dalla comunità medica e scientifica, e ad oggi presentano ancora molti punti interrogativi.

L'ipertensione arteriosa si è dimostrato essere associata all'insorgenza di danni strutturali delle arterie, alterazione del sistema renina-angiotensina-aldosterone e ipertrofia ventricolare sinistra (Kario, 2018).

Elevati livelli di acido urico invece possono promuovere l'infiammazione e la disfunzione endoteliale, predisponendo così alla formazione di placche aterosclerotiche e all'insorgenza di eventi cardiovascolari (Saadat et al., 2018). Alcuni studi longitudinali hanno inoltre indagato la relazione tra acido urico ed ipertensione, evidenziando come questo sia relato allo sviluppo di ipertensione

arteriosa (Bombelli et al., 2014). Dal 2018 infatti, le linee guida *dell'European Society of Hypertension* e l'*European Society of Cardiology* (ESH-ESC) includono il dosaggio di acido urico tra gli esami di screening consigliati per i pazienti ipertesi (Williams et al., 2018).

È opportuno ricordare infine che anche il diabete mellito svolge un ruolo determinante nello studio della relazione tra uricemia, ipertensione arteriosa ed eventi cardiovascolari. Infatti, nel diabete mellito l'infiammazione cronica e l'accumulo di prodotti avanzati della glicazione possono contribuire alla progressione dell'aterosclerosi e all'insorgenza di complicanze cardiovascolari (American Diabetes Association, 2020).

1.4 STORIA E SVILUPPO DELLO STUDIO DELLA RELAZIONE TRA URICEMIA, IPERTENSIONE ED EVENTI CARDIOVASCOLARI

La storia del legame tra uricemia, ipertensione arteriosa ed esiti cardiovascolari si sviluppa nella letteratura medica a partire dalla fine del 19^{esimo} secolo grazie al lavoro di medici come Frederick Akbar Mahomed, Alezander Haig e Nathan Smith Davis, i quali ipotizzarono una relazione di tipo causale tra l'acido urico e l'ipertensione arteriosa o il danno renale. Lo studio della relazione tra uricemia ed eventi cardiovascolari fu poi lasciata in secondo piano, per riacquisire un forte interesse da parte degli studiosi tra gli anni '50 e '60 (Cannon et al., 1966; Gertler, 1951). Durante questi ultimi decenni, un abbondante numero di studi epidemiologici ha dimostrato una relazione tra l'acido urico ed eventi cardiovascolari (Tuttle et al., 2001), considerando anche l'ipertensione arteriosa, la sindrome metabolica (Choi & Ford, 2007) e i disturbi legati alla funzionalità renale (Siu et al., 2006). Ad oggi, la natura della relazione e il rapporto tra queste variabili risulta ancora controversa e non completamente nota.

Comprendere questa complessa rete di interazioni è di grande rilevanza clinica per tre principali ragioni:

- Identificare i pazienti a rischio elevato di eventi cardiovascolari, consentendo un intervento preventivo mirato;

- Comprendere i meccanismi patogenetici sottostanti che potrebbero suggerire nuovi approcci terapeutici per la prevenzione e il trattamento delle malattie cardiovascolari;
- Migliorare la gestione clinica dei pazienti con condizioni correlate, ottimizzando il controllo dei fattori di rischio e monitorando da vicino la funzione cardiovascolare e renale.

Uno degli studi pionieristici in questo campo è stato condotto da Feig *et al*. nel 2008. Questo studio, raccogliendo le evidenze di diversi studi precedentemente condotti, ha sottolineato una forte associazione tra elevati livelli di acido urico, ipertensione arteriosa, malattie cardiovascolari e funzionalità renale in soggetti giovani, giustificando l'ipotesi che l'abbassamento dei livelli di acido urico possa apportare un beneficio clinico nella prevenzione o nel trattamento delle malattie cardiovascolari e renali.

A comprovare la complessità delle relazioni studiate, è bene sottolineare che solo pochi anni prima, il Framingham Heart Study insieme ad altri studi condotti sul tema, non erano giunti alla stessa conclusione, affermando che l'associazione fra livelli di acido urico ed eventi cardiovascolari fosse probabilmente dovuta all'associazione del livello di acido urico con altri fattori di rischio. (Chobanian et al., 2003; Culleton et al., 1999; F. S. Pearson et al., 2002).

Le evidenze emerse dallo studio di Feig et al. del 2008 hanno dunque gettato le basi per ulteriori indagini volte a esplorare il ruolo dell'uricemia nelle patologie cardiovascolari, con particolare interesse verso una possibile relazione di tipo causale. Negli anni successivi, numerosi studi hanno confermato e ampliato le asserzioni di Feig et al. in varie popolazioni selezionate e contesti clinici. Johnson et al., 2009 per esempio, hanno replicato l'associazione tra uricemia e ipertensione arteriosa, sottolineando il potenziale ruolo dell'infiammazione e della disfunzione endoteliale come meccanismi sottostanti. Come dichiarato anche sopra, ad oggi la relazione tra l'incremento di acido urico e l'ipertensione rimane controversa e non completamente nota.

Allo stesso tempo, altri ricercatori hanno esplorato il legame tra uricemia e diabete mellito, mettendo in rilievo l'importanza di comprendere come l'accumulo di acido

urico possa influenzare la patogenesi del diabete mellito di tipo 2 e le sue complicanze cardiovascolari (Xu et al., 2016).

La ricerca clinica ha anche approfondito i meccanismi patogenetici che sottendono alla relazione tra uricemia, ipertensione ed esiti cardiovascolari. Uno dei filoni di studio più significativi ha riguardato il ruolo dell'infiammazione e dello stress ossidativo nel promuovere la progressione delle malattie cardiovascolari in individui con alti livelli di acido urico (Sánchez-Lozada et al., 2008). Allo stesso tempo, altri studi hanno indagato le possibili vie metaboliche coinvolte nella regolazione della pressione arteriosa e della glicemia, evidenziando il potenziale impatto dell'uricemia su questi processi (Sharaf El Din et al., 2017).

Oltre a fornire una migliore comprensione dei meccanismi sottostanti, la ricerca sulla relazione tra uricemia, ipertensione ed esiti cardiovascolari ha anche aperto la strada a nuovi approcci terapeutici e a strategie di gestione clinica più efficaci. Ad esempio, l'identificazione di farmaci in grado di ridurre i livelli di acido urico ha suscitato interesse come potenziali trattamenti per ridurre il rischio cardiovascolare in individui ad alto rischio (Feig et al., 2008a). Allo stesso tempo, si è iniziata a delineare chiaramente l'importanza del controllo rigoroso dei fattori di rischio cardiovascolare come pressione arteriosa, glicemia e funzione renale, nell'ottimizzare l'*outcome* clinico in pazienti con uricemia elevata e patologie cardiovascolari correlate (Xu et al., 2016).

Nonostante i progressi significativi compiuti nella comprensione del tema di interesse, rimangono ancora oggi molte domande aperte e sfide da affrontare. Per la relazione tra acido urico ed ipertensione si rendono necessari ulteriori studi clinici su larga scala per determinare se la riduzione degli urati extracellulari e intracellulari possa apportare benefici all'ipertensione e alle malattie cardiometaboliche (Sánchez-Lozada et al., 2020). Inoltre, futuri studi vorrebbero concentrarsi sull'identificazione di biomarcatori predittivi e sulla valutazione dell'efficacia di interventi terapeutici mirati al fine di ridurre il rischio cardiovascolare in individui con alterazioni metaboliche e renali (Saadat et al., 2018).

1.5 LO STUDIO URRAH

In questo contesto articolato e multidimensionale, nel 2018 prende vita lo studio URRAH (*Uric Acid Right for Heart Health*), concepito come risposta alla crescente evidenza del coinvolgimento dell'acido urico nelle patologie cardiovascolari. Promosso da un gruppo di studio afferente alla Società Italiana di Ipertensione Arteriosa (SIIA), lo studio URRAH si propone come primo obiettivo quello di identificare il *cut-off* prognostico di acido urico in grado di discriminare al meglio l'insorgenza di eventi cardiovascolari in uomini e donne (Maloberti et al., 2020). Più in generale, lo studio si propone di valutare gli effetti dei livelli sierici di acido urico sul rischio e sulla mortalità cardiovascolare in individui ad alto rischio, come quelli affetti da ipertensione arteriosa e diabete mellito. In questi anni sono in corso ulteriori ricerche anche riguardanti le relazioni tra acido urico e funzionalità renale, e fra acido urico e sindrome metabolica (Maloberti et al., 2023).

Lo studio URRAH è organizzato come una grande raccolta di dati provenienti da 13 centri nazionali italiani e che già disponevano dei valori di acido urico, pressori, della fenotipizzazione cardiovascolare e dei successivi eventi cardiovascolari. L'unione di questi dati ha permesso di creare un database ad oggi composto da 27078 soggetti con un *follow-up* mediano di 11 anni.

Le prime due pubblicazioni dello studio URRAH hanno suscitato grande interesse nella comunità medica. La prima individua un *cut-off* di acido urico pari a 4.7 mg/dL e 5.6 mg/dL per discriminare al meglio rispettivamente la mortalità per tutte le cause e la mortalità cardiovascolare (Virdis et al., 2020); la seconda analisi identifica il *cut-off* di acido urico con riferimento all'insorgenza di infarto acuto del miocardio (Casiglia et al., 2020). Entrambi i risultati suggeriscono come un livello di uricemia non elevato possa rappresentare una strategia preventiva efficace nella gestione degli eventi cardiovascolari.

Le ultime scoperte pubblicate dallo studio URRAH confermano e ampliano i risultati precedenti. Un'analisi longitudinale ha evidenziato una correlazione significativa tra livelli più bassi di uricemia e ridotta incidenza di eventi cardiovascolari maggiori, tra cui infarto miocardico, ictus e scompenso cardiaco (URRAH *Research Group*, 2023). Questi risultati forniscono un solido supporto all'ipotesi che la gestione

dell'uricemia potrebbe rappresentare una nuova strategia terapeutica per la prevenzione e il trattamento delle malattie cardiovascolari.
Lo studio URRAH, con il suo approccio multidisciplinare e la sua ampia scala, si pone all'avanguardia nella ricerca cardiovascolare e offre importanti prospettive per il futuro della medicina preventiva e terapeutica.

1.6 OBIETTIVO

Il seguente lavoro si pone come domanda di ricerca clinica l'esplorazione del ruolo dell'ipertensione arteriosa nella relazione con acido urico ed eventi cardiovascolari, valutando parallelamente diverse modalità di definizione e misurazione dell'ipertensione arteriosa per determinarne l'adeguatezza nel contesto clinico esaminato.
Da un punto di vista statistico-metodologico, è di interesse discutere l'utilizzo del modello ad equazioni strutturali (SEM) come strumento di modellazione probabilistica a supporto di nuove intuizioni teorico-causali per la relazione clinica in oggetto, attraverso la valutazione degli effetti stimati e l'impiego di misure di bontà di adattamento delle relazioni modellate rispetto ai dati.
Infine, un ulteriore sviluppo metodologico si propone di esplorare l'utilizzo dei modelli grafici (GM) per la ricerca delle relazioni causali (*path* causali) a partire dal solo contesto osservazionale in esame (*causal discovery*), con particolare attenzione alle assunzioni che permettono di identificare una relazione causale e alle condizioni che garantiscono robustezza ai risultati nel contesto di dati trasversali.

2. I DATI

Il dataset in analisi è costituito da 27078 soggetti di età compresa tra i 18 e i 100 anni reclutati nel tempo su base comunitaria e regionale da tutto il territorio nazionale italiano, provenienti sia da studi prospettici di coorte osservazionali, sia da ambulatori regionali specializzati nella cura dell'ipertensione arteriosa e del diabete. Il periodo di campionamento si estende dai primi anni '90 fino al 31 luglio del 2017, per un arco temporale di circa trent'anni.

Per ogni paziente è disponibile al più una rilevazione di *follow-up*: il 16% di questi non presenta informazioni relative ad alcun controllo periodico programmato. Le informazioni di *follow-up* sono da riferirsi all'ultimo controllo effettuato su ciascun individuo o, qualora il soggetto avesse abbandonato lo studio, all'ultimo periodo in cui si sapeva essere ancora in vita. Il periodo nel quale sono stati monitorati i pazienti varia entro l'arco temporale di raccolta dei dati; la durata mediana del *follow-up* è di 11 anni, con un *range* che varia da un minimo di un mese fino ad un massimo di 29 anni; il 25^{mo} e 75^{mo} percentile sono pari a 6 e 14 anni rispettivamente.

2.1 MISURAZIONI

Per ciascun soggetto vengono rilevate le caratteristiche demografiche unitamente ad una serie di informazioni che fanno riferimento indicativamente a sei macrocategorie: abitudini e stile di vita, misure antropometriche, storia di pregressi eventi cardiaci, cerebrali o renali, anamnesi farmacologica, parametri bioumorali (metabolici e di funzionalità renale), eventi cardiovascolari. La pressione arteriosa è rilevata attraverso la misurazione della pressione arteriosa sistolica e diastolica. Infine, è presente una variabile che classifica i soggetti affetti da ipertensione arteriosa. Per quest'ultima variabile, verrà sviluppato nel prossimo paragrafo un approfondimento necessario a comprendere quali scelte e assunzioni stiano alla base di una definizione che offre diversi sviluppi a seconda dell'obiettivo e del contesto per cui viene utilizzata.

Si riporta ora una tabella contenente la distribuzione di frequenza univariata per le principali caratteristiche misurate. Per quanto concerne gli eventi cardiovascolari, alla sezione 2.5 verrà presentata separatamente la relativa tabella di frequenza.

Tabella 2.1 - Principali caratteristiche dell'insieme di dati in analisi (n=27078[1]).

Variabile	**Modalità**	**Frequenza assoluta**	**Frequenza percentuale**
Sesso	Donne	13539	50%
	Uomini	13539	50%
	Totale	27078	100%
Età (anni)	18-30	1423	5,3%
	31-40	2425	9%
	41-50	5215	19,3%
	51-60	6199	22,9%
	61-70	5613	20,7%
	71-80	4787	17,7%
	81-90	1312	4,8%
	91-100	100	0,4%
	Totale	27074	100%
Famigliarità per ipertensione arteriosa	Sì	10143	53,6%
	No	8795	46,4%
	Totale	18938	100%
Famigliarità per malattie cardiovascolari	Sì	6297	38,4%
	No	10086	61,6%
	Totale	16383	100%
Diabete mellito di tipo due	Sì	2921	11,3%
	No	22981	88,7%
	Totale	25902	100%
Fumatore attivo	Sì	6305	25,7%
	No	18221	74,3%
	Totale	24526	100%
Assunzione di alcolici	Sì	11751	60,2%
	No	7763	39,8%
	Totale	19514	100%
Attività fisica	Sì	6334	53,4%
	No	7271	46,6%
	Totale	13605	100%
Body Mass Index (BMI) (kg/m2)2	≤ 18,5 (Sottopeso)	281	1,2%
	18,5-24,9 (Normopeso)	8577	35,1%
	25-29,9 (Sovrappeso)	10816	44,2%
	≥ 29,9 (Obesità)	4776	19,5%
	Totale	24450	100%
Storia clinica ipertensiva	Sì	8573	55,9%
	No	10875	44,1%
	Totale	19448	100%

[1] Le frequenze e le percentuali riportate sono calcolate al netto dei valori mancanti presenti per ciascuna variabile. Di seguito si riporta la percentuale di valori mancanti per ciascuna variabile, ove presenti: età 0,01%; famigliarità per ipertensione 30,1%; famigliarità per malattie cardiovascolari 39,5%; diabete mellito due 4,3%; fumatore attivo 9,4%; assunzione di acolici 27,9%; attività fisica 49,8%; BMI 9,7%; storia clinica ipertensiva 28,2%; insufficienza renale cronica 4,4%; gotta 48,1%; pressione diastolica 2,2%; pressione sistolica 2,2%; dosaggio creatinina 8,8%; terapia farmacologica anti-ipertensiva 8,4%.

[2] Le soglie per le classi riportate fanno riferimento alla classificazione fornita dal Ministero della Salute (Ministero della Salute, 2022).

Variabile	Modalità	Frequenza assoluta	Frequenza percentuale
Insufficienza renale cronica	Sì	4299	16,6%
	No	21593	83,4%
	Totale	25892	100%
Gotta	Sì	247	1,8%
	No	13803	98,2%
	Totale	14050	100%
Classi di pressione arteriosa diastolica (mmHg) [3]	< 60	342	1,3%
	60-79	8503	31,4%
	80-89	7742	28,6%
	90-99	6367	23,5%
	100-119	3356	12,4%
	120+	184	0,7%
	Totale	26494	100%
Classi di pressione arteriosa sistolica (mmHg) [4]	70-89	19	0,1%
	90-119	3768	13,9%
	120-129	3839	14,2%
	130-139	5162	19,0%
	140-155	6811	25,2%
	155+	6890	25,5%
	Totale	26489	100%
Ipertensione arteriosa (si veda par 2.2)	Sì	19079	70,5%
	No	7999	29,5%
	Totale	27078	100%
Dosaggio plasmatico creatinina (mg/dL)[5]	< 0.5 (Basso)	74	0,3%
	0,5-1,2 (Normale)	22038	89,2%
	≥ 1,2 (Elevato)	2593	10,5%
	Totale	24705	100%
Dosaggio plasmatico acido urico (mg/dL)[6]	≤ 5,6	18066	66,7%
	> 5,6	9012	33,3%
	Totale	2707	100%
Terapia farmacologica antipertensiva	Sì	8645	34,9%
	No	16149	65,1%
	Totale	24794	100%

[3] Classi calcolate a partire dalla variabile continua.

[4] Classi calcolate a partire dalla variabile continua.

[5] Le soglie si riferiscono alla classificazione clinica fornita dall'Istituto Superiore di Sanità (ISS, 2022).

[6] La soglia adottata fa riferimento al cutoff prognostico per la mortalità cardiovascolare ottenuto nel lavoro di Virdis e al. (Virdis et al., 2020).

2.2 LA DEFINIZIONE DI IPERTENSIONE ARTERIOSA

Le ultime linee guida operative europee ESH-ESC del 2023 definiscono l'ipertensione arteriosa come una condizione caratterizzata dalla presenza di un livello di pressione arteriosa diastolica e/o sistolica rispettivamente ≥ 90 mmHg e ≥ 140 mmHg, rilevata per mezzo di almeno due misurazioni. Questa definizione empirica nasce dalla doppia necessità di ottimizzare il tempo alla diagnosi e ricondursi a linee guida unificate per la gestione e la cura del paziente iperteso (Mancia et al., 2023).

Recenti pubblicazioni scientifiche hanno discusso alcune problematiche associate a questa definizione operativa. In particolare, viene posta specifica attenzione sull'impiego degli operatori Booleani *e/o*, sottolineandone l'interpretazione spesso scorretta e dunque l'utilizzo improprio (Casiglia, 2024). In questa analisi, sposando le argomentazioni proposte nel lavoro del Prof. E. Casiglia, è stato deciso di fare riferimento alla seguente definizione operativa aggiornata: "*L'ipertensione arteriosa è definita come la rilevazione ripetuta di una pressione arteriosa diastolica ≥ 90 mmHg o di una pressione arteriosa sistolica ≥ 140 mmHg"*.

Le linee guida stesse, inoltre, specificano che "Tuttavia, esiste una relazione continua tra pressione arteriosa ed eventi cardiovascolari o renali morbosi o fatali a partire da valori di pressione arteriosa ambulatoriale >115 mmHg e una pressione arteriosa diastolica >75 mmHg [35 delle linee guida ESH 2023]. Pertanto, questa definizione è arbitraria e ha principalmente lo scopo pragmatico di semplificare la diagnosi e la decisione sulla gestione dell'ipertensione. In questo contesto, i valori pressori soglia ambulatoriali sopra indicati corrispondono al livello di pressoria al quale i benefici dell'intervento (interventi sullo stile di vita o trattamento farmacologico) superano quelli dell'inazione, come dimostrato da studi randomizzati basati sui risultati".

Per questo motivo, nell'ambito della ricerca clinica l'uso della definizione operativa delle linee guida ESH-ESC va integrato con altri elementi utili per identificare tutti i possibili soggetti affetti da ipertensione arteriosa all'interno di una coorte. La prassi internazionale dunque identifica nel soggetto *iperteso* una condizione più ampia, che non fa rifermento ai soli valori pressori rilevati, ma considera complessivamente tre aspetti, ove i rimanenti due riguardano l'assunzione di terapia farmacologica

ipertensiva e la storia clinica del paziente (Carson et al., 2013; Chobanian et al., 2003).

Riassumendo, nella ricerca clinica un soggetto viene classificato *iperteso* se è verificata almeno una delle seguenti condizioni:

- la rilevazione ripetuta della pressione arteriosa diastolica è ≥ 90 mmHg o la rilevazione ripetuta della pressione arteriosa sistolica è ≥ 140 mmHg; [7]
- è in corso una terapia farmacologica con farmaci utilizzati per il controllo della pressione arteriosa;
- è presente una storia clinica di ipertensione arteriosa certificata da un clinico esperto.

La definizione di soggetto *iperteso* nella ricerca clinica, dunque, è in espansione rispetto a quella di declinazione più empirica presentata nelle linee guida operative ESC-ESH volta ad identificare i soggetti affetti da ipertensione arteriosa, e tale scelta è motivata dalla volontà di includere nella categoria di ipertesi tutti quei soggetti che presentano un rischio clinico di sperimentare eventi cardiovascolari più alto rispetto ad un individuo sano.

In questa analisi verrà in primo luogo considerata l'ipertensione arteriosa tramite l'utilizzo della definizione clinica di soggetto *iperteso*, che considera unitamente i tre aspetti sopracitati e l'anamnesi clinica. Verrà poi ripetuta l'analisi utilizzando i soli valori pressori (mmHg), con l'obiettivo di comparare e discutere eventuali risultati differenti condizionatamente al modello causale ipotizzato per la relazione con le altre variabili di interesse.

Nel dataset a disposizione, la storia clinica dei soggetti è rintracciabile attraverso la variabile dicotomica, qui denominata *sapeva,* che indica se il soggetto fosse al corrente di essere affetto da ipertensione arteriosa al momento dell'arruolamento nello studio. L'anamnesi farmacologica è rilevata tramite una serie di variabili dicotomiche che identificano rispettivamente l'assunzione di ciascun farmaco di interesse clinico.

[7] Si faccia riferimento al lavoro di E. Casiglia, 2024 esposto poco sopra.

Da un punto di vista statistico, la differente formulazione del concetto di ipertensione arteriosa dalla sola misurazione dei valori pressori alla definizione clinica comporta un verosimile aumento della prevalenza di popolazione classificata come ipertesa, se si considera la definizione clinica più ampia. In questo studio, la prevalenza stimata di popolazione ipertesa aumenta di 8,9 punti percentuali, passando dal 56,7% (15360 soggetti) al 65,6% (17760 soggetti).

Di seguito si riporta la distribuzione di frequenza bivariata per la definizione clinica di ipertensione rispetto alla rilevazione congiunta dei valori pressori oltre il limite sopracitato, all'assunzione di terapia farmacologia antipertensiva, e alla storia clinica di ipertensione.

Tabella 2.2 - Distribuzione di frequenza bivariata dell'ipertensione clinica rispetto ai valori pressori (PAS ≥140 mmHg or PAD ≥ 90 mmHg)[8].

	No ipertensione valori pressori (%)	**Sì ipertensione valori pressori (%)**	**TOTALE (100 %)**
Non iperteso	8738 (78,5%)	0 (0%)	**8738 (33%)**
Iperteso	2400 (21,5%)	15360 (100%)	**17760 (67%)**
TOTALE	11138 (100%)	15360 (100%)	**26498 (100%)**

[8] Nel dataset sono presenti 580 soggetti (2,1%) con valore mancante sia nella rilevazione della pressione diastolica (PAD) che sistolica (PAS).

Tabella 2.3 - Distribuzione di frequenza bivariata dell'ipertensione clinica rispetto all'assunzione di almeno un farmaco ipertensivo[9].

	No farmaci antipertensivi (%)	**Sì farmaci antipertensivi (%)**	**TOTALE (100 %)**
Non iperteso	8820 (43,4%)	365 (0%)	**9185 (33,9%)**
Iperteso	9613 (56,6%)	8280 (100%)	**17893 (66,1%)**
TOTALE	18433 (100%)	8645 (100%)	**27078 (100%)**

Tabella 2.4 - Distribuzione di frequenza bivariata dell'ipertensione clinica secondo la presenza di una storia clinica ipertensiva[10].

	No storia clinica (%)	**Sì storia clinica (%)**	**TOTALE (100 %)**
Non iperteso	8158 (44,1%)	1027 (22%)	**9185 (33,9%)**
Iperteso	10347 (55,9%)	7546 (88%)	**17893 (66,1%)**
TOTALE	18505 (100%)	8573 (100%)	**27078 (100%)**

[9] 2284 soggetti (8,4%) presentano valore mancante congiuntamente in tutti i farmaci antiipertensivi. Per questi soggetti, si è imputato valore 0 in corrispondenza della variabile dicotomica creata.

[10] Ai soggetti con valore mancante nella rilevazione della storia clinica ipertensiva (n = 7630, 28,2%) è stato imputato valore pari a 0.

Dall'osservazione delle tabelle 2.3 e 2.4, si sottolinea la presenza di 365 soggetti sottoposti a terapia farmacologica anti-ipertensiva (in particolare, farmaci diuretici) classificati non ipertesi, e l'osservazione di 1027 soggetti con storia clinica di ipertensione e classificati come non ipertesi, per un totale di 1392 soggetti (5,1%) valutati non ipertesi a seguito di anamnesi clinica nonostante la presenza di almeno una delle tre caratteristiche peculiari per la definizione più ampia di ipertensione sopra riportata.
Volendo confrontare i risultati delle distribuzioni di frequenza bivariata tra le due definizioni di ipertensione con terapia farmacologica antipertensiva e storia clinica ipertensiva, è possibile affermare che:

- Tra i 15360 soggetti definiti ipertesi secondo le linee guida operative ESH, il 62,6% non risulta sottoposto a terapia farmacologica antipertensiva al momento del *follow-up*; inoltre, per il 63,2% di questi non si rileva una storia clinica di ipertensione arteriosa;
- Tra i 17893 soggetti ipertesi secondo la definizione adottata nella ricerca clinica, la percentuale di soggetti non sottoposti a terapia farmacologica si abbassa al 53,7%, e il 57,8% non presenta una storia clinica ipertensiva.

2.3 PRIVACY E TRATTAMENTO DATI

La gestione dei dati personali rispetta la legislazione europea sulla privacy. Tutte le informazioni raccolte, elaborate e conservate sono state anonimizzate, e tutti i documenti relativi allo studio sono custoditi in un ambiente sicuro. L'approvazione per condurre lo studio è stata ottenuta dal Comitato Etico del centro di coordinamento presso la Divisione di Medicina Interna dell'Università di Bologna (numero di protocollo 77/2018/Oss/AOUBo). Prima della partecipazione allo studio, ogni soggetto ha fornito il proprio consenso informato.

2.4 OUTCOME CARDIOVASCOLARI

Gli outcome cardiovascolari presi in considerazione in questa analisi sono tutti di natura dicotomica e rilevano se i soggetti hanno sperimentato rispettivamente infarto acuto del miocardio, eventi di tipo cerebrovascolare (ictus), scompenso cardiaco e rivascolarizzazione coronarica. Vi è poi una variabile che indica la morte del soggetto a causa del sopraggiungere di almeno un evento di tipo cardiovascolare, e un'ulteriore variabile che identifica qualora il soggetto risultasse deceduto al momento del *follow-up* per qualsiasi causa. Per gli esiti cardiovascolari che riguardano l'infarto acuto del miocardio, l'ictus e lo scompenso cardiaco è disponibile l'informazione condizionatamente al tipo di gravità dell'evento, facendo distinzione tra evento di tipo fatale e non fatale.

Il numero totale di *outcome* considerati dunque risulta pari a otto: mortalità cardiovascolare totale, infarto acuto del miocardio fatale, infarto acuto del miocardio non fatale, evento cerebrovascolare fatale, evento cerebrovascolare non fatale, scompenso cardiaco fatale, scompenso cardiaco non fatale, rivascolarizzazione coronarica.

Con l'obiettivo di esplorare le relazioni tra i diversi *outcome* cardiovascolari, si è deciso di procedere con il calcolo della matrice di associazione per l'indice Q di Yule e, successivamente, con la stima della matrice di correlazione tetracorica τ di Pearson. Ciascuno di questi indici, attraverso le rispettive formulazioni e assunzioni peculiari, rappresenta uno strumento utile e informativo per comprendere in maniera integrata le relazioni che sottendono i diversi eventi.

2.4.1 INDICE DI ASSOCIAZIONE Q DI YULE TRA OUTCOME CARDIOVASCOLARI

L'indice di associazione Q[11] di Yule, presentato da George Undy Yule nel 1900, rappresenta una misura statistica concepita per valutare la forza e la direzione dell'associazione tra due eventi dicotomici tramite l'utilizzo di tabelle di contingenza 2 x 2 (Yule, 1900). Utilizzando la notazione riportata di seguito:

[11] Q, in onore dello statistico belga Quetelet.

Tabella 2.5 – Notazione adottata per la costruzione delle tabelle di contingenza.

Evento	Y = 1	Y = 0
X = 1	a	b
X = 0	c	d

L'indice Q di Yule è definito come:

$$Q = \frac{ad-bc}{ad+bc} \qquad (2.1)$$

dove le lettere *a*, *b*, *c*, *d* identificano rispettivamente: il numero di casi in cui entrambi gli eventi X e Y si sono verificati, il numero di casi in cui si è verificato X ma non Y, il numero di casi in cui si è verificato Y ma non X, e il numero di casi in cui nessuno dei due eventi si è verificato.

Per N sufficientemente grande, la distribuzione di Q è Normale, con varianza σ_Q^2

$$\sigma_Q^2 = \frac{1}{4} \cdot (1 - Q^2)^2 \cdot \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)$$

e intervallo di confidenza di livello α pari a

$$Q \pm z_\alpha \cdot \sqrt{\sigma_Q^2}$$

con z_α quantile di livello α della distribuzione Normale standard.

L'indice Q può essere considerato un caso particolare dell'indice di cograduazione γ di Goodman e Kruskal per variabili ordinali. È possibile, inoltre, fornire un'interpretazione dell'indice Q di Yule in termini di Odd Ratio (OR), misura ben nota soprattutto in epidemiologia e statistica medica, come $Q = \frac{OR-1}{OR+1}$, con $OR = \frac{ad}{bc}$. Rispetto all'OR, la Q di Yule è una misura simmetrica, con supporto che varia tra -1 e 1.

Un valore di Q pari a 0 indica assenza di associazione tra i due eventi X, Y; un valore $Q = 1$ è osservabile quando una cella nella diagonale secondaria *b*-*c* è pari a 0- o quando si è in presenza di associazione perfetta tra X e Y, ossia $b = c = 0$.

I diagrammi di seguito illustrano i tre casi corrispondenti per $Q = 1$: $b = 0$, $c = 0$, $b = c = 0$.

Evento	Y = 1	Y = 0
X = 1	a	**0**
X = 0	c	d

Evento	Y = 1	Y = 0
X = 1	a	b
X = 0	**0**	d

Evento	Y = 1	Y = 0
X = 1	a	**0**
X = 0	**0**	d

Analogamente, $Q = -1$ si verifica quando almeno una cella della diagonale principale *a*-*d* è pari a 0, riducendosi quindi al calcolo $Q = \frac{-bc}{+bc} = -1$.

Nell'ambito della ricerca clinica, l'indice Q di Yule è comunemente utilizzato per misurare l'associazione tra variabili dicotomiche quali per esempio la presenza o assenza di sintomi clinici, la risposta o meno ad un trattamento terapeutico, l'incidenza di due condizioni patologiche rispetto ad un fattore di rischio (Altman, 1991).

Tra i principali punti di forza di questo indice vi sono la semplicità di calcolo e l'interpretazione intuitiva. Le maggiori criticità invece, sollevate già al tempo dal contemporaneo collega Karl Pearson, riguardano la limitazione applicativa al solo contesto in cui le variabili siano classificate come binarie, e la sensibilità dell'indice ai valori marginali della tabella di contingenza: qualora vi sia uno sbilanciamento nella frequenza di una cella e di conseguenza in un valore marginale di riga o di colonna, l'indice tende ad essere più instabile e assumere valori estremi, con il rischio di non riflettere la corretta relazione tra le variabili osservate. Inoltre, qualora si renda necessario aggregare una tabella I x J alla specificazione 2 x 2, la stima dell'indice Q dipenderà inevitabilmente dalla scelta di aggregazione fatta (Agresti, 2013).

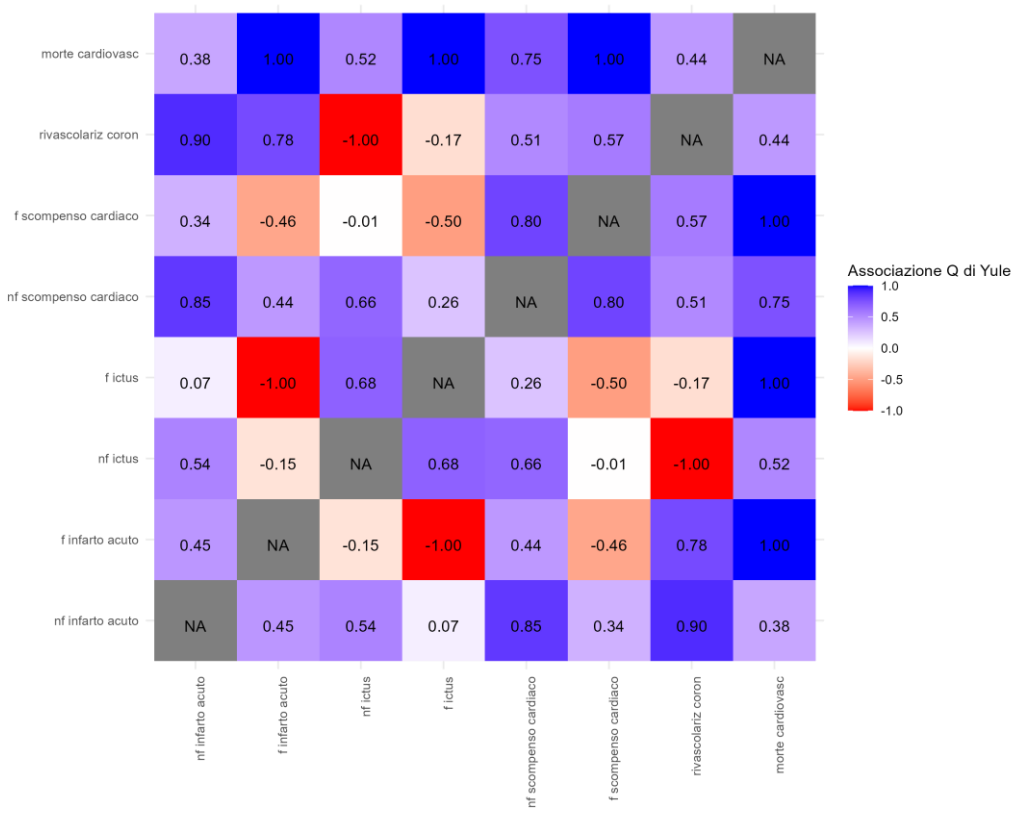


Figura 2.1 – Matrice di associazione Q di Yule per gli esiti cardiovascolari osservati. Con "nf" ed "f" si indicano rispettivamente gli eventi di tipo non fatale e fatale.

Dall'osservazione della matrice stimata in *Figura 2.1* si deduce, coerentemente con quanto ci si aspettava per costruzione, associazione perfetta positiva (Q = 1) tra la morte cardiovascolare e tutti gli eventi ad impatto fatale, e associazione positiva di magnitudo elevata tra eventi dello stesso tipo ma di diverso impatto: per esempio, si osserva associazione pari a 0,8 tra lo scompenso cardiaco fatale e lo scompenso cardiaco di tipo non fatale, associazione pari a 0,68 tra ictus di tipo fatale e non fatale. Si osserva inoltre associazione di magnitudo discreta tra eventi dello stesso impatto (fatale o non fatale), dove però la direzione dell'associazione cambia da positiva a negativa a seconda che l'impatto sia rispettivamente di tipo non fatale o fatale; si riportano a titolo esemplificativo la relazione tra scompenso cardiaco e ictus non fatale (Q = 0,66), e la relazione tra scompenso cardiaco e ictus fatale (Q=-0.50). Per quanto riguarda l'intervento di rivascolarizzazione coronarica, questo risulta positivamente e fortemente associato con l'infarto acuto del miocardio di tipo non fatale (Q = 0,90) e fatale (Q = 0,78). Le associazioni più deboli si riscontrano tra

ictus e infarto acuto del miocardio (a prescindere dall'impatto fatale o non fatale), e tra ictus non fatale e scompenso cardiaco fatale. L'indice Q di Yule restituisce associazione pari a -1 in corrispondenza delle relazioni bivariate in cui la diagonale principale *a-d* presenti una cella pari a 0: in questo caso, tutti i soggetti che sono stati sottoposti a rivascolarizzazione coronarica non hanno sperimentato ictus non fatale ($a = 0$), e tutti i soggetti che hanno sperimentato ictus fatale, non hanno sperimentato infarto acuto del miocardio fatale ($a = 0$).

2.4.2 INDICE DI CORRELAZIONE TETRACORICA TRA OUTCOME CARDIOVASCOLARI

Volendo approfondire le relazioni tra gli esiti cardiovascolari sotto una prospettiva che parte da assunti differenti sul meccanismo generatore dei fenomeni misurati, è possibile utilizzare il coefficiente di correlazione tetracorica τ. Questo, proposto da Pearson nel 1901, misura la forza e la direzione della relazione lineare tra due variabili misurate come binarie ma per le quali si assume distribuzione sottostante normale bivariata. Nel caso in esame, questo strumento si dimostra adeguato in quanto è plausibile ipotizzare che le variabili cardiovascolari misurate in forma dicotomizzata rappresentino il momento di rottura di un processo latente di modificazione del rischio (di sperimentare l'evento) inteso come funzione continua.

Le variabili osservate sono dunque messe in relazione tramite la tabella di contingenza 2 x 2, e il coefficiente di correlazione tetracorica τ di Pearson è definito come quel coefficiente τ che soddisfa

$$\int_{-\infty}^{z_2} \int_{-\infty}^{z_1} \phi(x_1, x_2, \tau)\, \mathrm{d}x_1\, \mathrm{d}x_2 = \frac{a}{N}$$

dove $\phi(x_1, x_2, \tau)$ indica la densità normale bivariata di media zero, varianza 1 e correlazione τ, mentre z_1 e z_2 indicano le deviazioni *standard* normali corrispondenti alle probabilità marginali $(a + c)/N$ e $(a + b)/N$.

Le probabilità di cella dei quattro quadranti formati dalla dicotomizzazione delle variabili per le righe x_1=z_1 e $x_2 = z_2$ corrispondono ad a/N, b/N, c/N, d/N (K. Pearson, 1901).

Un risultato metodologico importante che verrà sfruttato per il calcolo del coefficiente di correlazione tetracorica τ e per dimostrarne le proprietà statistiche riguarda il lavoro presentato da M.A. Hamdan nel 1970, il quale mostra come la stima del coefficiente di correlazione tetracorica $\hat{\tau}$ sia equivalente alla stima di massima verosimiglianza Fisheriana per la correlazione ρ in una matrice 2 x 2 (Hamdan, 1970). Si assuma che la tabella di contingenza sia risultante da una distribuzione bivariata normale con coefficiente di correlazione ρ e funzione di densità $f(x, y, \rho)$ data dall'identità di Mehler[12]

$$f(x, y, \rho) \; = \; \phi(x)\phi(y)\left\{\sum_{ii=1}^{+\infty} \frac{p^i}{i!} Hi(x)Hi(y)\right\}$$

dove $\phi(x)$ indica la funzione di densità normale per x e $H_i(x)$ indica l'insieme dei polinomi di Hermite-Chebyshev non normalizzati; si assuma inoltre h e k pari al valore dei punti standardizzati per la dicotomizzazione delle due variabili continue ottenuta adattando una distribuzione univariata normale per la variabile x marginale (totali di riga) e y marginale (totali di colonna) rispettivamente; si dimostra dunque che

$$w(\hat{\rho}) \; = \; N^{-1}(ad - bc) \; = w(\hat{\tau})$$

con

$$w(\rho) = \phi(h)\phi(k)\left\{\sum_{ii=1}^{+\infty} \frac{p^i}{i!} H_{i-1}(h)H_{i-1}(k)\right\}$$

Il risultato sopra esposto, grazie alle proprietà della funzione di verosimiglianza, permette di calcolare agilmente un'approssimazione asintotica per la varianza di τ, per la quale Pearson al contrario era riuscito a ricavare un'espressione piuttosto complicata e che risultava nella tabulazione per alcuni valori specifici. La deviazione standard $\hat{\sigma}(\tau)$ assume quindi la seguente forma asintotica compatta:

$$\hat{\sigma}(\tau) \; = \frac{1}{Nf(h, k, \tau)}\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)^{-1/2} \; = \; \hat{\sigma}(\hat{\rho}) \; = \; 1 \Bigg/ \left[\frac{\partial^2 logL}{\partial \rho^2}\right]_{\rho=\tau}$$

[12] Per ulteriori approfondimenti, si veda (Kendall, 1941).

È importante infine sottolineare come grazie al risultato di Hamdan sia possibile verificare le proprietà di consistenza, efficienza asintotica e normalità asintotica per lo stimatore $\hat{\tau}$.

Il coefficiente di correlazione tetracorica τ ha distribuzione simmetrica e assume valori tra -1 e 1, dove -1 indica relazione lineare perfetta negativa e +1 positiva. Un valore pari a 0 indica assenza di correlazione.

Nell'ambito della ricerca clinica, l'indice di correlazione tetracorica viene utilizzato, per esempio, per la valutazione della relazione tra due test diagnostici che producono risultati binari (positivo/negativo) e che si presume misurino la stessa condizione latente continua, o per studiare la relazione (lineare) tra la risposta a trattamenti terapeutici dicotomizzati (ad esempio, miglioramento sì/no) e la presenza di determinate caratteristiche cliniche o genetiche nei pazienti.

Tra i principali punti di forza di questo indice vi sono l'appropriatezza dell'utilizzo nel caso in cui si abbia a che fare con variabili dicotomiche che si presume derivino da un fenomeno latente di tipo continuo[13], la robustezza ed efficienza rispetto ad altre misure di correlazione binaria. Rispetto ai punti di debolezza, si citano la difficoltà di calcolo dell'indice e la sensibilità a campioni di piccole dimensioni (Bonett & Price, 2005; Holgado–Tello et al., 2010). Nel caso in esame, la numerosità campionaria risulta sufficientemente grande da garantire le proprietà asintotiche (robustezza ed efficienza) dello stimatore (N = 27078) e il calcolo per τ è stato eseguito tramite la funzione di massima verosimiglianza, come accennato sopra.

Dalla *Figura 2.2* rappresentata subito sotto è possibile osservare:

- correlazione tra eventi diversi ma dello stesso impatto, per esempio tra scompenso cardiaco non fatale e infarto acuto non fatale ($\tau = 0.49$);
- correlazione tra due eventi dello stesso tipo ma di diversa intensità, come per esempio se si considerano lo scompenso cardiaco fatale con lo scompenso cardiaco non fatale ($\tau = 0.44$), o l'ictus fatale e non fatale ($\tau = 0.32$);
- τ pari a 0.58 tra l'intervento di rivascolarizzazione coronarica e l'infarto acuto non fatale del miocardio;

[13] L'adeguatezza dell'assunzione di normalità bivariata per le coppie di variabili è stata motivata all'inizio del paragrafo.

- correlazione di direzione negativa tra scompenso cardiaco e infarto acuto fatali (τ = -0.15), scompenso cardiaco e ictus fatali (τ = -0.16), ictus non fatale e infarto acuto del miocardio fatale (τ = -0.05);
- τ = 1 per costruzione tra la morte per malattia cardiovascolare e tutti gli eventi ad impatto fatale.

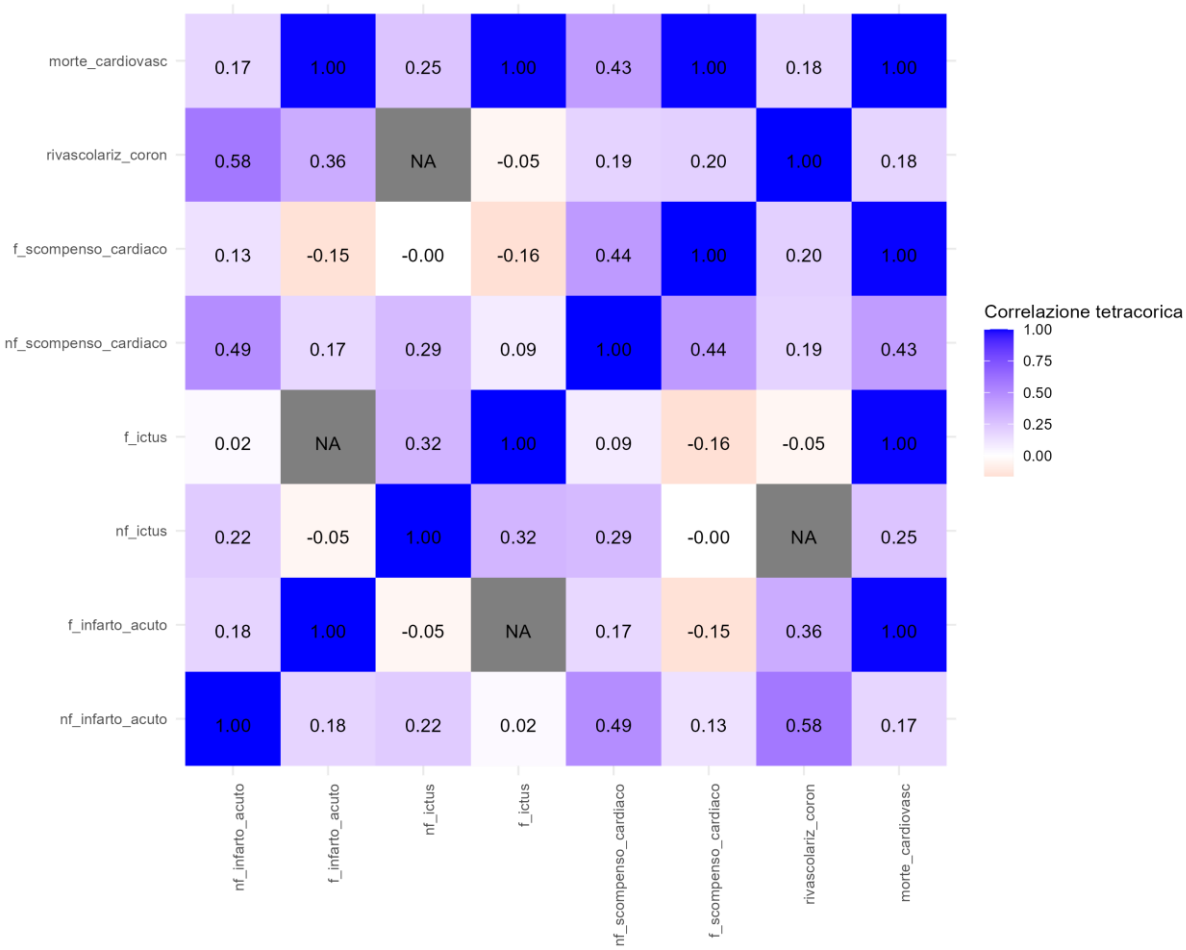


Figura 2.2 – Matrice per l'indice di correlazione tetracorica stimato $\hat{\tau}$ per gli esiti cardiovascolari tramite stima di massima verosimiglianza. Con "nf" ed "f" si indicano rispettivamente gli eventi di tipo non fatale e fatale[14].

[14] La stima dell'indice di correlazione tetracorica risulta pari a NA (teoricamente tendente ad infinito) per tutte le coppie di variabili in cui una variabile è estremamente rara, e l'altra rappresenta la quasi totalità dell'insieme di dati. Ciò accade perché l'algoritmo di stima cerca di adattare una distribuzione normale continua ai dati osservati, risultando in una pendenza estremamente ripida derivante dai margini delle tabelle di contingenza 2x2.

Confrontando i risultati ottenuti per l'indice di associazione Q di Yule e l'indice di correlazione tetracorica τ, si rileva concordanza sia nell'individuazione delle relazioni con maggior magnitudo per gli *outcome* cardiovascolari, sia nella direzione della relazione per tutte le coppie di variabili. Si sottolinea inoltre come l'indice di associazione Q di Yule presenti un coefficiente di magnitudo sistematicamente maggiore rispetto all'indice di correlazione τ: questo potrebbe essere dovuto al differente tipo di relazione stimata, che nel caso dell'associazione di Yule ha carattere più generale e dunque è in grado di catturare più tipi di relazioni tra variabili binarie, non limitandosi a quella lineare.

2.5 PECULIARITÀ DELL'INSIEME DI DATI

Nell'analisi dei dati, in particolare nella quantificazione dell'entità dei fenomeni, è importante tenere sempre a mente che il corpo di dati a disposizione non è un campione completamente rappresentativo della popolazione, per cui le stime riferite alle varie quantità (es. l'ipertensione) sono evidentemente maggiori rispetto ai dati di popolazione, essendo questi dati provenienti da centri per la cura dell'ipertensione.

Un secondo aspetto rilevante riguarda l'elevata presenza di valori mancanti in alcune variabili. A questo proposito, si presenta di seguito una breve introduzione sul tema, a cui seguirà un'analisi con lo scopo di comprendere la distribuzione e la natura dei valori mancanti all'interno del dataset, per proporre infine una strategia di gestione del problema rispetto all'obiettivo dell'analisi.

2.5.1 I DATI MANCANTI: TIPOLOGIE E DISTRIBUZIONE DELL'INSIEME DI DATI

Tra i principali ostacoli all'accuratezza e affidabilità negli studi osservazionali vi è la presenza di dati mancanti nelle variabili rilevate, che si verifica quando l'informazione per una data caratteristica di interesse non è osservata in alcuni soggetti. I dati mancanti possono verificarsi per diverse ragioni, tra cui per esempio la mancata risposta ad una domanda da parte dei soggetti considerati, errori di varia natura nella raccolta e codifica dei dati, o la perdita di contatto con il soggetto rispondente durante il periodo di *follow-up.* La presenza di dati mancanti, se non opportunamente gestita, può compromettere la qualità di uno studio

scientifico introducendo distorsione nelle stime e riducendo la validità interna ed esterna dei risultati (Graham, 2009).

I dati mancanti possono essere raggruppati in tre macro-categorie, ciascuna delle quali suggerisce a sua volta il metodo più opportuno per fronteggiare la problematica: i dati *mancanti completamente a caso* (MCAR, Missing Completely At Random) costituiscono valori mancanti in modo totalmente casuale rispetto alla matrice di dati a disposizione, che non dipendono dunque da alcuna caratteristica rilevata o non rilevata nello studio. Un dato si dice *mancante a caso* (MAR, Missing At Random) se la verosimiglianza per il dato mancante non è legata ad alcun valore della variabile stessa, condizionatamente ad una o più variabili rilevate nello studio; in questo secondo caso, dunque, un certo insieme di variabili può fungere da meccanismo a causa del quale un dato risulterà mancante. I dati *mancanti non a caso* (MNAR, Missing Not At Random) infine, identificano valori la cui probabilità di essere mancanti dipende da meccanismi non noti e associati verosimilmente al tipo di dato mancate in sé (Acock, 2005; Little & Rubin, 2019).

Gli approcci moderni all'analisi dei dati mancanti tendenzialmente presumono che i dati siano MCAR o MAR. Tuttavia, nella realtà non è possibile stabilire il meccanismo generatore con certezza, poiché solitamente -come anche in questo caso- non si dispone di informazioni sufficienti circa la natura dei valori mancanti e i fattori che ne determinano la presenza.

Tra le varie tecniche per la gestione dei dati mancanti si menzionano la *complete case analysis* o *listwise deletion*, la sostituzione con la media/moda/mediana, l'imputazione singola, l'imputazione multipla, la *Full-Information Maximum Likelihood (FIML)* (Little & Rubin, 2002). La scelta della tecnica più opportuna da adottare dipenderà dalla tipologia di dato mancante assunta, dalla quantità di osservazione mancante, nonché dall'obiettivo dello studio e il contesto di analisi.

Si riporta di seguito una rappresentazione grafica della matrice dei dati a disposizione, con lo scopo di visualizzarne intuitivamente la distribuzione e la quantità di valori mancanti. Subito sotto, viene rappresentata la distribuzione di frequenza percentuale dei valori mancanti per ciascuna variabile disponibile, così da porre maggiore enfasi sul quantitativo di informazione mancante e verificare quali siano le variabili maggiormente coinvolte nel problema considerato.

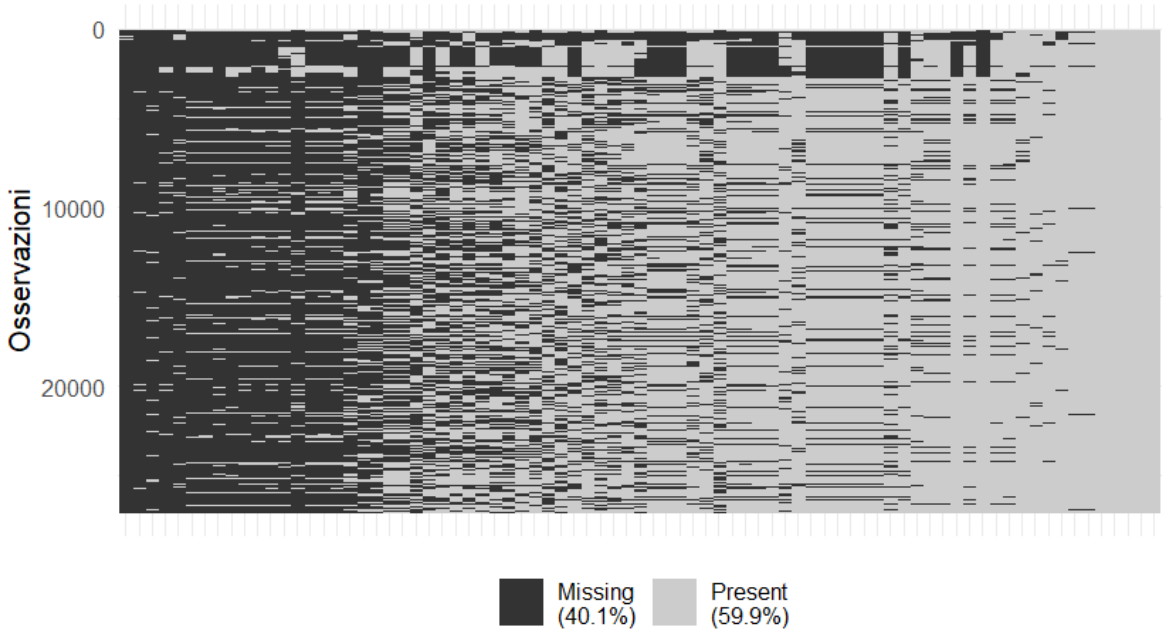


Figura. 2.3 – Rappresentazione grafica della matrice dei dati rispetto alla presenza di valori mancanti; percentuale totale di valori mancanti (Missing) e rilevati (Present).

A colpo d'occhio, la *Figura 2.3* mostra una considerevole quantità di valori mancanti, concentrati soprattutto in specifiche variabili (colonne) e nella prima porzione di unità statistiche (righe). Vista l'anonimizzazione delle rilevazioni, non disponendo di informazioni esplicite relative al centro e la modalità di raccolta dei dati per i vari soggetti, e considerato questo approfondimento necessario ma non di primario interesse rispetto all'obiettivo dell'analisi, non si ritiene prudente e opportuno formulare in questa fase assunzioni specifiche sul meccanismo generatore dei valori mancanti; sì propenderà dunque verso una soluzione metodologica a favore di un approccio cauto, garantendo allo stesso tempo l'affidabilità dei risultati.

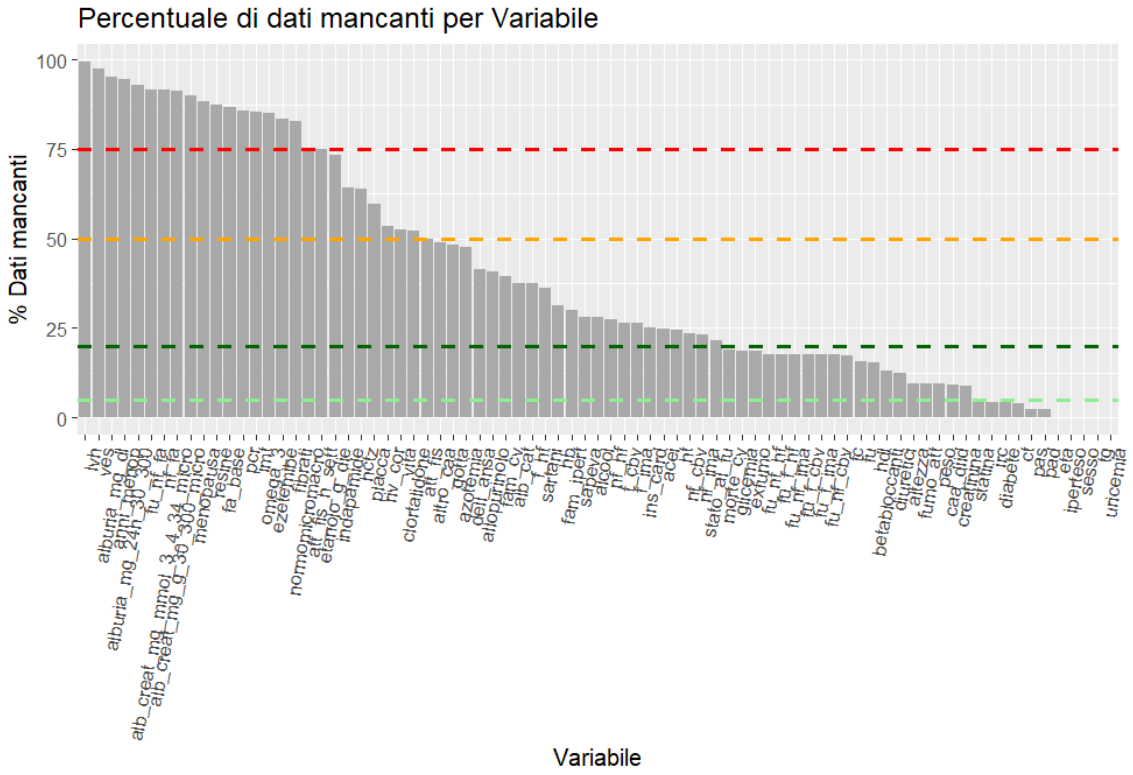


Figura 2.4 – Grafico a barre della frequenza percentuale ordinata in senso decrescente di dati mancanti rispetto alle variabili misurate.

Le variabili che presentano una quantità di valori mancanti uguale o superiore al 75% sono 18 (23%); queste fanno riferimento alla misurazione di fattori infiammatori («VES», «PCR»), alla funzionalità renale (livelli di albuminuria e rapporto tra albumina e creatinina), alla menopausa, all'anamnesi farmacologica per il trattamento del colesterolo e i trigliceridi, allo spessore della placca aterosclerotica carotidea («IMT»), e alla rilevazione dell'evento cardiovascolare non fatale di tipo fibrillazione atriale. A seguire, con una percentuale che varia tra il 50% e il 75% si trovano misurazioni relative allo stile di vita dei soggetti (assunzione di alcool, abitudine al fumo) e all'assunzione di specifici farmaci antipertensivi («indapamide», «hctz»); anche l'intervento di rivascolarizzazione coronarica si colloca in questa fascia, con il 54% di valori mancanti. Tra gli eventi cardiovascolari di interesse per l'analisi, soltanto l'intervento di rivascolarizzazione coronarica presenta una percentuale di valori mancanti superiore al 50%.

Solitamente, variabili che presentino una soglia di dati mancanti superiore al 20% sono considerate di difficile utilizzo, in quanto è necessario adottare tecniche

avanzate di gestione e analisi dei dati, con il rischio non trascurabile di incorrere in distorsione e risultati imprecisi per gli *standard error* delle stime. Qualora la percentuale sia inferiore al 5% invece, l'informazione mancante può considerarsi trascurabile o comunque di semplice gestione (Tabachnick & Fidell, 2013).
Osservando la *Figura 2.4*, si può notare come le variabili di maggior interesse per l'obiettivo clinico in esame presentino una quota minima (≤ 10%) o nulla di informazione mancante; in particolare, si fa riferimento ai valori pressori («PAD», «PAS», 2,2%) al dosaggio plasmatico di acido urico («uricemia»), alla variabile clinica «iperteso», al dosaggio plasmatico di creatinina («creatinina», 8,8%), per terminare con variabili demografiche di controllo come sesso ed età. Le variabili che non presentano alcun dato mancante sono il sesso, il dosaggio sierico dei trigliceridi («TG», mg/dL), il dosaggio plasmatico di acido urico («uricemia», mg/dL) e la misurazione dell'ipertensione secondo la definizione clinica *iperteso.*

2.5.2 ANALISI DEI DATI MANCANTI PER GLI OUTCOME CARDIOVASCOLARI

Con riferimento agli otto *outcome* cardiovascolari presi in considerazione in questa analisi, si riporta alla pagina seguente la tabella rappresentante la distribuzione di frequenza assoluta e percentuale per ciascuno di questi, unitamente alla percentuale di valori mancanti.

Tabella 2.6 - Distribuzione di frequenza univariata per gli *outcome* cardiovascolari: frequenza assoluta e percentuale, frequenza percentuale di valori mancanti nell'insieme di dati.

OUTCOME	**Frequenza assoluta**	**Frequenza percentuale**	**TOTALE**	**Valori mancanti (%)**
Morte cardiovascolare	1552	7,1%	21937	19%
Infarto acuto del miocardio, non fatale	456	2,2%	20307	23%
Infarto acuto del miocardio, fatale	432	2,2%	19500	26%
Ictus, non fatale	395	1,9%	20353	23%
Ictus, fatale	363	1,8%	19518	27%
Scompenso cardiaco, non fatale	272	1,4%	19415	27%
Scompenso cardiaco, fatale	420	2,5%	16512	38%
Rivascolarizzazione coronarica	173	1,4%	12419	54%

Come si osserva dalla Tabella 2.6, la percentuale di dati mancanti per ciascun *outcome* cardiovascolare è cospicua, passando da un valore minimo del 19% con la morte cardiovascolare, sino ad arrivare al 54% per l'intervento di rivascolarizzazione coronarica. Inoltre, studiando congiuntamente gli *outcome* cardiovascolari per ciascun *record* del dataset, si evince che solamente 9623 soggetti (35,5%) non mostrano alcun dato mancante nelle rilevazioni, mentre per 4695 soggetti (17,3%) non è presente alcuna informazione in corrispondenza degli eventi cardiovascolari considerati.

Dalle osservazioni appena esposte, risulta evidente come ci si trovi di fronte alla necessaria gestione di una significativa quantità di valori mancanti per i singoli eventi, oltre che ad una sostanziale difficoltà nell'impiegare per la fase di inferenza quel 17,3% di osservazioni che non presenta alcun evento cardiovascolare rilevato tra quelli di interesse. Diventa quindi di primaria importanza massimizzare l'utilizzo delle osservazioni disponibili preservandone il maggior numero possibile, così da garantire una maggior potenza statistica all'analisi, e stime più precise. Queste considerazioni vengono ulteriormente corroborate dalla presenza di *outcome* caratterizzati da una bassa prevalenza -compresa tra l'1% e il 3%, ad eccezione della

morte cardiovascolare-, per cui è necessario salvaguardare la rilevazione di più casi possibili.

2.5.3 CREAZIONE DELLA VARIABILE RISPOSTA OVERALL

Sulla base delle osservazioni fatte poco sopra, e dai risultati ottenuti dalle matrici di associazione e correlazione per gli *outcome* cardiovascolari (si veda la sezione 2.4), si è ritenuto ragionevole procedere nella fase di modellazione statistica creando un *outcome* composito denominato "almeno un evento cardiovascolare" (*overall*), che assume valore pari a 1 qualora il soggetto sperimenti almeno un evento cardiovascolare (compresa la morte cardiovascolare), e 0 altrimenti[15].

Questo approccio permette di ridurre notevolmente l'impatto dei valori mancanti con riferimento agli eventi cardiovascolari, mantenendo un adeguato numero di osservazioni (N = 22184) e supportando di conseguenza una buona potenza statistica per l'analisi. Al netto dei soggetti per cui non è presente alcuna rilevazione per gli *outcome* cardiovascolari, la prevalenza stimata per *l'outcome* composito *overall* nell'insieme di dati rimanente è pari al 10,8% (2395 soggetti).

Oltre alle motivazioni metodologico-statistiche a supporto della creazione *dell'outcome* composito *overall*, è importante sottolineare come questo possieda l'ulteriore vantaggio di rappresentare una misura semplice e intuitiva per la misurazione della presenza di eventi cardiovascolari, facilitando la comunicazione tra clinici e professionisti afferenti ad altri settori. In aggiunta, diverse fonti bibliografiche di letteratura medica supportano l'utilizzo di questo indicatore sottolineandone la rilevanza in ambito di ricerca clinica come misura del rischio complessivo di evento cardiovascolare, ma anche della salute cardiovascolare più in generale; ciascuno degli eventi cardiovascolari considerati è infatti accomunato agli altri per la propria intrinseca capacità di riduzione della qualità della vita per i soggetti che lo sperimentano.

Nonostante i vantaggi associati all'utilizzo *dell'outcome* composito, è importante riconoscerne le sfide annesse: combinando più *outcome* cardiovascolari, infatti,

[15] Se tra gli *outcome* cardiovascolari sono presenti valori mancanti, il conteggio per ciascun record rimane valido e risulterà pari a 1 purché il soggetto sperimenti almeno un evento; qualora un soggetto abbia registrato solo 0 e valori mancanti (NA), *l'outcome* osserverà valore 0; qualora un soggetto presenti solo NA, *l'outcome* sarà pari a NA.

diventa cruciale conoscerne la rilevanza clinica e la frequenza, così da prevenire interpretazioni fuorvianti o poter approfondire il risultato marginale. Quando, per esempio, le varie componenti dell'*outcome* composito presentano differenze in termini di importanza clinica, questo può ripercuotersi in complicazioni sia nell'interpretazione dei risultati, sia nella possibilità di confrontare sistematicamente diversi studi clinici, a meno che *l'outcome* scelto non sia riproducibile e già ampiamente accettato nel contesto di ricerca (Armstrong & Westerhout, 2017; Baracaldo-Santamaría et al., 2023; McCoy, 2018).

Considerata la cospicua quantità di dati disponibili in termini di numerosità campionaria (N = 27078), la difficoltà nello stabilire il/i meccanismo/i generatore/i dei valori mancanti, la modesta quota di valori mancanti con riferimento alle variabili di interesse per la specificazione dei modelli ad equazioni strutturali, e l'obiettivo dell'analisi, si è deciso di procedere nella gestione dei dati mancanti secondo due strategie:

- Utilizzare come prima scelta per la fase di inferenza la tecnica della *listwise deletion*, mantenendo per la stima dei modelli ad equazioni strutturali soltanto i *record* che non presentino alcun dato mancante tra le variabili di interesse considerate congiuntamente;
- Ripetere l'analisi stimando un modello ad equazioni strutturali tramite la tecnica della *Full Information Maximum Likelihood* (FIML) che, sotto opportune condizioni, permette di sfruttare tutta l'informazione disponibile senza eliminare alcun *record* e producendo stime non distorte, errori standard efficienti e talvolta maggiore potenza statistica rispetto alla tecnica di *listwise deletion* (C. Enders & Bandalos, 2001). Ulteriori approfondimenti in merito verranno forniti al Capitolo 3 del suddetto lavoro. I risultati saranno esposti nell'Appendice C.

3.I METODI

Nel capitolo 3, *I Metodi*, si presentano nel dettaglio i metodi statistici scelti per rispondere agli obiettivi di ricerca, sottolineandone le ragioni per cui se ne è ritenuto adeguato l'utilizzo nel contesto in esame, i punti di forza e le eventuali limitazioni. I principali strumenti metodologici a cui afferiscono i metodi impiegati sono i modelli ad equazioni strutturali (SEM: *Structural Equation Models*) e i modelli grafici (GM: *Graphical Models*) per la *causal discovery*.

I modelli SEM sono impiegati con lo scopo di discutere l'appropriatezza di diverse specificazioni teorico-causali (*path* causali) ipotizzate per il problema clinico e diverse misurazioni dell'ipertensione arteriosa, attraverso la valutazione della significatività delle interdipendenze stimate (β), la quantificazione degli effetti stimati, e l'utilizzo di indicatori per la bontà di adattamento dei modelli ai dati. Si è dunque all'interno dell'area metodologica della *causal inference* in quanto a partire da diverse strutture causali definita a priori, ci si concentra sulla quantificazione degli effetti causali ipotizzati. In questo contesto, viene sviluppato un ulteriore paragrafo a presentazione della tecnica di imputazione dei dati mancanti *Full Information Maximum Likelihood* (FIML), con l'obiettivo di valutare eventuali differenze tra il "miglior" modello SEM stimato tramite tecnica di *listwise deletion*, e la corrispondente specificazione stimata tramite tecnica FIML (si veda l'Appendice C).

A questo punto, poiché la conoscenza medica non permette ad oggi di stabilire quale tra i diversi *path* causali sia più appropriato, e la natura trasversale dei dati non consente di asserire nessi causali attraverso l'utilizzo dei SEM, si prosegue nell'analisi ricorrendo alla metodologia della *causal discovery,* con l'obiettivo di aggiungere plausibilità causale alle relazioni d'interesse.

La *causal discovery* è un'area di ricerca statistica e di *machine learning* che mira a identificare relazioni causali tra variabili a partire dai soli dati osservati[16] e servendosi dell'utilizzo dei modelli grafici. I modelli grafici rappresentano lo strumento matematico portante che permette (sotto opportune assunzioni) la stima

[16] Dunque, a partire da misure di associazione quali correlazione, indipendenza condizionata, verosimiglianza etc.

di una rete di relazioni causali a partire dalla sola informazione osservata. In questa seconda fase di analisi, si argomenta la validità del metodo (*causal discovery*) e la robustezza dei risultati nel contesto dello studio epidemiologico osservazionale in esame.

3.1 I MODELLI AD EQUAZIONI STRUTTURALI

I modelli ad equazioni strutturali (SEM) costituiscono uno strumento statistico multivariato che combina aspetti dell'analisi fattoriale con l'approccio di regressione multipla e la *path analysis* di S. Wright (1921), con lo scopo di esplorare relazioni complesse e dinamiche causali tra fenomeni che possono essere sia misurabili direttamente sia identificati tramite costrutti latenti. Questa metodologia è particolarmente impiegata nell'ambito delle scienze economiche e sociali così come nell'epidemiologia e la psicologia, ove solitamente le interazioni tra variabili sono particolarmente complicate e possono coinvolgere relazioni causali multiple.
Tra i principali vantaggi dei modelli SEM si annoverano la loro peculiare capacità di gestire simultaneamente più relazioni di dipendenza, facendo sì che una variabile dipendente in un'equazione possa risultare indipendente in altre parti del sistema[17], e la possibilità di contemplare l'utilizzo di variabili sia osservate (direttamente misurabili) che latenti; le variabili latenti, non direttamente osservabili, potranno essere altresì misurate tramite l'utilizzo di due o più variabili osservate, che assumono la funzione di indicatori (Byrne, 2016). I modelli ad equazioni strutturali consentono inoltre di condurre analisi di mediazione, risultando dunque un valido strumento per esplorare la relazione tra due variabili che (come spesso accade) si definiscono in relazione anche attraverso il rapporto con una terza variabile intermedia, chiamata appunto mediatore. Il mediatore ha dunque la peculiarità di poter generare un effetto indiretto influenzando la relazione tra le due variabili iniziali (Gunzler et al., 2013).
La capacità di gestire contemporaneamente più relazioni di dipendenza per il problema clinico in esame unitamente alla possibilità di stimare l'effetto diretto o

[17] In questo contesto, le variabili si definiscono endogene qualora fungano da variabili dipendenti all'interno del modello SEM per almeno un'equazione, mentre sono classificate come esogene se mantengono sempre il ruolo di variabile indipendente all'interno del modello.

indiretto di fattori causali sull'insorgenza di almeno un evento cardiovascolare -e più nello specifico, l'effetto di mediazione tra ipertensione, uricemia e variabile risposta *overall*- forniscono il razionale iniziale a supporto della scelta di questa famiglia di modelli.

L'applicazione dei modelli ad equazioni strutturali prevede diversi passaggi, che possono variare a seconda del metodo di stima scelto. Tra i più utilizzati si ricordano i *Covariance-Based Methods* (CB-SEM) e il *Partial Least Squares* (PLS) *method*. In questo lavoro si è scelto di utilizzare il metodo di stima parametrico basato sulla matrice di covarianza (CB-SEM) in quanto, a differenza del metodo PLS, permette di fare inferenza in modo diretto grazie all'utilizzo della funzione di verosimiglianza e dispone di indicatori per la valutazione della bontà di adattamento del modello ai dati; inoltre, la dimensione del campione in esame è sufficientemente ampia da garantirne un utilizzo appropriato. Con riferimento alle eventuali limitazioni associate al metodo di stima *covariance-based*, è dimostrato come questo sia robusto qualora i dati non soddisfino l'ordinaria assunzione di normalità multivariata (Bollen, 1989).

Di seguito si riassumono brevemente i passaggi necessari all'applicazione dei modelli SEM stimati tramite metodo *covariance-based* con assunzione di normalità multivariata dei dati. Per ulteriori approfondimenti a presentazione della metodologia statistica sottostante le cinque fasi elencate, si rimanda all'Appendice A:

- **Specificazione del modello:** questo primo passaggio prevede la definizione delle relazioni teoriche tra le variabili, attraverso lo studio di diverse fonti di letteratura e/o la conoscenza pregressa da parte dei ricercatori coinvolti;
- **Identificazione del modello:** una volta specificate le relazioni teoriche, si procede verificando la possibilità di ottenere una soluzione unica per la stima dei parametri, assicurando quindi che il modello sia *identificato* e che ci siano sufficienti dati per stimare tutti i parametri in modo accurato;
- **Stima dei parametri:** il metodo CB-SEM si basa sulla ricerca di un modello la cui matrice di covarianza $\sum$ si avvicini maggiormente alla matrice di covarianza osservata S. La stima è solitamente ottenuta (come in questa

analisi) tramite la funzione di verosimiglianza sotto assunzione di normalità multivariata dei dati.

- **Verifica del modello:** si valuta poi la bontà di adattamento del modello ai dati tramite vari indici (Chi-quadro, *Comparative Fit Index* (CFI), *Root Mean Square Error of Approximation* (RMSEA) etc.), con lo scopo di ottenere informazioni circa la ragionevolezza delle interdipendenze ipotizzate tra le variabili;
- **Modifica del modello:** sulla base dei risultati di stima e bontà di adattamento ottenuti, si valuta la possibilità di apportare eventuali correzioni al modello, con il fine di migliorarne l'adattamento e la coerenza con i dati empirici osservati (Bollen, 1989).

3.1.1 DESCRIZIONE DELLE VARIABILI UTILIZZATE PER IL MODELLO AD EQUAZIONI STRUTTURALI

Il *Capitolo 1* della presente tesi fornisce un'ampia panoramica sulle relazioni che intercorrono tra le variabili cliniche considerate -ovvero l'ipertensione, l'uricemia, gli eventi cardiovascolari, il diabete e la creatinina- e sulle principali evidenze scientifiche ad oggi riportate in letteratura.

Nei *paragrafi 3.1.1* e *3.1.2* si desidera riprendere brevemente quanto affrontato in precedenza, con lo scopo di riassumere il razionale sottostante le diverse specificazioni scelte per i modelli ad equazioni strutturali (SEM) che andranno stimati. Il centro del problema per i tre *path* causali in esame consiste nel poter comprendere cosa venga prima tra uricemia ed ipertensione arteriosa nella relazione con l'insorgenza di almeno un evento cardiovascolare. (*overall*); inoltre, per ciascun *path* causale verrà valutato l'utilizzo di tre differenti misurazioni per l'ipertensione arteriosa: la misurazione dicotomica *iperteso*, e le misurazioni continue PAD, PAS che rilevano rispettivamente la pressione arteriosa diastolica e sistolica. A queste variabili cliniche verranno aggiunte anche il sesso e l'età in qualità di variabili di "controllo", insieme al diabete mellito di tipo II e i livelli di creatinina. Si riporta una tabella descrittiva univariata per la distribuzione di frequenza delle variabili che verranno considerate in fase di modellistica, tenendo presente che per

la fase di stima saranno assunte come provenienti da una distribuzione continua gaussiana.

L'acido urico, la cui condizione di elevata concentrazione nel sangue è nota come iperuricemia, è dimostrato associato all'aumento del rischio di sperimentare eventi cardiovascolari. Studi recenti hanno inoltre mostrato come elevati livelli di acido urico possano comportare un aumento del livello di pressione arteriosa e del rischio di danni vascolari, ma su questo aspetto non è ancora dimostrabile un nesso causale. La presenza di iperuricemia è dunque riconosciuta come un predittore indipendente per diversi eventi cardiovascolari, tra cui in particolare l'infarto acuto del miocardico, l'ictus ischemico, lo scompenso cardiaco, mentre la sua relazione direzionata con la pressione arteriosa rimane ancora motivo di investigazione (Feig et al., 2008b; Sharaf El Din et al., 2017). Nei modelli specificati si assume sia la relazione ipertensione arteriosa → uricemia, sia la relazione direzionata in senso opposto: uricemia → ipertensione arteriosa. Rispetto all'insorgenza di eventi cardiovascolari, verrà sempre assunta la relazione diretta uricemia → *overall*.
L'ipertensione arteriosa, nelle sue diverse definizioni presentate al *paragrafo 2.2*[18], è riconosciuta un noto fattore di rischio per l'insorgenza di malattie cardiovascolari. Anche la sua relazione con l'acido urico è ben documentata: meccanismi fisiopatologici come la disfunzione endoteliale e l'attivazione del sistema renina-angiotensina-aldosterone si sono intatti dimostrati implicati nella relazione tra ipertensione arteriosa e iperuricemia (Whelton et al., 2018b; Williams et al., 2018); i meccanismi causali che regolano l'interdipendenza tra le due variabili però non sono ancora chiari.
Il diabete mellito di tipo II è fortemente correlato con l'aumento del rischio di eventi cardiovascolari, poiché provoca disfunzione endoteliale ed accelera il processo di aterosclerosi. Nella relazione tra diabete e i livelli sierici di acido urico, si evidenzia

[18] Si ricordano qui le due definizioni: le linee guida ESH-ESC definiscono l'ipertensione arteriosa come una condizione caratterizzata dalla presenza di un livello di pressione arteriosa diastolica e/o sistolica rispettivamente ≥ 90 mmHg e ≥ 140 mmHg, rilevata per mezzo di almeno due misurazioni. Nella ricerca clinica un soggetto è iperteso se è verificata almeno una delle seguenti condizioni: la rilevazione ripetuta della pressione arteriosa diastolica è ≥ 90 mmHg o la rilevazione ripetuta della pressione arteriosa sistolica è ≥ 140 mmHg; è in corso una terapia farmacologica con farmaci utilizzati per il controllo della pressione arteriosa; è presente una storia clinica di ipertensione arteriosa certificata da un clinico esperto.

come l'iperuricemia possa esacerbare l'insulino-resistenza, contribuendo all'infiammazione sistemica e aumentando il rischio di complicanze cardiovascolari nei pazienti diabetici (American Diabetes Association, 2020). In questa analisi, il diabete verrà sempre considerato come variabile esogena (esplicativa) per l'insorgenza di almeno un evento cardiovascolare (*overall*), e come variabile endogena (dipendente) rispetto al sesso e l'età dei soggetti.

La creatinina rappresenta un indicatore clinico della funzionalità renale (eGFR), che è a sua volta associata a un aumento del rischio di malattie cardiovascolari (Levey et al., 2006). La relazione direzionata tra creatinina, ipertensione arteriosa ed uricemia è tutt'ora incerta. Qui, si ipotizzano tre diversi *path* causali per la relazione tra creatinina, ipertensione ed uricemia; per ciascuno di questi, non viene ipotizzato un effetto diretto tra creatinina ed eventi cardiovascolari, in accordo con le evidenze di letteratura.

Le variabili demografiche, sesso ed età, in questa analisi ricoprono il ruolo di variabili di controllo esogene, permettendo di isolare gli effetti specifici delle altre variabili sul rischio cardiovascolare. Per quanto concerne il sesso, diversi studi epidemiologici ne hanno mostrato l'effetto in termini di differenze significative nella prevalenza, la presentazione clinica e la prognosi delle malattie cardiovascolari: in particolare, risulta che gli uomini tendano a sviluppare malattie cardiovascolari in età più giovane rispetto alle donne, le quali beneficiando della produzione fisiologica degli estrogeni, sono biologicamente protette dall'insorgenza di malattie cardiovascolari fino al termine della produzione degli stessi, ovvero nel periodo della menopausa (Benjamin et al., 2019); con l'inizio della menopausa invece, il rischio d'insorgenza di malattie cardiovascolari per le donne aumenta notevolmente, raggiungendo il valore stimato per gli uomini. Anche l'età rappresenta un importante fattore di rischio per l'insorgenza di malattie cardiovascolari. Con l'avanzare dell'età infatti, aumenta più che proporzionalmente anche la probabilità di sviluppare ipertensione, diabete e disfunzioni renali, i quali sono a loro volta fattori di rischio per l'insorgenza di eventi cardiovascolari (Mozaffarian et al., 2015).

Tabella 3.1 – Distribuzione di frequenza univariata per le variabili inserite nelle diverse specificazioni del modello SEM per l'insorgenza di almeno un evento cardiovascolare.

Variabile	**Modalità**	**Frequenza assoluta**	**Frequenza percentuale**	**Valori mancanti (%)**
Overall: Almeno un evento cardiovascolare	Sì	2395	10,8%	18,1%
	No	19789	89,2%	
	Totale	22184	100%	
Sesso	Donne	13539	50%	0%
	Uomini	13539	50%	
	Totale	27078	100%	
Età (anni)	18-30	1423	5,3%	0.0001 %
	31-40	2425	9%	
	41-50	5215	19,3%	
	51-60	6199	22,9%	
	61-70	5613	20,7%	
	71-80	4787	17,7%	
	81-90	1312	4,8%	
	91-100	100	0,4%	
	Totale	27074	100%	
Diabete mellito di tipo due	Sì	2921	11,3%	4,3%
	No	22981	88,7%	
	Totale	25902	100%	
Classi di pressione arteriosa diastolica (mmHg)[19]	< 60	342	1,3%	2,2%
	60-79	8503	31,4%	
	80-89	7742	28,6%	
	90-99	6367	23,5%	
	100-119	3356	12,4%	
	120+	184	0,7%	
	Totale	26494	100%	
Classi di pressione arteriosa sistolica (mmHg)[20]	70-89	19	0,1%	2,2%
	90-119	3768	13,9%	
	120-129	3839	14,2%	
	130-139	5162	19,0%	
	140-155	6811	25,2%	
	155+	6890	25,5%	
	Totale	26489	100%	
Ipertensione arteriosa (si veda par 2.2)	Sì	19079	70,5%	0%
	No	7999	29,5%	
	Totale	27078	100%	
Dosaggio plasmatico creatinina (mg/dL)[21]	< 0.5 (Basso)	74	0,3%	8,8%
	0,5-1,2 (Normale)	22038	89,2%	
	≥ 1,2 (Elevato)	2593	10,5%	
	Totale	24705	100%	
	≤ 5,6	18066	66,7%	0%
	> 5,6	9012	33,3%	

[19] Classi calcolate a partire dalla variabile continua.

[20] Classi calcolate a partire dalla variabile continua.

[21] Le soglie fanno riferimento alla classificazione clinica fornita dall'Istituto Superiore di Sanità (ISS, 2022).

Variabile	Modalità	Frequenza assoluta	Frequenza percentuale	Valori mancanti (%)
Dosaggio plasmatico acido urico (mg/dL)[22]	Totale	2707	100%	
	No	19789	89,2%	
	Totale	22184	100%	

Prima di procedere nella presentazione delle diverse specificazioni scelte, è importante sottolineare come in questa analisi si lavorerà solamente con variabili osservate, andando dunque ad ottenere un caso particolare del modello ad equazioni strutturali che consta solamente della parte strutturale ed è dunque riconducibile alla *path analysis*. Nel presente lavoro si è deciso di utilizzare comunque la dicitura "modelli ad equazioni strutturali" in quanto questi rappresentano la famiglia di modelli statistici più generale e godono di maggior flessibilità e sviluppi di letteratura rispetto alla sola *path analysis*[23].

3.1.2 DEFINIZIONE DELLE DIVERSE SPECIFICAZIONI PER IL MODELLO AD EQUAZIONI STRUTTURALI

Da quanto descritto poco sopra, appare evidente la complessità delle dinamiche multifattoriali che intercorrono tra ipertensione arteriosa, diabete, uricemia, creatinina ed eventi cardiovascolari, di cui ancora non sono note con chiarezza le interdipendenze causali. In particolare, il nucleo che presenta gli interrogativi di maggior interesse riguarda il rapporto tra uricemia ed ipertensione arteriosa nella relazione con gli eventi cardiovascolari.

La costruzione di diverse specificazioni per il modello SEM vuole servire a fornire nuove evidenze circa il ruolo dell'ipertensione nella relazione tra uricemia ed eventi cardiovascolari, valutandone allo stesso tempo eventuali differenze di impatto a seconda che sia utilizzata la sua misurazione clinica (*iperteso*) piuttosto che empirica ESC-ESH (PAD, PAS).

22 La soglia adottata fa riferimento al cutoff prognostico per la mortalità cardiovascolare ottenuto nel lavoro di Virdis e al. (Virdis et al., 2020).

23 L'utilizzo dei modelli SEM offre la possibilità di ipotizzare eventuali diverse strutture di dipendenza tra gli errori di misurazione e presenta una vasta gamma di indici di adattamento necessari all'obiettivo di ricerca (Kaplan, 2009).

Le diverse specificazioni ipotizzate nei modelli SEM specificati sono dunque la conseguenza di numerose evidenze (e allo stesso tempo questioni aperte) di letteratura clinica sul tema, discusse al *Capitolo 1* e qui riassunte brevemente.

Di seguito, nei *path diagram*, si riportano le tre diverse specificazioni causali ipotizzate per i modelli SEM, prendendo come riferimento la misurazione dell'ipertensione con la variabile clinica iperteso.

Per ogni specificazione verranno stimati tre modelli SEM: due che rilevano separatamente l'ipertensione come variabile continua riferita ai soli valori pressori diastolici o sistolici rispettivamente (≥ 90mmHg e ≥ 140mmHg rispettivamente), e uno che considera l'ipertensione come fenomeno più ampio, in cui sono valutate anche la storia clinica di ipertensione arteriosa e l'aderenza a terapia farmacologia antipertensiva. La numerosità campionaria per la modellazione SEM tramite *listwise deletion* è pari a N = 22184 nel caso si misuri l'ipertensione tramite la variabile dicotomica *iperteso*, N = 21879 altrimenti. I risultati sono presentati nel *Capitolo 4*.

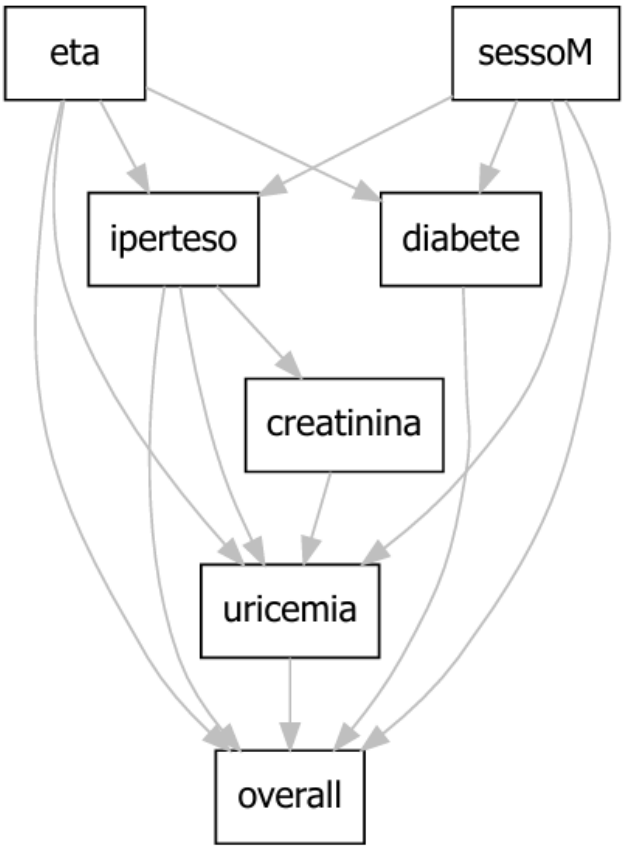


Figura 3.1 – *path diagram* per prima specificazione teorizzata M1

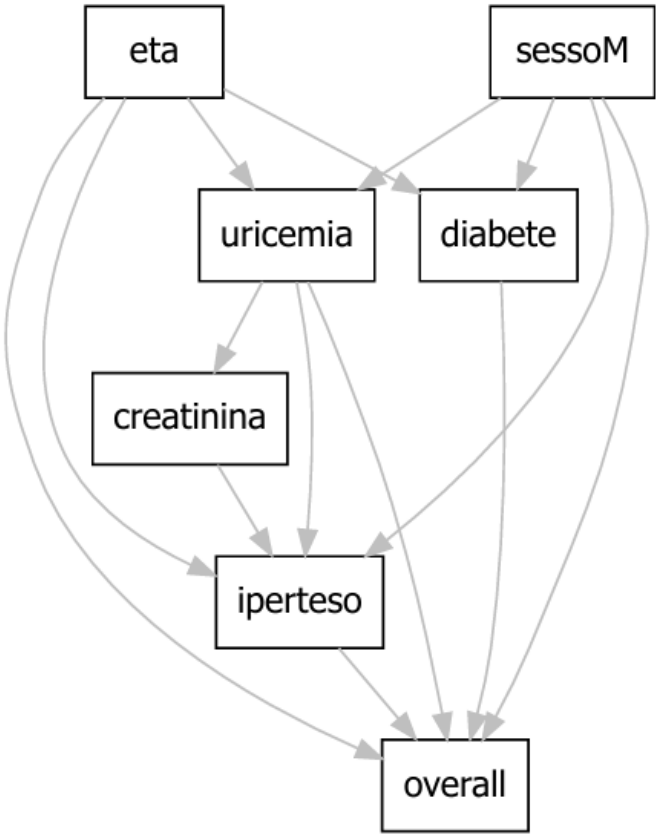


Figura 3.2 – *path diagram* per seconda specificazione teorizzata M2.

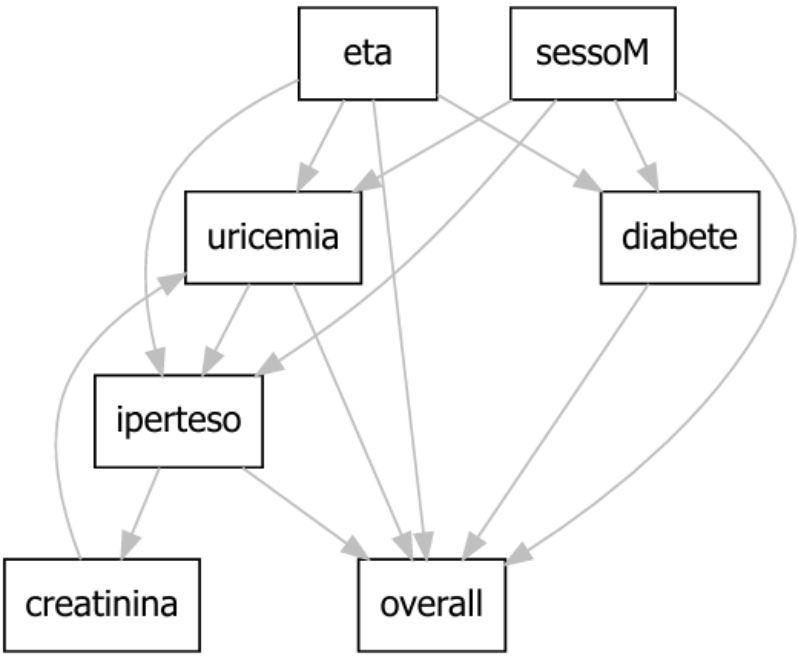


Figura 3.3 - *path diagram* per terza specificazione teorizzata M3.

Nel primo modello teorizzato (M1), si assumono il sesso e l'età come variabili endogene (esplicative) di controllo, con effetto causale diretto su ipertensione arteriosa, diabete, uricemia e l'insorgenza di almeno un evento cardiovascolare (*overall*); l'unica relazione indiretta che coinvolge sesso ed età, dunque, viene ipotizzata con la creatinina. Si assume poi il ruolo di mediatore della creatinina nella relazione tra ipertensione arteriosa e uricemia. L'ipertensione e l'uricemia a loro volta presentano un effetto diretto su *overall*, ponendo l'uricemia nella posizione di mediatore nella relazione iperteso → uricemia → *overall*. Il diabete viene trattato come ulteriore fattore di controllo, prevedendo solamente una relazione causale diretta con l'insorgenza di almeno un evento cardiovascolare (*overall*).
Rispetto al primo modello, nella specificazione M2 si assume un ruolo invertito tra uricemia ed ipertensione arteriosa, con la creatinina che rimane mediatore per la relazione direzionata (in senso opposto) uricemia → ipertensione e l'ipertensione che assume il ruolo di mediatore nel rapporto tra uricemia ed eventi cardiovascolari. Nel terzo modello invece viene stimata una relazione direzionata in senso opposto tra ipertensione arteriosa ed uricemia (rispetto a M1).

Di seguito si riporta a titolo esemplificativo la specificazione esplicita per il primo modello teorizzato M1:

$$y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \zeta_1$$

$$y_2 = \beta_{21}y_1 + \zeta_2$$

$$y_3 = \beta_{31}y_1 + \beta_{32}y_2 + \gamma_{31}x_1 + \gamma_{32}x_2 + \zeta_3$$

$$y_4 = \gamma_{41}x_1 + \gamma_{42}x_2 + \zeta_4$$

$$y_5 = \beta_{51}y_1 + \beta_{53}y_3 + \gamma_{31}x_1 + \gamma_{52}x_2 + \zeta_5$$

con $\zeta_i \sim N(0, \sigma_i^2)$ errori di misura indipendenti per le equazioni strutturali, e:

- **Variabili endogene:**
  - $x_1 = sesso\ (maschile)$
  - $x_2 = età$

- **Variabili esogene:**
  - $\boldsymbol{y_1} = ipertensione$
  - $\boldsymbol{y_2} = creatinina$
  - $\boldsymbol{y_3} = uricemia$
  - $\boldsymbol{y_4} = diabete$
  - $\boldsymbol{y_5} = overall\ (almeno\ un\ evento\ cardiovascolare)$

- *B:* matrice dei coefficienti β che misura l'impatto tra le variabili osservate endogene;
- Γ: matrice dei coefficienti γ che misura l'effetto tra le variabili osservate esogene ed endogene.

Per la stima dei modelli SEM con distribuzione normale multivariata è stata utilizzata la funzione sem del pacchetto lavaan (Rosseel, 2012) implementato nel software R (R Core Team, 2024), versione 4.3.2. L'identificabilità del modello è verificata tramite l'*output* prodotto dai comandi lanciati: in particolare, viene specificato se e dopo quante iterazioni l'algoritmo per la stima di massima verosimiglianza di $\widehat{\sum}$ (=$\sum(\hat{\theta})$) è arrivato a convergenza. Qualora non sia esplicitamente specificato, tutti i modelli di cui si mostrano i risultati sono da intendersi correttamente identificati.

3.1.3 FULL IMPUTATION MAXIMUM LIKELIHOOD PER I MODELLI AD EQUAZIONI STRUTTURALI (SEM)

Nel presente paragrafo si desidera sviluppare un approfondimento metodologico sul tema della stima di modelli SEM tramite tecniche che sfruttano la funzione di massima verosimiglianza nel caso di insiemi di dati con informazioni mancanti. In particolare, si desidera valutare l'approccio della *Full Imputation Maximum Likelihood* (FIML) come risoluzione alternativa alla problematica dei dati mancanti affrontata nei *paragrafi 2.5.1 - 2.5.3* e valutare eventuali differenze

ottenute rispetto all'utilizzo del metodo della *listwise deletion* in termini di magnitudo e precisione degli effetti stimati, e bontà di adattamento del modello[24].

Il metodo della *Full Imputation Maximum Likelihood* (FIML) è una tecnica avanzata di stima che permette di ottenere una funzione di verosimiglianza per θ basata sull'aggregazione (prodotto) delle verosimiglianze per ciascuna unità statistica, utilizzando di volta in volta tutta l'informazione disponibile (variabili) per quell'unità. La FIML permette di ottenere vantaggi significativi in presenza di dati mancanti, perché massimizza l'utilizzo dei dati senza necessità di imputazione e senza la perdita di unità statistiche. A differenza dei metodi tradizionali di Massima Verosimiglianza (ML) infatti, che richiedono l'applicazione su dataset completi e sono dunque soggetti a *bias* in contesti non MCAR, la FIML è in grado di garantire non distorsione per la stima dei parametri e proprietà di efficienza nel caso in cui le condizioni del meccanismo generatore dei dati mancanti siano assunte *Missing Completely at Random* (MCAR) o *Missing at Random* (MAR) (C. Enders & Bandalos, 2001).
La funzione di verosimiglianza FIML per il caso *i-mo* è data da:

$$L_i(\theta) = K_i \exp\left\{-\frac{1}{2}[\log|\Sigma_i| + (x_i - \mu_i)'\Sigma_i^{-1}(x_i - \mu_i)]\right\}$$

dove K_i è una costante dipendente dal numero di variabili osservate, x_i rappresenta i dati osservati per l'unità *i-ma*, e μ_i e Σ_i sono rispettivamente il vettore delle medie e la matrice di varianza-covarianza per le variabili osservate nel soggetto *i*-mo. La verosimiglianza complessiva si ottiene dunque come:

$$L(\mu, \sigma) = \prod_{I=1}^{N} L_i$$

La stima di massima verosimiglianza è usualmente ottenuta tramite procedure iterative basate sull'algoritmo di Newton-Raphson o l'algoritmo di Fisher Scoring.

[24] Si ricorda che verrà confrontato il "miglior" modello SEM stimato tramite massima verosimiglianza, con la corrispondente specificazione stimata tramite FIML.

Uno studio di simulazione Monte Carlo condotto da Enders e Bandalos (2001) ha dimostrato l'efficacia e la superiorità della FIML come metodo di stima per i modelli SEM rispetto a tecniche come la *listwise deletion*, la *pairwise deletion*, la *similar response pattern imputation* nel caso di dati MCAR o MAR: per diversi *setting* sperimentali, che prevedevano differenti livelli di dati mancanti, dimensione del campione, e magnitudine per i carichi fattoriali, lo studio ha rilevato come la FIML apportasse sistematicamente il minor *bias* e la massima efficienza nelle stime dei parametri unitamente alla minore proporzione di fallimenti in termini di convergenza dell'algoritmo di stima e tassi di errore di I^{o} Tipo quasi ottimali.
A supporto dell'utilizzo della tecnica di FIML nella gestione dei dati mancanti si citano inoltre il lavoro di Graham ed Enders (C. K. Enders, 2022; Graham, 2009).

In questa analisi, viste le considerazioni esposte nel Capitolo 2 circa l'esigua quota di valori mancanti per le variabili da modellare, l'ingente numerosità campionaria, e considerato che l'analisi specifica per identificare il meccanismo generatore dei dati mancanti non rientrava negli obiettivi del lavoro di tesi, si è ritenuto opportuno stimare in primis i modelli SEM tramite l'applicazione della tecnica di *listwise deletion*, per poi proporre in un secondo momento un confronto con il metodo di stima FIML, assumendo a priori che i dati siano almeno MAR. I risultati di questo confronto sono esposti nell'Appendice C.
Anche in questo caso, per la stima dei modelli SEM è stata utilizzata la funzione sem del pacchetto lavaan (Rosseel, 2012) con specificazione "fiml" nell'argomento missing della funzione.

3.1.4 ALCUNE NOTE SULLA STIMA DEI MODELLI SEM

Nel seguente lavoro di tesi sussistono due principali peculiarità dei dati che richiedono un'accurata gestione in fase di stima *covariance-based* tramite funzione di verosimiglianza: la non normalità dei dati, e i dati mancanti.
Per le osservazioni riguardanti la gestione dei dati mancanti si veda il paragrafo subito sopra e le osservazioni dei *paragrafi 2.5.1-2.5.3*.
Per quanto concerne la non normalità dei dati, è noto che la stima dei parametri per $\widehat{\sum} = \sum(\hat{\theta})$ rimane consistente (salvo corretta specificazione e identificazione del

modello) ma *gli standard error* stimati sono soggetti ad essere più grandi del dovuto, con la conseguenza di diminuire la probabilità di rifiutare i test di nullità per un singolo coefficiente e il test di significatività globale del modello χ^2**.** Per ovviare a questa difficoltà, si è deciso di utilizzare degli *standard error* robusti basati sull'utilizzo di una matrice di covarianza *sandwich,* del tipo:

$$Var(\vartheta) = I^{-1}(\theta)\ \widehat{U}\ I^{-1}(\theta)$$

con $I^{-1}(\theta)$ inversa della matrice di informazione per la matrice di covarianza $\widehat{\sum}$ e $\widehat{U}$ matrice dei prodotti dei residui $\hat{u}_i$.

Il test per la bontà di adattamento complessiva χ^2 verrà sostituito dalla statistica test di Satorra-Bentler (Satorra & Bentler, 1994, 2001), che riscala il valore del χ^2 di una quantità che riflette i gradi di curtosi della distribuzione multivariata. Diversi studi di simulazione hanno verificato e confermato l'efficacia di questa statistica corretta per dati non normali, anche in situazioni di moderata numerosità campionaria (Curran et al., 1996).

In lavaan, la statistica χ^2 con correzione di Satorra-Bentler è richiesta tramite l'argomento test = "satorra.bentler". Attraverso l'argomento estimator = "MLM" della funzione sem, si otterranno contemporaneamente *standard error* robusti e la statistica test scalata di Satorra-Bentler per il modello stimato tramite massima verosimiglianza. Gli stessi metodi sono impiegati efficacemente anche nel caso di dati non normali e informazione mancante (Rosseel, 2012; Yuan & Bentler, 2000).

3.2 I MODELLI GRAFICI

Nelle prossime quattro sezioni sono riassunti la terminologia e i concetti chiave necessari all'utilizzo dei modelli grafici nel contesto della *causal discovery*. Verrà poi presentato il problema della *causal discovery* proponendone un'applicazione ragionata nel contesto di studio in esame.

3.2.1 DEFINIZIONI E FONDAMENTI DEI MODELLI GRAFICI

I modelli grafici sono strumenti matematico-probabilistici utilizzati per identificare relazioni tra variabili casuali mediante distribuzioni di probabilità sui dati osservati, dove le relazioni di dipendenza (in probabilità) sono rappresentate tramite un grafo. Nel contesto della scoperta causale (*causal discovery*), i modelli grafici consentono sotto opportune assunzioni di identificare e rappresentare le relazioni causali tra variabili a partire dall'analisi delle dipendenze strutturali (associazioni) presenti nei dati.

In questa prima sezione viene riassunta la notazione essenziale associata alla teoria dei grafi. Successivamente, si presentano i concetti di indipendenza condizionata e le proprietà di Markov per grafi diretti e non diretti. Prima di procedere con la trattazione della *causal discovery*, viene fornita infine la definizione formale di grafo causale. Per un approfondimento più dettagliato, si faccia riferimento al lavoro di Lauritzen ed Edwards (Edwards, 2000; Lauritzen, 2000; Steffen Lauritzen, 1996).

Un grafo $\mathcal{G}$ = (V,E) è un oggetto matematico rappresentato dalla tupla di due insiemi: un insieme finito di vertici V (*vertices*), anche chiamati nodi, e un insieme finito di coppie ordinate di nodi E $\subseteq$ V x V (*edges*), chiamate archi, con $E = \{(x, y): x \neq y, (x, y) \in V \times V\}$. Gli archi si dicono *direzionati* se esattamente e solamente uno tra gli archi $\{(x, y), (y, x)\}$ è incluso in E; gli archi si dicono *non direzionati* o *adirezionati* se entrambi $\{(x, y), (y, x)\}$ sono inclusi nell'insieme degli archi E per $\mathcal{G}$. Un grafo $\mathcal{G}$ si dice *diretto* (DG) se tutti i suoi archi sono direzionati; al contrario, $\mathcal{G}$ si dice *indiretto* (UG) se tutti gli archi sono adirezionati. Un arco diretto (x, y) è rappresentato graficamente tramite una freccia $x \rightarrow y$, e induce un insieme di relazioni tra i vertici del grafo $\mathcal{G}$. Qualora esista un arco direzionato da x verso y, x viene identificato come il *genitore* per y, y come il *figlio* di x. L'insieme di vertici genitori è denotato con $pa(x)$ (dall'inglese, *parents*), mentre l'insieme dei vertici figli è denotato da $ch(y)$

(dall'inglese, *childrens*). Se esiste un arco tra i vertici x e y, questi sono detti *adiacenti*. o *vicini,* a seconda che siano connessi tra loro attraverso un arco direzionato o adirezionato. L'insieme dei vicini per un vertice (x) è denotato da $ne(x)$ (da *neighbors*), mentre l'insieme degli adiacenti è definito da $adj(x)$. La *frontiera (boundary)* $bd(x)$ di un vertice x è l'insieme dei vertici genitori e dei vicini per x; la frontiera per A ⊂ V è l'insieme dei vertici $V\backslash A$ che sono genitori o vicini ai vertici in A. La *chiusura* (*closure*) per A è definita da $cl(A) = A \cup bd(A)$. Un grafo $\mathcal{G}^{\sim}$ è detto la *versione indiretta* di $\mathcal{G}$ se si ottiene sostituendo tutti gli archi direzionati di $\mathcal{G}$ con archi non direzionati. Un grafo diretto (DG) si dice *moralizzato* in un grafo indiretto $\mathcal{G}_M$ quando gli archi direzionati presenti vengono resi adirezionati, e viene aggiunto un arco adirezionato in corrispondenza di ogni coppia di nodi genitori per uno stesso nodo figlio.

$\mathcal{G}_A$ = (A, E_A) è detto *subgrafo* di $\mathcal{G}$ = (V,E) se $A \subseteq V$ e $E_A \subseteq E \cap A \times A$. Se $E_A = E \cap A \times A$, allora $\mathcal{G}_A$ è detto *subgrafo* di $\mathcal{G}$ *indotto* dall'insieme di vertici A. Un grafo è detto *completo* se tutte le coppie di vertici sono unite da un arco. Un sottoinsieme di vertici per $\mathcal{G}$ è completo se induce subgrafo completo. Un *subgrafo* completo che è massimo (rispetto a ⊆) è detto *cricca* (*clique*, in inglese). Per esempio, nel grafo mostrato in figura 3.4 ci sono due cricche:$\{x, y, w\}$ e $\{x, z\}$.

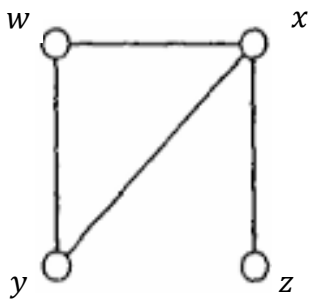


Figura 3.4 – Esempio di grafo con 4 nodi, 4 archi e due cricche.

In un grafo indiretto $\mathcal{G}$ si dice che un insieme di vertici S *separa* gli insiemi A e B se ogni percorso che porta da un nodo contenuto in A ad un nodo in B contiene almeno un nodo appartenente a S. Per esempio, nella figura 3.4, dato A = $\{y, w\}$, B = $\{z\}$ e C = $\{x\}$, C separa A da B.

Un *percorso* (*path*) $\pi = (x - \cdots - y)$ di lunghezza n da x a y è definito da una sequenza di vertici distinti e non ripetibili $x = x_0 \ldots, x_n = y$ tali per cui $(x_{i-1}, x_i) \in$ E, per ogni i $= 1,\ldots,$n. Un percorso *non direzionato* presenta tutti i vertici non direzionati. Al contrario, un percorso *direzionato* (*directed path*) $\pi = (x \to \cdots \to y)$ è definito come una tupla non ripetibile di vertici, in cui ciascun vertice è connesso al successivo nella sequenza dell'arco direzionato. Un percorso è detto *parzialmente direzionato (PD, Partially Directed)* se presenta sia archi direzionati che adirezionati. Un *ciclo* di lunghezza n è un percorso che inizia e finisce nello stesso vertice, ovvero $x = y$. Il ciclo è detto *diretto* se contiene almeno un arco direzionato. Un grafo diretto che non contiene cicli è chiamato *grafo diretto aciclico* (DAG, *Directed Acyclic Graph*). Dato un DAG, i vertici dell'insieme E tali per cui $x \to y$ ma non è verificato $y \to x$ si dicono *antenati di y an(y);* i *discendenti di x de(x)* invece sono tutti quei vertici y tali per cui $x \to y$ ma non viceversa. I *non-discendenti* di x *nd(x)* sono l'insieme di vertici $(V \backslash (de(x) \cup x)$.
Dato un grafo diretto $\mathcal{G}$, un percorso π per $\mathcal{G}$, e tre vertici $\{x, y, z\} \in V$ in π, si definiscono le seguenti relazioni:

- $x \leftarrow y \to z$ è detta *forchetta* (*fork*, in inglese) in π;
- $x \to y \to z$ è detta *catena* (*chain*) in π;
- $x \to y \leftarrow z$ è detta *collisore* (*collider*) in π.

Sia $A \subseteq V$. Il percorso π è *bloccato* da A se e solo se questo contiene:

- una forchetta $x \leftarrow y \to z$ o una catena $x \to y \to z$ tale che il vertice centrale y sia in A, oppure
- un collisore $x \to y \leftarrow z$ tale che il vertice centrale y o qualsiasi suo nodo discendente non sia in A.

L'insieme A *d-separa* due vertici x, z se e solo se blocca ogni percorso tra questi. Nella Figura 3.5 si riporta un esempio di d-separazione ripreso dal lavoro di (Zanga et al., 2022): x e z risultano d-separati poiché formano un collisore con y. x e u invece risultano d-separati attraverso il condizionamento con il vertice y che qui svolge il ruolo di vertice centrale per la catena $x \to y \to u$.

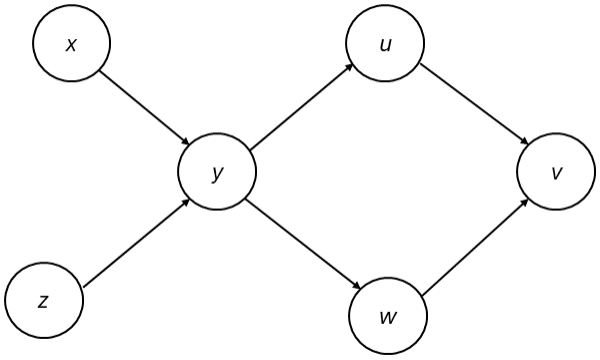


Figura 3.5 – Esempio di grafo diretto $\mathcal{G}$ con sei vertici.

3.2.1.1 INDIPENDENZA CONDIZIONATA

Il concetto di indipendenza condizionata rappresenta il punto di partenza fondamentale per lo sviluppo della teoria probabilistica e statistica associata ai modelli grafici (GM, *Graphical Models*). Dalla fine degli anni '70 infatti, l'indipendenza condizionata tra un insieme di variabili ha cominciato ad essere indagata tramite l'utilizzo di metodi grafici, identificando in ogni vertice una variabile e utilizzando gli archi per rappresentare le dipendenze condizionali (Pearl & Paz, 2022). Questa applicazione dei metodi grafici ha dato vita nel tempo ai cosiddetti *modelli grafici*.

Due eventi A, B sono detti *indipendenti* se $P(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$, o, equivalentemente $\mathrm{P}(A|B) = \mathrm{P}(A)$. Analogamente, due variabili casuali X, Y sono dette *indipendenti* se la loro distribuzione congiunta è pari a $p_{X,Y}(x, y) = p_X(x)p_Y(y)$, o, equivalentemente $p_{Y|X}(y|x) = p_Y(y)$, con p generica funzione di probabilità o densità.

Si considerino tre variabili casuali X, Y, Z. Se, per ciascun valore z, X e Y sono indipendenti nella distribuzione condizionale dato $Z = z$, allora X e Y si dicono *condizionatamente indipendenti* dato Z, e si scrive: $X \perp Y|\, Z$. A livello matematico, si scrive:

$$P(X \in A, Y \in B|Z) = P(X \in A|Z)P(Y \in B|Z)$$

con A e B misurabili nello spazio campionario per X e Y rispettivamente. Una caratterizzazione alternativa per la proprietà di indipendenza condizionata è proposta da A.P. Dawid (Dawid, 1979):

$$P(X \in A|Y, Z) = P(X \in A|Z).$$

Quest'ultima formulazione mostra chiaramente come al conoscere i valori osservati per Z, la distribuzione per X non dipenda più da Y, che diventa dunque irrilevante per l'interpretazione di X. Qualora X, Y, Z siano variabili casuali discrete, la relazione può essere espressa come:

$$P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z), \quad P(Z > 0)$$

Se X, Y, Z sono variabili casuali continue, la definizione di indipendenza condizionata si trasforma in:

$$X \perp Y|\, Z \;\Leftrightarrow f_{XY|Z}(x, y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z)$$

La relazione ternaria $X \perp Y|Z$ gode delle seguenti proprietà, dove h denota una funzione arbitraria misurabile nello spazio campionario X:

- (C1) se $X \perp Y|\, Z$, allora $Y \perp X|\, Z$;
- (C2) se $X \perp Y|\, Z$ e $U = h(X)$, allora $U \perp Y|\, Z$;
- (C3) se $X \perp Y|\, Z$ e $U = h(X)$, allora $X \perp Y|\, (Z, U)$;
- (C4) se $X \perp Y|\, Z$ e $X \perp W|\, (Y, Z)$, allora $X \perp (W, Y)|\, Z$.
  Con (C4) derivante dalle proprietà (C2) e (C3).

Riprendendo l'interpretazione proposta da (Lauritzen, 2000), le proprietà (C1)-(C5) possono essere tradotte attraverso il seguente esempio pratico, dove il termine "libro" e "capitolo" identificano rispettivamente una variabile casuale e una qualsiasi funzione misurabile della variabile casuale stessa:

- (I1) se, conoscendo Z, leggere Y è irrilevante per la lettura di X, allora ugualmente leggere X sarà inutile per la lettura di Y;

- (I2) se, conoscendo Z, leggere Y è irrilevante per la lettura del libro X, allora leggere Y sarà irrilevante per la lettura di un qualsiasi capitolo U di X;
- (I3) se, conoscendo Z, leggere Y è irrilevante per la lettura del libro X, allora questo rimarrà irrilevante anche dopo aver letto un qualsiasi capitolo U di X;
- (I4) se, conoscendo Z, leggere il libro Y è irrilevante per leggere X e anche dopo aver letto Y, W è irrilevante per leggere X, allora leggere sia Y che W è irrilevante per leggere X.

Un'ultima proprietà per la relazione di indipendenza condizionale che è spesso sfruttata nel campo applicativo implica che:

- (C5) se $X \perp Y | Z$ e $X \perp Z | Y$, allora $X \perp (Y, Z)$, a patto che non sussista alcuna relazione logica non banale tra Y e Z.

3.2.1.2 PROPRIETÀ MARKOVIANE PER GRAFI NON DIREZIONATI

Le proprietà di indipendenza condizionale viste sopra possono essere descritte in modo compatto anche tramite le proprietà di Markov per metodi grafici. Le proprietà di Markov sono particolarmente importanti in quanto costituiscono il punto di partenza per lo studio degli effetti causali nei modelli grafici: è noto, infatti, che gli effetti causali sono identificabili ogni qualvolta il modello è *Markoviano (*ovvero aciclico*)* e tutti i termini di errore sono congiuntamente indipendenti. Modelli non Markoviani, come quelli che coinvolgono per esempio errori correlati, consentono l'identificazione solo in condizioni specifiche (S. Lauritzen, 1996).

In questa sezione, si fa riferimento a grafi non direzionati, in cui dunque tutti gli archi sono adirezionati. Sia $\mathcal{G}$ = (V, E) un grafo non direzionato e sia **X** un vettore p-dimensionale di variabili casuali $(X_i)_{i \in V}$ a cui corrisponde un insieme di vertici $V = \{1, 2, \dots, p\}$. La distribuzione di probabilità P per **X** soddisfa

1. la *proprietà di Markov pairwise* per $\mathcal{G}$, se per ogni coppia di vertici non adiacenti (i, j)

$$X_i \perp X_j \mid \boldsymbol{X}_{V \setminus \{i,j\}}$$

2. la *proprietà di Markov locale* per $\mathcal{G}$, se per ogni vertice $i \in V$

$$X_i \perp \boldsymbol{X}_{V\setminus\{ne(i)\cup(i)\}} | \boldsymbol{X}_{ne(i)}$$

3. la *proprietà di Markov globale* per $\mathcal{G}$, se per ogni tupla (A, B, C) di sottoinsiemi di V tali che C separa A da B in $\mathcal{G}$,

$$\boldsymbol{X}_A \perp \boldsymbol{X}_B | \boldsymbol{X}_C$$

È dimostrato che la proprietà di Markov globale implica la proprietà di Markov locale, che a sua volta implica quella *pairwise* (Steffen Lauritzen, 1996). Utilizzando l'esempio proposto nel lavoro di (Banzato, 2023), si mostrano per chiarezza espositiva le proprietà di Markov applicate ad un grafo $\mathcal{G}$. Il primo grafo della Figura 3.6 rappresenta la proprietà *pairwise*, dove $X_1 \perp X_5 | \{X_2, X_3, X_4\}$. Il grafo centrale identifica la proprietà di Markov locale, con $X_3 \perp X_5 | \{X_2, X_4\}$ e $X_1 \perp \{X_3, X_5\} | X_4$. L'ultimo grafo, a destra, mostra la proprietà di Markov globale, tale per cui $X_1 \perp \{X_3, X_4\} | \{X_2, X_5\}$.

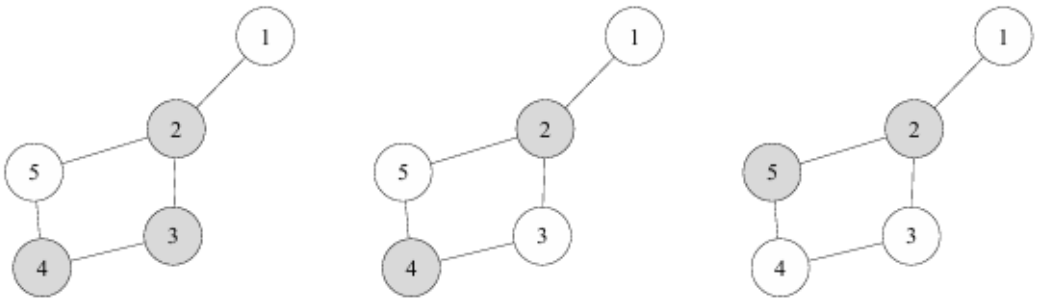


Figura 3.6 – Esempio di rappresentazione grafica delle proprietà di Markov. Da sinistra a destra: proprietà *pairwise, locale, globale*.

Si puntualizza infine che come l'indipendenza condizionale in probabilità è strettamente legata al concetto di fattorizzazione di densità congiunte, allo stesso modo lo sono anche le proprietà Markoviane presentate. Nello specifico, la

distribuzione di densità congiunta P tra più variabili casuali $(X_i)_{i\in V}$ può essere espressa come il prodotto di funzioni a cricca (chiuse).

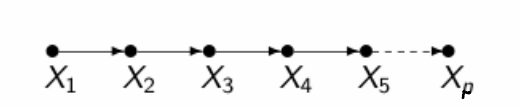
3.2.1.3 PROPRIETÀ MARKOVIANE NEI DIRECTED ACYCLIC GRAPHS (DAGs)

Sia $\mathcal{D}$ = (V, E) un DAG e sia **X** un vettore p-dimensionale di variabili casuali $(X_i)_{i\in V}$ a cui corrisponde un insieme di vertici $V = \{1, 2, \dots, p\}$. La distribuzione di probabilità P per **X** soddisfa:

1. La *proprietà di Markov locale (L)* per $\mathcal{D}$, se per ogni vertice $i \in V$

$$X_i \perp \{nd(X_i)\backslash pa(X_i)\}|pa(X_i)$$

con $pa(X_i)$ insieme dei genitori per il vertice *i-mo* e $nd(X_i)$ insieme dei non discendenti. Un esempio tipico di questa proprietà è descritto dalla seguente catena Markoviana:



con $X_{i+1} \perp | (X_1, \dots, X_{\mathrm{i}-1})$ per $i = 3, \dots, p$.

2. *La proprietà di fattorizzazione (F)* per $\mathcal{D}$ se la sua funzione di probabilità o densità congiunta p tra più variabili casuali $(X_i)_{i\in V}$ ha la forma:

$$p(x) = \prod_{v\in V} p\,(x_{\mathrm{v}}|x_{pa(v)})$$

La fattorizzazione in probabilità (F) si basa sull'assunto che le relazioni codificate dal grafo corrispondano esattamente alle relazioni di indipendenza condizionata in probabilità:

$$\mathrm{X} \perp_P \mathrm{Y} \mid \mathbf{Z} \implies \mathrm{X} \perp_G \mathrm{Y} \mid \mathbf{Z}$$

con $\perp_P$ e $\perp_G$ che indicano rispettivamente indipendenza in probabilità ed indipendenza grafica. In sostanza, si assume che l'indipendenza in probabilità (nei dati) implichi l'indipendenza grafica (indipendenza nel meccanismo causale). Questo assunto è noto come *directed-faithfulness* o *d-faithfulness (*tradotto: fedeltà

direzionata) e rappresenta l'assunzione causale portante per la metodologia della *causal discovery*.

3. La *proprietà di Markov globale (G)* per $\mathcal{D}$, se dati i sottoinsiemi di vertici (A, B) d-separati da un vertice S ∈ $(X_i)_{i\in V}$, si ottiene

$$A \perp_d B \mid S \rightarrow A \perp B \mid S.$$

Per qualsiasi DAG $\mathcal{D}$ e per qualunque distribuzione di probabilità P, le tre proprietà di Markov direzionate sono in relazione tra loro come

$$(G) \iff (L) \iff (F).$$

3.2.2 MODELLI GRAFICI CAUSALI E MODELLI AD EQUAZIONI STRUTTURALI (SEM)

Le definizioni riportate in questa sezione fanno riferimento alla formulazione di modello (grafico) causale proposta da Pearl e Zanga (Pearl, 2009; Zanga et al., 2022).

Sia $\mathcal{G}$ = (V, E) un grafo e sia **X** un vettore p-dimensionale di variabili casuali $(X_i)_{i\in V}$ a cui corrisponde un insieme di vertici $V = \{1, 2, \dots, p\}$. Un *grafo $\mathcal{G}$ è detto causale* se rappresenta la descrizione grafica di un sistema in termini di relazioni di causa-effetto, ovvero del *meccanismo causale*. Per ogni arco diretto $(X, Y) \in E$, X è detto *causa diretta* di Y, e Y è l'*effetto diretto* di X. Il valore assegnato a ciascuna variabile X è completamente determinato dalla funzione f dati i suoi genitori $pa(X)$:

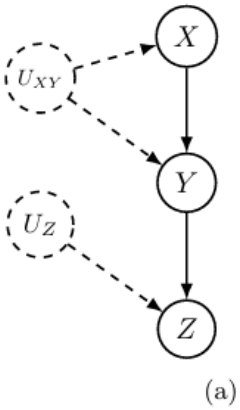
$$X_i := f(pa(X_i)) \qquad \forall x_i \in V$$

Un *modello strutturale causale* (SCM, *Structural Causal Model*), è definito dalla tupla M = (**V**, **U**, **F**, P), con:

- **V** insieme di variabili endogene, ovvero le variabili osservate;
- **U** insieme di variabili esogene –variabili non osservate- con $V \cap U = \emptyset$;
- **F** insieme di funzioni f, dove ciascuna funzione $f_i \in F$ è definita come $f_i: (V \cap U)^p \rightarrow V$, con p varietà f_i, così che f_i identifica completamente V_i;

- P distribuzione di probabilità congiunta per la variabile esogena $P(U) = \prod_i P(U_i)$.

I modelli causali strutturali posso essere pensati anche come modelli ad equazioni strutturali (SEM) non parametrici. Nella Figura 3.7 si riporta l'esempio di un grafo causale associato al corrispettivo modello causale strutturale, proposto nel lavoro di Zanga.



(a)

$$M = (\mathbf{V}, \mathbf{U}, \mathbf{F}, P)$$
$$\mathbf{V} = \{X, Y, Z\}$$
$$\mathbf{U} = \{U_{XY}, U_Z\}$$
$$\mathbf{F} = \begin{cases} f_X : X := 2U_{XY}, \\ f_Y : Y := X + U_{XY}, \\ f_Z : Z := 3Y + U_Z \end{cases}$$
$$P = \begin{cases} U_{XY} \sim \mathcal{N}(0,1), \\ U_Z \sim \mathcal{N}(0,1) \end{cases}$$

(b)

Figura 3.7 – Esempio di grafo causale $\mathcal{G}$ (a) e del corrispondente modello SCM (b).

3.2.3 CAUSAL DISCOVERY

In questa sezione si espone la metodologia scelta per lo sviluppo della *causal discovery* nel contesto di studio osservazionale. Vengono fornite le nozioni chiave necessarie alla comprensione dell'argomento, per poi passare alla presentazione dell'algoritmo *Constraint-Based* di Peter-Clark (PC) (Peter Spirtes et al., 1993) e delle motivazioni a supporto di tale scelta operativa.. I principali riferimenti bibliografici a supporto di notazione e terminologia sono Spirtes & Zhang (2016) e Zanga et al. (2022).

La *causal discovery* è un'area di ricerca che mira a identificare le relazioni causali tra variabili a partire dai dati, servendosi dell'utilizzo dei modelli grafici e tecniche di *machine learning*.

L'individuazione delle relazioni causali a partire esclusivamente da misure di associazione statistica (correlazione, indipendenza condizionata, verosimiglianza, etc.) rispetto che da strutture causali supposte a priori rappresenta una sfida significativa e di forte interesse contemporaneo. In particolare, questo campo ha guadagnato un'attenzione crescente a partire dall'inizio del XXI secolo, con lo sviluppo di importanti contributi metodologici tutt'ora in divenire.

Il concetto di causalità possiede radici antiche e vi è una vastissima letteratura al riguardo. La formalizzazione teorico-matematica e algoritmica del problema della *causal discovery* invece risale solamente agli ultimi trent'anni. Un punto di riferimento fondamentale in questo campo è rappresentato dal lavoro di J. Pearl (2000), il quale ha introdotto i grafi causali come strumenti utili all'analisi e alla rappresentazione delle relazioni causali tra variabili casuali (J. Pearl, 2000). Parallelamente, il lavoro di P. Spirtes, C. Glymour e R. Scheines (1993) ha contribuito allo sviluppo di metodi algoritmici per la *causal discovery* basati sui test di indipendenza condizionale (Peter Spirtes et al., 1993): tra i più noti e utilizzati tutt'oggi, si citano l'algoritmo PC e il FCI (Fast Causal Inference).

Il presupposto teorico-causale su cui si basa la metodologia della *causal discovery* è quello della *faithfulness (*in italiano: *fedeltà)*, in cui si assume che le dipendenze condizionali osservate nei dati (dipendenze in probabilità) siano esattamente le stesse indotte dalla struttura causale sottostante (dipendenza grafica tramite il criterio di d-separazione). In altre parole, si assume che le informazioni osservate sui soli dati siano rappresentative della vera struttura causale che intercorre tra i fenomeni misurati. Senza questa assunzione, si disporrebbe solamente di distribuzioni di probabilità congiunte per le variabili osservate (associazioni) e non ci sarebbero le premesse per stabilire relazioni di causazione.

Nel campo epidemiologico, la *causal discovery* può essere utilizzata per identificare relazioni causali tra fattori di rischio, patologie, ed esiti di salute, contribuendo quindi alla possibilità di pianificare interventi mirati per la prevenzione e la cura di diverse patologie, attraverso evidenze di tipo causale piuttosto che su semplici correlazioni.

Formalmente, sia $\boldsymbol{\mathcal{G}}$ l'insieme di grafi definito per le variabili $\boldsymbol{V}$ di un dataset $\boldsymbol{D}$, e sia $\mathcal{G}^* \in \boldsymbol{\mathcal{G}}$ il vero e ignoto grafo da cui $\boldsymbol{D}$ è stato generato**.** Il problema della *causal discovery* consiste nel risalire al vero grafo $\mathcal{G}^*$ a partire dall'insieme di dati disponibile $\boldsymbol{D}$. La rappresentazione e formulazione di un DAG non sempre permette di incorporare la mancanza di informazione tipicamente intrinseca in una procedura di *discovery,* specialmente con dati osservazionali *real world*. Per questo motivo, sono contemplate anche soluzioni più flessibili, che prevedono l'utilizzo dei *Partially-Directed Acyclic Graphs* (PDAG) -ovvero grafi che utilizzano archi direzionati e adirezionati- e *i Complete Partially Directed Acyclic Graphs* (CPDAG); questi ultimi si basano sulla ricerca di un grafo orientato appartenente alla classe di equivalenza markoviana osservazionale per $\mathcal{G}^*$, $MEC[\mathcal{G}^*]$[25].

Allo stesso modo, esistono delle applicazioni in cui è possibile lavorare sotto l'assunzione che l'insieme di variabili a disposizione $\boldsymbol{V}$ non sia *causalmente sufficiente*[26]*,* ipotizzando quindi la presenza di un insieme (non vuoto) di variabili non osservate (latenti) $\boldsymbol{U}$ che contenga almeno un meccanismo causale per la generazione del dataset $\boldsymbol{D}$**.** In questo caso, $\mathcal{G}$ sarà un subgrafo del *grafo aumentato* $\mathcal{G}^a$ definito in $\boldsymbol{V} \cup \boldsymbol{U}$**,** come si vede per esempio in Figura 3.6(a) (Bongers et al., 2016).

3.2.3.1 APPLICAZIONE DELL'ALGORITMO DI PETER-CLARK (PC) NEL CONTESTO OSSERVAZIONALE

Gli algoritmi per la *causal discovery* in *setting* osservazionali[27] si dividono principalmente in due categorie: algoritmi basati su vincoli (*constraint-based algorithms*) e algoritmi basati su una funzione punteggio (*score-based algorithms*). I *constraint-based algorithms* impiegano una successione di test statistici di indipendenza condizionata (in probabilità) con l'obiettivo di riportare i risultati ottenuti in un grafo causale che assuma *faithfulness* con la distribuzione

[25] Due PDAGs $\mathcal{G}$ e $\mathcal{H}$ si definiscono equivalenti a livello *Markoviano* osservazionale se hanno lo stesso scheletro (*skeleton*) e le stesse strutture a v; dove con struttura a *v* si denota una tripla $X \rightarrow Y \leftarrow Z$, X e Z non adiacenti, e con *scheletro* si intende la struttura non direzionata del PDAG. Questa proprietà si denota come $\mathcal{G} \equiv \mathcal{H}$.

[26] Un insieme di variabili **V** si dice *causalmente sufficiente* se e solo se ogni causa di ogni sottoinsieme di V è contenuta in V stesso.

[27] non si prende in considerazione né una componente temporale nei dati, né si interviene attivamente nella manipolazione delle variabili.

probabilistica sottostante e mostri le separazioni grafiche corrispondenti. Gli *score-based algorithms* invece si basano sulla massimizzazione di una misura di adattamento per un generico grafo $\mathcal{G}_i \epsilon \boldsymbol{\mathcal{G}}$ ai dati osservati $\boldsymbol{D}$, assegnando un punteggio di rilevanza (*scoring criterion*) a ciascun grafo candidato $S(\mathcal{G}, \boldsymbol{D})$ della forma

$$\mathcal{G}^* = \arg\max S(\mathcal{G}, \boldsymbol{D}), \qquad \mathcal{G} \epsilon \boldsymbol{\mathcal{G}}$$

In questa sezione, si approfondisce la formulazione dell'algoritmo di apprendimento non supervisionato *constraint-based* di Peter-Clark (Peter Spirtes et al., 1993) specificando le ragioni che ne hanno giustificato l'utilizzo nel contesto in esame.
La definizione di test di indipendenza condizionata per $X \perp Y | \boldsymbol{Z}$, identificato con $I(X, Y | \boldsymbol{Z})$, prevede che l'ipotesi nulla H_0 e alternativa H_1 siano definite rispettivamente come H_0: $X \perp Y | \boldsymbol{Z}$ e H_1: $\overline{H_0}$. Ne consegue che

$$\hat{I}(X, Y | \boldsymbol{Z}) > \alpha \implies \boldsymbol{X} \perp \boldsymbol{Y} | \boldsymbol{Z}$$

con α livello di significatività statistica. Si ribadisce che in questo contesto si assume che le indipendenze condizionali osservate nei dati siano rappresentate coerentemente dal grafo casuale stimato (*faithfulness*). Come la maggior parte dei metodi *constraint-based*, l'algoritmo PC consiste in due macro-fasi: la ricerca delle dipendenze (chiamata anche *skeleton phase*) e l'orientamento delle dipendenze trovate. La procedura iterativa è eseguita sotto l'assunzione di sufficienza causale per l'insieme di variabili osservate $\boldsymbol{V}$.
Volendo esplicitare la domanda di ricerca causale in questo contesto osservazionale, l'algoritmo PC permette di rispondere a: *"In che modo conoscere X mi permette di modificare le mie credenze per Y*?"[28].
Nella prima fase della procedura viene identificato un grafo non orientato completamente connesso (CUG, *Complete Undirected Graph*) $\mathcal{C}$, detto anche *skeleton*: per ogni coppia di variabili adiacenti X e Y, si testa l'indipendenza condizionale $X \perp$

[28] Qualsiasi domanda più specifica è da porsi solo in contesti di tipo interventistico (dati sperimentali) o controfattuale.

$Y|\boldsymbol{Z}$; la successione di test inizia con $\boldsymbol{Z} = \emptyset$ e procede iterativamente su insiemi di $\boldsymbol{Z}$ di dimensione sempre crescente. Se viene verificata l'indipendenza condizionale, l'arco adirezionato tra X e Y viene rimosso. Nella pratica ordinaria, si assume che le variabili seguano una distribuzione congiunta gaussiana multivariata; questa assunzione permette di semplificare il calcolo dei test poiché consente l'utilizzo di test di correlazione parziale per l'identificazione delle relazioni di indipendenza (per ulteriori approfondimenti a riguardo, si veda l'Appendice B). Rispetto a questa prima fase, è noto che la potenza dei test di indipendenza condizionata diminuisce all'aumentare della dimensione dell'insieme di condizionamento $\boldsymbol{Z}$ ($dim(\boldsymbol{Z})$), a causa della maledizione della dimensionalità. Per ovviare al problema, una soluzione comune è quella di fissare un limite superiore per la $dim(\boldsymbol{Z})$, riducendo così l'onere computazionale dell'algoritmo ed evitando test a bassa significatività (Li & Fan, 2020) Nel caso in esame, avendo a disposizione un esiguo numero di variabili, non verrà fissato alcun limite superiore per la dimensione dell'insieme di condizionamento $\boldsymbol{Z}$.

Nella fase di *orientamento* delle dipendenze vengono utilizzati i risultati ottenuti dai test $\hat{I}$ per ottenere degli archi direzionati attraverso una serie di regole che vertono sull'identificazione di strutture a *v* e sulle proprietà di aciclicità del grafo, a partire dal grafo CUG $\mathcal{C}$ trovato per l'insieme di variabili osservate $\boldsymbol{V}$. Più nello specifico, la procedura iterativa proposta da P. Spirtes et al. (1993) e che verrà adottata in questa analisi prevede i seguenti passaggi, qui riassunti brevemente:

1. sia $Sepset(X, Y)$ un sottoinsieme di variabili (nodi) $\in \boldsymbol{V}$ che, se rimosso dalla rete, permette di *d-separare* X e Y[29]. Si considerano le variabili X, Y, W collegate in modo non direzionato $X - Y - W$, tali per cui le coppie X, Y e Y, W sono adiacenti in $\mathcal{C}$, ma X e W non sono adiacenti in $\mathcal{C}$. Se e solo se $Y \notin Sepset(X, Z)$, allora gli archi vengono direzionati in modo da formare una struttura a *v* del tipo $X \rightarrow Y \leftarrow W$, con (X, Y, W) struttura a *v*;

[29] Per il concetto di *d-separazione*, si rimanda a pg.54.

2. se esiste una relazione diretta $X \rightarrow Y$, con Y e W adiacenti ($Y - W$), ma non $X - W$, e non vi è un arco direzionato del tipo $W \rightarrow Y$, allora si istituisce un arco direzionato per $Y - W$ del tipo $Y \rightarrow W$;
3. se esiste un percorso *diretto* tra X e Y, e un arco non direzionato tra X e Y, allora viene costituito un arco direzionato del tipo $X \rightarrow Y$.

Il risultato finale è un CPDAG per la classe di equivalenza Markoviana $MEC[\mathcal{G}^*]$, da cui consegue dunque che gli archi rimasti non direzionati non comporteranno perdita di informazione rispetto alla distribuzione probabilistica generata dalla sola osservazione dei dati. Nel grafo finale stimato, gli archi non direzionati (associazioni) sono identificati con il simbolo $\leftrightarrow$.

Per ovviare al problema della dipendenza dei risultati dall'ordine in cui vengono analizzate le variabili nella *skeleton phase* (*order dependence*), sono state proposte nel tempo alcune varianti dell'algoritmo PC: si vedano per esempio l'algoritmo *PC-stable* e il *conservative PC*, proposti da Colombo & Maathius (Colombo & Maathuis, 2014), Ramsey et al. (2012). L'algoritmo *PC-stable* si focalizza esclusivamente sul garantire che l'ordine di eliminazione degli archi nella *skeleton phase* non influenzi il risultato finale, garantendo di conseguenza più stabilità e robustezza dei risultati anche nella fase di orientamento degli archi. L'algoritmo PC conservativo invece si concentra *in primis* sulla riduzione dell'errore di II tipo per la fase di identificazione delle dipendenze (mancata identificazione di una dipendenza), e si dimostra anche prudente nell'orientamento degli archi attraverso l'incrocio di conferme multiple. Per il problema della maledizione della dimensionalità, qui non affrontato vista l'esigua dimensione di $\boldsymbol{V}$, si rimanda al lavoro di Tsamardinos et al. (2006).

Esistono poi algoritmi *constraint-based* che non partono dall'assunto di sufficienza causale per il grafo stimato, ipotizzando al contrario sistemi che prevedano l'utilizzo di variabili latenti (non osservate). L'esempio più rappresentativo di questo tipo di algoritmi è il *Full Causal Inference* (FCI) proposto da Spirtes et al. (1995).

In questa analisi si è deciso di testare il processo di *causal discovery* partendo dallo stesso insieme di variabili osservate $\boldsymbol{V}$ scelte per la modellazione tramite SEM (si veda la *sezione 3.1.1*), così da poter effettuare un confronto diretto tra il grafo causale stimato con le diverse misurazioni dell'ipertensione e le specificazioni SEM

ipotizzate a priori, con particolare attenzione rispetto alla specificazione che gode di maggior bontà di adattamento ai dati. Lavorando quindi con un numero esiguo di variabili (*low-dimensional setting*), e non prevedendo la presenza di fattori latenti (sufficienza causale per $\boldsymbol{V}$), si è deciso di applicare la versione conservativa dell'algoritmo PC, in quanto possiede i presupposti teorici per essere ragionevolmente adeguato rispetto all'obiettivo in esame, sia gestendo il problema di *order-dependence* (Colombo & Maathuis, 2014), e sia adottando un approccio prudente all'orientamento degli archi. . La distribuzione congiunta per le variabili è assunta normale multivariata, ugualmente a quanto fatto per i SEM. Per un approfondimento sui modelli grafici gaussiani si rimanda all'Appendice B. Si precisa infine che verranno considerate solo le unità che non presentino alcun valore mancante in $\boldsymbol{V}$, per una numerosità totale pari a N = 22184 e N = 21894 qualora si utilizzino rispettivamente la variabile clinica "iperteso" o la pressione arteriosa diastolica/sistolica (PAD, PAS). Per la stima dei *Complete Partially Directed Acyclic Graphs* (CPDAG) con distribuzione normale multivariata è stata utilizzata la funzione `pc()` del pacchetto `pcalg` (Kalisch et al., 2012) implementato nel software R (R Core Team, 2023), versione 4.3.2.

4. I RISULTATI

In questo capitolo si presentano i risultati ottenuti dall'applicazione delle due aree metodologiche adottate: i modelli ad equazioni strutturali (SEM: *Structural Equation Models*) e i modelli grafici per la *causal discovery*. Si ricorda che in questa tesi i modelli SEM sono impiegati con lo scopo di discutere l'utilizzo di diverse misurazioni dell'ipertensione arteriosa e l'appropriatezza di diversi *path causali* per la modellazione dell'insorgenza di eventi cardiovascolari. Nel processo di *causal discovery* invece, si mira a stimare un *path* causale robusto per le relazioni tra le variabili osservate, con lo scopo di aggiungere ulteriore validità alle relazioni ipotizzate a priori nei modelli SEM.

4.1 RISULTATI PER I MODELLI AD EQUAZIONI STRUTTURALI (SEM)

4.1.1 VALUTAZIONE DELLE STIME DEI COEFFICIENTI NEI MODELLI SEM

Di seguito si riporta la rappresentazione dei vari modelli SEM stimati tramite massima verosimiglianza, con il valore delle stime dei coefficienti di regressione standardizzati $\hat{\beta}^{st}$ e la relativa significatività statistica. Formalmente, dato un modello di regressione lineare semplice del tipo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_k X_k + \varepsilon, \qquad \varepsilon \sim N(0,1)$$

il generico coefficiente di regressione standardizzato ${\beta_j}^{st}$ è definito come

$${\beta_j}^{st} = \beta_j \cdot \frac{\sigma_{X_j}}{\sigma_Y}$$

dove σ_{X_j}indica la deviazione standard per X_j e σ_Y la deviazione standard per la variabile dipendente Y. La scelta di rappresentare i coefficienti standardizzati è dettata dalla volontà di fornire una panoramica in ottica comparativa tra gli effetti misurati, così da poter identificare le relazioni più importanti in termini di forza e valutare la consistenza dei risultati tra i vari modelli. I coefficienti standardizzati sono dunque utilizzati per confrontare in modo diretto i fattori clinici misurati, al netto della loro unità di misura originaria.

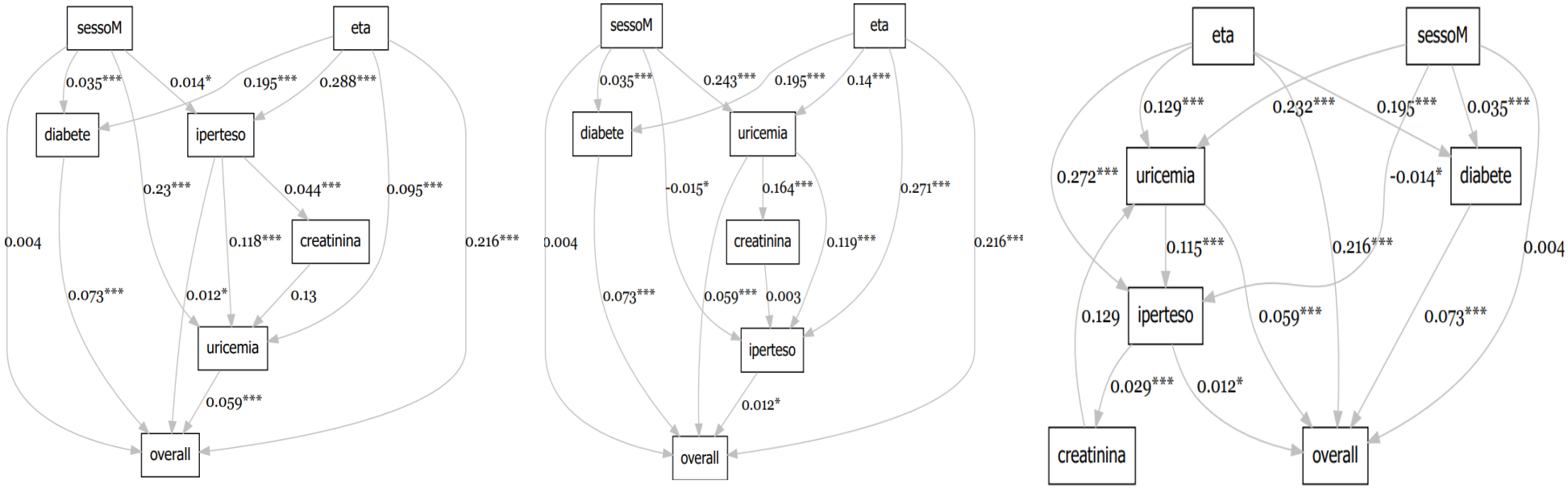


Figura 4.1 – Rappresentazione grafica dei modelli SEM stimati, con misurazione dell'ipertensione tramite variabile *iperteso*, specificazione (da sinistra) M1, M2, M3, coefficienti standardizzati; N = 22184. ***p≤ 0,001; **p ≤ 0,01; *p ≤ 0,05.

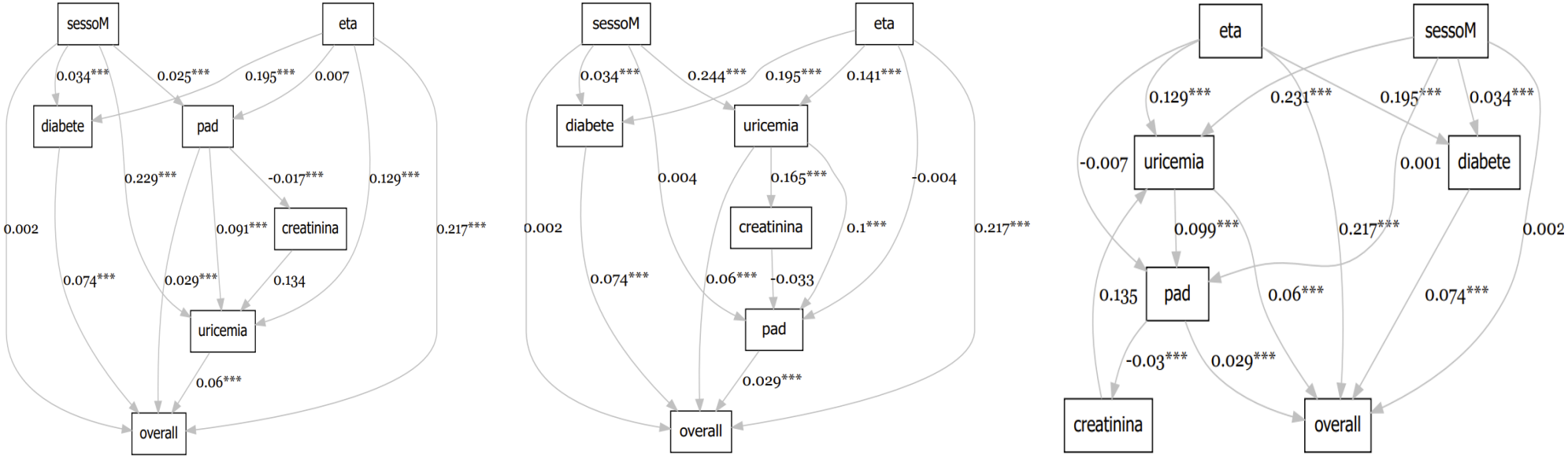


Figura 4.2 – Rappresentazione grafica dei modelli SEM stimati, con misurazione dell'ipertensione tramite variabile *PAD*, specificazione (da sinistra) M1, M2, M3, coefficienti standardizzati; N = 21894. ***p≤ 0,001; **p ≤ 0,01; *p ≤ 0,05.

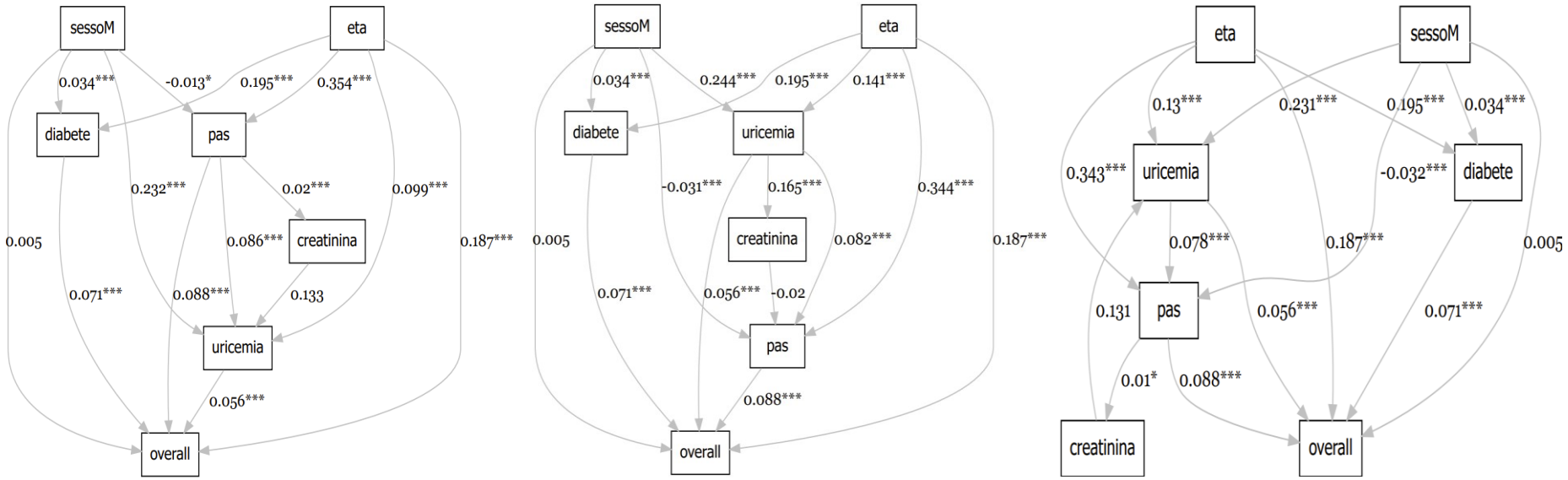


Figura 4.3 – Rappresentazione grafica dei modelli SEM stimati, con misurazione dell'ipertensione tramite variabile *PAS*, specificazione (da sinistra) M1, M2, M3, coefficienti standardizzati; N = 21894. ***p≤ 0,001; **p ≤ 0,01; *p ≤ 0,05.

4.1.1.1 RUOLO DEL SESSO NELLA RELAZIONE CON EVENTI CARDIOVASCOLARI, DIABETE, URICEMIA ED IPERTENSIONE ARTERIOSA

Partendo dall'analisi della variabile di controllo che identifica il sesso dei soggetti, si osserva un coefficiente per l'effetto diretto su *overall* $\widehat{\beta^{st}}_{sessoM \rightarrow overall}$ sempre[30] non significativo e tra i più piccoli complessivamente stimati, con un valore che varia tra 0,002 e 0,005. Per quanto riguarda la non significatività dell'effetto diretto tra il sesso e l'insorgenza di almeno un evento cardiovascolare (*overall)*, è noto come questo generalmente ne influenzi l'insorgenza; tuttavia, è altresì comprovato che il rischio di eventi cardiovascolari tende a convergere tra i due sessi con l'avanzare dell'età e/o in presenza di altre comorbidità, in particolare dopo la menopausa femminile (Maas & Appelman, 2010). Nell'insieme di dati in esame, l'età mediana per donne e uomini è pari rispettivamente a 58 e 55 anni, e sapendo che l'età media per l'entrata in menopausa femminile corrisponde a circa 50 anni, ecco dunque, che il risultato inerente la non significatività del coefficiente $\widehat{\beta^{st}}_{sessoM \rightarrow overall}$ può essere almeno intuitivamente giustificato.

Il coefficiente per la stima dell'effetto diretto del sesso sulla presenza di diabete mellito di tipo II, uricemia, e l'ipertensione arteriosa (in tutte le sue misurazioni) si dimostra invece sempre significativo ($p \leq 0{,}001$), con un conseguente effetto indiretto di fattore di rischio sullo sviluppo di almeno un evento cardiovascolare (*overall*).

Nella relazione con il diabete di tipo II, il sesso presenta un coefficiente $\widehat{\beta^{st}}_{sessoM \rightarrow diabete}$ pari a 0,035 nei modelli che utilizzando la variabile clinica *iperteso*, e 0,034 altrimenti, senza differenze tra le specificazioni M1, M2, M3.

Nella relazione con l'uricemia invece, i coefficienti $\widehat{\beta^{st}}_{sessoM \rightarrow ur}$ variano sia rispetto alla definizione di ipertensione arteriosa, sia rispetto ai *path* causali, con un valore che tra si estende da 0,23 (specificazione M1, *iperteso*) a 0,244 (M2, *PAD-PAS*); la maggior variabilità del coefficiente è associata alla specificazione causale rispetto che alla definizione di ipertensione arteriosa.

[30] in tutte le specificazioni causali, e per tutte le misurazioni di ipertensione.

Concentrandosi sull'effetto diretto del sesso (variabile esplicativa) sull'ipertensione arteriosa, si individua un effetto protettivo del sesso maschile sull'aumento della pressione arteriosa sistolica (PAS), e un effetto protettivo del sesso femminile qualora si consideri l'ipertensione tramite la misurazione della pressione arteriosa diastolica (PAD) o la variabile clinica *iperteso*[31]; in termini di forza dell'effetto diretto stimato $\widehat{\beta^{st}}_{sessoM \rightarrow ipert}$, questo è maggiore nelle specificazioni M2 e M3 rispetto al *path* causale M1 in cui si assume che la pressione arteriosa preceda uricemia e creatinina.

Rispetto alla significatività della relazione che vede il sesso come variabile esplicativa per diabete mellito II, uricemia ed ipertensione arteriosa, i risultati sono coerenti con quanto riscontrato in letteratura medica (Halperin Kuhns & Woodward, 2020; Maas & Appelman, 2010; Wang et al., 2019).

4.1.1.2 RUOLO DELL'ETÀ NELLA RELAZIONE CON EVENTI CARDIOVASCOLARI, IPERTENSIONE ARTERIOSA E URICEMIA

L'età si dimostra complessivamente il determinante con più impatto diretto sullo sviluppo di almeno un evento cardiovascolare (*overall*), con un coefficiente $\widehat{\beta^{st}}_{età \rightarrow overall}$ ($p \leq 0{,}001$) che varia da un minimo di 0,187 nei modelli che utilizzano la misurazione della pressione arteriosa sistolica (PAS), fino ad una forza di 0,217 nei modelli che misurano la pressione arteriosa diastolica (PAD); utilizzando la variabile *iperteso*, $\widehat{\beta^{st}}_{età \rightarrow overall_{ip}} = 0{,}216$ ($p \leq 0{,}001$).

Nella relazione diretta con l'ipertensione arteriosa, l'età risulta un fattore di rischio significativo sia rispetto all'aumento della pressione arteriosa sistolica (PAS), sia -seppur con minor forza- rispetto all'identificazione di soggetti ipertesi con definizione clinica (*iperteso*); in relazione alla pressione arteriosa diastolica (PAD) invece, l'età presenta coefficiente non significativo ($p \geq 0{,}05$) e registra un coefficiente di segno negativo in due specificazioni (M2, M3), ed un coefficiente $\widehat{\beta^{st}}_{età \rightarrow pad}$ pari a 0,007 ($p \geq 0{,}05$) nella specificazione M1[32]. I risultati riguardanti la

[31] Che, si ricorda, considera anche la storia clinica e l'assunzione di terapia farmacologica antiipertensiva.

[32] Si ricorda che nella specificazione M1, rispetto ad M2 e M3, l'ipertensione arteriosa precede l'uricemia nella relazione causale.

relazione tra età ed ipertensione arteriosa sono in linea con le evidenze di letteratura medica, in cui è noto che se da un lato l'età rappresenta un fattore di rischio per l'aumento della pressione arteriosa sistolica, dall'altra parte assume un ruolo meno marcato o addirittura protettivo in termini di rischio nei confronti dell'aumento della pressione arteriosa diastolica (Pinto, 2007).

Nella relazione diretta con l'uricemia, i coefficienti $\widehat{\beta^{st}}_{età \rightarrow ur}$ risultano sempre significativi, con una variazione di magnitudo riscontrabile prevalentemente tra le diverse misurazioni dell'ipertensione arteriosa all'interno della specificazione causale M1: qui, nel caso in cui si utilizzi la variabile clinica, il coefficiente $\widehat{\beta^{st}}_{età \rightarrow ur_{ip}}$ risulta pari a 0,095, mentre con l'utilizzo della pressione arteriosa diastolica (PAD) e sistolica (PAS) il coefficiente assume rispettivamente valore pari a 0,129 e 0,099. Per la specificazione M2 ed M3 i coefficienti $\widehat{\beta^{st}}_{età \rightarrow ur}$ risultano qualitativamente equivalenti al variare della misurazione dell'ipertensione arteriosa, con valore pari a $\cong$ 0,14 e $\cong$ 0,13 rispettivamente. Il ruolo dell'età come fattore di rischio per l'iperuricemia è un risultato che trova conferma nelle evidenze di letteratura clinica. A supporto di tale affermazione, si veda per esempio il lavoro di Saadat et al. (2018) (Saadat et al., 2018).

4.1.1.3 RUOLO DEL DIABETE NELLA RELAZIONE CON GLI EVENTI CARDIOVASCOLARI

L'aumento del rischio di sperimentare eventi cardiovascolari in presenza di diabete mellito di tipo II è ben documentato e riconosciuto dalla comunità scientifica (American Diabetes Association, 2020). Nei modelli stimati, i risultati per il coefficiente $\widehat{\beta^{st}}_{diab \rightarrow overall}$ si dimostrano coerenti con quanto noto e comparabili tra di loro: si osserva infatti un effetto $\widehat{\beta^{st}}_{diab \rightarrow overall_{ip}}$ pari a 0,073 ($p \leq 0{,}001$) nella misurazione dell'ipertensione con *iperteso*, $\widehat{\beta^{st}}_{diab \rightarrow overall_{pad}}$ = 0,074 ($p \leq 0{,}001$) e $\widehat{\beta^{st}}_{diab \rightarrow overall_{pas}}$ = 0,071 ($p \leq 0{,}001$) nella misurazione con PAD e PAS rispettivamente. Essendo stato assunto il diabete mellito di tipo II sempre e solo in relazione diretta con il sesso, l'età e gli eventi cardiovascolari (*overall)* nei vari *path* causali, non è possibile in questo caso formulare ulteriori considerazioni rispetto al confronto tra le specificazioni M1, M2, M3.

4.1.1.4 RUOLO DELLA CREATININA NELLA RELAZIONE CON URICEMIA ED IPERTENSIONE ARTERIOSA

Spostandosi verso la relazione che coinvolge creatinina, ipertensione arteriosa, e uricemia, si osserva per la specificazione M2 (uricemia → creatinina → ipertensione) un effetto diretto non significativo ($p \geq 0{,}005$) per $\widehat{\beta^{st}}_{creat \to ipert}$ in tutte le modellazioni dell'ipertensione arteriosa, con il coefficiente $\widehat{\beta^{st}}_{creat \to ipert}$ che presenta segno negativo (effetto protettivo) in relazione con la pressione arteriosa diastolica (-0,033) o sistolica (-0,02); altrimenti, con l'utilizzo della variabile clinica *iperteso*, il coefficiente risulta pari a 0,003. Nelle specificazioni M1 ed M3 invece, in cui si assume una relazione invertita tra creatinina ed ipertensione, (ipertensione → creatinina → uricemia) il coefficiente $\widehat{\beta^{st}}_{ipert \to creat}$ risulta sempre significativo ($p \leq 0{,}001$), con un valore di segno positivo nelle modellazioni con pressione arteriosa sistolica e *iperteso*, negativo altrimenti. Rimanendo nelle specificazioni M1, M3, il coefficiente $\widehat{\beta^{st}}_{creat \to ur}$ risulta sempre positivo, non significativo, e di valore pressoché equivalente per tutte le misurazioni dell'ipertensione arteriosa, pari a $\widehat{\beta^{st}}_{creat \to ur} \cong 0{,}13$.

Riassumendo, la creatinina presenta mancanza di effetto diretto (variabile esplicativa) significativo sia nella relazione con l'ipertensione arteriosa, sia con l'uricemia. Da un punto di vista statistico, la non significatività del coefficiente $\widehat{\beta^{st}}_{creat \to ipert}$ nella specificazione M2 potrebbe essere dovuta al fatto che l'effetto di mediazione stimato è molto piccolo, e dunque non riesce ad essere rilevato con i dati a disposizione. Da un punto di vista medico, il risultato è coerente con quanto noto in letteratura: se infatti l'ipertensione arteriosa cronica rappresenta una causa accertata per il danno renale (il quale è associato ad un aumento dei livello di creatinina), dall'altro lato vi sono meno evidenze che suggeriscono un effetto diretto e causale della creatinina sull'insorgenza dell'ipertensione arteriosa (Hanratty et al., 2011; Weiner et al., 2004).

Per quanto riguarda la non significatività dell'effetto diretto tra creatinina (variabile esplicativa) ed uricemia $\widehat{\beta^{st}}_{creat \to ur}$ nelle specificazioni M1 ed M3, il risultato è inatteso rispetto alle evidenze di letteratura; è comprovato infatti che una funzione

renale compromessa (elevati livelli di creatinina) possa portare ad un'inefficienza nell'eliminazione dell'acido urico, con un conseguente aumento della sua concentrazione nel sangue (Johnson et al., 2013; Kang & Nakagawa, 2005). Da un punto di vista statistico, è possibile che il risultato sia dovuto all'esigenza di ipotizzare ulteriori fattori da includere nella modellazione, la cui assenza altrimenti potrebbe inficiare i risultati ottenuti.

In termini di confronto della forza degli effetti stimati nella relazione tra uricemia, creatinina ed ipertensione arteriosa, le relazioni più importanti si rilevano tra l'uricemia come variabile esplicativa per l' ipertensione, e tra la creatinina come variabile esplicativa per l'uricemia, a seconda del modello considerato. Per la creatinina non è stato previsto un effetto diretto con il rischio di sperimentare almeno un evento cardiovascolare, in accordo con la letteratura medica.

4.1.1.5 ANALISI DELLA RELAZIONE TRA URICEMIA, IPERTENSIONE ARTERIOSA ED EVENTI CARDIOVASCOLARI

Si riporta ora, per una lettura più agevole, la rappresentazione grafica della stima tramite SEM per la relazione causale di maggior interesse tra ipertensione arteriosa, uricemia, eventi cardiovascolari nelle diverse specificazioni (M1, M2, M3) e con le diverse misurazioni dell'ipertensione.

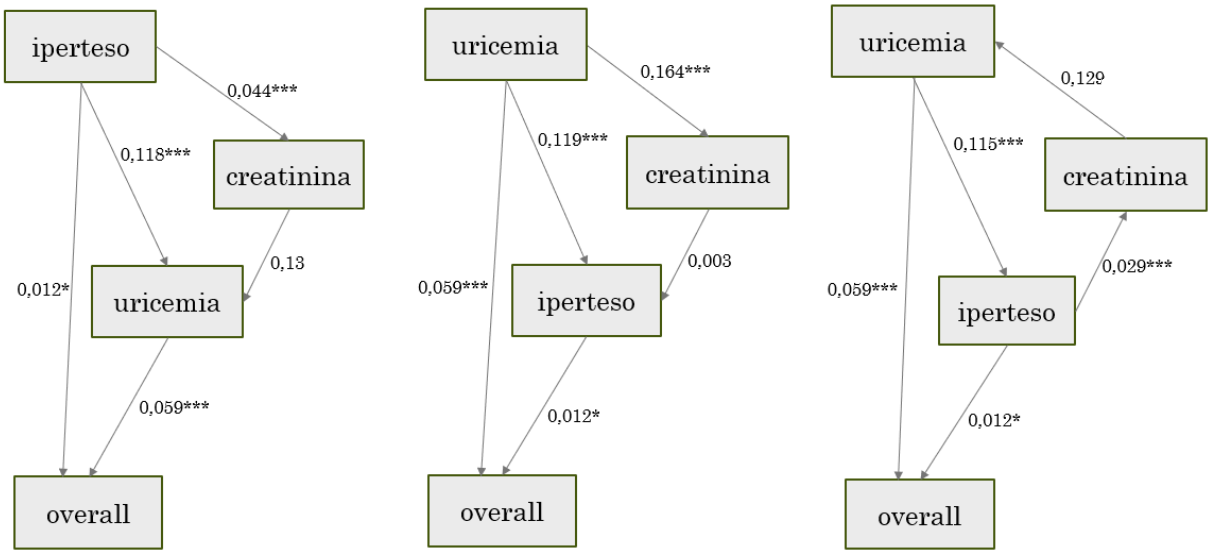


Figura 4.4 – Confronto grafico dei *path* causali (da sinistra: M1, M2, M3) stimati per la relazione tra ipertensione, uricemia, *overall*, con misurazione dell'ipertensione

tramite variabile *iperteso*; coefficienti standardizzati; N = 22184. ***p≤ 0,001; **p ≤ 0,01; *p ≤ 0,05.

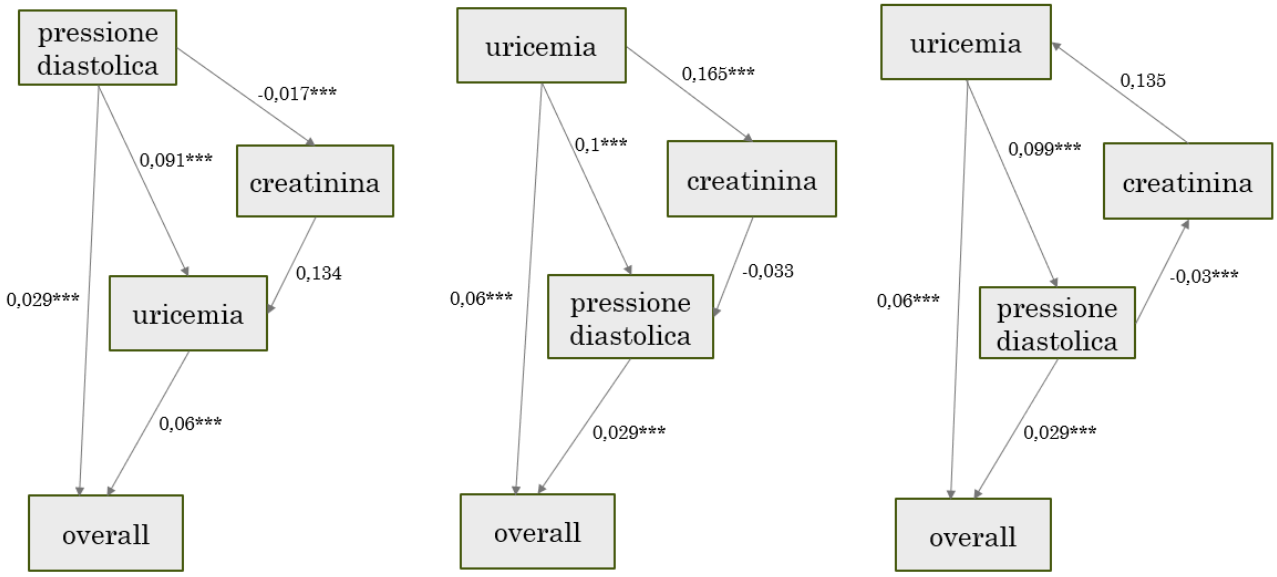


Figura 4.5 – Confronto grafico dei *path* causali (da sinistra: M1, M2, M3) stimati per la relazione tra ipertensione, uricemia, *overall*, con misurazione dell'ipertensione tramite la pressione diastolica; coefficienti standardizzati; N = 21894. ***p≤ 0,001; **p ≤ 0,01; *p ≤ 0,05.

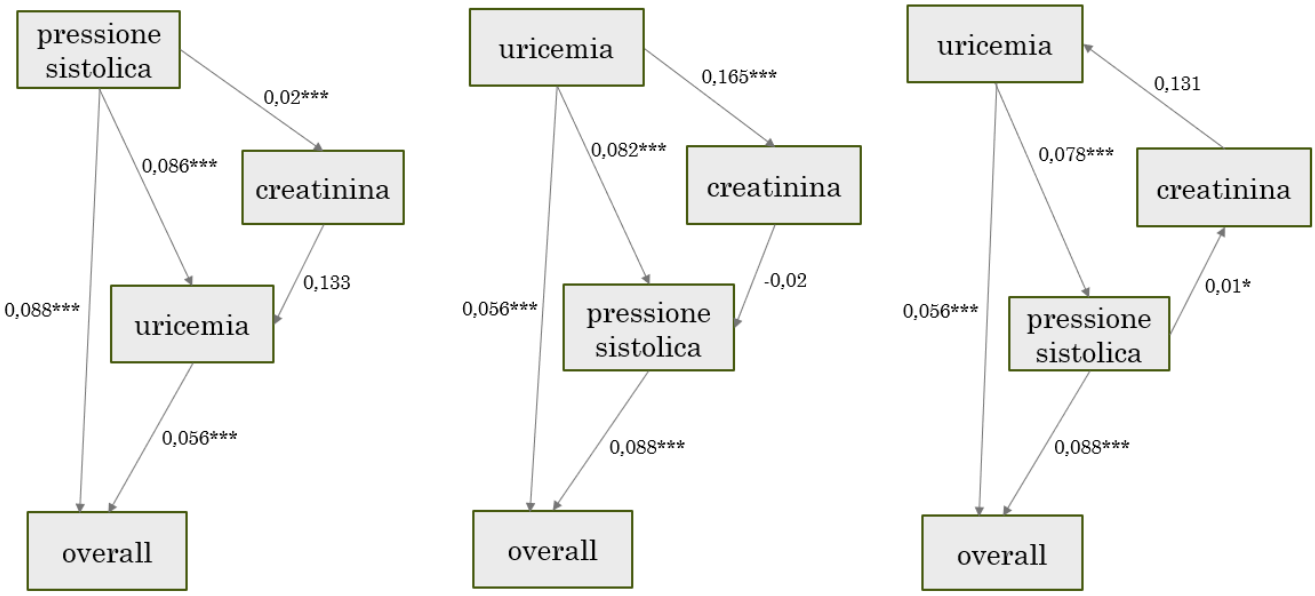


Figura 4.6 – Confronto grafico dei *path* causali (da sinistra: M1, M2, M3) stimati per la relazione tra ipertensione, uricemia, *overall*, con misurazione dell'ipertensione tramite la pressione sistolica; coefficienti standardizzati; N = 21894. ***p≤ 0,001; **p ≤ 0,01; *p ≤ 0,05.

Dall'osservazione delle figure 4.4, 4.5, 4.6 si evidenzia una relazione sempre significativa tra pressione arteriosa, uricemia, e l'insorgenza di almeno un evento cardiovascolare (*overall*), per qualsiasi *path* causale e misurazione dell'ipertensione arteriosa. Questo risultato conferma l'utilizzo dell'ipertensione arteriosa come fattore di rischio per l'insorgenza di eventi cardiovascolari, e ne sottolinea l'importante relazione con l'uricemia, a sua volta relata in modo significativo con l'insorgenza di eventi cardiovascolari.

Rispetto alle diverse misurazioni dell'ipertensione arteriosa, si osserva come queste influiscano nella forza per la relazione diretta stimata tra quest'ultima e l'insorgenza di almeno un evento cardiovascolare (*overall)*: la pressione arteriosa sistolica (PAS) si dimostra la più legata allo sviluppo di eventi cardiovascolari, con un effetto diretto stimato pari a 0,088 ($p \leq 0{,}001$); a seguire decrescendo, vi è l'effetto della pressione diastolica (PAD), pari a 0,029 ($p \leq 0{,}001$) e la variabile clinica *iperteso* (0,012, $p \leq 0{,}05$). Il maggior impatto della pressione arteriosa sistolica sul rischio di eventi cardiovascolari rispetto alla pressione arteriosa diastolica è un risultato che trova ampia conferma nella letteratura medica, specialmente nel caso di individui di età avanzata (Haider et al., 2003). Per quanto riguarda la variabile *iperteso*, è possibile che l'effetto sull'insorgenza di eventi cardiovascolari sia mitigato dal fatto che la definizione clinica identifica come soggetti ipertesi anche coloro che al momento della rilevazione non presentano valori pressori oltre i limiti stabiliti. Anche la significatività associata al coefficiente $\widehat{\beta^{st}}_{ipert \rightarrow overall}$ si presenta più debole rispetto a quella associata alla pressione arteriosa diastolica e sistolica, con un *p-value* pari a 0,048 per la specificazione M1 e 0,047 per le specificazioni M2 ed M3. Rispetto alla modellazione della variabile *iperteso*, è importante infine sottolineare che i risultati potrebbero essere inficiati dalla natura dicotomica della variabile stessa, in contrasto con l'assunzione di normalità multivariata ipotizzata.

Considerando l'effetto diretto dell'uricemia sul rischio di eventi cardiovascolari, questo è pari a 0,056 ($p \leq 0{,}001$) nei modelli che includono la pressione arteriosa sistolica, 0,059 ($p \leq 0{,}001$) nei modelli che includono *iperteso*, e 0,06 ($p \leq 0{,}001$) nei modelli che misurano la pressione arteriosa diastolica; la forza dell'effetto si dimostra dunque equiparabile al netto delle diverse misurazioni dell'ipertensione.

Confrontano la forza dell'effetto diretto dell'uricemia rispetto all'effetto diretto dell'ipertensione arteriosa sull'insorgenza di eventi cardiovascolari, si riscontra nell'uricemia un fattore di rischio più impattante nelle modellazioni che utilizzano la pressione arteriosa diastolica e la variabile *iperteso*. Solamente la pressione arteriosa sistolica risulta più predittiva per *overall* nel confronto con l'uricemia.
Infine, dall'osservazione delle stime ottenute per la relazione causale diretta tra uricemia (come variabile esplicativa) ed ipertensione arteriosa, si evince che anche cambiando la direzione della relazione, la forza degli effetti rimane comparabile: a supporto di quanto appena detto, si riporta che la maggior variabilità nella stima dell'effetto tra uricemia ed ipertensione arteriosa si riscontra tra le specificazioni causali con la pressione arteriosa sistolica, con una deviazione standard (*sd*) pari a $sd = 0{,}004$.

4.1.2 CONFRONTO TRA I VARI INDICI DI BONTÀ DI ADATTAMENTO

Nella Tabella 4.1 sono riportati gli indici di bontà di adattamento scelti per la valutazione dei modelli SEM stimati nelle diverse specificazioni (M1, M2, M3) e con le diverse misurazioni per l'ipertensione arteriosa (*iperteso*, PAD, PAS). Per una presentazione formale degli indici di bontà di adattamento si veda l'Appendice A, sezione 4.

Tabella 4.1 – Indici di bontà di adattamento per i modelli SEM stimati nelle tre specificazioni causali (M1, M2, M3) e con la misurazione dell'ipertensione arteriosa tramite variabile *iperteso*, PAD, PAS[33].

Modello	χ^2_{S-B}	**Df corretti**	***p-value* corretto**	**CFI robusto**	**TLI robusto**	**RMSEA robusto**	**SRMR**	**BIC**
M1 iperteso	285,136	6	,000	,935	,785	,060	,030	162077,188
M2 iperteso	**190,481**	6	,000	,957	,857	**,049**	**,024**	**161917,716**
M3 iperteso	295,675	6	,000	,933	,777	,061	,030	162094,230

[33] In grassetto sono evidenziati i valori di adattamento migliori secondo ciascun indice.

Modello	χ^2_{S-B}	Df corretti	*p-value* corretto	CFI robusto	TLI robusto	RMSEA robusto	SRMR	BIC
M1 PAD	324,476	6	,000	,895	,651	,065	,032	304806,500
M2 PAD	204,565	6	,000	,934	,780	,052	,025	304601,681
M3 PAD	325,527	6	,000	,895	,649	,065	,033	304808,275
M1PAS	306,228	6	,000	,935	,784	,064	,032	328413,014
M2 PAS	195,417	6	,000	**,959**	**,863**	,051	,025	328217,712
M3 PAS	314,542	6	,000	,934	,779	,065	,032	328427,364

Procedendo nella lettura della tabella da sinistra verso destra, si osserva *in primis* il valore per la statistica test χ^2 con correzione di Satorra-Bentler (χ^2_{S-B}), insieme ai rispettivi gradi di libertà e al *p-value* per il test che verifica H_0: $S = \hat{\Sigma} = \Sigma(\hat{\theta})$ vs H_1: $S \neq \hat{\Sigma}$. L'obiettivo del test consiste nel misurare la discrepanza complessiva tra la matrice di covarianza osservata S e quella stimata dal modello $\hat{\Sigma}$. I valori osservati per χ^2_{S-B} variano da un minimo di 190,481 fino a 325,527, con la specificazione M2, variabile *iperteso*, che identifica il miglior adattamento ai dati (minor differenza tra $\hat{\Sigma}$ed S). Il valore più alto di χ^2_{S-B} è associato alla modellazione M3 con variabile PAD. I gradi di libertà associati alla distribuzione χ^2_{S-B} sono uguali tra tutti i modelli, pertanto non ne influenzano la comparabilità. Il *p-value* associato alla statistica test χ^2_{S-B} è sempre significativo, come previsto a causa della sensibilità della statistica χ^2 alla dimensione campionaria, qui molto elevata[34].

Il *Comparative Fit Index* (CFI) e il *Tucker-Lewis Index* (TLI) rappresentano degli indici comparativi (con supporto tra 0 e 1) volti a confrontare il modello specificato ($\hat{\Sigma}$) rispetto al modello nullo in cui si assume assenza di relazioni tra le variabili ($\Sigma_{H_0} = diag(\sigma_1^2, \ldots, \sigma_p^2)$). Questo viene fatto attraverso l'utilizzo della statistica χ^2, qui adottata nella versione robusta χ^2_{S-B}. La principale differenza tra i due indici consiste nel fatto che il TLI penalizza il rapporto $\frac{\chi^2_{model}/df_{model}}{\chi^2_{null}/df_{null}}$ per la quantità χ^2_{null} / df_{null}, che identifica la complessità del modello. I valori osservati per il CFI variano tra .895 (M1, PAD) e .959 (M2, PAS), suggerendo un adattamento complessivamente molto buono per tutti i modelli. La specificazione M2 con *iperteso* rappresenta

[34] N = 22184 per le specificazioni (M1, M2, M3) con variabile *iperteso* e N = 21879 per le specificazioni con PAD e PAS.

quantitativamente il secondo miglior adattamento, con un CFI = 0,957. Se si considera il TLI, tutti i valori sono traslati più in basso, su un *range* che varia tra 0,649 per la specificazione M3 con PAD e 0,863 (M2, PAS); anche in questo caso, il modello M2 *iperteso* misura il secondo miglior adattamento, con un valore assoluto complessivamente più che buono (0,857).

L'errore quadratico medio di approssimazione (RMSEA) robusto utilizza la statistica χ^2_{S-B} per valutare l'errore di approssimazione tra S e $\hat{\Sigma}$ tenendo in considerazione i gradi di libertà del modello adattato. Valori inferiori a .06 indicano un buon adattamento, mentre tra 0,06 e 0,08 l'adattamento è considerato solo accettabile. In queste analisi, i valori variano da 0,049 per la specificazione M2 con *iperteso*, fino a 0,065 osservato sia per la specificazione M1 con PAD che M3 con PAD.

L'errore quadratico medio standardizzato (SRMR) misura la covarianza residua standardizzata tra gli elementi s_{ij} della matrice di covarianza osservata S e gli elementi $\hat{\sigma}_{ij}$ di $\hat{\Sigma}$. In questo caso, i valori osservati per le modellazioni sono tutti a favore di un buon adattamento (SRMR ≤ 0.08), ma il minor SRMR è raggiunto dalla specificazione M2 con variabile *iperteso*. Il valore più alto si osserva (nuovamente) in corrispondenza della modellazione M3 con misurazione di PAD.

Il *Bayesian Information Criterion* (BIC) identifica un altro criterio di valutazione del modello che penalizza la funzione di verosimiglianza -indicatore di qualità di adattamento- per una quantità legata alla complessità del modello ($p\,ln(n)$). In questa analisi viene riportato il BIC in quanto rispetto ad altri criteri d'informazione è più rigoroso nel caso di dimensioni campionarie elevate ed è particolarmente adeguato per evitare di incorrere in sovra-parametrizzazioni.

Il confronto tra i valori di BIC per i vari modelli SEM vede nella specificazione M2 con variabile *iperteso* il miglior modello in termini di compromesso tra varianza e distorsione, con un valore BIC = 161917,716.

Dall'analisi degli indici di bontà di adattamento, emerge come la specificazione causale M2 con misurazione dell'ipertensione arteriosa tramite la variabile clinica *iperteso* sia complessivamente preferibile rispetto ai nove modelli analizzati. Questa affermazione è supportata sia rispetto all'adattamento complessivo in termini di confronto tra S e $\hat{\Sigma}$, sia rispetto ad un principio di parsimonia per i modelli adattati.

Effettuando un confronto tra l'efficacia delle tre diverse misurazioni dell'ipertensione arteriosa nelle modellazioni proposte, la variabile clinica *iperteso* si dimostra associata ai migliori adattamenti per $\frac{2}{3}$ degli indici, con $\frac{1}{3}$ degli indici rimanenti (CFI e TLI) associati al *path* causale M2 con variabile PAS; il valore degli indici CFI e TLI nella specificazione M2 con *iperteso* e M2 con PAS si dimostra comunque qualitativamente comparabile. La peggiore aderenza dei modelli ai dati si riscontra in corrispondenza dell'utilizzo della pressione diastolica (PAD). Quanto appena espresso è in linea con i riferimenti di letteratura medica, in cui si supporta l'utilizzo di una definizione clinica di ipertensione che includa oltre ai valori pressori anche la valutazione della storia clinica e l'anamnesi farmacologica dei pazienti. Tra gli studi che hanno dimostrato il beneficio di un approccio integrato all'ipertensione arteriosa nella predizione degli eventi cardiovascolari si citano i lavori di S. Lewington e del SHEP *Cooperative Research Group* (Sarah Lewington & Robert Clarke, 2002; SHEP *Cooperative Research Group*., 1991). Inoltre, rispetto al confronto tra l'utilizzo della pressione arteriosa diastolica e sistolica, è comprovato come la pressione sistolica rifletta maggiormente la rigidità delle arterie e il carico di lavoro a cui il cuore è sottoposto, con la conseguenza di risultare spesso un predittore più potente per il rischio di sperimentare eventi cardiovascolari, soprattutto negli individui più anziani (Haider et al., 2003). La stessa considerazione circa l'utilizzo della pressione arteriosa sistolica rispetto che diastolica è stata dedotta in modo concorde anche attraverso l'analisi degli effetti causali stimati[35].

Confrontando le tre specificazioni causali, la modellazione di tipo M2 si dimostra sistematicamente più adeguata rispetto alle altre. Si ricorda che la specificazione M2 prevede una modellazione che parte dall'uricemia per arrivare all'ipertensione arteriosa attraverso la mediazione della creatinina. L'uricemia e l'ipertensione arteriosa sono poi a loro volta relate in modo diretto con la variabile endogena (dipendente) *overall*.

In ultima analisi, volendo identificare se influisca di più la specificazione del *path* causale o la misurazione dell'ipertensione arteriosa nell'adattamento ai dati, si osserva maggior variabilità degli indici di adattamento al variare del *path* causale, a

[35] Si veda il paragrafo 4.1.1, pagina 76.

suggerire dunque che l'impostazione delle relazioni causali abbia maggior peso rispetto al modo in cui i fattori coinvolti vengono misurati.

Come già esposto nel *Capitolo 1*, la relazione causale tra uricemia, ipertensione arteriosa ed eventi cardiovascolari rappresenta un'area di ricerca complessa e tuttora in divenire, con risultati talvolta tra loro non concordanti. Rispetto alla seguente analisi, i risultati ottenuti vogliono servire come ulteriore evidenza a supporto di un approccio di valutazione integrato che necessita di successivi approfondimenti. Per poter asserire l'esistenza di un effettivo legame causale infatti, è necessario che lo studio epidemiologico soddisfi dei criteri metodologici ben precisi e qui non tutti verificatisi, quali per esempio la sequenzialità temporale tra i fattori di interesse (longitudinalità dei dati), la plausibilità biologica e la consistenza della relazione, l'evidenza sperimentale, la forza dell'associazione. Per un approfondimento sul tema, si veda il lavoro di A. B. Hill (Austin Bradford Hill, 1965).

4.2 RISULTATI PER LA CAUSAL DISCOVERY

Di seguito si riporta la rappresentazione dei grafi causali stimati attraverso l'algoritmo PC conservativo con assunzione di normalità multivariata dei dati. Per ogni grafo è utilizzata rispettivamente una diversa misurazione dell'ipertensione arteriosa: la definizione clinica (*iperteso*), la pressione arteriosa diastolica (PAD) e la pressione arteriosa sistolica (PAS). Si ricorda infine che i grafi ottenuti appartengono alla classe dei *Complete Partially Directed Acyclic Graphs* (CPDAGs).

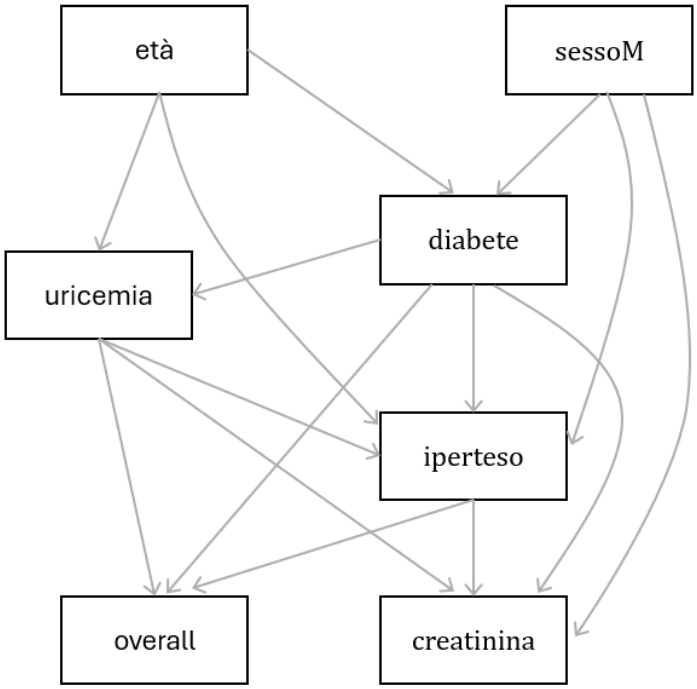


Figura 4.7 – Rappresentazione del grafo causale stimato tramite algoritmo PC-Conservativo; variabile *iperteso*.

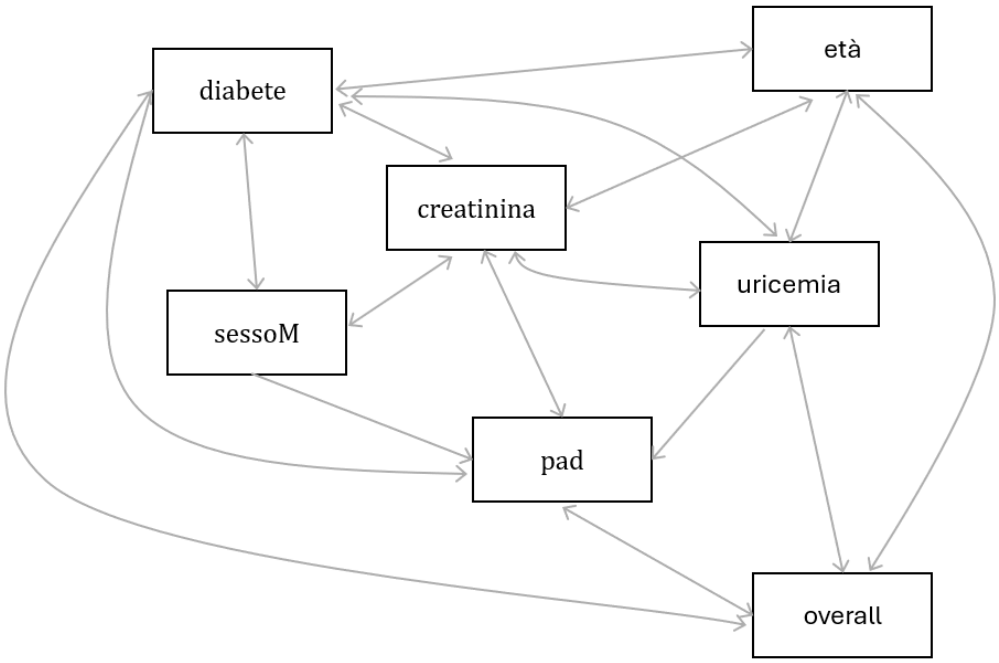


Figura 4.8– Rappresentazione del grafo causale stimato tramite algoritmo PC-Conservativo; variabile *PAD (pressione arteriosa diastolica)*.

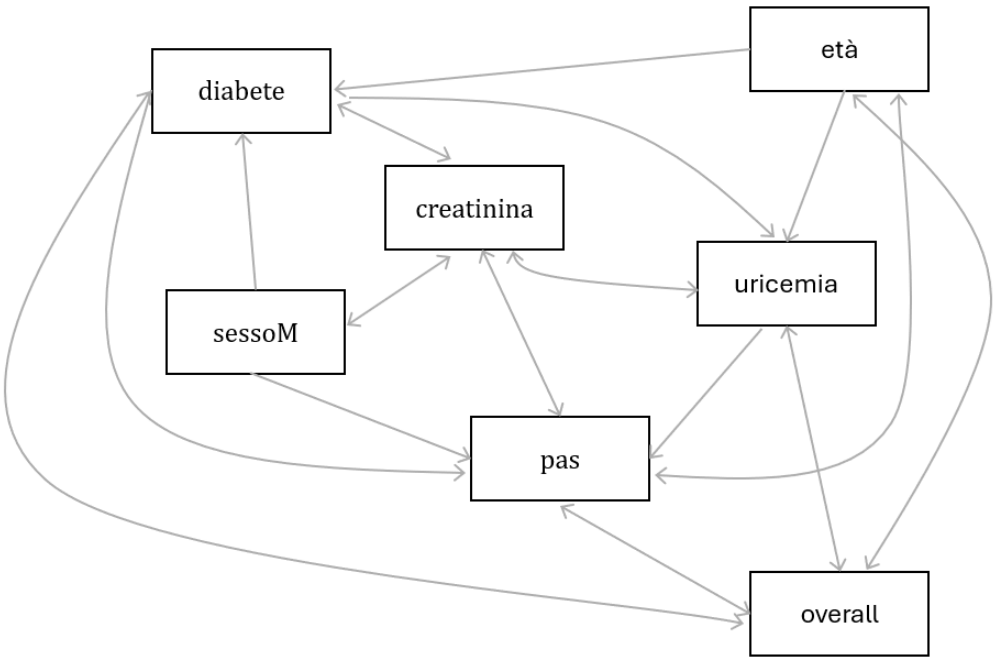


Figura 4.9 – Rappresentazione del grafo causale stimato tramite algoritmo PC-Conservativo; variabile *PAS (pressione arteriosa sistolica)*.

Dal primo grafo causale stimato in *Figura 4.7* -che utilizza la variabile clinica *iperteso*- emerge a colpo d'occhio la presenza di soli archi orientati. Questo accade quando nessuno tra i test di indipendenza condizionata[36] conclude a favore dell'assenza di una connessione diretta tra variabili. Nel *path* causale emerge il ruolo centrale dell'uricemia, che risulta variabile dipendente nella relazione con il sesso, l'età e il diabete, e mostra a sua volta un'influenza direzionata (variabile esplicativa) su ipertensione clinica, creatinina e l'insorgenza di eventi cardiovascolari. Anche l'ipertensione clinica assume una posizione chiave, influenzando direttamente i livelli di creatinina e *overall*, e risultando a sua volta influenzata (variabile dipendente) da uricemia, diabete, sesso ed età. Il ruolo esplicativo dell'uricemia su ipertensione e creatinina, e l'effetto diretto (variabile esplicativa) di uricemia ed ipertensione sull'insorgenza di almeno un evento cardiovascolare (*overall*) rappresentano risultati concordi con la specificazione SEM M2, risultata la migliore in termini di bontà di adattamento ai dati.

[36] Qui test di correlazione parziale H_0: $\rho_{X_i X_j|\mathbf{Z}} = 0$ *vs* H_1: $\rho_{X_i X_j|\mathbf{Z}} \neq 0$.

Nel secondo grafo, che modella la pressione arteriosa diastolica (PAD), sono presenti molti archi non direzionati, evidenziando dunque maggior difficoltà dell'algoritmo nell'identificazione delle relazioni causali. Si ricorda che l'algoritmo di Peter-Clark (PC) si prefigge la stima di un grafo CPDAG appartenente alla classe di equivalenza Markoviana $MEC[\mathcal{G}^*]$ per il vero DAG. Poiché la classe di equivalenza Markoviana è definita in termini di associazioni (*skeleton*) e strutture a *v*, gli archi sono da intendersi non direzionati rispetto a quest'ultimo tipo di relazione specifica, e non comportano alcuna perdita di informazione rispetto alla distribuzione probabilistica congiunta osservata nei dati. La freccia bi-direzionata indica dunque associazione. Concluso l'inciso, in *Figura 4.8* si osserva che l'uricemia esercita un'influenza (variabile esplicativa) sulla pressione arteriosa diastolica, e risulta associata con creatinina, diabete, età e l'insorgenza di almeno un evento cardiovascolare (*overall*). Per quanto riguarda la pressione arteriosa diastolica (PAD), si registra un'associazione con *overall* e creatinina, e un ruolo di variabile dipendente nella relazione con il sesso e il diabete (oltre che con l'uricemia, come specificato poco sopra). Rispetto alle specificazioni SEM, si riscontra concordanza con i *path* causali M2 ed M3 che prevedono un effetto dell'uricemia e del sesso sul livello della pressione arteriosa diastolica. In aggiunta, il grafo causale in *Figura 4.8* identifica un ruolo di variabile esplicativa del diabete sulla pressione arteriosa diastolica.

Anche nell'ultimo grafo stimato in *Figura 4.9* (pressione arteriosa sistolica), molte frecce non sono orientate, indicando nuovamente la difficoltà dell'algoritmo nel determinare le direzioni causali nel CPDAG. L'uricemia mantiene un ruolo centrale sia come variabile dipendente nella relazione con età e diabete, sia come variabile indipendente nella relazione con la pressione arteriosa sistolica; nei confronti dell'insorgenza di almeno un evento cardiovascolare (*overall*) e dell'aumento di creatinina si rileva una relazione non direzionata (associazione). La pressione arteriosa sistolica risulta associata ad *overall*, età e creatinina, e influenzata dal sesso ed i livelli di uricemia. Anche in quest'ultimo grafo si riconferma il ruolo di variabile dipendente dell'uricemia nella relazione con la pressione arteriosa. Le associazioni riscontrate con l'outcome *overall* sono tutte ritrovate anche nelle modellazioni SEM.

Confrontando i tre grafi causali stimati, emerge coerente il ruolo dell'uricemia come variabile "causa" (variabile esplicativa) in relazione all'ipertensione arteriosa in tutte le sue configurazioni misurate (*iperteso*, PAD, PAS). Questo risultato va a supporto delle specificazioni causali a priori di tipo M2 ed M3. Al contrario, le dipendenze specifiche tra uricemia, ipertensione arteriosa e l'insorgenza di almeno un evento cardiovascolare variano a seconda che si rilevi la variabile clinica piuttosto che i valori della pressione arteriosa diastolica o sistolica. In particolare, la variabile clinica *iperteso* consente l'identificazione di un modello CPDAG più robusto e orientato rispetto agli altri due casi. Questi ultimi infatti presentano più archi non direzionati (associazioni), suggerendo una maggiore complessità o difficoltà nell'identificazione delle relazioni causali. Le differenze tra i grafi dunque suggeriscono che la scelta della misura dell'ipertensione incida significativamente sulle relazioni causali identificate, valorizzando l'utilizzo della variabile clinica "iperteso". Il ruolo della creatinina nella relazione tra uricemia ed ipertensione arteriosa rimane complessivamente poco chiaro: i risultati dei modelli grafici non permettono di supportare maggiormente la specificazione M2 piuttosto che M3.

4.3 CONFRONTO DEI RISULTATI OTTENUTI TRAMITE MODELLI AD EQUAZIONI STRUTTURALI (SEM) E CAUSAL DISCOVERY

Nel presente paragrafo si confrontano le relazioni causali ipotizzate nel miglior modello SEM in termini di bontà di adattamento ai dati -specificazione M2, con misurazione dell'ipertensione arteriosa tramite variabile clinica *iperteso*- e il miglior modello grafico ottenuto tramite il processo di *causal discovery* -ovvero il CPDAG che ha portato all'identificazione di più relazioni direzionate-. Questo, con il fine di evidenziare similitudini e differenze tra i risultati più performanti per ciascun metodo applicato, e per stabilire il modello causale la cui specificazione e misurazione dell'ipertensione risulta preferibile alla luce delle evidenze raccolte.
Per una più agevole lettura, si riportano una a fianco all'altra la specificazione del modello SEM che gode di miglior adattamento ai dati (sinistra) e il grafo causale CPDAG risultato più robusto ed efficace (destra):

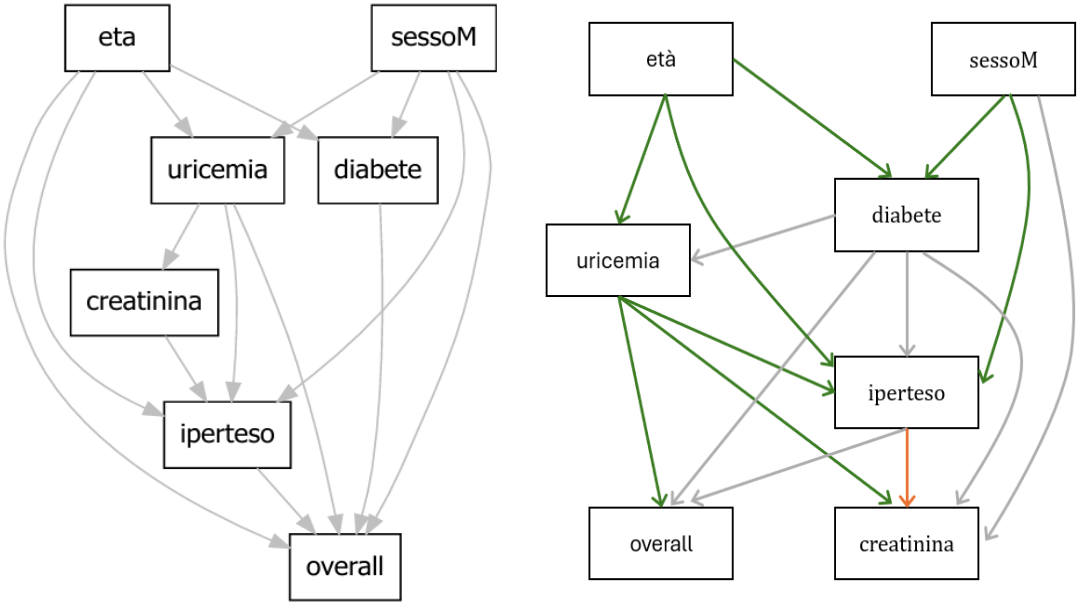


Figura 4.10 – Da sinistra: rappresentazione grafica del modello SEM con miglior adattamento ai dati (M2, *iperteso*) e del miglior modello grafico stimato tramite algoritmo PC conservativo. In verde: relazioni concordanti, arancione: relazioni con direzione invertita, grigio: relazioni presenti solo nel modello grafico.

Partendo dal confronto tra il *path* causale M2 ipotizzato a priori e il *path* causale (grafo CPDAG) stimato tramite algoritmo PC conservativo, si osserva per lo più concordanza tra le relazioni rappresentate (archi verdi), in particolare nella relazione di primario interesse tra uricemia, ipertensione arteriosa e l'insorgenza di almeno un evento cardiovascolare (*overall*). Nello specifico, l'uricemia assume il ruolo di variabile esplicativa per creatinina, ipertensione arteriosa e l'insorgenza di almeno un evento cardiovascolare (*overall)*, con l'ipertensione che è a sua volta variabile esplicativa per l'insorgenza di almeno un evento cardiovascolare. Questo risultato è positivo e rassicurante in quanto mostra coerenza tra le relazioni supportate dal miglior adattamento ai dati in un contesto di inferenza causale, e le relazioni trovate a partire esclusivamente dall'informazione osservata nei dati (*causal discovery*). Inoltre, si sottolinea come le relazioni appena citate godano tutte di plausibilità clinica.

Per quanto riguarda la relazione direzionata trovata tramite *causal discovery* tra ipertensione arteriosa e creatinina, il modello grafico in figura 4.10 (destra) risulta in conflitto con la specificazione M2 e concorde, al contrario, con le specificazioni M1 ed M3 che vedono l'ipertensione come variabile esplicativa dei livelli di

creatinina. Come riportato in precedenza nei *paragrafi 1.2.4, 4.1 e 4.2*, da un punto di vista medico vi sono ad oggi maggiori evidenze a favore della relazione direzionata del tipo ipertensione → creatinina[37]: se l'ipertensione arteriosa cronica rappresenta una causa accertata per il danno renale (il quale è associato ad un aumento dei livello di creatinina), al contrario vi sono meno evidenze a supporto di un effetto diretto e causale della creatinina sull'insorgenza dell'ipertensione arteriosa (Hanratty et al., 2011; Weiner et al., 2004).

Le quattro relazioni trovate in aggiunta tramite l'approccio di *causal discovery* (archi direzionati grigi) vedono il diabete mellito di tipo II come variabile esplicativa in relazione diretta con creatinina, uricemia ed ipertensione arteriosa, ed il sesso come variabile esplicativa per i livelli di creatinina. Essendo qui il sesso e il diabete trattati come variabili "di controllo" rispetto all'obiettivo clinico in esame, si può concludere che complessivamente i risultati ottenuti tramite la valutazione di più modelli SEM e il processo di *causal discovery* siano qualitativamente comparabili. Volendo adottare un approccio prudente nell'asserire la presenza di relazioni causali nel contesto epidemiologico osservazionale in esame e a partire esclusivamente dall'informazione contenuta nei dati, è ragionevole pensare di utilizzare le nuove relazioni identificate (archi grigi) come punto di partenza per ulteriori approfondimenti di letteratura medica e di analisi.

Passando ora alle considerazioni circa la misurazione dell'ipertensione arteriosa, si osserva che entrambi i metodi (inferenza causale e *causal discovery*) performano meglio nel caso in cui si utilizzi la variabile clinica *iperteso*; questo risultato è coerente con molti riferimenti di letteratura medica che appoggiano l'utilizzo di una definizione integrata di ipertensione arteriosa nell'ambito della ricerca clinica, che includa oltre ai valori pressori anche la valutazione della storia clinica e l'anamnesi farmacologica dei pazienti (Carey et al., 2018; US Preventive Services Task Force et al., 2021; Whelton et al., 2018b).

Sulla base di tutte le considerazioni fatte, il *path* causale finale scelto globalmente dunque è il M2 con misurazione dell'ipertensione arteriosa tramite la variabile clinica *iperteso*.

[37] Dunque, a supporto della relazione specificata nelle modellazioni M1 ed M3.

5 CONCLUSIONI

In Italia gli eventi cardiovascolari costituiscono la principale causa di mortalità, morbosità ed invalidità. Nel 2021 si stima che il 30,8% dei decessi sia da attribuire alle patologie cardiovascolari, con una prevalenza più marcata tra la popolazione anziana (Ministero della Salute, 2024). In Europa, gli eventi cardiovascolari sono altrettanto preponderanti: nel 2022 la Società Europea di Cardiologia ha evidenziato la loro rilevanza su scala continentale, riportando come questi siano responsabili del 45% di tutti i decessi (ESC, 2022). A livello globale, l'Organizzazione Mondiale della Sanità (OMS) stima che 17,9 milioni di persone muoiano ogni anno per eventi cardiovascolari, rappresentando il 31% di tutte le morti mondiali (OMS, 2021). L'impatto delle patologie cardiovascolari va oltre l'elevata prevalenza in termini di mortalità, gravando pesantemente sui sistemi sanitari e sulle dinamiche sociali: i pazienti sopravvissuti ad eventi acuti spesso diventano cronici, affrontando significative ripercussioni in termini di qualità della vita, costi economici e disabilità di tipo fisico e cognitivo. Questi effetti si traducono in un notevole onere per la sanità pubblica, influenzando non solo la gestione clinica dei soggetti, ma anche l'economia e il benessere sociale delle famiglie coinvolte (OMS, 2021). È evidente dunque come gli eventi cardiovascolari rappresentino una sfida complessa e contemporanea, che necessita di approfondimenti ed interventi integrati sia a livello clinico che socio-economico.

Tra i principali fattori di rischio clinici associati agli eventi cardiovascolari, ipertensione arteriosa ed uricemia svolgono un ruolo centrale: elevati livelli di uricemia infatti, risultano associati ad una serie di condizioni patologiche tra cui l'ipertensione arteriosa, che a sua volta rappresenta uno dei più importanti fattori di rischio per eventi cardiovascolari quali l'infarto miocardico e l'ictus (Feig et al., 2008b; Sánchez-Lozada et al., 2020; Whelton et al., 2018a). Ad oggi, nonostante l'abbondante numero di studi epidemiologici condotti al riguardo, la natura della relazione e il rapporto causale tra uricemia, ipertensione arteriosa ed eventi cardiovascolari risulta ancora controversa e non completamente nota. Inoltre, insieme ad uricemia ed ipertensione arteriosa, altri fattori come il diabete mellito di tipo II, la creatinina, il sesso e l'età influenzano significativamente l'incidenza degli

eventi cardiovascolari, rendendo necessario un approccio multidimensionale per la valutazione dei rischi (Benjamin et al., 2019).

In questa situazione articolata, si pone un'ulteriore complessità legata al concetto di ipertensione arteriosa, il quale può assumere diverse declinazioni a seconda del contesto di ricerca e delle scelte concettuali sottostanti alla definizione. Le linee guida operative ESH-ESC del 2023 definiscono l'ipertensione arteriosa come una condizione caratterizzata da valori di pressione arteriosa diastolica e/o sistolica rispettivamente ≥ 90 mmHg e ≥ 140 mmHg, basata su almeno due misurazioni (Mancia et al., 2023). Tuttavia, recenti studi hanno evidenziato problematiche legate all'uso dei criteri "*e/o*" nella definizione, suggerendo un'interpretazione più accurata che sostituisca l'operatore booleano "*e/o"* con "*o"* (Casiglia, 2024). Nella ricerca clinica invece, la definizione di soggetto *iperteso* è più ampia, e include nella diagnosi non solo i valori pressori ma anche informazioni riguardanti l'uso di farmaci ipertensivi e la storia clinica del paziente (Carson et al., 2013; Chobanian et al., 2003).

L'obiettivo della presente tesi è stato quello di esplorare il ruolo dell'ipertensione arteriosa nella relazione causale con uricemia ed eventi cardiovascolari, tenendo in considerazione ulteriori fattori di rischio e caratteristiche demografiche necessarie ad una rappresentazione più accurata della relazione d'interesse (diabete mellito di tipo II, creatinina, sesso, età).

Parallelamente, è di interesse valutare l'impatto di diverse misurazioni dell'ipertensione arteriosa nel contesto epidemiologico in esame, confrontando l'utilizzo della variabile clinica dicotomica *iperteso* con la misurazione su scala continua della sola pressione arteriosa diastolica (PAD) o sistolica (PAS) rispettivamente.

I dati utilizzati derivano dallo studio URRAH (*Uric Acid Right for Heart Health),* che fornisce un ampio database epidemiologico (N= 27078) composto dall'aggregazione di dati di natura trasversale provenienti da 13 centri nazionali italiani e contenenti le misurazioni dei valori di acido urico, pressione arteriosa, fenotipizzazione cardiovascolare ed eventi cardiovascolari, oltre che le caratteristiche demografiche dei soggetti. Per ridurre la perdita di informazione derivante dai valori mancanti in corrispondenza degli eventi cardiovascolari, si è deciso creare un *outcome*

composito denominato "almeno un evento cardiovascolare" (*overall*), che assume valore pari a 1 qualora il soggetto sperimenti almeno un evento cardiovascolare (compresa la morte cardiovascolare), e 0 altrimenti.

Lo strumento statistico utilizzato per la valutazione delle interdipendenze tra le variabili cliniche e per la quantificazione degli effetti causali ipotizzati a priori sono i modelli ad equazioni strutturali (SEM) *covariance-based* con funzione di verosimiglianza normale multivariata e coefficienti standardizzati $\widehat{\beta^{st}}$. L'approccio tramite SEM è particolarmente efficace nel rappresentare le relazioni tra fenomeni reali complessi; inoltre, l'utilizzo specifico della funzione di verosimiglianza permette di fare inferenza in modo diretto e disporre di indicatori per la valutazione della bontà di adattamento dei modelli ai dati.

Per l'analisi tramite modelli SEM sono state formulate a priori tre specificazioni teoriche causali (*path* causali M1, M2, M3) secondo evidenze di letteratura clinica, che mettono in discussione ipertensione arteriosa, uricemia e creatinina scambiandone il ruolo nella relazione con l'insorgenza di eventi cardiovascolari.

Per ciascuna specificazione causale si è ripetuta l'analisi utilizzando l'ipertensione arteriosa nelle sue tre diverse misurazioni (*iperteso*, PAD, PAS), per un totale di 9 modelli SEM.

Per valutare quale modellazione fosse più supportata dai dati sono stati analizzati alcuni indicatori di bontà di adattamento (TLI, RMSEA, BIC etc.). I risultati ottenuti tramite l'utilizzo dei modelli SEM confermano la plausibilità di tutte e tre le specificazioni causali sia in termini di significatività dei coefficienti stimati, sia in termini di bontà di adattamento ai dati, offrendo tuttavia maggior supporto in termini di adattamento al percorso causale M2 che vede la creatinina come variabile mediatrice nella relazione tra uricemia (variabile esplicativa) ed ipertensione arteriosa (variabile dipendente), dove uricemia ed ipertensione sono a loro volta in relazione diretta con l'insorgenza di almeno un evento cardiovascolare.

Nel confronto tra la forza degli effetti stimati nella relazione tra uricemia, creatinina ed ipertensione arteriosa, le relazioni più importanti si rilevano tra uricemia ed ipertensione, e tra creatinina ed uricemia, a seconda del modello considerato. In particolare, la relazione positiva e significativa tra uricemia ed ipertensione arteriosa, a sua volta legata a un aumento del rischio di eventi cardiovascolari,

rafforza l'importanza di considerare l'ipertensione arteriosa non solo come fattore di rischio, ma anche come variabile chiave nelle interazioni con altre condizioni patologiche, confermando le evidenze disponibili in letteratura clinica.

La creatinina, nel ruolo di variabile esplicativa in relazione con ipertensione arteriosa o uricemia, non risulta mai significativa: questo risultato, se per la specificazione M2 può trovare un razionale clinico, nelle altre configurazioni causali risulta inaspettato, a significare dunque la potenziale esigenza di considerare altre variabili o ipotizzare altre relazioni causali nella modellazione.

Sempre in termini di quantificazione degli effetti causali, l'età si dimostra complessivamente il determinante con maggior impatto diretto sullo sviluppo di almeno un evento cardiovascolare, con un coefficiente standardizzato che varia da un minimo di 0,187 nei modelli che utilizzano la misurazione della pressione arteriosa sistolica (PAS), fino ad una forza di 0,217 nei modelli che misurano la pressione arteriosa diastolica (PAD).

Rispetto alle diverse definizioni dell'ipertensione arteriosa, la variabile clinica *iperteso* è risultata associata alla miglior modellazione M2 stimata in termini di bontà di adattamento ai dati, suggerendo che un approccio integrato all'ipertensione arteriosa possa migliorare la capacità predittiva per l'insorgenza di eventi cardiovascolari. Tuttavia, dal punto di vista statistico, l'effetto diretto stimato per la variabile *iperteso* sull'insorgenza di almeno un event cardiovascolare (*overall)* risulta sistematicamente più piccolo rispetto a quello della pressione arteriosa diastolica (PAD) o sistolica (PAS), e il livello di significatività osservato per il coefficiente è sistematicamente più alto e molto vicino alla soglia $\alpha = 0.05$ ($p = 0{,}047$ / $p = 0{,}048$); è possibile che questi risultati siano dovuti alla definizione clinica di ipertensione, che include al suo interno anche soggetti che non presentano valori pressori anomali in quanto ipertesi controllati tramite assunzione di terapia farmacologica, o in quanto pazienti con storia clinica ipertensiva ma che al momento non presentano sintomatologia.

A questo punto, a causa della natura trasversale dei dati, i modelli SEM presentano il limite di non poter asserire relazioni causali, fermandosi dunque ad un approccio di tipo ipotetico-deduttivo basato sull'entità e la significatività delle relazioni

ipotizzate a priori, unitamente ad una valutazione degli indici di bontà di adattamento.

Al fine di provare a superare questa limitazione e ottenere nuove evidenze causali per le relazioni indagate, sono stati impiegati i modelli grafici nel contesto della *causal discovery*. La *causal discovery* è un'area metodologica che combina statistica e apprendimento non supervisionato con il fine di identificare relazioni causali tra variabili a partire dai dati. In particolare, l'algoritmo di *causal discovery* di PC (Peter-Clark) conservativo è stato impiegato per stimare una rete causale direttamente a partire dai soli dati osservazionali, fornendo un'evidenza empirica indipendente a integrazione delle ipotesi causali formulate a priori.

Per ogni misurazione dell'ipertensione arteriosa è stato stimato un modello grafico, per un totale di tre *Complete Partially Acyclic Graphs* (CPDAG).

Dall'analisi dei tre grafi ottenuti, emerge in tutte le configurazioni il ruolo centrale dell'uricemia nella relazione con la pressione arteriosa e gli eventi cardiovascolari. In particolare, in tutti i grafi è sempre presente una relazione direzionata dall'uricemia (variabile esplicativa) all'ipertensione arteriosa (variabile dipendente). Il grafo causale basato sulla variabile clinica *iperteso* mostra complessivamente relazioni causali definite grazie alla presenza di soli archi direzionati, ed identifica l'uricemia come variabile esplicativa non solo in relazione con l'ipertensione clinica, ma anche con creatinina ed eventi cardiovascolari. Il ruolo di variabile esplicativa dell'uricemia in relazione a creatinina, ipertensione arteriosa ed eventi cardiovascolari (*overall*) è coerente con il *path* causale M2 che ha ottenuto il miglior adattamento ai dati in fase di inferenza, andando dunque a costituire un risultato potenzialmente significativo e solido verso la comprensione delle relazioni causali oggetto di studio.

Al contrario, nei grafi che utilizzano la pressione arteriosa diastolica (PAD) e sistolica (PAS), molti archi risultano non direzionati, ad evidenziare la difficoltà dell'algoritmo nell'identificare direzioni causali qualora si utilizzino le sole misurazioni pressorie separatamente. Per i grafi che utilizzano i soli valori pressori, le uniche relazioni causali orientate si trovano in corrispondenza della relazione tra uricemia ed ipertensione arteriosa, e tra fattori di controllo come sesso e diabete mellito di tipo II in relazione con i valori pressori.

Una limitazione riscontrata durante l'applicazione dell'algoritmo PC è l'impossibilità di specificare a priori il ruolo (endogeno o esogeno) di alcune variabili; questa opportunità permetterebbe di ottenere grafi causali più coerenti con le conoscenze ad oggi disponibili circa le relazioni causa-effetto tra i fenomeni studiati, portando di conseguenza a risultati più potenti e accurati. Si sottolinea quindi la necessità di futuri sviluppi in questa direzione.

Da un punto di vista statistico-metodologico, la sostanziale concordanza tra il *path* causale M2 e il più potente modello grafico stimato (tramite ipertensione clinica) rappresenta un risultato positivo, a supporto della consistenza delle relazioni stimate tra uricemia, ipertensione arteriosa ed eventi cardiovascolari. Inoltre, i risultati ottenuti mostrano come le differenze concettuali sottostanti alla definizione di ipertensione arteriosa possano influire significativamente sulle relazioni identificate; tuttavia, nel caso della modellazione SEM questo aspetto ha influito meno rispetto all'impiego di diverse specificazioni causali M1, M2, M3 in termini di bontà di adattamento.

Infine, per quanto riguarda il ruolo della creatinina nella relazione tra uricemia e ipertensione arteriosa, questo rimane poco chiaro, sia attraverso l'applicazione dei modelli SEM che dei modelli grafici CPDAG, che non favoriscono in modo univoco una direzione causale rispetto all'altra; tuttavia, da un punto di vista medico, vi sono ad oggi maggiori evidenze a favore della relazione direzionata del tipo ipertensione → creatinina.

In conclusione, questo lavoro mostra come l'integrazione tra modelli SEM e tecniche di *causal discovery* possa fornire un'analisi più accurata e consistente per lo studio delle relazioni causali tra uricemia, ipertensione arteriosa ed eventi cardiovascolari nell'ambito di dati clinici osservazionali, attraverso la combinazione di un approccio di tipo ipotetico-deduttivo con tecniche di apprendimento non supervisionato. L'uricemia si rivela una variabile esplicativa determinante per l'osservazione dell'ipertensione arteriosa, la creatinina, e gli eventi cardiovascolari.

L'ipertensione arteriosa conferma a sua volta il suo ruolo centrale nella relazione con uricemia ed eventi cardiovascolari, rappresentando rispettivamente una variabile dipendente ed un fattore di rischio significativo.

I risultati ottenuti evidenziano inoltre l'efficacia dell'utilizzo della definizione clinica di ipertensione, manifestando allo stesso tempo come la scelta della definizione comporti diverse implicazioni per le relazioni stimate. La scelta della definizione e della conseguente misurazione dell'ipertensione arteriosa dipenderà dunque in ultima battuta dalle esigenze di ricerca e dal ricercatore stesso.

È importante infine sottolineare che i risultati presentati non si prefiggono di asserire conclusioni definitive circa le relazioni causali indagate, bensì vanno considerati come un ulteriore evidenza a supporto di un processo integrato di analisi delle dinamiche che intercorrono tra ipertensione arteriosa, uricemia ed eventi cardiovascolari. Affermare l'esistenza di un effettivo legame causale richiederebbe il soddisfacimento di numerosi e precisi criteri metodologici, che in questo studio non sono verificati: primo tra tutti, la sequenzialità temporale tra i fattori di interesse. Risulta quindi evidente la necessità di condurre ulteriori studi e approfondimenti volti a consolidare i risultati ottenuti e sviluppare nuovi approcci di gestione per gli eventi cardiovascolari, una problematica su scala globale di impatto significativo sia sul piano clinico che socio-economico.

APPENDICE A

I MODELLI AD EQUAZIONI STRUTTURALI

A.1 LA SPECIFICAZIONE DEL MODELLO

La prima fase necessaria all'applicazione dei modelli ad equazioni strutturali consiste nella specificazione del modello multivariato. Qui, il modello teorico ipotizzato viene formalizzato attraverso un insieme di equazioni strutturali. A questa prima fase appartengono dunque l'identificazione e la scelta delle variabili osservate e latenti (ove presenti) da includere nel modello, e la conseguente definizione delle relazioni e i legami che intercorrono tra queste. Seguendo la notazione più comunemente utilizzata, proposta da (Bollen, 1989), il modello ad equazioni strutturali può essere formalizzato nelle tre seguenti componenti lineari:

$$\eta = B\eta + \Gamma\xi + \zeta \tag{A.1}$$

$$y = \Lambda_y\eta + \varepsilon, \qquad \varepsilon \sim N(0, \Theta_\varepsilon) \tag{A.2}$$

$$x = \Lambda_x\xi + \delta, \qquad \delta \sim N(0, \Theta_\delta) \tag{A.3}$$

dove in ordine di presentazione:

- η: vettore m x 1 di variabili latenti endogene;
- *B:* matrice *m* x *m* dei coefficienti che misura l'impatto tra le variabili latenti endogene η;
- Γ: matrice *m* × *n* dei coefficienti che misura l'effetto tra le variabili esogene ξ ed endogene η;
- ξ: vettore *n* × 1 delle variabili latenti esogene;
- ζ: vettore *m* × 1 degli errori di misura per le equazioni strutturali delle variabili latenti endogene η;
- *y*: vettore *p* x 1 delle variabili osservate endogene;
- Λ_y: matrice *p* x *m* dei coefficienti per la relazione tra le variabili *y* e η;
- ε*:* vettore *p* x 1 degli errori di misura per le variabili osservate endogene *y;*

- x: vettore q x 1 delle variabili osservate esogene;
- Λ_x: matrice q x n dei coefficienti per la relazione tra le variabili x e ξ;
- δ: vettore q x 1 degli errori di misura per le variabili osservate esogene x.

La prima equazione A.1 identifica il modello strutturale, mentre le equazioni A.2 e A.3 delineano il modello di misura, il quale stabilisce le relazioni che intercorrono tra le variabili latenti (i costrutti) e le variabili osservate. Le assunzioni usuali alla base della formulazione del modello SEM prevedono indipendenza tra le osservazioni e distribuzione normale multivariata per le variabili dipendenti η, y, x. Per la specificazione del modello, è necessaria inoltre la definizione di quattro matrici di covarianza (Φ, Ψ, Θ_δ e Θ_ε) quadrate e simmetriche, costruite come segue:

- Φ matrice n x n di varianza-covarianza tra due variabili latenti esogene ξ;
- Ψ matrice m x m di varianza-covarianza tra due variabili latenti endogene ξ;
- Θ_ε matrice p x p di varianza-covarianza tra due errori di misura per le variabili osservate endogene y;
- Θ_δ matrice q x q di varianza-covarianza tra due errori di misura per le variabili osservate esogene x.

I modelli ad equazioni strutturali possono essere formalizzati anche attraverso una rappresentazione diagrammatica, comunemente identificata con il nome di *path diagram*. Il *path diagram* è composto da due elementi principali: i nodi e le frecce. I nodi rappresentano le variabili coinvolte nel modello, che possono assumere forma rettangolare o ellittica a seconda che misurino rispettivamente fenomeni direttamente osservati o costrutti latenti. Le frecce rappresentano le relazioni tra le variabili e possono essere direzionate sia unilateralmente che bilateralmente, ad indicare rispettivamente un percorso causale e una relazione di associazione lineare tra le variabili (covarianza o correlazione).

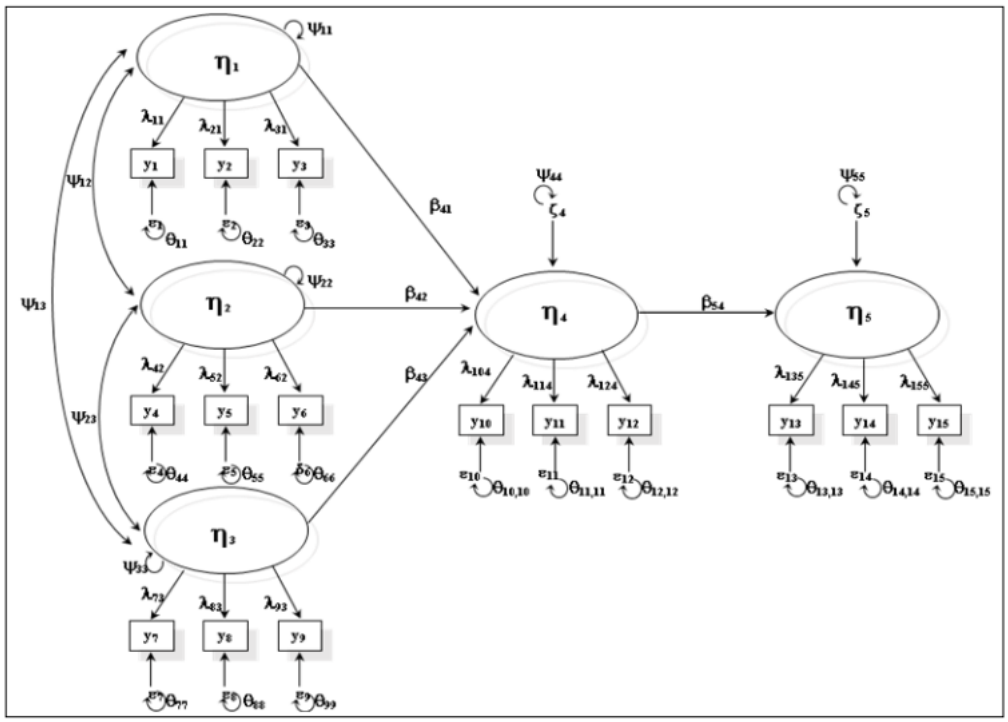


Figura A.1 – Esempio di *path diagram* per un modello ad equazioni strutturali, con la notazione sopra esposta.

A.2 L'IDENTIFICAZIONE DEL MODELLO

L'identificazione di un modello ad equazioni strutturali (SEM) è una fase necessaria a verificare che per tutti i parametri del modello esista una soluzione numerica unica e significativa a partire dai dati disponibili. In altre parole, un modello si dice identificato quando per la matrice delle covarianze stimata è disponibile soltanto un *unico* insieme di parametri stimati che indicizzano il modello. Senza identificazione, le stime dei parametri potrebbero dunque risultare infinite o indeterminate, rendendo il modello non efficace per l'analisi.

Seguendo la metodologia proposta da Bollen (1989), esistono diverse regole e concetti chiave da considerare per garantire l'identificabilità di un modello SEM:

- **T-rule, o condizione minima per l'identificabilità:** il numero *t* dei parametri da stimare deve essere inferiore al numero di elementi non ridondanti nella matrice delle covarianze, ovvero

$$t < \frac{1}{2}(p+q)(p+q-1)$$

con p + q numero di variabili osservate e t numero di elementi da stimare per il vettore dei parametri $\boldsymbol{\theta}$. Qualora il modello presenti tante incognite quanti parametri noti, questo è detto "appena identificato" (*just identified*) e presenterà 0 gradi di libertà, risultando altresì non efficace per le procedure inferenziali di verifica d'ipotesi.

- **Two-step rule:** La regola dei due passi, sufficiente ma non necessaria per l'identificazione, stabilisce che l'identificabilità del modello è valutabile attraverso la verifica dell'identificabilità separatamente per la parte di misura -secondo le regole dei modelli di Analisi Fattoriale Confermativa- e la parte strutturale. L'identificabilità del modello è garantita qualora entrambe le parti del modello si dimostrino identificabili.
- **Null B rule:** Se nessuna variabile latente endogena η influenza alcuna variabile latente endogena, la matrice B è nulla. Qualora B sia nulla, questo rappresenta condizione sufficiente ma non necessaria per l'identificabilità.
  Rispetto allo studio della matrice B nel contesto di modelli ricorsivi, è dimostrato che questi sono identificabili qualora B sia triangolare inferiore, con Ψ matrice diagonale; anche quest'ultima condizione rappresenta condizione sufficiente ma non necessaria per l'identificabilità.
- **Regola d'ordine:** Tale regola, insieme alla *t-rule*, rappresenta una condizione necessaria ma non sufficiente per l'identificabilità di un modello e asserisce che se la matrice Ψ di varianza-covarianza tra due variabili latenti endogene ξ non ha alcun vincolo, allora il modello è identificabile.
- **Regola di rango:** Questa condizione è necessaria e sufficiente per modelli senza restrizioni sulla matrice Ψ: se per ogni equazione del modello la matrice C = [I −B|Γ] ha rango ≥ p−1, allora il modello è identificabile.

A.3 LA STIMA DEL MODELLO

Una volta effettuata la specificazione e assicurata l'identificabilità del modello, è necessario procedere alla stima dello stesso, ovvero alla stima dei parametri che indicizzano il modello ad equazioni strutturali ipotizzato.

Il principio su cui si basano i metodi di stima *covariance-based,* ovvero quelli scelti per l'analisi in esame, prevede la minimizzazione della differenza tra la matrice osservata S di varianza-covarianza e la matrice di varianza-covarianza stimata $\widehat{\sum}$ (=$\sum(\hat{\theta})$). Questo avviene tipicamente attraverso un processo iterativo che sfrutta la funzione di massima verosimiglianza per trovare quei valori di $\hat{\theta}$ che minimizzano la differenza tra $\sum(\vartheta)$ ed S, con $\hat{\theta}$ stimatore di massima verosimiglianza.

La funzione di verosimiglianza da massimizzare è:

$$F_{ML} = \log|\Sigma(\theta)| + \mathrm{tr}\left(s\Sigma^{-1}(\theta)\right) - \log|S| - (\rho + q)$$

con $\Sigma(\theta)$ e S definite positive, x, y normali multivariate e S che segue distribuzione di Wishart (Bollen, 1989).

L'utilizzo di un processo iterativo si rende necessario poiché la funzione di massima verosimiglianza è una funzione non lineare complessa. $\hat{\theta}$ rappresenta quindi il vettore più credibile per ϑ stante l'osservazione della matrice campionaria di varianza-covarianza S.

Altri metodi utilizzabili per il processo di stima del modello ad equazioni strutturali consistono nell'utilizzo dei minimi quadrati parziali (PLS: *Partial Least Squares*), i minimi quadrati generalizzati (GLS: *Generalized Least Squares*) o i minimi quadrati pesati (WLS, *Weighted Least Squares*).

A.4 LA VERIFICA DEL MODELLO

Nel caso di modelli ad equazioni strutturali *covariance-based* stimati tramite funzione di verosimiglianza per dati che seguono una distribuzione normale multivariata, è possibile disporre di alcuni indici che misurano la bontà di adattamento del modello ai dati. Questi indici di basano sul confronto tra la matrice di covarianza osservata S e la matrice di covarianza stimata dal modello $\widehat{\sum}$.

Di seguito si presentano le principali statistiche impiegate nella pratica comune:

- **Test Chi-Quadro χ^2:** il test chi-quadrato verifica la bontà di adattamento complessiva del modello ai dati osservati, confrontando il modello stimato con il modello nullo, ovvero quello in cui tutti gli elementi della matrice di covarianza sono ipotizzati pari a S. Maggiori sono i valori assunti dal test statistico χ^2, maggiore sarà la covarianza residua tra $\widehat{\sum}$ ed S. La distribuzione del test statistico sotto H_0 è pari a $\boldsymbol{\chi^2_{\frac{1}{2}q(q+1)-t}}$, dove q è pari al numero di elementi distinti nella matrice di varianza-covarianza S e t corrisponde al numero di parametri da stimare. Valori di ($p > 0.05$) suggeriscono compatibilità tra il modello stimato e i dati osservati (un buon adattamento), misurando poca discrepanza tra $\widehat{\sum}$ ed S. Il principale limite di questo indice consiste nella sua sensibilità alla dimensione campionaria. Per campioni numerosi infatti, il test tende a concludere in favore dell'ipotesi alternativa H_1: $S \neq \widehat{\sum}$ anche di fronte a minimi scostamenti, risultando dunque *conservativo*. Gli indici elencati di seguito nascono dall'esigenza di risponde proprio alla problematica del Test di adattamento complessivo Chi-Quadro χ^2.
- ***Tucker-Lewis Index*** *(TLI)* e ***Comparative Fit Index*** *(CFI):* l'indice di Tucker-Lewis costituisce insieme al *Comparative Fit Index* (CFI) una misura volta a valutare il miglioramento che il modello stimato apporta nello spiegare i dati osservati rispetto al modello nullo che assume l'assenza di qualsiasi relazione tra le variabili. Per questo motivo, il TLI e il CFI sono detti indici incrementali o comparativi. Entrambe le misure mantengono l'utilizzo della statistica chi-quadrato: L'indice TLI è definito come

$$TLI = \frac{\chi^2_{model}/df_{model} - \chi^2_{null} \,/\, df_{null}}{\chi^2_{null}/df_{null} - 1}$$

mentre l'indice CFI è pari a

$$CFI = 1 - \frac{\chi^2_{model}/df_{model}}{\chi^2_{null}/df_{null}}$$

dove df_{model} e df_{null} rappresentano rispettivamente i gradi di libertà nel modello adattato (ipotizzato) e nel modello nullo. Per entrambi gli indici, valori superiori a 0.90 indicano buon adattamento del modello ai dati.

- ***Root Mean Square Error of Approximation*** (RMSEA): L' errore quadratico medio di approssimazione si avvale sempre della statistica chi-quadrato per valutare quanto il modello adattato ($\hat{\sum}$) approssimi bene la realtà osservata (S). Se l'errore di approssimazione è basso, il modello può ritenersi verosimile nell'approssimare la realtà; altrimenti, vi è indicazione a favore di una differente modellazione dei dati di interesse. L'RMSEA è definito come

$$\text{RMSEA} = \sqrt{\frac{\chi^2_{model} - df_{model}}{(N-1)\,df_{model}}}$$

con N numerosità campionaria. Seguendo le linee guida empiriche fornite da (T. D. Little & Card, 2023), è possibile affermare che il modello approssima sufficientemente bene la realtà per valori dell'RMSEA inferiori a 0.05. Valori ≤ 0.2 sono associati ad un adattamento eccellente, mentre valori compresi tra 0.06 e 0.08 identificano un adattamento accettabile.
Alla stima puntuale dell'RMSEA è possibile associare un intervallo di confidenza approssimato: qualora l'intervallo di confidenza al 90% presenti il limite inferiore e superiore rispettivamente ≤ 0.05 e ≤ 0.08, il modello è da considerarsi un buon adattamento ai dati. Infine, è possibile eseguire una verifica di ipotesi unilaterale al 5% per H_0: RMSEA ≤ 0.05 vs. H_1: RMSEA > 0.05.

- ***Standardized Root Mean Square Residuals*** (SRMSR): L' errore quadratico medio standardizzato fornisce una stima della covarianza residua non spiegata dal modello, ovvero la differenza media tra gli elementi s_{ij} della matrice osservata S e gli elementi $\hat{\sigma}_{ij}$ della matrice $\hat{\sum}$ Più l'SRMSR è piccolo, migliore sarà l'adattamento del modello ai dati.

$$SRMSR = \sqrt{\mathbf{2} \cdot \boldsymbol{\Sigma}_{\boldsymbol{i=1}}^{\boldsymbol{q}} \boldsymbol{\Sigma}_{\boldsymbol{j=1}}^{\boldsymbol{i}} \frac{(\boldsymbol{s_{ij}} - \hat{\sigma}_{\boldsymbol{ij}})^{\mathbf{2}}}{q(q+1)}}$$

dove q è il numero di variabili impiegate nel modello.

Solitamente, per questo indice si identifica un buon adattamento al di sotto del valore soglia stimato 0.08.

A.5 LA MODIFICA DEL MODELLO

Qualora gli indici di bontà di adattamento indichino una scarsa compatibilità tra il modello stimato e i dati osservati, è necessario intraprendere un processo di modifica del modello. Questo processo può includere l'aggiunta o la rimozione di parametri, la revisione delle relazioni tra variabili e di eventuali vincoli imposti, e il riesame delle assunzioni alla base del modello.

In questa fase iterativa di miglioramento, è importante sottolineare la necessità di mantenere sempre per le modifiche un equilibrio tra migliorie metodologico-statistiche e giustificazioni teorico-cliniche. A questo proposito, per la fase di valutazione della bontà di adattamento ed eventuale modifica dei modelli proposti, si farà riferimento principalmente al lavoro proposto da Hu e Bentler (1998).

APPENDICE B

IL MODELLO GRAFICO NORMALE MULTIVARIATO

NOTAZIONE E DEFINIZIONE

Si consideri il vettore di variabili casuali $\boldsymbol{X} = (X_1, X_2, \ldots, X_P)$ che segue una distribuzione normale multivariata con media $\boldsymbol{\mu} \in \mathbb{R}^p$ e matrice di covarianza $\boldsymbol{\Sigma}$. La densità della distribuzione normale multivariata per $\boldsymbol{X}$ è data da:

$$f(x) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{-1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right)$$

Un modello grafico normale multivariato è rappresentato da un grafo $\mathcal{G} = (V, E)$, con V insieme dei nodi di distribuzione normale p-variata, ed E insieme degli archi che rappresentano le relazioni di dipendenza condizionata tra le variabili normali univariate X_i. Essendo $f(x)$ continua e strettamente positiva, le tre proprietà di fattorizzazione, come anche quelle di Markov, si equivalgono.

INTERPRETAZIONE DELLA MATRICE DI PRECISIONE

Nei modelli grafici normali, le relazioni di indipendenza condizionata implicite nel grafo $\mathcal{G}$ sono usualmente rappresentate attraverso l'utilizzo della matrice di precisione (detta anche matrice di concentrazione) $\boldsymbol{K}$, corrispondente all'inverso della matrice di covarianza $\boldsymbol{\Sigma}$:

$$\boldsymbol{K} = \sum^{-1}$$

Se $\boldsymbol{K_{ij}} = 0$, allora X_i e X_j sono da considerarsi condizionatamente indipendenti date tutte le altre variabili contenute in $\boldsymbol{X}$, ovvero

$$X_i \perp X_j | \boldsymbol{X}_{-ij}$$

Gli elementi diagonali k_{ii} della matrice di precisione $\boldsymbol{K}$ possono essere interpretati come il reciproco delle varianze condizionate rispetto ai rimanenti vertici V:

$$k_{ii} = Var\left(X_i | \boldsymbol{X}_{V\setminus\{i\}}\right)^{-1}$$

mentre gli elementi k_{ij} al di fuori della diagonale misurano la relazione lineare tra le variabili:

$$k_{ij} = Cov\left(X_i | \boldsymbol{X}_{V\setminus\{i\}}\right)^{-1} = \frac{-k_{ij}}{k_{ii}k_{jj} - k^2{}_{ij}}$$

Un valore $k_{ij} = 0$, indica l'assenza di relazione tra le variabili $\boldsymbol{X_i}$ **e** $\boldsymbol{X_j}$, il che si traduce nell'assenza assenza di un arco che colleghi i rispettivi vertici del grafo $\mathcal{G}$. In questo caso, utilizzando il criterio di fattorizzazione, la densità congiunta può essere scomposta in due parti distinte: una contenente $\boldsymbol{X_i}$ e l'altra che include $\boldsymbol{X_j}$**.** Questa relazione fondamentale sta alla base della costruzione dei modelli grafici gaussiani e deriva dall'interpretazione della matrice di precisione $\boldsymbol{K}$.

TESI DI INDIPENDENZA CONDIZIONATA PER DATI NORMALI MULTIVARIATI

La definizione di test di indipendenza condizionata per $X_\mathrm{i} \perp X_\mathrm{j} | \boldsymbol{X}_{-\mathrm{ij}}$, identificato con $I(X_\mathrm{i}, X_\mathrm{j} \big| \boldsymbol{X}_{-\mathrm{ij}})$, prevede che l'ipotesi nulla H_0 e alternativa H_1 siano definite rispettivamente come H_0: $X_\mathrm{i} \perp X_\mathrm{j} \big| \boldsymbol{X}_{-\mathrm{ij}}$ e H_1: $\overline{H_0}$. Ne consegue che

$$\hat{I}(X_i, X_j \big| \boldsymbol{X}_{-ij}) > \alpha \implies X_i \perp X_j | \boldsymbol{X}_{-ij}$$

con α livello di significatività statistica.

Per semplicità di notazione, si identifica $\boldsymbol{Z} = \boldsymbol{X}_{-ij}$**.** Nel caso di un modello normale multivariato, H_0 può essere calcolato attraverso l'utilizzo del coefficiente di correlazione parziale $\rho_{X_iX_j|\boldsymbol{Z}}$ derivato dalla matrice di covarianza $\boldsymbol{\Sigma}$ e definito come:

$$\rho_{X_iX_j|\boldsymbol{Z}} = \frac{\rho_{X_iX_j} - \rho_{X_i\boldsymbol{Z}\cdot}\rho_{X_j\boldsymbol{Z}\cdot}}{\sqrt{(1 - \rho^2{}_{X_i\boldsymbol{Z}\cdot})(1 - \rho^2{}_{X_j\boldsymbol{Z}\cdot})}}$$

Una volta calcolato il coefficiente di correlazione parziale $\rho_{X_iX_j|\boldsymbol{Z}}$, è possibile applicare il *test t* per verificare l'ipotesi H_0: $\rho_{X_iX_j|\boldsymbol{Z}}$ *vs* H_1: $\rho_{X_iX_j|\boldsymbol{Z}} \neq 0$.

La statistica per il test è la seguente:

$$t = \rho_{X_i X_j | \boldsymbol{Z}} \cdot \sqrt{\frac{n - |\boldsymbol{Z}| - 2}{1 - \rho^2_{X_i X_j | \boldsymbol{Z}}}}$$

con *n* numero di campioni e $|\boldsymbol{Z}|$ dimensione dell'insieme di condizionamento Z. Sotto H_0, $t \sim t_{n-|\boldsymbol{Z}|-2}$.

Dato il livello di significatività α, si rifiuta H_0 per valori di $|t| > t_{n-|\boldsymbol{Z}|-2,\frac{\alpha}{2}}$, o, equivalentemente, per valori del *p-value* $p = \alpha^{oss} < \alpha$.

APPENDICE C

CONFRONTO TRA MODELLI AD EQUAZIONI STRUTTURALI STIMATI TRAMITE MASSIMA VEROSIMIGLIANZA E FULL-INFORMATION MAXIMUM LIKELIHOOD

Nel presente paragrafo si presentano i risultati derivanti dal confronto tra il miglior modello SEM stimato tramite massima verosimiglianza[38] (N = 22184) e la medesima specificazione stimata tramite l'approccio di *Full Imputation Maximum Likelihood*[39] (N = 27074), con l'obiettivo di offrire una diversa risoluzione alla problematica dei dati mancanti e valutare eventuali differenze ottenute in termini di magnitudo e precisione degli effetti stimati, e bontà di adattamento del modello. Per entrambi i metodi di stima viene assunta una distribuzione normale multivariata per i dati, e viene fornita una versione robusta per gli *standard error* associati alle stime dei coefficienti di regressione standardizzati $\widehat{\beta^{st}}$.

[38] In termini di bontà di adattamento; per un approfondimento, si veda il paragrafo 4.1.2.
[39] Si veda il paragrafo 3.1.3.

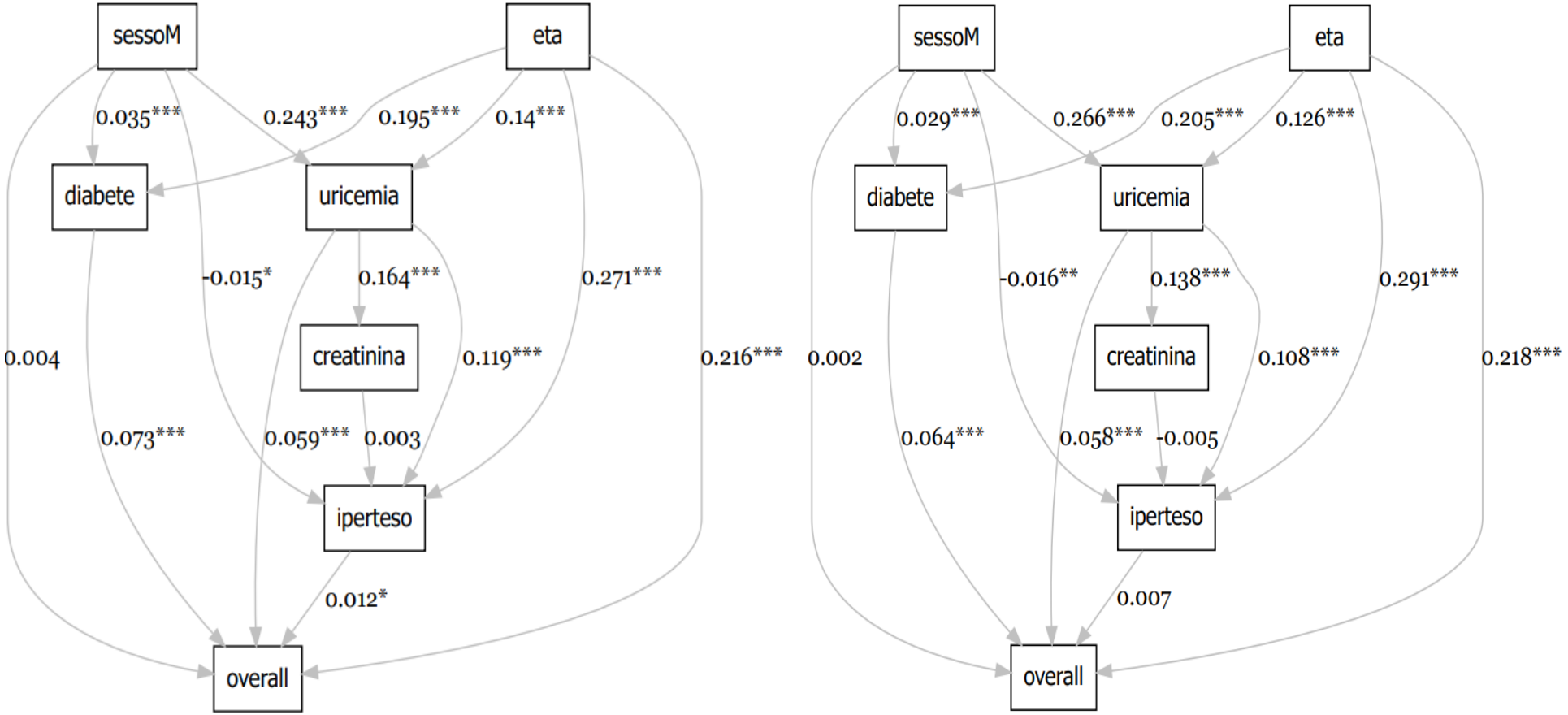


Figura C.1 – Rappresentazione grafica dei modelli SEM stimati tramite ML (sinistra) e FIML (destra); ***p≤ 0,001; **p ≤ 0,01; *p ≤ 0,05.

Concentrandosi principalmente nella relazione di interesse che coinvolge uricemia, ipertensione arteriosa ed eventi cardiovascolari, si può osservare che:

- La significatività della relazione diretta tra l'ipertensione arteriosa e l'insorgenza di almeno un evento cardiovascolare cambia in base al metodo di stima adottato; passando da $p = 0.047$ a $p = 0.243$. Il cambiamento nella significatività può essere associato ai presupposti stessi dei due metodi di stima: la FIML permette di ottenere (sotto assunzione di dati mancanti MCAR o MAR) una stima più robusta, efficiente, e meno influenzata dai dati mancanti rispetto alla massima verosimiglianza, comportando una migliore rappresentazione della reale relazione tra ipertensione arteriosa ed eventi cardiovascolari e riducendo quindi il *bias* di selezione indotto dalla tecnica di *listwise deletion.* Allo stesso tempo, diversi fattori come la distribuzione dei dati mancanti o la presenza di fattori di confondimento latenti potrebbero influenzare la capacità di rilevare associazioni significative tramite FIML. Qualora si decida di scegliere il metodo di stima FIML diventa dunque fondamentale eseguire un'accurata analisi preliminare volta a verificare l'aderenza dei dati con le assunzioni alla base del metodo, così da giustificarne l'utilizzo.

  A livello interpretativo, tenendo in considerazione quanto discusso nel *paragrafo 2.2* e *4.1.1*. circa la costruzione e il comportamento della variabile *iperteso*, e volendo adottare un approccio prudente, si può dire che il *p-value* di 0.047 indichi una relazione significativa al 5% ma che potrebbe non essere robusta vista la vicinanza con il valore soglia nominale α[40]. Sarebbero dunque necessari ulteriori approfondimenti per stabile se i risultati trovati con i due metodi di stima siano sostanzialmente in contrasto tra di loro;
- La significatività delle relazioni tra uricemia ed ipertensione arteriosa, e uricemia e *overall*, rimane invariata, ma la magnitudo dei coefficienti nel modello FIML è tendenzialmente più bassa rispetto alle stime di massima verosimiglianza per il campione di dati ridotto;

[40] Piccole variazioni nei dati o nelle condizioni dell'analisi potrebbero modificare il *p-value* sopra o sotto questa soglia.

- Anche le altre relazioni che coinvolgono le variabili di controllo (sesso, età, diabete) rimangono invariate in termini di significatività;
- L'età risulta complessivamente il determinante con maggior impatto nello sviluppo di eventi cardiovascolari; questo risultato è in linea con quanto emerso anche dal *paragrafo 4.1.1*.

Rispetto alla bontà di adattamento per i due modelli, il modello stimato tramite FIML mostra un'ottima capacità esplicativa, con un CFI robusto pari a 0,969 e un TLI robusto = 0,898. L'RMSEA robusto di 0,041 e lo SRMR di 0,018 confermano l'adattamento eccellente. Il modello stimato tramite massima verosimiglianza invece, sebbene presenti misure più che buone, mostra un adattamento leggermente inferiore, con CFI e TLI robusti rispettivamente pari a 0,957 e 0,857. L'RMSEA robusto è di 0,049 e lo SRMR è di 0,024. I principali indicatori di adattamento per le due modellazioni sono riassunti nella tabella sottostante.

Tabella C.1 – Indici di bontà di adattamento per il modello SEM con specificazione M2 *iperteso* stimata tramite tecnica di massima verosimiglianza (ML, N = 22184) e *Full Information Maximum Likelihood* (FIML, N = 27078)[41].

Modello	χ^2_{S-B}	**Df corretti**	***p-value* corretto**	**CFI robusto**	**TLI robusto**	**RMSEA robusto**	**SRMR**	**BIC**
ML	190.481	6	.000	.957	.857	.049	.024	**161917.716**
FIML	**150.815**	6	.000	**0.969**	**0.898**	**0.041**	**0.018**	197132.887

Gli *standard error* per i coefficienti stimati tramite *Full-Imputation Maximum Likelihood* risultano uguali o poco più piccoli rispetto al modello con massima verosimiglianza, indicando una precisione talvolta migliore per le stime degli effetti causali a favore del metodo della FIML. Questo risultato è concorde alle proprietà di efficienza della stima FIML (C. Enders & Bandalos, 2001).

[41] In grassetto sono evidenziati i valori di adattamento migliori rispetto a ciascun indice.

Tabella C.2 – Coefficienti di regressione standardizzati e relativi *standard error* per la stima del modello SEM (M2, *iperteso*) tramite massima verosimiglianza (ML, N = 22184) e *Full Information Maximum Likelihood* (FIML, N = 27078).

Relazione	Coefficiente (ML)	Std. Error (ML)	Coefficiente (FIML)	Std. Error (FIML)
creatinina ~				
uricemia	0.164	0.002	0.138	0.001
iperteso ~				
creatinina	0.003	0.004	-0-005	0.004
uricemia	0.119	0.002	0.108	0.002
età	0.271	0.000	0.291	0.000
sesso (M)	-0.015	0.006	-0.016	0.006
uricemia ~				
età	0.140	0.001	0.126	0.001
Sesso (M)	0.266	0.018	0.243	0.017
overall ~				
iperteso	0.012	0.004	0.007	0.004
uricemia	0.059	0.002	0.058	0.001
diabete	0.073	0.008	0.064	0.007
sesso (M)	0.004	0.004	0.002	0.004
età	0.216	0.000	0.218	0.000
diabete ~				
età	0.195	0.000	0.205	0.000
sesso (M)	0.035	0.000	0.029	0.000

Riassumendo quanto emerso dal seguente approfondimento, si può concludere che i modelli stimati tramite massima verosimiglianza e *Full Information Maximum Likelihood* si presentano perlopiù coerenti e comparabili nei risultati ottenuti con i due metodi di stima, e offrono un'ottima adattabilità ai dati. Tuttavia, il metodo della *Full Information Maximum Likelihood* dimostra talvolta una maggiore precisione nella stima dei coefficienti di regressione -come evidenziato dagli *std. error* più piccoli- e una migliore capacità esplicativa in termini di bontà di adattamento. La scelta di adottare il metodo di *listwise deletion* basato sulla funzione di verosimiglianza "semplice" rispetto al metodo della *Full Imputation Maximum Likelihood* dipenderà dunque dalle caratteristiche distributive riscontrate pe l'insieme di dati mancanti in esame, le esigenze operative e l'obiettivo dell'analisi.

BIBLIOGRAFIA

Acock, A. C. (2005). Working With Missing Values. *Journal of Marriage and Family*, *67*(4), 1012–1028. https://doi.org/10.1111/j.1741-3737.2005.00191.x

Agresti, A. (2013). *Categorical data analysis* (3rd ed). Wiley.

Altman, D. G. (1991). Practical statistics for medical research. Douglas G. Altman, Chapman and Hall, London, 1991. No. of pages: 611. Price: £32.00. *Statistics in Medicine*, *10*(10), 1635–1636. https://doi.org/10.1002/sim.4780101015

American Diabetes Association. (2020). 2. Classification and Diagnosis of Diabetes: *Standards of Medical Care in Diabetes—2020*. *Diabetes Care*, *43*(Supplement_1), S14–S31. https://doi.org/10.2337/dc20-S002

Armstrong, P. W., & Westerhout, C. M. (2017). Composite End Points in Clinical Research: A Time for Reappraisal. *Circulation*, *135*(23), 2299–2307. https://doi.org/10.1161/CIRCULATIONAHA.117.026229

Arnett, D. K., Blumenthal, R. S., Albert, M. A., Buroker, A. B., Goldberger, Z. D., Hahn, E. J., Himmelfarb, C. D., Khera, A., Lloyd-Jones, D., McEvoy, J. W., Michos, E. D., Miedema, M. D., Muñoz, D., Smith, S. C., Virani, S. S., Williams, K. A., Yeboah, J., & Ziaeian, B. (2019). 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*, *140*(11). https://doi.org/10.1161/CIR.0000000000000678

Austin Bradford Hill. (1965). The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295–300.

Banzato, E. (2023). *Two-sample inference for graphical models* (https://hdl.handle.net/11577/3487882). Università degli Studi di Padova.

Baracaldo-Santamaría, D., Feliciano-Alfonso, J. E., Ramirez-Grueso, R., Rojas-Rodríguez, L. C., Dominguez-Dominguez, C. A., & Calderon-Ospina, C. A. (2023). Making Sense of Composite Endpoints in Clinical Research. *Journal of Clinical Medicine*, *12*(13), 4371. https://doi.org/10.3390/jcm12134371

Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Chamberlain, A. M., Chang, A. R., Cheng, S., Das, S. R., Delling, F. N., Djousse, L., Elkind, M. S. V., Ferguson, J. F., Fornage, M., Jordan, L. C., Khan, S. S., Kissela, B. M., Knutson, K. L., … On behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. (2019). Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. *Circulation*, *139*(10). https://doi.org/10.1161/CIR.0000000000000659

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Bombelli, M., Ronchi, I., Volpe, M., Facchetti, R., Carugo, S., Dell'Oro, R., Cuspidi, C., Grassi, G., & Mancia, G. (2014). Prognostic value of serum uric acid: New-onset in and out-of-office hypertension and long-term mortality. *Journal of Hypertension*, *32*(6), 1237–1244. https://doi.org/10.1097/HJH.0000000000000161

Bonett, D. G., & Price, R. M. (2005). Inferential Methods for the Tetrachoric Correlation Coefficient. *Journal of Educational and Behavioral Statistics*, *30*(2), 213–225. https://doi.org/10.3102/10769986030002213

Bongers, S., Forré, P., Peters, J., & Mooij, J. M. (2016). *Foundations of Structural Causal Models with Cycles and Latent Variables*. https://doi.org/10.48550/ARXIV.1611.06221

Bos, M. J., Koudstaal, P. J., Hofman, A., Witteman, J. C. M., & Breteler, M. M. B. (2006). Uric Acid Is a Risk Factor for Myocardial Infarction and Stroke: The Rotterdam Study. *Stroke*, *37*(6), 1503–1507. https://doi.org/10.1161/01.STR.0000221716.55088.d4

Byrne, B. M. (2016). *Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming, Third Edition* (0 ed.). Routledge. https://doi.org/10.4324/9781315757421

Cannon, P. J., Stason, W. B., Demartini, F. E., Sommers, S. C., & Laragh, J. H. (1966). Hyperuricemia in Primary and Renal Hypertension. *New England Journal of Medicine*, *275*(9), 457–464. https://doi.org/10.1056/NEJM196609012750902

Carey, R. M., Whelton, P. K., & for the 2017 ACC/AHA Hypertension Guideline Writing Committee. (2018). Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Synopsis of the 2017 American College of Cardiology/American Heart Association Hypertension Guideline. *Annals of Internal Medicine*, *168*(5), 351. https://doi.org/10.7326/M17-3203

Carson, A. P., Lewis, C. E., Jacobs, D. R., Peralta, C. A., Steffen, L. M., Bower, J. K., Person, S. D., & Muntner, P. (2013). Evaluating the Framingham Hypertension Risk Prediction Model in Young Adults: The Coronary Artery Risk Development in Young Adults (CARDIA) Study. *Hypertension*, *62*(6), 1015–1020. https://doi.org/10.1161/HYPERTENSIONAHA.113.01539

Casiglia, E. (2024). AND, OR, AND/OR in hypertension guidelines. *Journal of Hypertension*, *42*(5), 934–935. https://doi.org/10.1097/HJH.0000000000003700

Casiglia, E., Tikhonoff, V., Virdis, A., Masi, S., Barbagallo, C. M., Bombelli, M., Bruno, B., Cicero, A. F. G., Cirillo, M., Cirillo, P., Desideri, G., D'Elia, L., Ferri, C., Galletti, F., Gesualdo, L., Giannattasio, C., Iaccarino, G., Lippa, L., Mallamaci, F., … Borghi, C. (2020). Serum uric acid and fatal myocardial infarction: Detection of prognostic cut-off values: The URRAH (Uric Acid Right for Heart Health)

study. *Journal of Hypertension*, *38*(3), 412–419. https://doi.org/10.1097/HJH.0000000000002287

Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo, J. L., Jones, D. W., Materson, B. J., Oparil, S., Wright, J. T., Roccella, E. J., & the National High Blood Pressure Education Program Coordinating Committee. (2003). Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension*, *42*(6), 1206–1252. https://doi.org/10.1161/01.HYP.0000107251.49515.c2

Choi, H. K., & Ford, E. S. (2007). Prevalence of the Metabolic Syndrome in Individuals with Hyperuricemia. *The American Journal of Medicine*, *120*(5), 442–447. https://doi.org/10.1016/j.amjmed.2006.06.040

Colombo, D., & Maathuis, M. H. (2014). *Order-independent constraint-based causal structure learning* (Versione 2). arXiv. https://doi.org/10.48550/ARXIV.1211.3295

Culleton, B. F., Larson, M. G., Kannel, W. B., & Levy, D. (1999). Serum Uric Acid and Risk for Cardiovascular Disease and Death: The Framingham Heart Study. *Annals of Internal Medicine*, *131*(1), 7. https://doi.org/10.7326/0003-4819-131-1-199907060-00003

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16–29. https://doi.org/10.1037/1082-989X.1.1.16

Dawid, A. P. (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *41*(1), 1–15. https://doi.org/10.1111/j.2517-6161.1979.tb01052.x

Edwards, D. (2000). *Introduction to graphical modelling* (2. ed). Springer.

Enders, C., & Bandalos, D. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 430–457. https://doi.org/10.1207/S15328007SEM0803_5

Enders, C. K. (2022). *Applied missing data analysis* (Second Edition). The Guilford Press.

ESC. (2022). *ESC Atlas of Cardiology*. https://www.escardio.org/Research/ESC-Atlas-of-cardiology

Fang, J., & Alderman, M. H. (2000). Serum Uric Acid and Cardiovascular Mortality: The NHANES I Epidemiologic Follow-up Study, 1971-1992. *JAMA*, *283*(18), 2404. https://doi.org/10.1001/jama.283.18.2404

Feig, D. I., Kang, D.-H., & Johnson, R. J. (2008a). Uric Acid and Cardiovascular Risk. *New England Journal of Medicine*, *359*(17), 1811–1821. https://doi.org/10.1056/NEJMra0800885

Feig, D. I., Kang, D.-H., & Johnson, R. J. (2008b). Uric Acid and Cardiovascular Risk. *New England Journal of Medicine*, *359*(17), 1811–1821. https://doi.org/10.1056/NEJMra0800885

Gertler, M. M. (1951). SERUM URIC ACID IN RELATION TO AGE AND PHYSIQUE IN HEALTH AND IN CORONARY HEART DISEASE. *Annals of Internal Medicine*, *34*(6), 1421. https://doi.org/10.7326/0003-4819-34-6-1421

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, *60*(1), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Grundy, S. M., Stone, N. J., Bailey, A. L., Beam, C., Birtcher, K. K., Blumenthal, R. S., Braun, L. T., De Ferranti, S., Faiella-Tommasino, J., Forman, D. E., Goldberg, R., Heidenreich, P. A., Hlatky, M. A., Jones, D. W., Lloyd-Jones, D., Lopez-Pajares, N., Ndumele, C. E., Orringer, C. E., Peralta, C. A., … Yeboah, J. (2019). 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*, *139*(25). https://doi.org/10.1161/CIR.0000000000000625

Gunzler, D., Chen, T., Wu, P., & Zhang, H. (2013). Introduction to mediation analysis with structural equation modeling. *Shanghai Archives of Psychiatry*, *25*(6), 390–394. https://doi.org/10.3969/j.issn.1002-0829.2013.06.009

Haider, A. W., Larson, M. G., Franklin, S. S., & Levy, D. (2003). Systolic Blood Pressure, Diastolic Blood Pressure, and Pulse Pressure as Predictors of Risk for Congestive Heart Failure in the Framingham Heart Study. *Annals of Internal Medicine*, *138*(1), 10. https://doi.org/10.7326/0003-4819-138-1-200301070-00006

Halperin Kuhns, V. L., & Woodward, O. M. (2020). Sex Differences in Urate Handling. *International Journal of Molecular Sciences*, *21*(12), 4269. https://doi.org/10.3390/ijms21124269

Hamdan, M. A. (1970). The Equivalence of Tetrachoric and Maximum Likelihood Estimates of ρ in 2 × 2 Tables. *Biometrika*, *57*(1), 212. https://doi.org/10.2307/2334955

Hanratty, R., Chonchol, M., Havranek, E. P., Powers, J. D., Dickinson, L. M., Ho, P. M., Magid, D. J., & Steiner, J. F. (2011). Relationship between Blood Pressure and Incident Chronic Kidney Disease in Hypertensive Patients. *Clinical Journal of the American Society of Nephrology*, *6*(11), 2605–2611. https://doi.org/10.2215/CJN.02240311

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., & Vila–Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *44*(1), 153–166. https://doi.org/10.1007/s11135-008-9190-y

ISS. (2022, gennaio). *Il dosaggio di creatinina negli adulti*. https://www.issalute.it/index.php/la-salute-dalla-a-alla-z-menu/a/analisi-cliniche/creatinina#:~:text=Il%20valore%20normale%20per%20la,l'ampiezza%20della%20massa%20muscolare.

J. Pearl. (2000). CAUSALITY: MODELS, REASONING, AND INFERENCE. *Econometric Theory*, *19*(04). https://doi.org/10.1017/S0266466603004109

Johnson, R. J., Nakagawa, T., Jalal, D., Sanchez-Lozada, L. G., Kang, D.-H., & Ritz, E. (2013). Uric acid and chronic kidney disease: Which is chasing which? *Nephrology Dialysis Transplantation*, *28*(9), 2221–2228. https://doi.org/10.1093/ndt/gft029

Kalisch, M., M. Maechler, D. Colombo, M. H. Maathius, & P. Buhlmann. (2012). Causal Inference Using Graphical Models with the R Package pcalg. *Journal of Statistical Software*, 1–26.

Kang, D.-H., & Nakagawa, T. (2005). Uric acid and chronic renal disease: Possible implication of hyperuricemia on progression of renal disease. *Seminars in Nephrology*, *25*(1), 43–49. https://doi.org/10.1016/j.semnephrol.2004.10.001

Kaplan, D. (2009). *Structural Equation Modeling (2nd ed.): Foundations and Extensions*. SAGE Publications, Inc. https://doi.org/10.4135/9781452226576

Kario, K. (2018). Nocturnal Hypertension: New Technology and Evidence. *Hypertension*, *71*(6), 997–1009. https://doi.org/10.1161/HYPERTENSIONAHA.118.10971

Kendall, M. G. (1941). (Iv) Proof of Relations connected with the Tetrachoric Series and its Generalization. *Biometrika*, *32*(2), 196–198. https://doi.org/10.1093/biomet/32.2.196

Lauritzen, S. (2000). *Causal Inference from Graphical Models*.

Levey, A. S., Coresh, J., Greene, T., Stevens, L. A., Zhang, Y. (Lucy), Hendriksen, S., Kusek, J. W., & Van Lente, F. (2006). Using Standardized Serum Creatinine Values in the Modification of Diet in Renal Disease Study Equation for Estimating Glomerular Filtration Rate. *Annals of Internal Medicine*, *145*(4), 247–254. https://doi.org/10.7326/0003-4819-145-4-200608150-00004

Li, C., & Fan, X. (2020). On nonparametric conditional independence tests for continuous variables. *WIREs Computational Statistics*, *12*(3), e1489. https://doi.org/10.1002/wics.1489

Libby, P., Buring, J. E., Badimon, L., Hansson, G. K., Deanfield, J., Bittencourt, M. S., Tokgözoğlu, L., & Lewis, E. F. (2019). Atherosclerosis. *Nature Reviews Disease Primers*, *5*(1), 56. https://doi.org/10.1038/s41572-019-0106-z

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (1ª ed.). Wiley. https://doi.org/10.1002/9781119013563

Little, R., & Rubin, D. (2019). *Statistical Analysis with Missing Data, Third Edition* (1ª ed.). Wiley. https://doi.org/10.1002/9781119482260

Little, T. D., & Card, N. A. (2023). *Longitudinal structural equation modeling* (Second edition). The Guilford Press.

Maas, A. H. E. M., & Appelman, Y. E. A. (2010). Gender differences in coronary heart disease. *Netherlands Heart Journal*, *18*(12), 598–603. https://doi.org/10.1007/s12471-010-0841-y

Maloberti, A., Giannattasio, C., Bombelli, M., Desideri, G., Cicero, A. F. G., Muiesan, M. L., Rosei, E. A., Salvetti, M., Ungar, A., Rivasi, G., Pontremoli, R., Viazzi, F., Facchetti, R., Ferri, C., Bernardino, B., Galletti, F., D'Elia, L., Palatini, P., Casiglia, E., … Working Group on Uric Acid and Cardiovascular Risk of the Italian Society of Hypertension (SIIA). (2020). Hyperuricemia and Risk of Cardiovascular Outcomes: The Experience of the URRAH (Uric Acid Right for Heart Health) Project. *High Blood Pressure & Cardiovascular Prevention*, *27*(2), 121–128. https://doi.org/10.1007/s40292-020-00368-z

Maloberti, A., Mengozzi, A., Russo, E., Cicero, A. F. G., Angeli, F., Agabiti Rosei, E., Barbagallo, C. M., Bernardino, B., Bombelli, M., Cappelli, F., Casiglia, E., Cianci, R., Ciccarelli, M., Cirillo, M., Cirillo, P., Desideri, G., D'Elia, L., Dell'Oro, R., Facchetti, R., … Working Group on Uric Acid and Cardiovascular Risk of the Italian Society of Hypertension (SIIA). (2023). The Results of the URRAH (Uric Acid Right for Heart Health) Project: A Focus on Hyperuricemia in Relation to Cardiovascular and Kidney Disease and its Role in Metabolic Dysregulation. *High Blood Pressure & Cardiovascular Prevention*, *30*(5), 411–425. https://doi.org/10.1007/s40292-023-00602-4

Mancia, G., Kreutz, R., Brunström, M., Burnier, M., Grassi, G., Januszewicz, A., Muiesan, M. L., Tsioufis, K., Agabiti-Rosei, E., Algharably, E. A. E., Azizi, M., Benetos, A., Borghi, C., Hitij, J. B., Cifkova, R., Coca, A., Cornelissen, V., Cruickshank, J. K., Cunha, P. G., … Kjeldsen, S. E. (2023). 2023 ESH Guidelines for the management of arterial hypertension The Task Force for the management of arterial hypertension of the European Society of Hypertension: Endorsed by the International Society of Hypertension (ISH) and the European Renal

Association (ERA). *Journal of Hypertension*, *41*(12), 1874–2071. https://doi.org/10.1097/HJH.0000000000003480

McCoy, C. (2018). Understanding the Use of Composite Endpoints in Clinical Trials. *Western Journal of Emergency Medicine*, *19*(4), 631–634. https://doi.org/10.5811/westjem.2018.4.38383

Ministero della Salute. (2022, marzo). *Calcolo Indice massa corporea—IMC (BMI - Body mass index)*. https://www.salute.gov.it/portale/nutrizione/dettaglioIMCNutrizione.jsp?lingua=italiano&id=5479&area=nutrizione&menu=vuoto

Ministero della Salute. (2024, aprile). *Malattie cardiovascolari in Italia*. https://www.salute.gov.it/portale/donna/dettaglioContenutiDonna.jsp?lingua=italiano&id=4490&area=Salute%20donna&menu=patologie

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., De Ferranti, S., Després, J.-P., Fullerton, H. J., Howard, V. J., Huffman, M. D., Judd, S. E., Kissela, B. M., Lackland, D. T., Lichtman, J. H., Lisabeth, L. D., Liu, S., Mackey, R. H., Matchar, D. B., … Turner, M. B. (2015). Heart Disease and Stroke Statistics—2015 Update: A Report From the American Heart Association. *Circulation*, *131*(4). https://doi.org/10.1161/CIR.0000000000000152

Niskanen, L. K., Laaksonen, D. E., Nyyssönen, K., Alfthan, G., Lakka, H.-M., Lakka, T. A., & Salonen, J. T. (2004). Uric Acid Level as a Risk Factor for Cardiovascular and All-Cause Mortality in Middle-aged Men: A Prospective Cohort Study. *Archives of Internal Medicine*, *164*(14), 1546. https://doi.org/10.1001/archinte.164.14.1546

OMS. (2021, giugno). *Cardiovascular Diseases (CVDs) Fact Sheets*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Patel, S. S., Molnar, M. Z., Tayek, J. A., Ix, J. H., Noori, N., Benner, D., Heymsfield, S., Kopple, J. D., Kovesdy, C. P., & Kalantar-Zadeh, K. (2013). Serum creatinine as a marker of muscle mass in chronic kidney disease: Results of a cross-sectional study and review of literature. *Journal of Cachexia, Sarcopenia and Muscle*, *4*(1), 19–29. https://doi.org/10.1007/s13539-012-0079-1

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*(none). https://doi.org/10.1214/09-SS057

Pearl, J., & Paz, A. (2022). GRAPHOIDS: Graph-Based Logic for Reasoning about Relevance Relations Or When Would X Tell You More about Y If You Already Know Z? In H. Geffner, R. Dechter, & J. Y. Halpern (A c. Di), *Probabilistic and Causal Inference* (1ª ed., pp. 189–200). ACM. https://doi.org/10.1145/3501714.3501729

Pearson, F. S., Lipton, D. S., Cleland, C. M., & Yee, D. S. (2002). The Effects of Behavioral/Cognitive-Behavioral Programs on Recidivism. *Crime & Delinquency*, *48*(3), 476–496. https://doi.org/10.1177/0011128702048003006

Pearson, K. (1901). I. Mathematical contributions to the theory of evolution. —VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *195*(262–273), 1–47. https://doi.org/10.1098/rsta.1900.0022

Peter Spirtes, Clark Glymour, & Richard Scheines. (1993). *Causation, prediction, and search*. Springer.

Pinto, E. (2007). Blood pressure and ageing. *Postgraduate Medical Journal*, *83*(976), 109–114. https://doi.org/10.1136/pgmj.2006.048371

Ponikowski, P., Voors, A. A., Anker, S. D., Bueno, H., Cleland, J. G. F., Coats, A. J. S., Falk, V., González-Juanatey, J. R., Harjola, V.-P., Jankowska, E. A., Jessup, M., Linde, C., Nihoyannopoulos, P., Parissis, J. T., Pieske, B., Riley, J. P., Rosano, G. M. C., Ruilope, L. M., Ruschitzka, F., … Van Der Meer, P. (2016). 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *European Heart Journal*, *37*(27), 2129–2200. https://doi.org/10.1093/eurheartj/ehw128

Powers, W. J., Rabinstein, A. A., Ackerson, T., Adeoye, O. M., Bambakidis, N. C., Becker, K., Biller, J., Brown, M., Demaerschalk, B. M., Hoh, B., Jauch, E. C., Kidwell, C. S., Leslie-Mazwi, T. M., Ovbiagele, B., Scott, P. A., Sheth, K. N., Southerland, A. M., Summers, D. V., Tirschwell, D. L., & on behalf of the American Heart Association Stroke Council. (2019). Guidelines for the Early Management of Patients With Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. *Stroke*, *50*(12). https://doi.org/10.1161/STR.0000000000000211

R Core Team. (2023). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* [Software]. https://www.R-project.org/

Ramsey, J., Zhang, J., & Spirtes, P. L. (2012). *Adjacency-Faithfulness and Conservative Causal Inference* (Versione 1). arXiv. https://doi.org/10.48550/ARXIV.1206.6843

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Rule, A. D., Larson, T. S., Bergstralh, E. J., Slezak, J. M., Jacobsen, S. J., & Cosio, F. G. (2004). Using Serum Creatinine To Estimate Glomerular Filtration Rate: Accuracy in Good Health and in Chronic Kidney Disease. *Annals of Internal Medicine*, *141*(12), 929. https://doi.org/10.7326/0003-4819-141-12-200412210-00009

Saadat, P., Ahmadi Ahangar, A., Babaei, M., Kalantar, M., Bayani, M. A., Barzegar, H., Gholinia, H., Zahedi Tajrishi, F., Faraji, S., & Frajzadeh, F. (2018). Relationship of Serum Uric Acid Level with Demographic Features, Risk Factors, Severity, Prognosis, Serum Levels of Vitamin D, Calcium, and Magnesium in Stroke. *Stroke Research and Treatment*, *2018*, 1–8. https://doi.org/10.1155/2018/6580178

Sánchez-Lozada, L. G., Rodriguez-Iturbe, B., Kelley, E. E., Nakagawa, T., Madero, M., Feig, D. I., Borghi, C., Piani, F., Cara-Fuentes, G., Bjornstad, P., Lanaspa, M. A., & Johnson, R. J. (2020). Uric Acid and Hypertension: An Update With Recommendations. *American Journal of Hypertension*, *33*(7), 583–594. https://doi.org/10.1093/ajh/hpaa044

Sánchez-Lozada, L. G., Soto, V., Tapia, E., Avila-Casado, C., Sautin, Y. Y., Nakagawa, T., Franco, M., Rodríguez-Iturbe, B., & Johnson, R. J. (2008). Role of oxidative stress in the renal abnormalities induced by experimental hyperuricemia. *American Journal of Physiology-Renal Physiology*, *295*(4), F1134–F1141. https://doi.org/10.1152/ajprenal.00104.2008

Sarah Lewington & Robert Clarke. (2002). Age-specific relevance of usual blood pressure to vascular mortality: A meta-analysis of individual data for one million adults in 61 prospective studies. *The Lancet*, *360*(9349), 1903–1913. https://doi.org/10.1016/S0140-6736(02)11911-8

Satorra, A., & Bentler, P. M. (1994). *Corrections to Test Statistics and Standard Errors in Co variance Structure Analysis.* 399–419.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

Sharaf El Din, U. A. A., Salem, M. M., & Abdulazim, D. O. (2017). Uric acid in the pathogenesis of metabolic, renal, and cardiovascular diseases: A review. *Journal of Advanced Research*, *8*(5), 537–548. https://doi.org/10.1016/j.jare.2016.11.004

SHEP Cooperative Research Group. (1991). Prevention of Stroke by Antihypertensive Drug Treatment in Older Persons With Isolated Systolic Hypertension: Final Results of the Systolic Hypertension in the Elderly

Program (SHEP). *JAMA*, *265*(24), 3255. https://doi.org/10.1001/jama.1991.03460240051027

Siu, Y.-P., Leung, K.-T., Tong, M. K.-H., & Kwan, T.-H. (2006). Use of Allopurinol in Slowing the Progression of Renal Disease Through Its Ability to Lower Serum Uric Acid Level. *American Journal of Kidney Diseases*, *47*(1), 51–59. https://doi.org/10.1053/j.ajkd.2005.10.006

Spirtes, P., Glymour, C., & Scheines, R. (1995). *Causation, Prediction, and Search* (Vol. 81). Springer New York. https://doi.org/10.1007/978-1-4612-2748-9

Spirtes, P., & Zhang, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, *3*(1), 3. https://doi.org/10.1186/s40535-016-0018-x

Steffen Lauritzen. (1996). *Graphical Models*. Oxford University PressOxford. https://doi.org/10.1093/oso/9780198522195.001.0001

Stevens, B., Allen, N. J., Vazquez, L. E., Howell, G. R., Christopherson, K. S., Nouri, N., Micheva, K. D., Mehalow, A. K., Huberman, A. D., Stafford, B., Sher, A., Litke, A. M., Lambris, J. D., Smith, S. J., John, S. W. M., & Barres, B. A. (2007). The Classical Complement Cascade Mediates CNS Synapse Elimination. *Cell*, *131*(6), 1164–1178. https://doi.org/10.1016/j.cell.2007.10.036

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6. ed., international ed). Pearson.

Thygesen, K., Alpert, J. S., Jaffe, A. S., Chaitman, B. R., Bax, J. J., Morrow, D. A., White, H. D., & The Executive Group on behalf of the Joint European Society of Cardiology (ESC)/American College of Cardiology (ACC)/American Heart Association (AHA)/World Heart Federation (WHF) Task Force for the Universal Definition of Myocardial Infarction. (2018). Fourth Universal Definition of Myocardial Infarction (2018). *Circulation*, *138*(20). https://doi.org/10.1161/CIR.0000000000000617

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, *65*(1), 31–78. https://doi.org/10.1007/s10994-006-6889-7

Tuttle, K. R., Short, R. A., & Johnson, R. J. (2001). Sex differences in uric acid and risk factors for coronary artery disease. *The American Journal of Cardiology*, *87*(12), 1411–1414. https://doi.org/10.1016/S0002-9149(01)01566-1

US Preventive Services Task Force, Krist, A. H., Davidson, K. W., Mangione, C. M., Cabana, M., Caughey, A. B., Davis, E. M., Donahue, K. E., Doubeni, C. A., Kubik, M., Li, L., Ogedegbe, G., Pbert, L., Silverstein, M., Stevermer, J., Tseng, C.-W., & Wong, J. B. (2021). Screening for Hypertension in Adults: US Preventive Services Task Force Reaffirmation Recommendation Statement. *JAMA*, *325*(16), 1650. https://doi.org/10.1001/jama.2021.4987

Virdis, A., Masi, S., Casiglia, E., Tikhonoff, V., Cicero, A. F. G., Ungar, A., Rivasi, G., Salvetti, M., Barbagallo, C. M., Bombelli, M., Dell'Oro, R., Bruno, B., Lippa, L., D'Elia, L., Verdecchia, P., Mallamaci, F., Cirillo, M., Rattazzi, M., Cirillo, P., … from the Working Group on Uric Acid and Cardiovascular Risk of the Italian Society of Hypertension. (2020). Identification of the Uric Acid Thresholds Predicting an Increased Total and Cardiovascular Mortality Over 20 Years. *Hypertension*, *75*(2), 302–308. https://doi.org/10.1161/HYPERTENSIONAHA.119.13643

Wang, Y., O'Neil, A., Jiao, Y., Wang, L., Huang, J., Lan, Y., Zhu, Y., & Yu, C. (2019). Sex differences in the association between diabetes and risk of cardiovascular disease, cancer, and all-cause and cause-specific mortality: A systematic review and meta-analysis of 5,162,654 participants. *BMC Medicine*, *17*(1), 136. https://doi.org/10.1186/s12916-019-1355-0

Weiner, D. E., Tighiouart, H., Amin, M. G., Stark, P. C., MacLeod, B., Griffith, J. L., Salem, D. N., Levey, A. S., & Sarnak, M. J. (2004). Chronic Kidney Disease as a Risk Factor for Cardiovascular Disease and All-Cause Mortality: A Pooled Analysis of Community-Based Studies. *Journal of the American Society of Nephrology*, *15*(5), 1307–1315. https://doi.org/10.1097/01.ASN.0000123691.46138.E2

Whelton, P. K., Carey, R. M., Aronow, W. S., Casey, D. E., Collins, K. J., Dennison Himmelfarb, C., DePalma, S. M., Gidding, S., Jamerson, K. A., Jones, D. W., MacLaughlin, E. J., Muntner, P., Ovbiagele, B., Smith, S. C., Spencer, C. C., Stafford, R. S., Taler, S. J., Thomas, R. J., Williams, K. A., … Wright, J. T. (2018a). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults. *Journal of the American College of Cardiology*, *71*(19), e127–e248. https://doi.org/10.1016/j.jacc.2017.11.006

Whelton, P. K., Carey, R. M., Aronow, W. S., Casey, D. E., Collins, K. J., Dennison Himmelfarb, C., DePalma, S. M., Gidding, S., Jamerson, K. A., Jones, D. W., MacLaughlin, E. J., Muntner, P., Ovbiagele, B., Smith, S. C., Spencer, C. C., Stafford, R. S., Taler, S. J., Thomas, R. J., Williams, K. A., … Wright, J. T. (2018b). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Hypertension*, *71*(6). https://doi.org/10.1161/HYP.0000000000000065

Williams, B., Mancia, G., Spiering, W., Agabiti Rosei, E., Azizi, M., Burnier, M., Clement, D. L., Coca, A., De Simone, G., Dominiczak, A., Kahan, T., Mahfoud, F., Redon, J., Ruilope, L., Zanchetti, A., Kerins, M., Kjeldsen, S. E., Kreutz, R., Laurent, S., … Brady, A. (2018). 2018 ESC/ESH Guidelines for the management of arterial hypertension. *European Heart Journal*, *39*(33), 3021–3104. https://doi.org/10.1093/eurheartj/ehy339

Xu, Y., Xu, K., Bai, J., Liu, Y., Yu, R., Liu, C., Shen, C., & Wu, X. (2016). Elevation of serum uric acid and incidence of type 2 diabetes: A systematic review and meta-analysis. *Chronic Diseases and Translational Medicine*, *2*(2), 81–91. https://doi.org/10.1016/j.cdtm.2016.09.003

Yuan, K.-H., & Bentler, P. M. (2000). 5. Three Likelihood-Based Methods for Mean and Covariance Structure Analysis with Nonnormal Missing Data. *Sociological Methodology*, *30*(1), 165–200. https://doi.org/10.1111/0081-1750.00078

Yule. (1900). VII. On the association of attributes in statistics: With illustrations from the material of the childhood society, &c. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *194*(252–261), 257–319. https://doi.org/10.1098/rsta.1900.0019

Zanga, A., Ozkirimli, E., & Stella, F. (2022). A Survey on Causal Discovery: Theory and Practice. *International Journal of Approximate Reasoning*, *151*, 101–129. https://doi.org/10.1016/j.ijar.2022.09.004