Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per l'Economia e l'Impresa

Relazione Finale

Uno studio di meta-analisi per lo screening dei disturbi alimentari

Relatore: Prof.ssa Annamaria Guolo
Dipartimento di Scienze Statistiche

Laureando: Silvia Chiurchiù
Matricola 2015829

Anno Accademico 2022/2023

Indice

Introduzione

I disturbi alimentari rappresentano un complesso insieme di condizioni psicologiche e comportamentali che influenzano in modo significativo il rapporto di un individuo con il cibo, il suo corpo e la sua salute. Questi disturbi comprendono una serie di condizioni ben note, tra cui l'anoressia nervosa, la bulimia nervosa e il disturbo da alimentazione incontrollata, ed altre meno note riconosciute solo recentemente. Una serie di fattori negli ultimi decenni ha contribuito negativamente sulla loro prevalenza, tra cui una crescente esposizione ai social media ed alle pressioni sociali legate all'immagine corporea, l'accesso facilitato ad informazioni online che possono promuovere comportamenti alimentari disordinati, e ultimamente la pandemia da Covid-19. I disordini alimentari costituiscono così un preoccupante fenomeno in crescita che interessa la salute pubblica e rappresenta una sfida significativa per i professionisti della salute mentale e medica. Infatti, un ostacolo aggiuntivo risiede nella difficoltà del diagnosticarli, spesso chi ne soffre tende a nasconderne i sintomi e a minimizzare la gravità problema. Una prima valutazione può essere effettuata con l'utilizzo di test di screening, come la somministrazione di questionari che mirano a rilevare eventuali presenze di disturbi del comportamento alimentare. Questi test rappresentano uno strumento rapido e facilmente accessibile che può aiutare nella diagnosi precoce e nell'intervenire tempestivamente con un trattamento adeguato.

L'obiettivo di questo elaborato è di valutare la validità del questionario SCOFF, un test di screening composto da 5 domande e proposto nel 1999 da John F. Morgan e colleghi. Per valutarne l'accuratezza si procederà con uno studio di meta-analisi, trattando il caso specifico delle tecniche sviluppate per considerare congiuntamente le due principali misure di validità di un test diagnostico: la sensibilità e la specificità. Brevemente, la meta-analisi è una tecnica statistica che combinando i risultati derivanti da studi diversi condotti su una stessa domanda di interesse, fornisce un risultato robusto e generalizzabile che integra quanto emerso dai singoli studi.

L'analisi che segue riprende gli studi selezionati nello studio di Kutz et al. (2020),

si applicheranno le tecniche della meta-analisi che verranno precedentemente descritte. In particolare si tratterà il caso dei modelli bivariati ad effetti misti, che permettono di considerare insieme entrambe le misure utilizzate per valutare le caratteristiche dei test di screening. Infine si valuterà con la meta-regressione il possibile ruolo che possono avere altre variabili, come l'età, nell'accuratezza del test e in modo da approfondire le cause di eterogeneità tra gli studi. Le analisi verranno effettuate utilizzando il linguaggio di programmazione R (R Core Team, 2022).

Una nota, il colore lilla sarà ricorrente nelle figure di questa tesi non per una scelta di preferenza personale, ma come riferimento al simbolo della lotta contro i disturbi alimentari, un fiocchetto lilla. Un pensiero di forza e vicinanza per chi sta lottando contro questa malattia.

Capitolo 1

I disturbi alimentari

1.1 Caratteristiche di un fenomeno in crescita

L'Organizzazione Mondiale della Sanità categorizza i disturbi alimentari come una forma di disturbo mentale, ritenendo che essi consistano in un comportamento alimentare anomalo che si manifesta con preoccupazioni per il cibo, accompagnate nella maggior parte dei casi anche da preoccupazione per il peso e la forma del corpo (World Health Organization, 2022).

I disturbi del comportamento alimentare sono un fenomeno in crescita, l'incremento negli anni dei soggetti che sperimentano malattie come anoressia nervosa (AN), bulimia nervosa (BN) e disturbo da alimentazione incontrollata (BED) fa emergere una situazione allarmante. I dati raccolti dall'indagine GBD 2019 (Global Burden of Disease Study, 2020) evidenziano come dal 1990 al 2019, a livello mondiale, si è riscontrato un aumento di circa il 60% dei soggetti con AN o BN, come si può vedere dalla Figura 1.1. E' da considerare inoltre che i dati rilevati dall'indagine GBD 2019 non tengono conto dei casi di disturbo da alimentazione incontrollata, e di altre tipologie di disturbi alimentari, ma esclusivamente di AN e BN. Ampliando le tipologie di disturbi considerati, i numeri potrebbero raggiungere i 55.5 milioni di soggetti rispetto ai 13.6 rilevati dall'indagine (Santomauro et al., 2021).

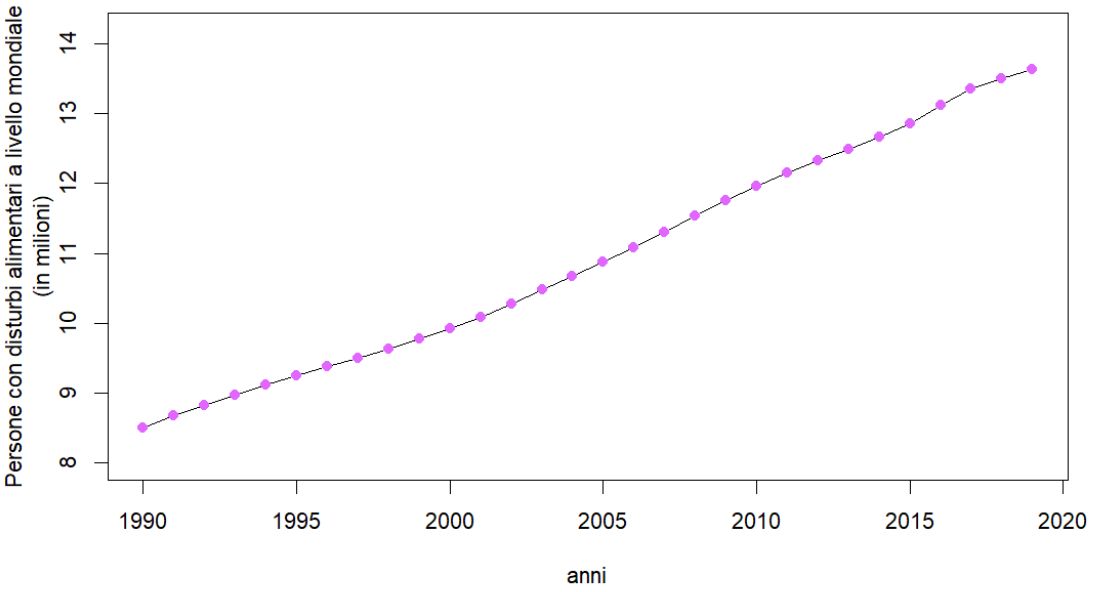


FIGURA 1.1: Aumento dei soggetti che soffrono di disturbi alimentari, a livello mondiale, dal 1990 al 2019.
Fonte dati: Global Burden of Disease Study 2019 (GBD 2019) Results.

Dal 2020, un peggioramento della situazione si è verificato con lo sviluppo della pandemia da Covid-19. Secondo uno studio condotto principalmente negli USA su soggetti di età giovane e inferiore ai 30 anni, è stato rilevato un aumento dei disturbi alimentari del 15.3% nel 2020 rispetto all'anno precedente (Taquet et al., 2022). In un contesto di maggiore stress, ansia e solitudine, tra le cause che hanno contribuito ad aumentare il rischio di incorrere in disordini alimentari hanno avuto ruolo fondamentale proprio l'aumento dello stress e l'isolamento sociale, insieme al maggior tempo trascorso in casa, alle limitazioni dell'attività fisica all'aperto, e alla diminuzione del supporto sociale (Rodgers et al., 2020). La sfera emotiva e psicologica è infatti di primaria importanza nel comprendere lo sviluppo di questi disturbi.

I disordini alimentari possono svilupparsi in risposta a diversi fattori di rischio, tra questi troviamo le influenze familiari, come la presenza di parenti con disturbi simili o problemi di depressione e abuso di sostanze. Altri principi di causa della malattia possono essere esperienze negative vissute, come episodi di abusi fisici o psicologici, traumi, carico delle pressioni sociali per mantenere un certo aspetto fisico, e tratti di personalità come l'ansia e il perfezionismo, bassi livelli di autostima ed elevato stress. Oltre ai fattori specifici dell'individuo e alla propria storia familiare, fattori generici

come il sesso e l'età comportano un maggiore rischio di presentare disordini alimentari, essi sono infatti più frequenti tra le donne e gli adolescenti. Le pressioni e i canoni di bellezza imposti dalla società, amplificati dalla diffusione dei social media, esercitano una forte pressione sulle donne, spingendole a cercare di adattarsi a corpi ideali spesso caratterizzati da magrezza eccessiva. Dalla Figura 1.2 si può notare il gap di genere nella crescita dei casi di disturbi alimentari rilevati dall'indagine GBD 2019. Mentre gli adolescenti sono maggiormente esposti al rischio dal momento che l'età adolescenziale segna una fase di transizione. Tale fase è caratterizzata da cambiamenti ormonali che comportano modifiche del proprio corpo e del fabbisogno energetico richiesto, andando a determinare possibili difficoltà per i ragazzi nel gestire il proprio consumo alimentare. L'adolescenza è inoltre un periodo delicato ed il cibo, o la privazione di esso, può rappresentare un modo con cui rispondere alle sfide che si incontrano. In questa fase possono quindi svilupparsi comportamenti disfunzionali che potrebbero in seguito concretizzarsi in disturbi alimentari.

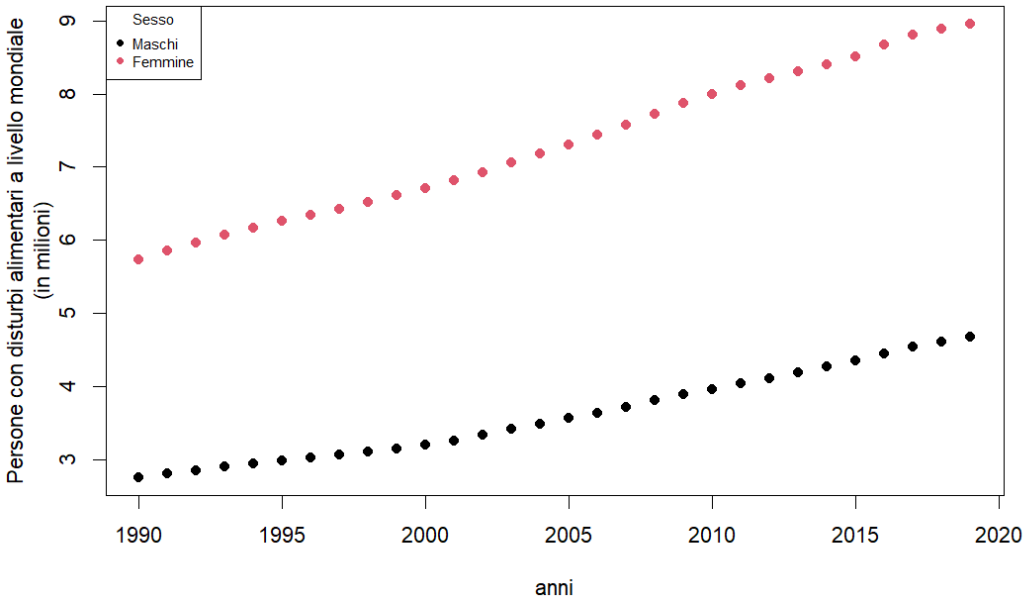


FIGURA 1.2: Differenza di genere nell'aumento dei soggetti che soffrono di disturbi alimentari, a livello mondiale, dal 1990 al 2019.
Fonte dati: Global Burden of Disease Study 2019 (GBD 2019) Results.

Secondo l'Organizzazione Mondiale della Sanità i disturbi alimentari sono la seconda causa di morte, dopo gli incidenti stradali, per le ragazze tra i 12 e i 25 anni. Solo in Italia ne soffrono oltre 3 milioni di persone, su 70 milioni totali stimati nel mondo, il 70% sono adolescenti, di cui il 95,9% sono donne. La situazione, come riportano i dati, è preoccupante e cresce sempre di più l'importanza di sensibilizzare sul tema dei

disturbi alimentari, in modo tale da accrescere la consapevolezza generale del problema, e soprattutto promuoverne la prevenzione, è fondamentale infatti aiutare chi ne soffre ad aprirsi e parlare delle proprie difficoltà, per favorire l'eventuale rilevazione della malattia. La difficoltà nel diagnosticare i disturbi alimentari rappresenta una sfida significativa nel campo della salute mentale. I sintomi dei disturbi alimentari non sempre sono evidenti e spesso i soggetti tendono a nasconderne i segnali, l'atteggiamento di negazione riguardo la gravità della loro condizione rappresenta un ulteriore ostacolo alla diagnosi precoce e al trattamento.

1.2 I metodi di screening e il questionario SCOFF

La diagnosi può avvenire attraverso diversi approcci di valutazione da parte dei professionisti della salute mentale e degli specialisti in disturbi alimentari. Psicologi o psichiatri effettuano una valutazione psicologica per rilevare sintomi correlati ai disordini alimentari come ansia, depressione, ossessione per il cibo ed il peso. Controlli fisici come misurazione del peso e dell'altezza consentono di calcolare l'indice di massa corporea (BMI), questo indice può aiutare a identificare il sottopeso o il sovrappeso, mentre gli esami del sangue, tra cui la valutazione dei livelli di elettroliti, zucchero nel sangue, funzione epatica, renale e ormoni, possono rivelare squilibri chimici causati dai disturbi alimentari.

Per lo screening di questi disturbi è inoltre ampiamente utilizzata la somministrazione di questionari per raccogliere informazioni sui comportamenti alimentari e sulle preoccupazioni legate al peso e all'immagine corporea. Esistono vari tipi di questionari, tra i più utilizzati troviamo l'EAT-26 (Eating Attitude Test) composto da 26 domande la cui risposta è una frequenza che va da "mai", associato a 0 punti, a "sempre", associato a 3 punti, un punteggio almeno pari a 20 suggerisce il rischio di avere un disturbo dell'alimentazione. L'EDE-Q (Eating Disorder Examination Questionnaire), disponibile anche in una versione specifica per gli adolescenti ed una per la compilazione da parte dei genitori, è invece un questionario auto-somministrato di 28 domande che valuta la frequenza e la gravità di comportamenti tipici dei disturbi alimentari, più elevato è il punteggio e maggiore è la gravità del possibile disturbo. Esistono molti altri questionari e varie versioni degli stessi, in questo elaborato si analizzerà nello specifico il questionario SCOFF.

Presentato nel 1999 dallo Psichiatra inglese John F. Morgan e colleghi, si articola nelle seguenti domande (Morgan et al., 1999):

*1) "Do you make yourself **Sick** because you feel uncomfortably full?"*

*2) "Do you worry you have lost **Control** over how much you eat?"*

*3) "Have you recently lost **One** stone (14 lb) in a 3-month period?"*

*4) "Do you believe yourself to be **Fat** when others say you are too thin?"*

*5) "Would you say that **Food** dominates your life?"*

Viene assegnato un punto per ogni risposta affermativa, un punteggio ≥ 2 indica un possibile caso di anoressia o bulimia nervosa.

E' rilevante tenere presente come negli anni nuove categorie di disturbi alimentari sono state prese in considerazione, l'ultima versione del *Manuale diagnostico e statistico dei disturbi mentali* (DSM-5) presenta nuove categorie di DCA rispetto alle tre principali riportate nella versione DSM-4: anoressia nervosa (AN), bulimia nervosa (BN) e disturbo alimentare non altrimenti specificato (EDNOS). Il test SCOFF, sviluppato negli anni della versione DSM-4, si concentra su AN e BN ciononostante è uno strumento ampiamente utilizzato per rilevare una possibile situazione di disordine alimentare. Gli stessi autori del questionario specificano tuttavia come tali domande siano solo un primo strumento di screening, che risulta semplice, rapido e utilizzabile anche dai non specialisti. Essi affermano infatti che il questionario ha il ruolo di far emergere il sospetto della presenza di un disturbo alimentare, deve seguire poi un'approfondita valutazione clinica da parte di uno specialista, che ne valuterà la conferma diagnostica.

1.3 Valutazione dei metodi di screening tramite meta-analisi

La valutazione dell'efficacia del questionario SCOFF nel rilevare in modo accurato i disturbi alimentari tra i pazienti della popolazione generale è l'obiettivo dello studio presentato da Kutz et al. (2020). La tecnica utilizzata nello studio per verificare la validità del questionario è la meta-analisi, una tecnica statistica che si approfondirà in questa tesi sia dal punto di vista teorico che applicativo. Lo studio di Kutz et al. dopo aver eseguito una particolare versione di revisione sistematica specifica per i test diagnostici, le cui linee guida sono presentate in (McInnes et al., 2018), ha selezionato 25 studi internazionali. I parametri di interesse sono la sensibilità e la specificità, misure di riferimento nella valutazione dell'accuratezza dei test diagnostici. Ogni studio presenta inoltre altre variabili che descrivono caratteristiche dello studio stesso e dei soggetti che lo compongono. Dall'analisi svolta è emerso che la validità del test SCOFF si dimostra

elevata negli studi campionati e si sostiene l'utilità del questionario come strumento di screening, in particolare per la rilevazione di AN e BN in giovani donne (Kutz et al., 2020).

L'analisi sviluppata in questa tesi ha l'obiettivo di valutare l'accuratezza del questionario SCOFF, lavorando sugli stessi dati di Kutz et al. e approfondendo gli aspetti statistici della meta-analisi, in particolare quella per i test diagnostici. I metodi utilizzati possono però essere applicati a qualsiasi altro strumento di screening per la rilevazione dei disturbi alimentari, e più in generale a test diagnostici di vario tipo.

Capitolo 2

Meta-analisi e test diagnostici

2.1 Introduzione alla meta-analisi

"*La meta-analisi si riferisce all'analisi delle analisi*" (Glass, 1976). Con queste parole Gene V. Glass, statistico e ricercatore statunitense pioniere dell'approccio meta-analitico, presenta la meta-analisi nel 1976. Glass introduce il termine meta-analisi per riferirsi "*all'analisi statistica di un'ampia raccolta di risultati di analisi provenienti da singoli studi allo scopo di integrare i risultati*". La vasta e crescente quantità di articoli e studi disponibili in letteratura su uno stesso argomento, porta alla necessità di sviluppare una tecnica che consenta di analizzare insieme le evidenze disponibili, per arrivare a risultati di sintesi generali ed accurati. Ogni studio primario ha le proprie caratteristiche, che ne condizionano i risultati. Gli studi si differenziano per diversi aspetti, tra cui la diversa popolazione di riferimento e la numerosità del campione, il disegno dello studio, gli strumenti e le metodologie utilizzate nel condurre l'analisi. La meta-analisi si rende quindi necessaria per ottenere una misura che sintetizzi quanto ottenuto dai singoli e differenti studi condotti in modo indipendente, così da poter arrivare ad avere risultati robusti e generali in merito ad uno stesso fenomeno di interesse.

Il primo passo è quello di formulare la domanda di ricerca, così da condurre le analisi con l'obiettivo di rispondere al preciso quesito di interesse. Formulare la domanda in modo adeguato condurrà a una corretta progettazione dello studio, a una strategia di ricerca letteraria appropriata e a un'analisi statistica che produrrà le evidenze necessarie per orientare le decisioni pratiche (Gogtay & Thatte, 2017). La meta-analisi è poi preceduta da un'accurata revisione sistematica della letteratura al fine di individuare tutti gli studi potenzialmente utili per rispondere alla domanda di ricerca che guida l'analisi

(Bioletto et al., 2022). La ricerca si effettua principalmente ricorrendo a database specifici come ad esempio PubMed e Cochrane, utilizzati nello studio di Kutz et al. (2020). Il vantaggio della revisione sistematica è quello di essere generalmente più rapida e meno costosa rispetto alla creazione di un nuovo studio (Mulrow, 1994). Il processo di selezione degli studi può contribuire ad introdurre errore sistematico nella meta-analisi, attraverso la *selection bias* e la *publication bias*. La *selection bias* è collegata ai criteri di inclusione degli studi, ad esempio nel contesto epidemiologico si verifica quando esiste una differenza sistematica tra le caratteristiche delle persone incluse nello studio e quelle invece non considerate. La selezione casuale rappresenta un modo efficace per ridurla (Henderson & Page, 2007). La *publication bias* può manifestarsi in quanto gli studi con risultati negativi o non significativi vengono pubblicati in misura minore rispetto a quelli che presentano risultati positivi, questo bias si traduce nel rischio di sovrastimare l'effetto positivo.

Selezionati gli studi da includere nell'analisi, per ognuno di essi si estrae la stima del parametro che misura l'effetto di interesse, ed una misura della precisione associata alla stima. L'*effect-size*, così viene generalmente nominata la misura dell'effetto di interesse, può consistere in una differenza tra medie standardizzate, una proporzione, un rapporto di probabilità (odds ratio), un rapporto di rischio (risk ratio), o altre misure opportune a rilevare quanto si vuole analizzare. Tramite la meta-analisi gli *effect-size* dei vari studi vengono combinati per giungere ad una misura riassuntiva dell'effetto totale, una stima ottenuta dal contributo dei risultati di ciascuno studio incluso nell'analisi. L'effetto totale viene ottenuto come media pesata degli *effect-size* dei singoli studi, i pesi utilizzati dipendono dal modello che si sta considerando. Sono due i modelli statistici che vengono utilizzati per combinare gli *effect-size* degli studi primari: il modello ad effetti fissi ed il modello ad effetti casuali. La selezione del modello più appropriato tra i due viene stabilita analizzando la possibile presenza di eterogeneità tra gli studi considerati, diverse statistiche e quantità permettono di quantificarne l'entità.

I due modelli a effetti fissi e ad effetti casuali si distinguono per le assunzioni fatte sulla natura degli studi e sui veri *effect-size* di tali studi. Il principale obiettivo della meta-analisi è infatti quello di fare inferenza su un vero parametro μ, che rappresenta l'effetto totale ottenuto combinando i risultati dei vari studi in analisi. Nel modello ad effetti fissi si assume che tutti gli studi condividano lo stesso vero *effect-size* μ_i pari a μ, e che le differenze osservate tra le misure siano dovute ad errori causali, come l'errore di campionamento (Borenstein et al., 2010). Tale modello assume quindi l'omogeneità degli studi inclusi nella meta-analisi. Quando invece gli studi presentano eterogeneità tra loro, e quindi il vero *effect-size* è diverso per ogni studio, risulterà opportuno ricorrere

al modello ad effetti casuali, che consente di ottenere l'effetto totale μ considerando che il vero valore μ_i differisce in ogni studio.

2.1.1 Modello ad effetti fissi

Si considerino N studi inclusi nella meta-analisi a seguito della revisione sistematica della letteratura, che seleziona quelli validi e di interesse per rispondere alla domanda di ricerca. Con Y_i, $i = 1, ..., N$ si indicano le variabili che generano i valori osservati y_i dell'*effect-size* per ogni studio i-esimo. Il vero valore del parametro che misura l'effetto di interesse in ogni studio viene rappresentato da μ_i, $i = 1, ..., N$.

Il modello ad effetti fissi (FE) viene chiamato anche modello ad effetti comuni (CE), questo per sottolineare che l'*effect-size* μ è comune a tutti gli studi, ovvero $\mu = \mu_1 = ... = \mu_N$. Gli effetti osservati y_i saranno distribuiti tutti attorno alla stesso stesso valore μ, con una varianza σ_i^2 che dipende principalmente dalla dimensione campionaria dei singoli studi (Borenstein et al., 2007). Il modello lineare ad effetti fissi (o ad effetti comuni) assume la forma

$$Y_i = \mu + \epsilon_i \ , \quad \text{con } \epsilon_i \sim N(0, \sigma_i^2), \tag{2.1}$$

dove i termini di errore ϵ_i sono indipendenti e σ_i^2 misura la varianza interna agli studi, *within-study*, unica fonte di variabilità prevista da questo modello, attribuita all'errore casuale che discosta le stime dal vero valore μ.

La stima dell'effetto totale di interesse si ottiene come

$$\hat{\mu} = \frac{\sum_{i=1}^{N} \omega_i y_i}{\sum_{i=1}^{N} \omega_i} \ , \qquad \text{con } \ \omega_i = \frac{1}{\hat{\sigma}_i^2}.$$

I pesi ω_i che entrano nella media pesata sono pari all'inverso della varianza *within-study* stimata nell'i-esimo studio. In questo modo viene associato un peso maggiore agli studi che presentano una varianza ridotta, e quindi una stima di μ più precisa. I pesi così ottenuti sollevano però un limite nel modello CE. Infatti assumendo che l'effetto sia comune a tutti gli studi, i pesi si basano esclusivamente sulle informazioni che ciascuno di essi acquisisce, allora se uno studio è di grandi dimensioni avrà la maggior parte del peso, mentre se ridotto potrebbe essere pressoché ignorato. Con il modello RE gli studi di grandi dimensioni possono sì fornire stime più precise rispetto ai piccoli studi, ma ogni studio stima una diversa dimensione dell'effetto, ognuno di essi rappresenta un campione della popolazione da cui vogliamo stimare la media μ, tenendo conto di ciò i pesi previsti dal modello RE sono più equilibrati (Borenstein et al., 2007).

2.1.2 Modello ad effetti casuali

Nella realtà non c'è generalmente motivo di assumere che gli studi inclusi nella meta-analisi possano essere identici in modo tale da poter ritenere che l'*effect-size* sia lo stesso in tutti loro (Borenstein et al., 2010). Gli studi possono differire per diversi aspetti, come le caratteristiche di partecipanti, le modalità con cui viene condotto, gli strumenti di rilevazione dell'effetto di interesse, e altri. Se presente, l'eterogeneità tra gli studi può essere considerata ricorrendo all'utilizzo del modello ad effetti casuali. Il modello, proposto da DerSimonian e Laird nel 1986, è specificato come

$$Y_i = \mu_i + \epsilon_i \ , \quad \text{con} \ \ \mu_i = \mu + \delta_i.$$

In modo più compatto

$$Y_i = \mu + \delta_i + \epsilon_i, \tag{2.2}$$

con $\epsilon_i \sim N(0, \sigma_i^2)$ e $\delta_i \sim N(0, \tau^2)$, dove ϵ_i e δ_i sono indipendenti.

Questo approccio assume che esista una distribuzione per l'effetto di interesse e gli effetti osservati dai singoli studi vengono utilizzati per stimare la media μ di tale distribuzione (DerSimonian & Laird, 1986). Si considera che gli studi selezionati siano campionati da una popolazione di possibili studi con effetto medio del trattamento μ e varianza τ^2. L'*effect-size* y_i è l'osservazione ottenuta nell'i-esimo studio campionato, la distribuzione dell'effetto è caratterizzata dal vero valore μ_i, diverso per ogni studio. Il termine di errore δ_i rappresenta la differenza tra la media μ e il vero valore μ_i dello studio i-esimo. La varianza τ^2 del termine di errore δ_i misura l'eterogeneità presente tra gli studi, quantifica quindi la varianza *between-study*.

La stima dell'effetto totale si ottiene come

$$\hat{\mu} = \frac{\sum_{i=1}^{N} {\omega_i}^{\mathrm{RE}} y_i}{\sum_{i=1}^{N} {\omega_i}^{\mathrm{RE}}} \ , \qquad \text{con} \ \ {\omega_i}^{\mathrm{RE}} = \frac{1}{(\hat{\sigma}_i^2 + \hat{\tau}^2)}.$$

Per stimare la varianza τ^2 DerSimonian e Laird hanno proposto uno stimatore basato sul metodo dei momenti, la stima è ottenuta da

$$\hat{\tau}^2 = \frac{q - (N-1)}{\sum_{i=1}^{N} \omega_i - \frac{\sum_{i=1}^{N} \omega_i^2}{\sum_{i=1}^{N} \omega_i}}, \tag{2.3}$$

dove q è il valore della statistica Q di Cochran, che vedremo in seguito. Per i valori $\hat{\tau}^2 < 0$ viene assegnato un valore $\hat{\tau}^2 = 0$.

Altri numerosi stimatori sono stati proposti per stimare τ^2, nello specifico si rimanda

a Viechtbauer (2005), López-López et al. (2014), Sidik & Jonkman (2005), ad Hardy & Thompson (1996) per i metodi legati alla verosimiglianza.

2.1.3 Analisi dell'eterogeneità tra gli studi

La valutazione della presenza di eterogeneità tra gli studi costituisce una fase molto importante quando si lavora con la meta-analisi. L'importanza di tale valutazione risiede nel contribuire a determinare quale tra il modello ad effetti fissi e quello ad effetti casuali sia il più adeguato per condurre l'analisi. Negli anni sono state presentate varie misure che consentono di rilevare la diversità tra gli studi e quantificarne l'eterogeneità. Come vedremo, tali misure seppur utili presentano anche dei limiti. Esse non dovrebbero quindi rappresentare l'unico determinante nella scelta tra i due modelli, sono rilevanti sia per le indagini che per la modellazione anche l'analisi di grafici e l'approfondimento delle caratteristiche degli studi (Hardy & Thompson, 1998).

A livello pratico l'analisi è basata su τ^2, la presenza di significativa eterogeneità corrisponde a rifiutare l'ipotesi nulla del sistema d'ipotesi

$$\begin{cases} H_0 : \tau^2 = 0 \\ H_1 : \tau^2 > 0 \end{cases}$$

Nel 1954 in Cochran (1954) si introduce la statistica Q, che permette di risolvere il sistema d'ipotesi riportato sopra, ed è calcolata come

$$Q = \sum_{i=1}^{N} \omega_i (y_i - \hat{\mu})^2, \tag{2.4}$$

dove la stima $\hat{\mu}$ è quella ottenuta con il modello CE. La statistica Q sotto H_0 segue una distribuzione χ^2_{N-1}. Sono i valori elevati della statistica Q che portano a rifiutare l'ipotesi nulla, constatando la presenza di significativa differenza tra gli studi.

Un limite della statistica Q è lo scarso potere che ha nel rilevare l'eterogeneità quando il numero di studi inclusi nella meta-analisi è relativamente basso, potere che risulta invece eccessivo ed in grado di catturare anche differenze trascurabili tra gli studi quando il loro numero è elevato (Huedo-Medina et al., 2006).

Successivamente alla statistica Q sono state introdotte altre misure per valutare la presenza della varianza tra gli studi, con l'obiettivo di superare anche i limiti che presentava la Q. Le statistiche I^2 e H^2 proposte in Higgins & Thompson (2002) hanno trovato ampio utilizzo nella valutazione dell'eterogeneità in meta-analisi.

I^2 rappresenta un indice di eterogeneità che quantifica quanto della varianza totale è conferibile all'eterogeneità tra gli studi, calcolabile come

$$I^2 = \frac{q - (N-1)}{q}. \tag{2.5}$$

$I^2 \in [0, 1]$, e i valori di $I^2 < 0$ che possono essere ottenuti dal numeratore $q - (N-1)$ vengono posti pari a 0. Moltiplicando per 100 il valore I^2 se ne ottiene il valore in percentuale. Percentuali superiori al 75% suggeriscono alta eterogeneità secondo la classificazione proposta dagli stessi Higgins e Thompson.

H^2 è la statistica che descrive l'eterogeneità misurando l'eccesso del valore della statistica Q relativamente ai suoi gradi di libertà. Il valore di H^2 si ottiene infatti da

$$H^2 = \frac{q}{N-1}. \tag{2.6}$$

Se il valore H^2 ottenuto è pari ad 1 il suggerimento che tale statistica apporta è la non presenza di differenza sostanziale tra gli studi, mentre per $H^2 > 1$ l'eterogeneità è significativa tra gli studi.

H^2 e I^2 sono inoltre legate dalla relazione $I^2 = (H^2 - 1)/H^2$. Entrambe però continuano ad essere influenzate dal numero N di studi inclusi nell'analisi, con il rischio di giungere a conclusioni inesatte sulla presenza dell'eterogeneità.

Le misure descritte per identificare la variabilità tra gli studi sono quindi utili come guida nella valutazione dell'eterogeneità ma è importante considerare altri aspetti che possono aiutare a giungere a conclusioni più esatte. Per esplorare in modo più dettagliato l'eterogeneità si può ad esempio ricorrere alla meta-regressione o ad un analisi in sottogruppi, in modo da considerare caratteristiche più specifiche dei vari studi che possano aver portato ad una differenza significativa tra loro.

Nei successivi paragrafi di questo capitolo si introdurranno inizialmente i test diagnostici e le metodologie utilizzate per valutarne l'affidabilità, si presenteranno poi le tecniche di meta-analisi specificamente sviluppate per adattarsi alle peculiarità di tali test.

2.2 Test diagnostici e valutazione della loro accuratezza

Il test SCOFF è uno strumento di screening, ha quindi l'obiettivo di classificare correttamente come positivi al test i soggetti che presentano un disturbo alimentare. I test che hanno la funzione di classificare i soggetti in base al loro stato di salute, quindi stabilire la presenza o la non presenza di una determinata malattia, si definiscono test diagnostici. La validità di un test diagnostico è associata alla sua performance nel classificare correttamente i pazienti in base alla loro effettiva condizione di salute. Per valutare l'accuratezza del test sono disponibili diverse misure, in particolare si fa riferimento a due importanti indicatori della qualità del test diagnostico: la sensibilità e la specificità. Il diagnostic odds ratio, l'area sotto la curva ROC e l'indice di accuratezza rientrano anche loro tra le possibili misure per valutare la validità di uno strumento diagnostico.

Per condurre uno studio sull'accuratezza del test, i dati vengono comunemente rappresentati in una tabella a doppia entrata che include il numero di casi di corretta o errata classificazione del test, un esempio di tale struttura è riportato nella Tabella 2.1.

Tabella 2.1: Tabella di errata classificazione

		Status del paziente	
		Paziente malato	Paziente non malato
Risultato del test	Positivo	Veri Positivi (VP)	Falsi Positivi (FP)
	Negativo	Falsi Negativi (FN)	Veri Negativi (VN)
	Totale	n_1 = VP + FN	n_2 = FP + VN

Dalla tabella possiamo estrarre i dati per calcolare le misure di interesse, come l'indice di accuratezza del test, che misura la proporzione di diagnosi corrette, quantificando la corretta classificazione dei pazienti

$$accuratezza = \frac{VP+VN}{n},$$

dove $n = n_1+n_2$ è il numero totale dei pazienti sottoposti al test.

La specificità e la sensibilità costituiscono però i parametri di maggiore rilevanza, saranno infatti di primaria importanza anche nella meta-analisi che si andrà a condurre. La sensibilità misura la capacità del test di identificare come positivo un soggetto malato. Questa misura corrisponde al tasso di veri positivi (TPR), e in termini probabilistici identifica la probabilità di ottenere un test positivo dato che è stato effettuato su un soggetto malato

$$\textit{sensibilità} = Pr(\text{Test positivo} \mid \text{Paziente malato}),$$

e viene stimata come il tasso di veri positivi sui pazienti malati

$$TPR = \frac{VP}{VP+FN}.$$

La specificità misura la capacità del test di identificare come negativo un soggetto non malato, corrisponde al tasso di veri negativi (TNR) ed è la probabilità di ottenere un test negativo dato che è stato eseguito su un soggetto non malato

$$\textit{specificità} = Pr(\text{Test negativo} \mid \text{Paziente non malato}),$$

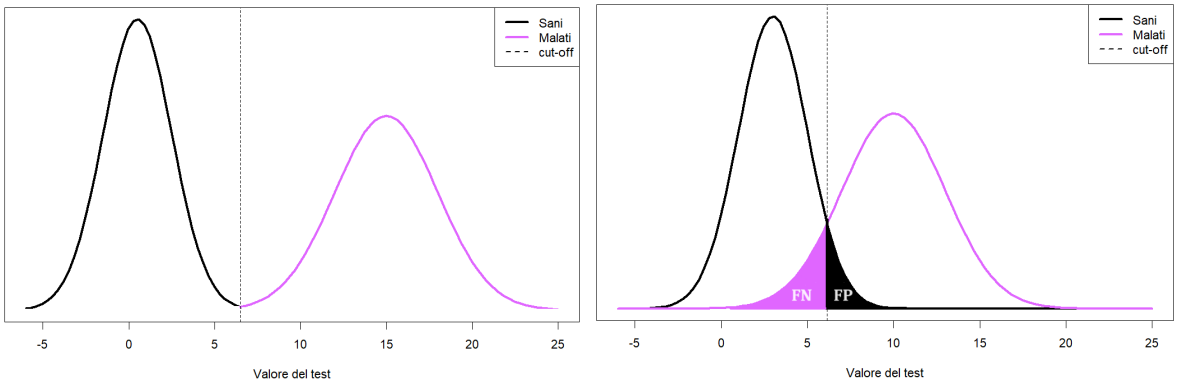
e viene stimata come il tasso di veri negativi sui pazienti non malati

$$TNR = \frac{VN}{VN+FP}.$$

I valori di sensibilità e specificità sono correlati, in genere negativamente, e dipendono dal valore di soglia scelto per differenziare tra risultato del test positivo e negativo. Variando il valore di cut-off varia la classificazione dei pazienti e con essa i valori delle due misure di accuratezza del test. Un aumento della specificità del test, che si ottiene abbassandone il valore di soglia, corrisponde ad una maggiore correttezza nel classificare i pazienti non malati come negativi al test, riducendo quindi il numero di falsi negativi. Ma tale aumento corrisponde anche ad un valore della sensibilità che diminuisce, incrementando il rischio di avere falsi positivi. Una diagnosi inaccurata, come quella che viene conferita ai falsi negativi o positivi, è un serio problema dato che il paziente rischia di essere sottoposto ad una terapia non necessaria o di ricevere in ritardo le cure opportune (Jackson, 2008).

Un test ideale avrebbe la capacità di identificare correttamente tutti i pazienti, non producendo quindi falsi positivi e falsi negativi, tale situazione è rappresentata nella parte sinistra della Figura 2.1, ma nella realtà quello che emerge è quanto si verifica nella parte destra. Il test porta in alcuni casi a conclusioni errate, infatti le distribuzioni si sovrappongono e considerando ad esempio un soggetto appartenente alla popolazione sana che effettua un test con risultato sopra la soglia, tale soggetto viene ritenuto positivo pur non essendolo.

Il diagnostic odds ratio verrà trattato nello specifico in seguito. Si descrive adesso un approccio grafico per valutare l'affidabilità di un test diagnostico, ovvero la curva ROC. La curva ROC (Receiver Operating Characteristic) permette di valutare le performance dei test diagnostici analizzando graficamente le relazioni tra sensibilità e tasso di falsi positivi (1-specificità) al variare dei valori di soglia che il test assume per discriminare tra positivi e negativi (Mandrekar, 2010). Come si è detto precedentemente, i valori



(a) Test ideale (b) Test reale

FIGURA 2.1: Distribuzione del test nei pazienti sani e malati e cut-off del test, a destra di tale soglia il risultato è negativo, a sinistra positivo. Esempio nel caso di test ideale e di test reale.

di sensibilità e specificità variano al variare del valore di cut-off scelto, la curva ROC raffigura le possibili coppie di sensibilità ed 1-specificità ottenibili dal test. La situazione ideale, cioè un test che abbia sensitività e specificità pari al 100%, si concretizza in una curva ROC che raggiunge il vertice nel punto (0,1) del piano. Al contrario, una curva ROC che combacia con la bisettrice del primo e terzo quadrante è associata ad un test che non ha la capacità di assegnare l'esito corrispondente al vero stato di salute dei soggetti, è un test che lavora assegnando casualmente l'esito ai pazienti. Nella realtà i test avranno una curva ROC compresa tra le due situazioni limite appena descritte, e più la curva è vicina al vertice in alto a sinistra maggiore sarà l'accuratezza del test.

La misura che traduce a livello numerico quanto emerge dalla curva è l'area sottesa alla curva ROC (AUC, *Area Under the ROC Curve*). Il valore che l'AUC può assumere varia tra 0 e 1, dove il valore centrale 0.5 indica un test non informativo, mentre 1 è associato al test ideale. Se l'AUC ricade tra 0.5 e 0.7 il test è ritenuto poco accurato, tra 0.7 e 0.9 il test è moderatamente accurato, tra 0.9 e 1 il test è altamente accurato (Swets, 1988).

Le curve ROC e i valori AUC risultano utili anche per confrontare diversi test diagnostici, in modo da rilevare quale sia il migliore nel classificare i pazienti che possono essere affetti da una determinata patologia. Un esempio del confronto tra due curve ROC, con i rispettivi valori AUC, è raffigurato in Figura 2.2. La curva ROC in lilla è consistentemente migliore della curva in nero, come indicato anche dai valori dell'AUC, con 0.95 che suggerisce che il test associato alla curva ROC lilla è altamente accurato.

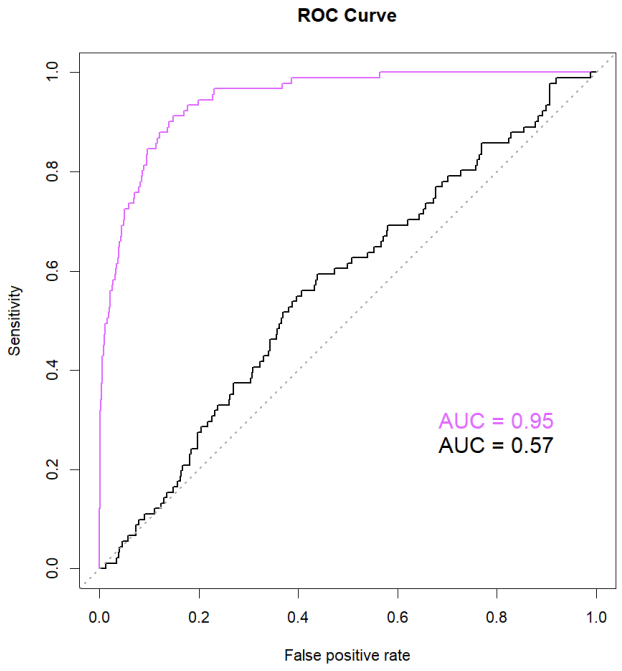


FIGURA 2.2: Confronto di due curve ROC

2.3 Modelli di meta-analisi per test diagnostici

Abbiamo discusso della meta-analisi come un'efficace strumento che, attraverso la combinazione dei risultati derivanti da un insieme di studi diversi e indipendenti condotti su un medesimo argomento, consente di giungere ad una stima più robusta per la misura di interesse ed estendibile ad una popolazione più generale rispetto a quella dei singoli studi. Nel contesto dei test diagnostici, la meta-analisi assume un ruolo fondamentale nel processo decisionale clinico e di politica sanitaria riguardo il loro utilizzo, ed è fondamentale inoltre per guidare il processo di sviluppo e valutazione della tecnologia in medicina diagnostica (Rutter & Gatsonis, 2001). Una delle applicazioni si trova nel valutare ad esempio un nuovo test, che spesso costituisce un'alternativa meno invasiva, e talvolta anche più economica, rispetto ai test di riferimento, noti per la loro accreditata validità.

La particolarità che caratterizza gli studi sui test diagnostici risiede nel fatto che ognuno riporta una coppia di statistiche riassuntive correlate, come sensibilità e specificità, questo rende necessario rivedere i metodi classici di meta-analisi con cui si analizzavano singoli *effect-size* (Deeks, 2001). Prima dei modelli bivariati proposti da Rutter & Gatsonis (2001) e Reitsma et al. (2005) erano di maggiore utilizzo approcci univariati per le meta-analisi di accuratezza diagnostica (Doebler & Holling, 2015). Lavorare con le due misure di interesse separatamente presenta il limite di non considerare la correlazione che generalmente si verifica tra di esse e che è dovuta all'effetto soglia,

cioè alla variazione di sensibilità e specificità in base alla scelta del valore di cut-off che discrimina tra positivi e negativi al test. La scelta del valore di soglia può avere un impatto significativo sulle stime delle prestazioni del test, ed inoltre una considerevole variabilità nei risultati può essere dovuta ai diversi valori adottati nei vari studi. Analizzare i parametri di sensibilità e specificità in modo distinto non permette inoltre di cogliere la naturale struttura bivariata dei risultati sull'accuratezza del test diagnostico.

Varie opzioni sono state proposte per la valutazione meta-analitica dei test diagnostici. Inizialmente si lavorava seguendo l'approccio univariato e valutando separatamente sensibilità e specificità, o combinandole in un'unica misura come il diagnostic odds ratio. Successivamente venne proposto in Littenberg & Moses (1993) un metodo per stimare una curva ROC globale (SROC, *Summary Receiver Operating Characteristics*) che sintetizzasse le evidenze dei vari studi. Tale metodo soffre però di diverse limitazioni, tra cui il fatto di basarsi su un modello ad effetti fissi. Per superare le mancanze riscontrate nei vari metodi, sono stati introdotti i modelli bivariati ad effetti casuali, (Reitsma et al. (2005); Arends et al. (2008)), il cui utilizzo è ad oggi fortemente raccomandato. Prima di questi, Rutter & Gatsonis (2001) svilupparono il modello HSROC (*Hierarchical Summary Receiver Operating Characteristic*), un modello gerarchico che includesse gli effetti casuali, tenendo conto dell'eterogeneità presente tra gli studi.

Di seguito si presenteranno nello specifico i vari modelli, passando dall'approccio univariato a quello bivariato.

2.3.1 Approccio univariato

Meta-analisi classica per le proporzioni

Per ottenere stime distinte per la sensibilità e la specificità dagli studi analizzati, è possibile utilizzare i metodi della meta-analisi classica, con delle modifiche applicate per gestire il caso specifico delle proporzioni.

Si consideri la stima della proporzione di interesse $\hat{p}_i = d_i/n_i$ con d_i numero di esposti al rischio sulla popolazione totale n_i, dove $\hat{p}_i$ può essere la proporzione di veri positivi o quella di veri negativi, cioè sensibilità o specificità. Come primo passo generalmente si applicano delle trasformazioni alle proporzioni grezze, in questo modo si ottengono risultati maggiormente conservativi (Shim, 2022). Le trasformazioni possibili sono il logaritmo, il logit, l'arcsine e il double-arcsine, ci si focalizza sul logit perché maggiormente utilizzata, ma l'utilizzo delle altre trasformazioni è chiaramente possibile.

La proporzione $\hat{p}_i$ viene trasformata come

$$logit(\hat{p}_i) = log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right), \tag{2.7}$$

la cui varianza è pari a

$$\hat{\nu}_i^2 = \frac{1}{d_i} + \frac{1}{(n_i - d_i)}. \tag{2.8}$$

Si stima la misura di interesse $\hat{\mu}$ applicando la meta-analisi classica sulle proporzioni trasformate, utilizzando quindi come pesi $w_i{=}1/\hat{\nu}_i^2$ nel caso del modello CE e $w_i{=}1/(\hat{\nu}_i^2{+}\tau^2)$ nel caso di quello RE. Una volta ottenuta la stima, $\hat{\mu}$ ed il relativo intervallo di confidenza vengono ri-trasformati in modo da poter interpretare i risultati nella stessa misura della proporzione iniziale. Avendo applicato la trasformazione logit, la stima viene riportata alla scala originale tramite la trasformazione $f^{\text{-}1}(\hat{\mu}){=}e^{\hat{\mu}}/(1{+}e^{\hat{\mu}})$.

Questo metodo soffre però di limitazioni significative, ad esempio le trasformazioni con l'utilizzo del logaritmo o del logit trattano le varianze *within-study* ${\nu_i}^2$ come fissate e note, questo incide sull'accuratezza della successiva inferenza statistica, anche i risultati delle trasformazioni basate sull'arcoseno potrebbero risultare poco chiari da interpretare, e correzioni specifiche devono essere applicate nel caso di valori pari a zero per evitare di avere quantità indefinite (Lin & Chu, 2020). Una valida alternativa è rappresentata dall'utilizzo di modelli lineari generalizzati, con funzione di legame logit, che possono poi diventare modelli misti una volta introdotti gli effetti casuali per includere l'eterogeneità tra gli studi. Questi modelli hanno il vantaggio di non utilizzare la varianza inversa come metodo per aggregare i risultati ed ottenere la stima totale.

Modello lineare generalizzato misto

Una particolare specificazione del modello lineare generalizzato permette di ricavare le stime per valutare univariatamente i valori di sensibilità e specificità. Nello specifico viene adattato un modello di regressione logistica, in cui è presente solo l'intercetta, ed in cui gli effetti casuali vengono poi inseriti per tenere conto dell'eterogeneità tra gli studi (Harrer et al., 2021), passando così ad un modello misto.

Definiamo il modello come segue

1.
$$d_i \sim Bi(n_i, p_i),$$

2.
$$logit(p_i) = log(p_i/(1-p_i)) = \mu, \tag{2.9}$$

3.
$$logit(p_i) = log(p_i/(1-p_i)) = \mu + \delta_i \quad \text{, con } \delta_i \sim \mathrm{N}(0, \tau^2).$$

Dove d_i sono il numero di *VP* nel caso in cui p_i rappresenti la sensibilità oppure sono il numero di *VN* quando p_i rappresenta la specificità. La stima di μ, misura dell'effetto totale, si ottiene tramite massima verosimiglianza. La specificazione 2. differisce dalla 3. poiché inizialmente si tiene conto solo delle variazioni casuali interne agli studi, mentre per incorporare la variabilità tra gli studi si aggiungono gli effetti casuali rappresentati dai termini δ_i, che seguono una distribuzione $N(0,\ \tau^2)$, dove τ^2 è la varianza stimata tra gli studi.

Come dimostra lo studio di Lin & Chu (2020), l'utilizzo della regressione logistica per effettuare una meta-analisi sulle proporzioni ha il vantaggio di basarsi su meno assunzioni, di tenere pienamente conto delle incertezze all'interno dello studio, e di avere generalmente performance migliori rispetto all'utilizzo dei modelli classici di meta-analisi adattati per le proporzioni.

Diagnostic odds ratio

Il diagnostic odds ratio (DOR) è un indicatore che consente di quantificare la performance di un test diagnostico in un'unica misura. Utilizzare delle coppie di indicatori, come sensibilità e specificità, può risultare svantaggioso soprattutto per confrontare le performance di test diversi, in particolare nel caso in cui un test non supera il concorrente su entrambi gli indicatori (Glas et al., 2003). Il DOR è definito come il rapporto tra l'odds di ottenere un test positivo nei soggetti malati e l'odds di ottenere un risultato positivo nei soggetti senza malattia, dove l'odds di un evento è il rapporto tra la probabilità che si verifichi ed il suo complemento a 1, ovvero la probabilità che non si verifichi.

$$DOR = \frac{Pr(Test\ positivo|Paziente\ malato)}{1-Pr(Test\ positivo|Paziente\ malato)} \Big/ \frac{Pr(Test\ positivo|Paziente\ non\ malato)}{1-Pr(Test\ positivo|Paziente\ non\ malato)},$$

e viene stimato come

$$\widehat{DOR} = \frac{\frac{VP}{FN}}{\frac{FP}{VN}} = \frac{sens/(1-sens)}{(1-spec)/spec}.$$

Più alto è il suo valore, migliore sarà la capacità del test di discriminare tra soggetti positivi e negativi. Sono però sinonimi di un test non accurato i valori del DOR ≤ 1, infatti se inferiore ad 1 suggerisce che il test fornisce un'errata interpretazione, come maggiori risultati negativi del test tra i pazienti malati, se pari ad 1 il test non è in grado di distinguere tra le due categorie di pazienti (Glas et al., 2003).

Nella meta-analisi la stima del DOR può essere ottenuta dai vari studi inclusi nell'analisi seguendo i metodi classici della meta-analisi. In particolare per il modello ad effetti fissi si fa riferimento al metodo Mantel-Haenszel (MH). Il modello sottostante presenta la forma consueta del *fixed-effects*, $\mathrm{DOR_i} = \mu + \epsilon_\mathrm{i}$, con $\epsilon_i \sim N(0, \sigma_i^2)$, dove σ_i^2 sono le varianze specifiche dei vari studi. Mentre cambiano i pesi utilizzati dallo stimatore per combinare i vari $\mathrm{DOR_i}$ degli studi primari. Lo stimatore per μ ed i pesi ω_i Mantel-Haenszel sono

$$\hat{\mu} = \frac{\sum_{i=1}^{N} {\omega_i}^{\mathrm{MH}} DOR_i}{\sum_{i=1}^{N} {\omega_i}^{\mathrm{MH}}} \ , \qquad \text{con} \ \ {\omega_i}^{\mathrm{MH}} = \frac{FP(n_1 - VP)}{(n_1 + n_2)}. \tag{2.10}$$

Per ottenere il DOR seguendo il modello ad effetti casuali si fa riferimento all'approccio di DerSimonian and Laird, considerando però il logaritmo del DOR, il modello sottostante è infatti il seguente: $log(DOR_i) = \mu + \epsilon_i + \delta_i$, con $\epsilon_i \sim N(0, \sigma_i^2)$ e $\delta_i \sim N(0, \tau^2)$. In questo caso σ_i^2 è stimata come $\hat{\sigma}_i^2 = \frac{1}{VP} + \frac{1}{(n_1 - VP)} + \frac{1}{FP} + \frac{1}{(n_2 - FP)}$, mentre la varianza τ^2 è stimata con il metodo DSL (2.3). Lo stimatore per μ ed i pesi ω_i di DerSimonian and Laird sono

$$\hat{\mu} = \frac{\sum_{i=1}^{N} {\omega_i}^{\mathrm{DSL}} DOR_i}{\sum_{i=1}^{N} {\omega_i}^{\mathrm{DSL}}} \ , \qquad \text{con} \ \ {\omega_i}^{\mathrm{DSL}} = \frac{1}{(\hat{\sigma}_i^2 + \hat{\tau}^2)}. \tag{2.11}$$

2.3.2 Approccio bivariato

Metodo SROC di Littenberg e Moses

La curva ROC, come abbiamo visto in precedenza, è uno strumento utile per valutare le performance di un test diagnostico, che tiene conto simultaneamente dei valori di 1-specificità e sensibilità. Quando si hanno N studi riferiti allo stesso test, per ognuno di essi si ha a disposizione la coppia dei due tassi TPR e FPR, la rispettiva soglia di tale misura di precisione spesso non è menzionata o differisce in maniera sostanziale tra gli operatori (Littenberg & Moses, 1993). L'interesse è quello di sintetizzare le misure fornite dalle varie coppie di valori e riferite a studi diversi, eseguire quindi un analisi

che consenta di stimare una curva che si adatti ai vari punti nello spazio ROC, ognuno associato al rispettivo studio (Littenberg & Moses, 1993).

Il primo passo di tale metodo consiste nell'applicare una trasformazione per rendere la relazione tra le variabili coinvolte più lineare, al fine di poter adattare un modello di regressione anch'esso lineare. Si applica allora la trasformazione logit come segue

$$\xi = logit(FPR) = ln\left(\frac{FPR}{1 - FPR}\right),$$
$$\eta = logit(TPR) = ln\left(\frac{TPR}{1 - TPR}\right).$$

Successivamente si definiscono le quantità D ed S

$$D = \eta - \xi,$$
$$S = \eta + \xi.$$

D è inoltre pari al ln(DOR), è quindi una misura della potenza che il test ha nel discriminare correttamente tra pazienti positivi e negativi. S è una misura che si riferisce alla soglia del test, ha valore positivo negli studi in cui la sensibilità è superiore alla specificità, negativo quando avviene il contrario, e pari a 0 quando le due misure si equivalgono (Reitsma et al., 2005).

La relazione lineare di interesse è esplicitata come

$$D = \alpha + \beta S,$$

dove i parametri α e β del modello sono stimati ai minimi quadrati utilizzando una regressione lineare ponderata o non ponderata (Littenberg & Moses, 1993). Nel caso con ponderazione, i pesi W sono proporzionali all'inverso della varianza di D stimata

$$W = \frac{1}{\frac{1}{(FP+0.5)} + \frac{1}{(n_2 - FP+0.5)} + \frac{1}{(VP+0.5)} + \frac{1}{(n_1 - VP+0.5)}}.$$

Si stima quindi una retta di regressione attraverso i punti (S, D) esplicitando come varia il potere discriminatorio con il variare della soglia implicita del test. La curva SROC è poi ottenuta ritrasformando le stime di α e β nello spazio ROC originale.

Il metodo SROC presenta il vantaggio di essere relativamente semplice, ma soffre di diverse limitazioni. In primo luogo, è un modello ad effetti fissi, ovvero assume che i valori di α e β non differiscano tra gli studi, la variabilità è quindi dovuta all'effetto soglia ed alla varianza interna agli studi (Arends et al., 2008). Questo trascura la possibile

varianza tra gli studi dovuta alle diverse caratteristiche degli stessi, portando a stime distorte ed a sottostimare i valori degli standard errors. Le misure D ed S sono inoltre correlate all'interno dei vari studi, e con il modello ad effetti fissi questa correlazioni viene ignorata (Arends et al., 2008).

Modelli bivariati ad effetti casuali

I modelli gerarchici, come il modello bivariato e il modello gerarchico riassuntivo delle caratteristiche operative del ricevitore (HSROC), sono considerati approcci consolidati per condurre la meta-analisi di studi che valutano l'accuratezza dei test diagnostici (Lee et al., 2015). In Reitsma et al. (2005) venne proposto il modello ad effetti casuali bivariato, ripreso successivamente da Arends et al. (2008).

L'utilizzo dei modelli bivariati ad effetti casuali per condurre la meta-analisi sulla valutazione dei test diagnostici, consente di integrare la possibile eterogeneità tra gli studi. Eterogeneità che si verifica molto spesso negli studi che riguardano i test diagnostici, questo per via del differire degli aspetti tecnici del test, tra cui il valore di *cut-off* scelto, della selezione dei pazienti, delle modalità con cui gli studi vengono condotti Arends et al. (2008). Anche la correlazione, che molto spesso si osserva tra sensibilità e specificità, è prevista nel modello. Soprattutto poi questo modello è fedele alla natura bivariata dei due parametri di interesse.

Il modello proposto da Reitsma et al. (2005) ha una struttura gerarchica, articolandosi in un livello interno agli studi ed uno tra gli studi che lascia spazio anche alla correlazione tra sensibilità e specificità (Riley et al., 2007). Si riprendono le trasformazioni logit dei parametri di interesse come visto prima in Littenberg & Moses (1993), allora

$$\eta_i = logit(sens_i),$$
$$\xi_i = logit(1 - spec_i).$$

Le misure η_i e ξ_i seguono una distribuzione normale bivariata con vettore delle medie $\boldsymbol{\mu} = (\bar{\eta}, \bar{\xi})'$ e matrice di varianza-covarianza tra gli studi $\boldsymbol{\Sigma}$. Quindi il livello relazionato alla varianza tra gli studi presenta una distribuzione congiunta degli effetti casuali η_i e ξ_i che si esplicita come

$$\begin{pmatrix} \eta_i \\ \xi_i \end{pmatrix} \sim N_2\left(\boldsymbol{\mu} = \begin{pmatrix} \bar{\eta} \\ \bar{\xi} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\eta^2 & \rho\sigma_\eta\sigma_\xi \\ \rho\sigma_\eta\sigma_\xi & \sigma_\xi^2 \end{pmatrix}\right), \tag{2.12}$$

dove $\bar{\eta}$ e $\bar{\xi}$ sono i veri valori delle medie di sensibilità ed 1-specificità tra gli studi, in

scala logit, σ_η^2 e σ_ξ^2 sono le varianze *between-study* e ρ è il coefficiente di correlazione, che viene così incluso nel modello.

Il livello che considera la variabilità interna agli studi, e descrive la relazione tra le stime e i veri valori di η_i e ξ_i, specifica il modello *within-study* come

$$\begin{pmatrix} \hat{\eta}_i \\ \hat{\xi}_i \end{pmatrix} \sim N_2\left(\begin{pmatrix} \eta_i \\ \xi_i \end{pmatrix}, \begin{pmatrix} s_{\eta_i}^2 & 0 \\ 0 & s_{\xi_i}^2 \end{pmatrix} \right), \tag{2.13}$$

con

$$s_{\eta_i}^2 = \frac{1}{VP} + \frac{1}{(n_1 - VP)} \quad , s_{\xi_i}^2 = \frac{1}{VN} + \frac{1}{(n_2 - VN)},$$

dove $\hat{\eta}_i$ e $\hat{\xi}_i$ sono le stime per le trasformate logit di sensibilità e 1-specificità per ogni studio, i cui veri valori corrispondono a η_i e ξ_i, mentre $s_{\eta_i}^2$ e $s_{\xi_i}^2$ sono le varianze osservate all'interno degli studi.

Infine, combinando insieme i due livelli (2.12) e (2.13) si ottiene il modello finale, un modello lineare bivariato ad effetti misti specificato come segue

$$\begin{pmatrix} \hat{\eta}_i \\ \hat{\xi}_i \end{pmatrix} \sim N_2\left(\begin{pmatrix} \bar{\eta} \\ \bar{\xi} \end{pmatrix}, \begin{pmatrix} \sigma_\eta^2 + s_{\eta_i}^2 & \rho\sigma_\eta\sigma_\xi \\ \rho\sigma_\eta\sigma_\xi & \sigma_\xi^2 + s_{\xi_i}^2 \end{pmatrix} \right). \tag{2.14}$$

La stima dei parametri del modello, $\bar{\eta},\ \bar{\xi},\ \sigma_\eta^2,\ \sigma_\xi^2,\ \rho$, può essere ottenuta sia tramite massima verosimiglianza che massima verosimiglianza ristretta, ed è facilmente computabile utilizzando software statistici, come R.

Un limite che si verifica sia nel modello Reitsma et al. (2005) che in quello di Rutter & Gatsonis (2001) è la necessità di applicare una correzione di continuità, convenzionalmente fissata a 0.5, ai valori pari a 0 nelle tabelle di errata classificazione, questo per evitare che le trasformazioni logit non siano definite. Il problema di tali correzioni è che può avere un effetto importante sulla curva ROC, allontanandola dal vertice in alto a sinistra (Arends et al., 2008). Per ovviare a questo problema, ed ottenere inoltre una distribuzione esatta del modello interno agli studi, in quanto quello in equazione (2.13) è un modello con distribuzione normale approssimata, si suggerisce una distribuzione binomiale per il numero di test positivi osservati nel gruppo dei malati e in quello dei non malati (Arends et al., 2008).

$$\begin{aligned} VP_i &= Binomial(n_{1i}, TPR_i), \\ FP_i &= Binomial(n_{2i}, FPR_i). \end{aligned} \tag{2.15}$$

I modelli (2.12) e (2.15) specificano ora un modello lineare generalizzato misto, lo svantaggio è la maggiore complessità per la stima dei parametri, la cui implementazione non è inoltre comune nei software quanto quella del modello approssimato. Un problema aggiuntivo di questo modello riguarda la possibilità di stime inaffidabili della matrice di covarianza e che essa non sia positiva, questi problemi di convergenza incrementano quando il numero di studi è limitato (Hamza et al., 2008).

Il modello HSROC proposto da Rutter & Gatsonis (2001) segue un approccio di regressione bayesiana gerarchica, questo implica però la difficoltà nel calcolo delle stime utilizzando la simulazione Markov Chain Monte Carlo (MCMC) (Arends et al., 2008). In assenza di covariate, il modello HSROC di Rutter e Gatsonis è equivalente al modello bivariato genralizzato misto specificato dalle equazioni (2.12) e (2.15), i due modelli forniscono stime equivalenti per i valori di sensibilità e specificità (Harbord et al., 2007). Si rimanda a Rutter & Gatsonis (2001) per approfondire le caratteristiche specifiche del modello.

2.4 La meta-regressione

Come emerso da quanto visto in questo capitolo, l'eterogeneità gioca un ruolo fondamentale nelle analisi, guidando la scelta tra modello ad effetti fissi e casuali, e valutando quanto gli studi inclusi nell'analisi differiscono tra loro. La meta-regressione è uno strumento che fonde i principi della meta-analisi e quelli della regressione casuale, permettendo di esplorare l'eterogeneità e verificare se una o più covariate abbiano un effetto significativo sulla misura di outcome (Baker et al., 2009). Le covariate permettono di valutare sia caratteristiche degli individui che quelle degli studi, in modo da poter valutare come queste contribuiscano ad incrementare l'eterogeneità.

Seguirà la trattazione della meta-regressione nel caso bivariato, che nel contesto dei test diagnostici abbiamo visto essere più appropriato rispetto a quello univariato, per la meta-regressione nel caso di singoli outcome si rimanda a Thompson & Higgins (2002).

Possiamo lavorare con la meta-regressione nel caso bivariato estendendo i modelli (2.14) e (2.15) con l'inclusione delle covariate. Indicando con X_i il vettore delle p covariate incluse nel modello, e con B la matrice dei coefficienti di regressione, allora il modello bivariato ad effetti misti di meta-regressione si specifica come

$$\begin{pmatrix} \hat{\eta}_i \\ \hat{\xi}_i \end{pmatrix} \sim N(BX_i,\ \mathbf{\Sigma} + \mathbf{C}_i),$$

dove

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_\eta^2 & \rho\sigma_\eta\sigma_\xi \\ \rho\sigma_\eta\sigma_\xi & \sigma_\xi^2 \end{pmatrix} \text{ e } \mathbf{C}_i = \begin{pmatrix} s_{\eta_i}^2 & 0 \\ 0 & s_{\xi_i}^2 \end{pmatrix}.$$

La stima di questo modello può sempre essere ottenuta con i metodi della massima verosimiglianza, la parte impegnativa risiede nell'interpretazione dei parametri ottenuti (Van Houwelingen et al., 2002). È importante evidenziare che nella meta-regressione multivariata in discussione, si presume che la covarianza dell'effetto casuale tra gli studi sia costante per tutte le combinazioni di covariate, e nel caso di variabili categoriali questo vuol dire che la covarianza è identica in tutti i gruppi. E' possibile mitigare questa assunzione introducendo una covarianza specifica tra gli studi per ciascun sottogruppo (Doebler et al., 2018).

Capitolo 3

Applicazione ai dati

3.1 Descrizione del dataset

Lo studio di Kutz et al. (2020) è partito da una ricerca sistematica della letteratura interrogando i database PubMed e Cochrane su parole chiavi collegate alla domanda di ricerca, come "SCOFF" o "eating disorders". La revisione sistematica è stata svolta seguendo le procedure del metodo *PRISMA-DTA*, una guida di reporting specifica per le revisioni sistematiche diagnostico-terapeutiche. Dalle ricerche condotte, sono stati selezionati 25 studi, per un totale di 11531 soggetti, che soddisfacessero gli standard di qualità richiesti e contenessero le informazioni necessarie per condurre l'analisi. I dati fondamentali in ogni studio, ed indispensabili per svolgere la meta-analisi, sono il numero di veri positivi, veri negativi, falsi positivi e falsi negativi, dai quali vengono poi ottenute le stime di sensibilità e specificità per ogni studio. Gli standard di riferimento che si utilizzano come paragone, sono riportati nello specifico per ogni studio. Gli studi inclusi sono stati svolti in ambiti diversi, come scuole, ambiente medico o comunità generale, e per ogni studio è riportata la nazione in cui è stato condotto. Le altre informazioni riportano il range di età, i soggetti presentano un'età compresa tra i 10 e i 95 anni, ed i range differiscono notevolmente nella maggior parte dei gruppi. Si riporta anche la percentuale femminile nei soggetti campionati, con 12 studi composti da sole donne. Alcuni studi riportano parzialmente la percentuale per ogni specifico disturbo alimentare rilevato nel campione. Nel corso della meta-analisi, tra le variabili a disposizione, ci si focalizzerà su quelle relative al range di età, alla nazione e alla percentuale femminile nel campione.

3.2 Analisi grafiche

Una prima analisi dei dati può essere condotta con l'utilizzo di strumenti grafici tipici della meta-analisi, alcuni specifici per la natura bivariata dei test diagnostici. Questi contribuiscono a descrivere le misure riassuntive dei singoli studi, come specificità e sensibilità, e l'eterogeneità presente tra gli studi.

Il forest plot è uno strumento grafico primario che raffigura, per ogni studio, la stima dell'effetto medio e il relativo intervallo di confidenza. Più nello specifico, sulla sinistra del grafico è riportato il riferimento allo studio corrispondente e per ognuno di questi è rappresentata, nel nostro grafico con un quadrato, la stima dell'effetto di interesse. Con i segmenti orizzontali vengono rappresentati gli intervalli di confidenza di livello 0.95 associati alle rispettive stime. I valori delle stime e degli estremi degli intervalli si possono visualizzare nella parte destra. Per interpretare questo grafico, focalizzandoci sui singoli studi, si deve considerare che maggiore è la lunghezza dei segmenti e maggiore è la variabilità, interna allo studio, associata alla stima dell'effetto. Come abbiamo visto nella teoria presentata al Capitolo 2, i pesi utilizzati per ottenere l'effetto medio di tutti gli studi inclusi nell'analisi sono inversamente proporzionali alla varianza, *within-study* per il modello ad effetti fissi e *within-study* + *between-study* per il modello ad effetti casuali. Quindi gli studi che nel forest plot presentano intervalli maggiormente concentrati attorno alla stima avranno un maggiore impatto nella stima dell'effetto medio aggregato. Mentre a livello complessivo, il grafico consente di visionare la possibile eterogeneità tra gli studi. L'eterogeneità è evidente quando si può osservare che le stime dell'effetto e i relativi intervalli di confidenza differiscono sensibilmente tra i vari studi.

Prendendo in considerazione la Figura 3.1 e la Figura 3.2 possiamo analizzare i forest plot ottenuti per la sensibilità e la specificità del nostro dataset. A differenza della meta-analisi classica in cui ci si focalizza sulla misura di un solo effetto, e si ha quindi un singolo forest plot, nell'analisi dei test diagnostici si costruiscono due forest plot, per considerare entrambi i parametri di accuratezza diagnostica. Si può osservare come, soprattutto per la sensibilità, la lunghezza degli intervalli varia in modo rilevante tra i singoli studi. Così come è evidente una notevole discrepanza nei valori delle stime, e dei rispettivi intervalli di confidenza, tra i vari studi, sia per la sensibilità che per la specificità. Quanto osservato suggerisce la presenza di eterogeneità, che potrà poi essere confermata con opportuni test statistici, quali il test Q di Cochran (2.4), e le statistiche I^2 (2.5) e H^2 (2.6) di Higgins e Thompson.

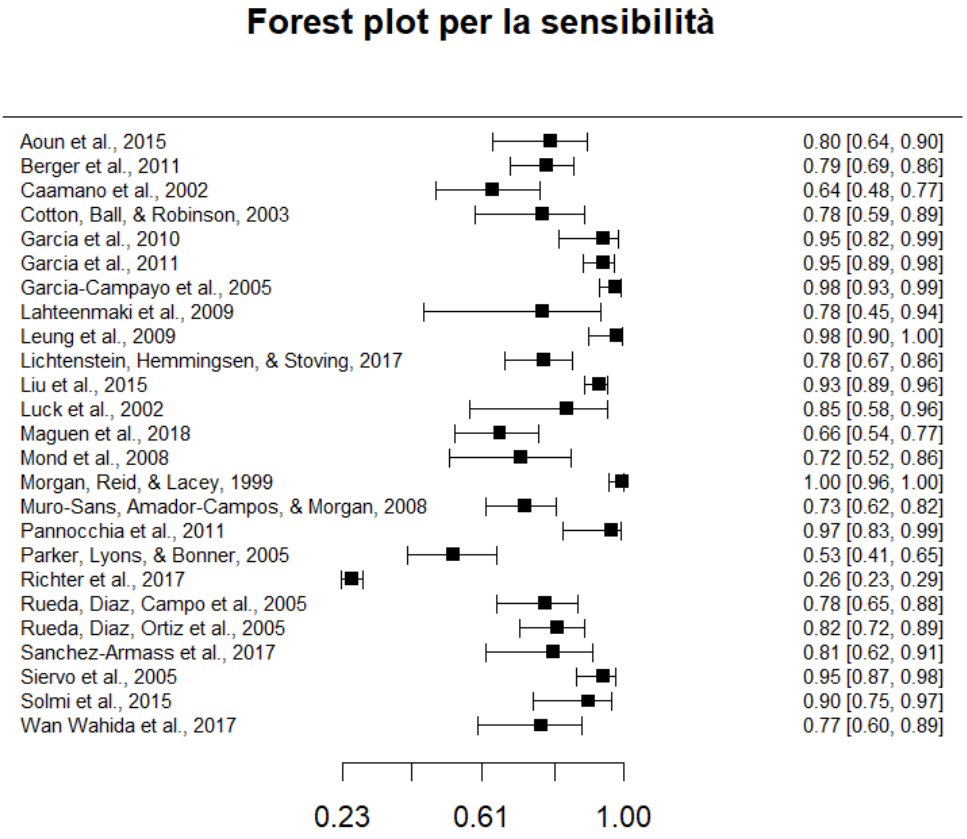


FIGURA 3.1: Forest plot per la sensibilità del dataset in analisi

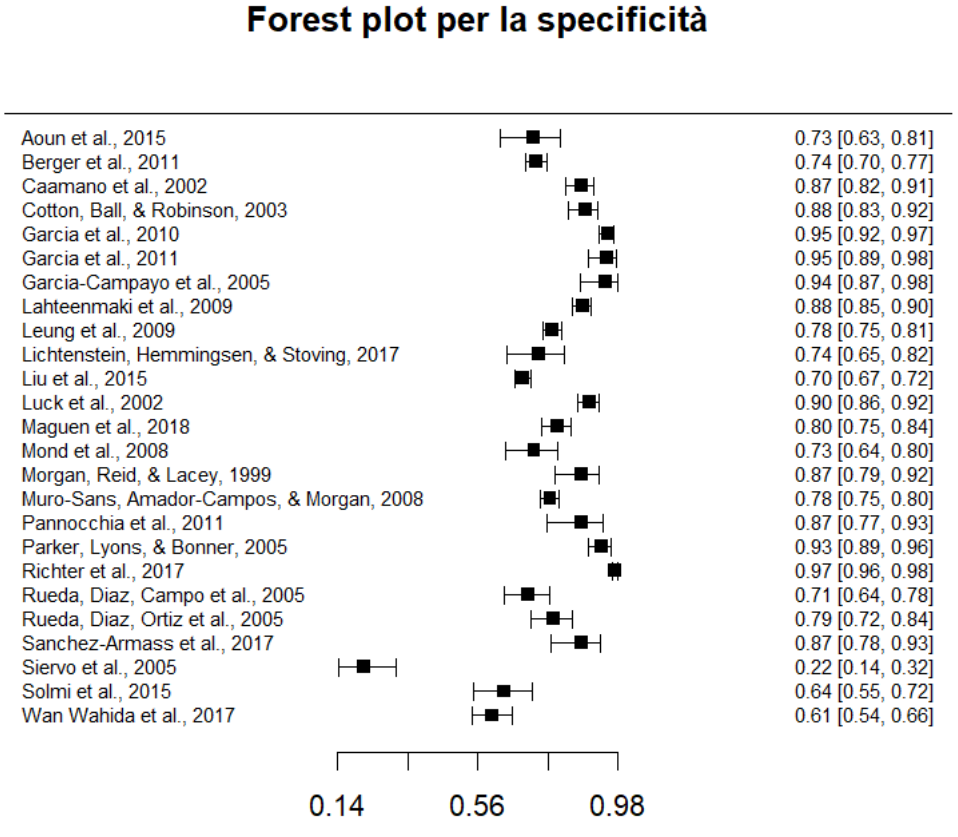


FIGURA 3.2: Forest plot per la specificità del dataset in analisi

Due grafici che consentono di risaltare la natura bivariata della misura di accuratezza di un test diagnostico sono il crosshair e il ROCellipse. Lo spazio in cui si sviluppano i due grafici è lo spazio ROC, quello in cui si raffigura l'omonima curva. Infatti l'asse delle ascisse riporta il tasso di falsi positivi (*FPR*), che ricordiamo corrispondere ad *1-specificità*, e l'asse delle ordinate si riferisce ai valori della *sensibilità*. In questo modo vengono messe in relazione contemporaneamente sensibilità e specificità, quest'ultima collegata al *FPR*. Nel grafico crosshair ogni croce raffigura gli intervalli di confidenza per la *sensibilità* e il *FPR*, il punto di intersezione delle due linee corrisponde alla stima della coppia di tali parametri, lo spessore delle linee è direttamente proporzionale alla numerosità campionaria dello studio. Da questo grafico emerge il livello di eterogeneità sia tra i due parametri presi in considerazione e sia quello tra i vari studi. Sempre nello spazio ROC si sviluppa il grafico ROCellipse, che permette di visualizzare le regioni di confidenza costruite per le coppie (*1-specificità, sensibilità*) e le stime di questi punti. Una maggiore espansione della regione è associata ad una maggiore variabilità all'interno dello studio. Una sensibile discrepanza tra le varie regioni descrive invece la presenza di eterogeneità tra gli studi.

Nel nostro caso, come possiamo vedere dalla Figura 3.3 e Figura 3.4, risulta evidente una notevole eterogeneità tra gli studi, a conferma di quanto si era visto analizzando la coppia di forest plot. Inoltre due studi in particolare si discostano marcatamente da tutti gli altri, quello di Siervo et al. (2005) per via di un elevato valore del tasso di falsi positivi (0.78), mentre quello di Richter et al. (2017) per via di un basso valore del parametro sensibilità (0.26).

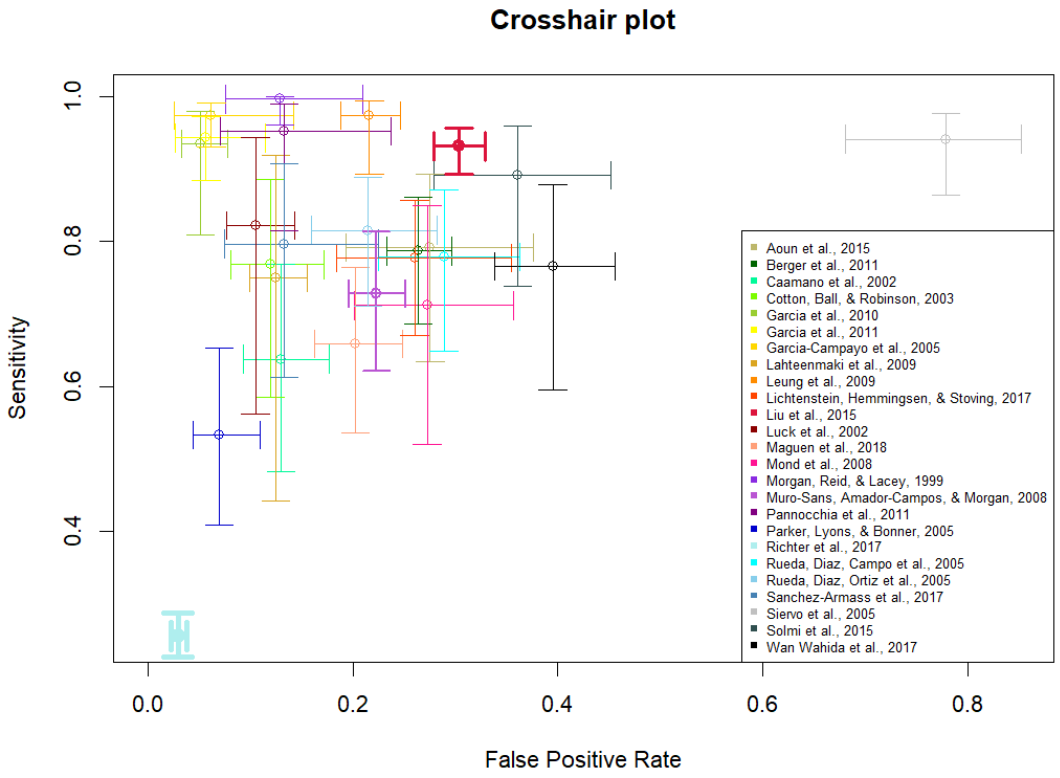


FIGURA 3.3: Grafico crosshair. Intervalli di confidenza per le stime di sensibilità e tasso di falsi positivi

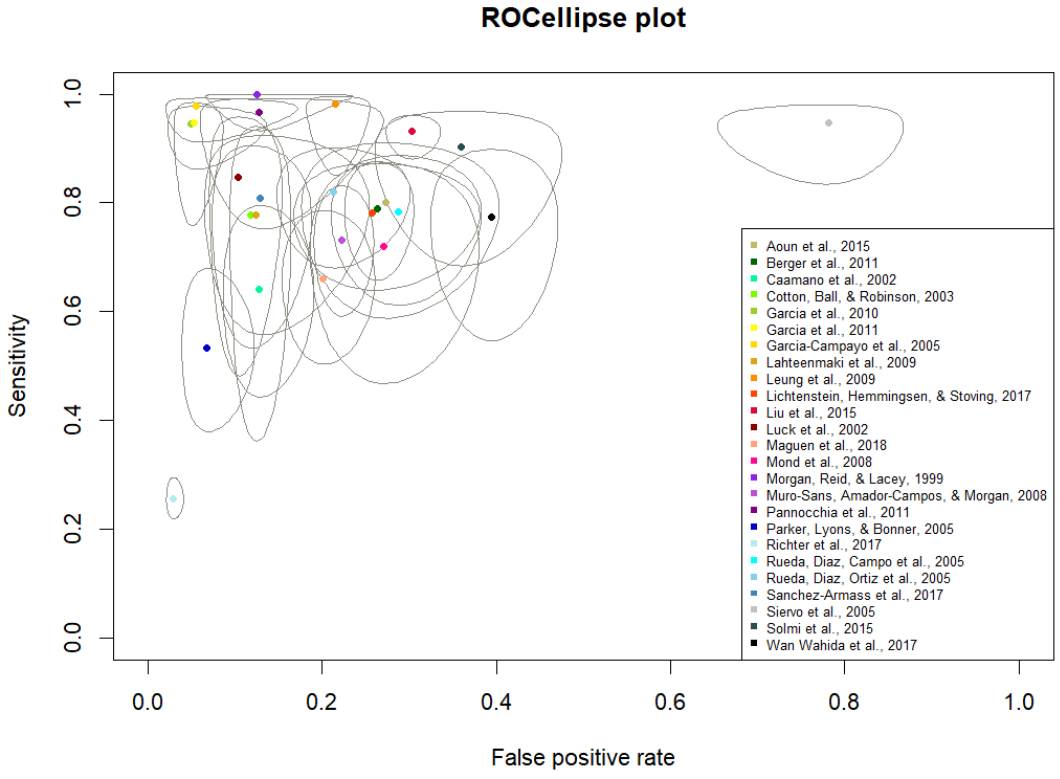


FIGURA 3.4: Grafico ROCellipse. Stime delle coppie di sensibilità e tasso di falsi positivi con le rispettive regioni di confidenza

3.3 Applicazione

Approccio univariato

Esaminati i grafici, si conduce inizialmente l'analisi seguendo l'approccio univariato, considerando perciò singolarmente le misure di interesse: sensibilità, specificità e diagnostic odds ratio (DOR).

Per ottenere le stime aggregate di sensibilità e specificità, considerando tutti gli studi inclusi nell'analisi, si utilizza il comando *metaprop*() della libreria *meta*, accurato per elaborare dati relativi a proporzioni, potendo così considerare in modo univariato i parametri di interesse. I principali valori di output, relativi alle stime ed alle misure di eterogeneità, sono riportati in Tabella 3.1. Nella Figura 3.5 e Figura 3.6 sono riportati i forest plot con le misure di effetto totale ottenute, i pesi con le quali sono state ottenute, ed i valori (eventi su totale) con i quali sono state calcolate le proporzioni.

Metodo	Modello	sens	95% CI	τ^2	I^2	H	spec	95% IC	τ^2	I^2	H
CLASSIC	CE	0.556	[0.528; 0.583]	1.216	96.1%	5.04	0.783	[0.773; 0.792]	0.931	96.4%	5.26
	RE	0.838	[0.762; 0.893]				0.826	[0.763; 0.875]			
GLMM	CE	0.653	[0.634; 0.673]	1.537	96.0%	4.98	0.819	[0.811; 0.827]	0.918	96.4%	5.26
	RE (mixed)	0.858	[0.781; 0.911]				0.828	[0.766; 0.877]			

TABELLA 3.1: Stime di sensibilità e specificità, relativi intervalli di confidenza 0.95, e misure di eterogeneità τ^2, I^2, H. Risultati ottenuti sia con il metodo classico che con il GLMM, e sia per il modello CE che quello RE.

Utilizzando il modello ad effetti fissi (2.1) e quello ad effetti casuali (2.2) sulle proporzioni logit trasformate (2.7), otteniamo come stima della sensibilità rispettivamente i valori di 0.556 e 0.838. E' importante notare come i pesi che contribuiscono ad aggregare i vari studi con il modello CE, avendo utilizzato la trasformazione logit delle proporzioni, sono dati da da $1/{\nu_i}^2$, con ${\nu_i}^2$ calcolata come riportato in (2.8). I pesi dipendono quindi dalla numerosità dei casi VP e dalla differenza tra il numero totale dei malati effettivi ed il numero di VP con i quali si calcola la proporzione. La numerosità dei gruppi ha quindi un'influenza rilevante nella ponderazione che fornisce poi la misura dell'effetto totale. Questo porta ad avere, ad esempio con i nostri dati, che lo studio Richter et al. (2017) contribuisce alla stima dell'effetto totale in misura del 48.1%. Vale a dire che quasi metà della stima è determinata da quanto osservato in questo studio, che essendo caratterizzato da un basso valore di sensibilità (0.26), porta ad avere complessivamente una sensibilità notevolmente inferiore a quella che si ottiene con il modello RE, e a quella osservata singolarmente negli altri studi che risulta variare tra un minimo di 0.53 ed un massimo di 1. Mentre il modello ad effetti casuali tenendo conto anche della

varianza tra gli studi τ^2, porta ad un restringimento dei pesi, in modo che gli studi con varianza ${\nu_i}^2$ minore ricevano pesi più bassi rispetto al modello ad effetti fissi. Nel nostro esempio, il peso dello studio Richter et al. (2017) è ancora il più alto, ma in linea con i pesi degli altri studi. In merito all'eterogeneità, il valore stimato di τ^2 è pari a 1.22, secondo la statistica I^2 il 96.1% della varianza totale è dovuto all'eterogeneità tra i gruppi, e anche la statistica H pari a 5.04 ci suggerisce la presenza di un'elevata variabilità tra gli studi. Il test Q di Cochran con un valore della statistica Q pari a 608.87 ed un p-value inferiore a 0.01 conferma quanto indicato dalle altre statistiche. Risulta quindi opportuno e preferibile l'utilizzo del modello ad effetti casuali rispetto a quello ad effetti fissi, in modo da poter considerare la notevole eterogeneità presente tra gli studi in analisi.

Per quanto riguarda la specificità, i valori stimati dal modello CE e da quello RE sono pari rispettivamente a 0.783 e 0.826. Nel modello CE lo studio a cui è associato il maggior peso, pari al 22.2%, è quello di Liu et al. (2015), questo è determinato dalla considerevole numerosità dei soggetti sani e dal ridotto gap tra *VN* e soggetti sani. A differenza però di quanto riscontrato per la sensibilità, lo studio non presenta una misura anomala per il valore della specificità, ed il peso non è ingente tanto quanto lo era sopra, pertanto le stime ottenute con i due differenti modelli risultano più vicine. Lavorando con il modello RE, nella ponderazione degli studi si aggiunge alla varianza ${\nu_i}^2$ il valore τ^2, stimato tramite la massima verosimiglianza ristretta, pari a 0.931. Tenendo conto dell'eterogeneità tra gli studi si ottengono dei pesi più equilibrati, ed un intervallo per la stima dell'effetto totale evidentemente più ampio rispetto a quello ottenuto con il modello CE. Questo accade anche nel caso della sensibilità, perché in presenza di evidente eterogeneità tra i risultati degli studi individuali, le misure di sintesi ottenute con l'approccio a effetti casuali tendono ad avere varianze stimate più elevate e, di conseguenza, intervalli di confidenza più estesi rispetto alle sintesi basate sull'approccio ad effetti fissi (Poole & Greenland, 1999). La presenza di elevata eterogeneità tra gli studi è confermata dal valore della statistica test Q di Cochran pari a 664.18 (p-value < 0.01) e dalle altre misure di eterogeneità, riportate in alto a destra della Tabella 3.1.

L'altro modo per ottenere separatamente le stime per la sensibilità e la specificità, aggregando quanto osservato dai vari studi, è l'utilizzo del modello lineare generalizzato ad effetti misti, con funzione di legame logit, (2.9). L'applicazione di tale modello è anch'essa disponibile con il comando *metaprop* della libreria *meta*, più precisamente è il metodo di default utilizzato da tale comando per ottenere le stime delle proporzioni. Come si può notare anche dai forest plot in Figura 3.7, utilizzando un modello di regressione i pesi non hanno ruolo nella stima dei parametri, questi vengono stimati tramite

massima verosimiglianza. Adattando il modello di regressione logistica con unica fonte di variabilità data dall'errore casuale, otteniamo 0.653 come stima per la sensibilità e 0.819 per la specificità. Inserendo nel modello gli effetti casuali u_{i}, entra in gioco τ^2, infatti $u_{\mathrm{i}} \sim N(0, \tau^2)$, che stimato anch'esso tramite massima verosimiglianza, risulta pari a 1.537 per la sensibilità e 0.918 per la specificità. Di conseguenza, il modello di regressione logistica misto, più appropriato data l'elevata eterogeneità presente tra gli studi, fornisce i valori di 0.858 e 0.828 per la sensibilità e la specificità.

I metodi esaminati finora considerano separatamente la sensibilità e la specificità, una misura che permette di legarle insieme è il diagnostic odds ratio. Utilizzando il metodo Mantel-Haenszel (2.10), che segue il modello ad effetti fissi, si ottiene una stima per il DOR pari a 16.539 , CI(0.95)=[14.372; 19.033]. L'utilizzo del modello ad effetti casuali seguendo l'approccio di DerSimonian and Laird (2.11) attribuisce al DOR una stima pari a 22.058 , CI(0.95)=[14.840; 32.786]. Ovvero è 22 volte più probabile che il test dia un risultato positivo quando il paziente soffre di disturbi alimentari rispetto a quando il paziente non ne soffra. In entrambi i casi, maggiormente se si include l'eterogeneità con l'approccio DSL, il DOR per il test diagnostico in esame suggerisce una notevole capacità dello stesso nell'identificare correttamente i pazienti affetti da disturbi alimentari, riducendo al contempo i falsi positivi, e sostenendo così l'affidabilità del test nell'ambito della diagnosi preliminare.

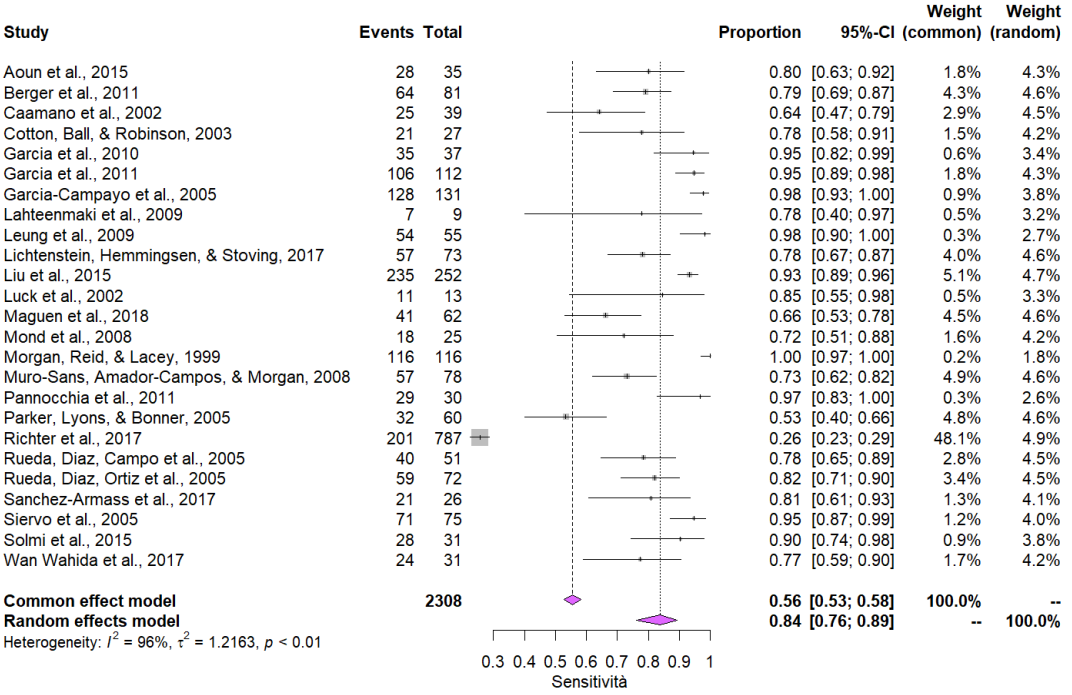


FIGURA 3.5: Forest plot per la sensibilità del dataset in analisi, con effetto totale (common e random) e pesi utilizzati per la stima. Ottenuto con il metodo classico.

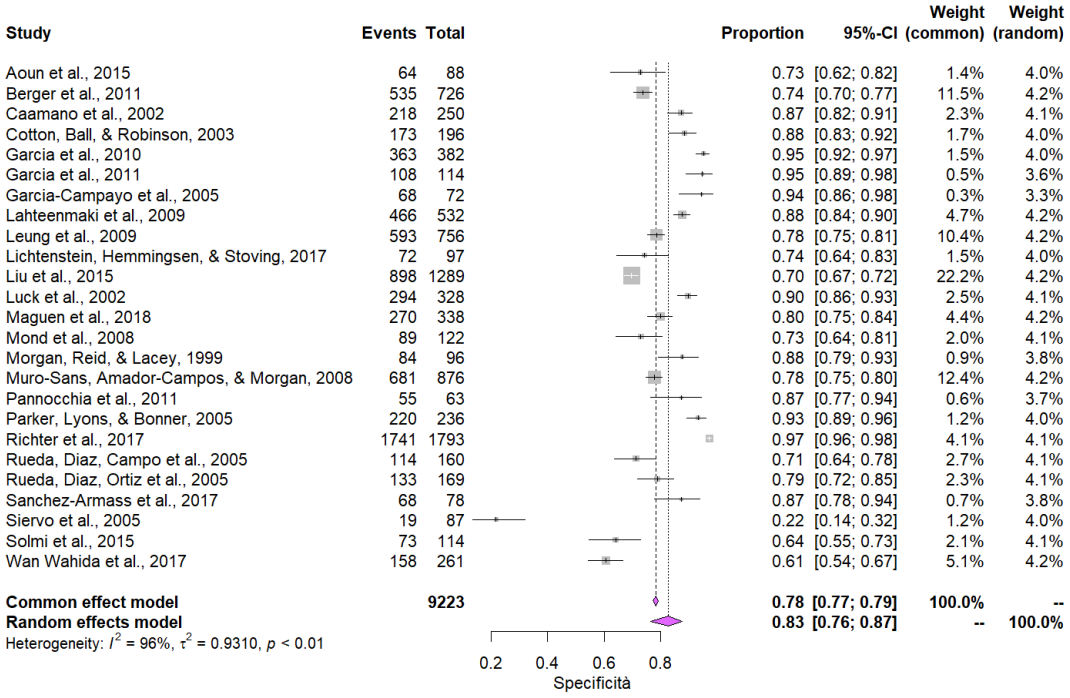
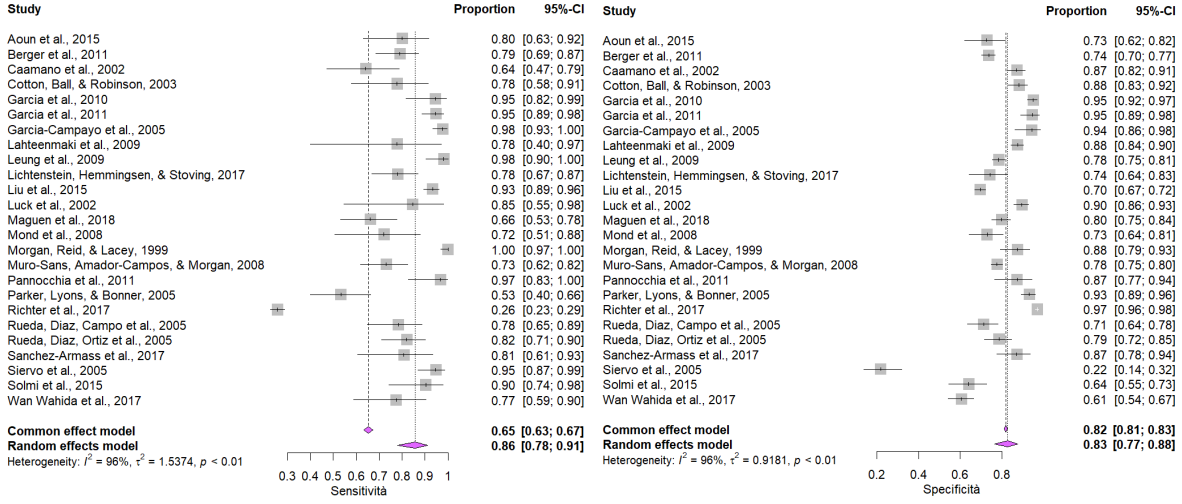


FIGURA 3.6: Forest plot per la specificità del dataset in analisi, con effetto totale (common e random) e pesi utilizzati per la stima. Ottenuto con il metodo classico.



(a) Forest plot per la sensibilità

(b) Forest plot per la specificità

FIGURA 3.7: Forest plot ottenuti utilizzando il modello misto lineare generalizzato (GLMM)

Approccio bivariato

La tecnica di analizzare separatamente le misure di sensibilità e specificità, seppur ancora oggi utilizzata, come dagli stessi autori in Kutz et al. (2020), presenta molti limiti ed è stata superata con l'introduzione dei modelli di Reitsma et al. (2005) e Rutter & Gatsonis (2001). Si è discusso di come l'approccio univariato non consideri congiuntamente i due parametri di accuratezza e la possibile correlazione tra gli stessi. Dai diagrammi di dispersione per i nostri dati, in Figura 3.8, si nota una correlazione negativa tra sensibilità e specificità che diventa naturalmente di segno opposto quando si considera il *FPR* anziché la specificità.

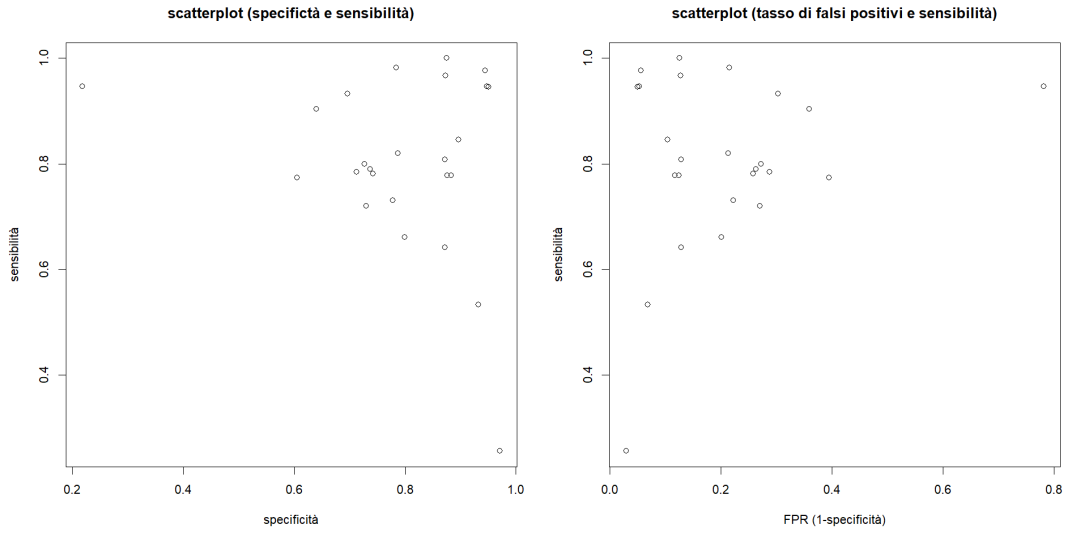


FIGURA 3.8: Diagramma di dispersione per i punti di specificità e sensibilità. Diagramma di dispersione per i punti di FPR e sensibilità

Per considerare tale correlazione e lavorare congiuntamente con sensibilità e specificità, si utilizza il modello di Reitsma et al. (2005) implementato in R nella libreria *mada* attraverso il comando *reitsma*(). Il modello (2.14), stimato tramite massima verosimiglianza ristretta, fornisce le stime dei coefficienti $\bar{\eta}$ e $\bar{\xi}$, riportate in Tabella 3.2, che essendo trasformazioni logit sono meno immediate da interpretare, per questo motivo l'output fornisce anche i valori dei parametri ritrasformati su scala originale. Allora

	stima	std. error	z	pval	95% CI	
intercetta $\bar{\eta}$	1.624	0.239	6.804	0.000	[1.156; 2.092]	***
intercetta $\bar{\xi}$	-1.547	0.197	-7.855	0.000	[-1.934; -1.161]	***
sensibilità	0.835	-	-	-	[0.761; 0.890]	
FPR	0.175	-	-	-	[0.126, 0.238]	
Codici di significatività: 0 '***' , 0.001 '**', 0.01 '*', 0.05 '.'						

TABELLA 3.2: Stime dei coefficienti con il modello Reitsma

le stime riassuntive della sensibilità e della specificità per i nostri studi in analisi sono pari rispettivamente a 0.835 e 0.825. Inoltre i valori stimati delle deviazioni standard *between-study* σ_η e σ_ξ sono pari a 1.076 e 0.957, mentre il coefficiente di correlazione ρ assume valore pari a 0.258. Il modello fornisce inoltre un valore per l'AUC pari a 0.896. Possiamo visualizzare a livello grafico quanto stimato dal modello ottenendo una curva SROC di tipo gerarchico come descritto in Rutter & Gatsonis (2001). Nel nostro caso la curva SROC ottenuta è rappresentata in Figura 3.9, in cui è possibile notare anche la coppia di valori (*1-specificità*, *sensibilità*) stimata e la relativa regione di confidenza di livello 0.95.

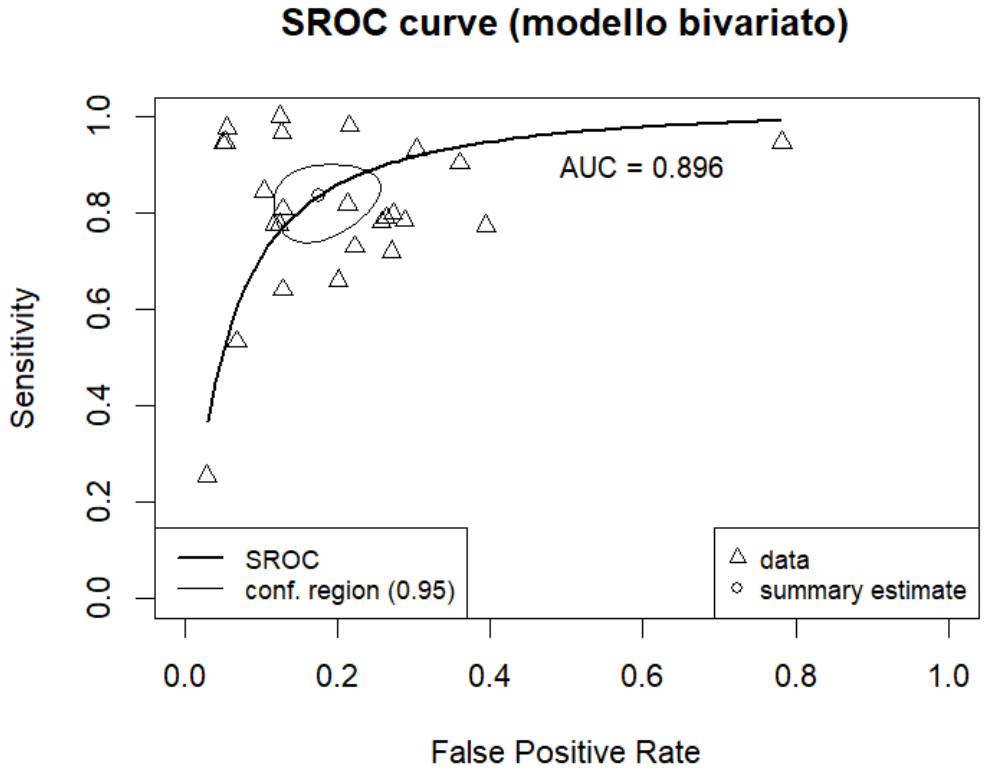


FIGURA 3.9: Curva SROC stimata dal modello Reitsma

Come emerso dalle analisi grafiche nella sezione 3.2, due studi si discostano notevolmente dagli altri. Si procede allora a stimare il modello (2.14) sui 23 studi rimanenti a seguito dell'esclusione degli outliers Siervo et al. (2005) e Richter et al. (2017). Possiamo in questo modo valutare l'effetto, e le distorsioni, che la presenza o meno di questi studi hanno sulle stime. I valori stimati dal modello sono riportati in Tabella 3.3, da cui si ottiengono le misure 0.845 e 0.828 per la sensibilità e la specificità dei vari studi. Seppur leggermente superiori, le stime non si discostano molto da quelle ottenute in precedenza

considerando tutti gli studi. Per questo motivo le analisi che seguiranno saranno svolte considerando tutti i 25 studi di partenza.

	stima	std. error	z	pval	95% CI	
intercetta $\bar{\eta}$	1.694	0.212	7.987	0.000	[1.278; 2.110]	***
intercetta $\bar{\xi}$	-1.574	0.152	-10.376	0.000	[-1.871; -1.277]	***
sensibilità	0.845	-	-	-	[0.782; 0.892]	
FPR	0.172	-	-	-	[0.133, 0.218]	

Codici di significatività: 0 '***' , 0.001 '**', 0.01 '*', 0.05 '.'

Tabella 3.3: Stime dei coefficienti con il modello Reitsma escludendo dagli studi i due outliers

Si vuole ora indagare sull'impatto che le variabili che si hanno a disposizione, nello specifico il range di età, nazione e percentuale di sesso femminile nel campione, possano avere sulle variabili risposta. Si andranno quindi ad inserire le covariate nel modello, lavorando ad una meta-regressione bivariata.

Come primo passo sono state modificate le variabili originali. Questo è stato necessario per la variabile che riportava l'informazione sull'età in quanto espressa in un range delimitato dall'osservazione più piccola e da quella più grande, e i range per i vari studi differivano in modo consistente non permettendo un raggruppamento degli stessi. Si è quindi scelto di prendere per ogni studio il valore centrale del range. Successivamente però, nell'analisi condotta inserendo le covariate, si è visto che la variabile *età* ottenuta non era in genere rilevante. Data la consistente distinzione della presenza di disturbi alimentari tra i giovani/adolescenti e gli adulti, come abbiamo visto nella presentazione di questa malattia, si è pensato di modificare ulteriormente la variabile *età*, trovando dei risultati interessanti. La variabile *età* definitiva corrisponde ad una variabile categoriale ("AgeAd") che assume come livelli "giovane" se lo studio presentava valore centrale del range di età ≤ 26, e "adulto" se invece il valore era > 26. La variabile, categoriale, che indicava la nazione in cui è stato svolto lo studio è stata modificata raggruppando le varie nazioni in aree geografiche, avendo altrimenti troppe categorie rispetto al numero di studi. Alla variabile che si riferiva alla percentuale del sesso femminile nello studio si è affiancata anche una variabile dicotomica che distinguesse gli studi completamente femminili, ma per entrambe non sono poi emersi nell'analisi effetti rilevanti.

Nella meta-regressione due studi sono stati esclusi dall'analisi in quanto per loro non erano disponibili nel dataset i valori di tutte le covariate. La meta-regressione bivariata viene quindi condotta sui 23 studi rimanenti.

Riprendendo il modello proposto da Reitsma et al. (2005), senza covariate, e stimandolo adesso sui 23 studi in considerazione, le stime risultanti sono riportate nella

Tabella 3.4. I valori di $\bar{\eta}$ e $\bar{\xi}$ forniti dal modello sono pari a 0.836 e 0.82. Rispetto a quanto ottenuto considerando l'insieme compelto di studi, la sensibilità risulta pressoché la stessa, la specificità è leggermente superiore. Anche il valore stimato per l'AUC risulta praticamente lo stesso e pari a 0.895.

	stima	std. error	z	pval	95% CI	
intercetta $\bar{\eta}$	1.627	0.248	6.572	0.000	[1.142; 2.113]	***
intercetta $\bar{\xi}$	-1.518	0.213	-7.136	0.000	[-1.935; -1.101]	***
sensibilità	0.836	-	-	-	[0.758; 0.892]	
FPR	0.180	-	-	-	[0.126, 0.250]	

Codici di significatività: 0 '***' , 0.001 '**', 0.01 '*', 0.05 '.'

TABELLA 3.4: Stime dei coefficienti con il modello Reitsma sui 23 studi che presentano valori per tutte le variabili

Si passa ora alla meta-regressione, considerando come possibili covariate le variabili descritte in precedenza. Sono stati adattati vari modelli, inserendo le diverse covariate nel modello e cambiandone le combinazioni. Il modello che presenta dei risultati interessanti e parametri maggiormente significativi tra gli altri, è quello che prevede come unica covariata "AgeAd", differenziando tra gli studi che si distinguono per un campione di età giovane e quelli di età adulta. La Tabella 3.5 mostra le stime risultanti da tale modello.

	stima	std. error	z	pval	95% CI	
intercetta $\bar{\eta}$	1.617	0.317	5.099	0.000	[0.995; 2.238]	***
tsens.AgeAdAdo	0.054	0.526	0.103	0.918	[-0.976; 1.085]	***
intercetta $\bar{\xi}$	-1.929	0.226	-8.542	0.000	[-2.371; -1.486]	
tfpr.AgeAdAdo	1.164	0.377	3.089	0.002	[0.426, 1.902]	**

Codici di significatività: 0 '***' , 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ', 1

TABELLA 3.5: Stime dei coefficienti del modello Reitsma con l'aggiunta della covariata "AgeAd"

Tutti i parametri risultano ampiamente significativi tranne quello riferito alla trasformata logit del tasso di falsi positivi per il gruppo degli adulti. I valori di sensibilità e specificità, ri-trasformando le stime in scala originale, per la categoria adulti sono stimati pari a 0.834 e 0.873. Per la categoria dei giovani i parametri di interesse si ottengono come $sens = \frac{e^{(1.617+0.054)}}{1+e^{(1.617+0.054)}} = 0.842$ e $spec = 1 - \frac{e^{(-1.929+1.164)}}{1+e^{(-1.929+1.164)}} = 0.682$. Il test si caratterizza quindi per una sensibilità leggermente superiore nel gruppo dei giovani, mentre la specificità è fortemente minore per questo gruppo rispetto a quanto osservato in quello degli adulti. Un valore così ridotto per la specificità del test nel gruppo dei giovani può essere dovuto ad un maggiore numero di falsi positivi. Tuttavia,

nel contesto in cui ci stiamo riferendo, un valore più basso della specificità può essere tollerato, se la sensibilità risulta al contempo elevata. Questo perché stiamo parlando di un test di screening, che non è invasivo ed è facilmente accessibile, il cui risultato deve essere inoltre confermato da una valutazione diagnostica più approfondita. Non risulta allora un problema troppo rilevante se il numero di falsi positivi che esso produce è consistente. Nell'ambito dei disturbi alimentari, ed in una fase delicata come quella adolescenziale/giovanile, è preferibile avere una diagnosi positiva poi smentita piuttosto che una mancata rilevazione del problema.

Possiamo continuare l'analisi a livello statistico confrontando con il test log-rapporto di verosimiglianza i due modelli annidati, quello che presenta "AgeAd" come covariata e quello con le sole intercette. Il risultato del test è riportato nella Tabella 3.6, assieme ai valori dei criteri di informazione AIC e BIC. Sia questi ultimi, che il test log-rapporto di verosimiglianza, indicano come migliore il modello con la covariata, che spiega quindi parte dell'eterogeneità presente nei dati.

Likelihood-ratio test	Criteri di informazione
Model 1: cbind(tsens, tfpr) ∼ AgeAd	AIC: -55.956 BIC: -43.156
Model 2: cbind(tsens, tfpr) ∼ 1	AIC: -50.871 BIC: -41.728
ChiSquared Df Pr(>ChiSquared) 9.188 2 0.0101 * Codici di significatività: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	

Tabella 3.6: Confronto tra il modello con sole intercette ed il modello con l'aggiunta della covariata "AgeAd"

Per evidenziare le differenti performance del test nelle due categorie, si costruisce un grafico che sovrapponga la curva SROC del modello base, e le due curve SROC distinte per i giovani e per gli adulti. Il grafico ottenuto è riportato in Figura 3.10.

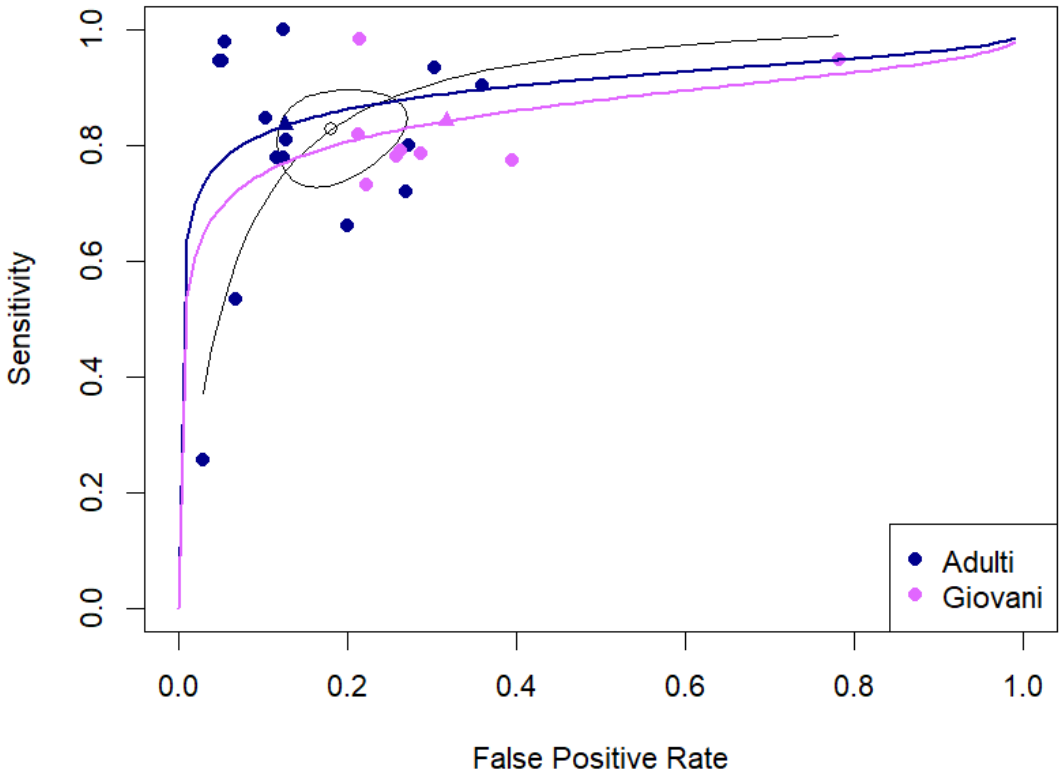


FIGURA 3.10: Confronto delle curve SROC. In nero la curva SROC ottenuta con il modello Reitsma base, in blu quella per gli adulti, in lilla quella per i giovani.

Il grafico mostra la differenza tra le due curve. La curva SROC per gli adulti, in blu, è più alta e più vicina al vertice sinistro rispetto a quella dei giovani, raffigurata in lilla, che è posizionata sotto quella blu. Questo riflette quanto visto precedentemente con le stime risultanti dai due modelli. Il test di screening, in base a quanto rilevato dagli studi in analisi, risulta più accurato per gli adulti.

Capitolo 4

Conclusioni

L'obiettivo di questa relazione è quello di valutare la validità del questionario SCOFF, uno dei test di screening proposti per rilevare la presenza dei disturbi del comportamento alimentare. Per valutarne la validità, si è svolta una meta-analisi sugli stessi studi utilizzati in Kutz et al. (2020). Come primo passo un'analisi grafica è stata eseguita per rilevare la possibile eterogeneità tra gli studi in esame. Di seguito, l'approccio meta-analitico utilizzato è stato quello di considerare separatamente le misure di sensibilità e specificità, che insieme permettono di valutare l'accuratezza di un test diagnostico. In questo modo sono stati applicati i metodi classici della meta-analisi, che solitamente lavora su un singolo *effect-size*. Avendo riscontrato importante eterogeneità tra gli studi, il modello più adatto per combinare i dati si identifica con quello ad effetti casuali, che produce le stime 0.838 e 0.826 per la sensibilità e la specificità. Considerando il modello lineare generalizzato misto le stime ottenute tenendo conto dell'eterogeneità sono pari rispettivamente a 0.858 e 0.828. I limiti del considerare separatamente le due misure, e non tenere conto della loro correlazione, vengono superati con l'utilizzo dei modelli bivariati. Si procede quindi l'analisi con l'utilizzo del modello Reitsma et al. (2005), le cui stime risultanti per la sensibilità e la specificità sono 0.835 e 0.825. Sono questi ultimi i valori a cui si fa riferimento per stabilire la validità del test, considerando che i modelli bivariati ad effetti casuali rappresentano il metodo più adeguato, soprattutto rispetto agli altri presi in considerazione, per valutare attraverso meta-analisi l'accuratezza di un test diagnostico. Il questionario SCOFF risulta un valido strumento di screening per la diagnosi preliminare dei disturbi alimentari, con valori buoni sia per la sensibilità che per la specificità. Uno studio di meta-regressione è stato poi utile per riscontrare che il test risulta più accurato per i soggetti di età adulta piuttosto che per quelli nella fascia giovane e adolescenziale.

Possibili sviluppi futuri possono affrontare il problema di stima del modello bivariato ad effetti casuali tramite i metodi di verosimiglianza, che nel caso di campioni poco numerosi possono portare a conclusioni inferenziali inesatte e problemi di convergenza. In merito a questo problema viene proposta in Guolo (2017) la metodologia SIMEX, basata sulle tecniche di simulazione. Si possono considerare inoltre approcci di modellazione bayesiana bivariata, metodologia che amplia notevolmente l'ambito delle tecniche bivariate tradizionali, consentendo alla distribuzione strutturale degli effetti casuali di essere influenzata da diverse fonti di variabilità (Verde, 2010).

Appendice

.1 Codice R

```
library("meta")
library("mada")

dati <- readxl::read_excel("dati_tesi.xlsx")

#ANALISI GRAFICHE
#forest plot
forest(madad(dati, correction.control = "single"), type= "sens",
       main="Forest plot per la sensibilità",cex=0.7,
       snames=dati$Article)
forest(madad(dati, correction.control = "single"), type= "spec",
       main="Forest plot per la specificità",cex=0.7,
       snames=dati$Article)

#crosshair plot
my_colors <- c("#BDB76B","#006400","#00FA9A","#7CFC00","#9ACD32",
               "#FFFF00","#FFD700","#DAA520","#FF8C00","#FF4500",
               "#DC143C","#8B0000","#FFA07A","#FF1493","#8A2BE2",
               "#BA55D3","#800080","#0000CD","#AFEEEE","#00FFFF",
               "#87CEEB","#4682B4","#C0C0C0","#2F4F4F","#000000")
rs <- rowSums(dati[,2:5])
weights <- 4 * rs / max(rs)
crosshair(dati, xlim = c(0,0.85), ylim = c(0.25,1),col=my_colors,
          lwd = weights, main="Crosshair plot")
legend('bottomright', legend=dati$Article, pch = 15,
       col=my_colors, cex = 0.6)
```

```
#ROCellipse
ROCellipse(dati, pch = "", main="ROCellipse plot", ellipsecol = "ivory4")
points(fpr(dati), sens(dati), pch=20, col=my_colors)
legend('bottomright', legend=dati$Article, pch = 15,
       col=my_colors, cex = 0.6)


#ANALISI CON APPROCCIO UNIVARIATO
#meta-analisi su proporzioni
sens_sum <- metaprop(dati$TP, dati$TP+dati$FN, comb.fixed = T,
                     comb.random = T, sm="PLOGIT", method.ci = "CP",
                     studlab = dati$Article, method = "Inverse",
                     method.incr = "only0")
sens_sum

spec_sum <- metaprop(dati$TN, dati$TN+dati$FP, comb.fixed = T,
                     comb.random = T, sm="PLOGIT", method.ci = "CP",
                     studlab = dati$Article, method = "Inverse",
                     method.incr = "only0")
spec_sum

forest.meta(sens_sum, xlab = "Sensitività", col.diamond.common
            = "mediumorchid1", col.diamond.random = "mediumorchid1")
forest.meta(spec_sum, xlab = "Specificità", col.diamond.common
            = "mediumorchid1", col.diamond.random = "mediumorchid1")

#modello lineare generalizzato misto (GLMM)
sens_sum_GLMM <- metaprop(dati$TP, dati$TP+dati$FN, comb.fixed = T,
                          comb.random = T, sm="PLOGIT", method.ci = "CP",
                          studlab = dati$Article, data=dati,
                          method = "GLMM", method.incr = "only0")
sens_sum_GLMM

spec_sum_GLMM <- metaprop(dati$TN, dati$TN+dati$FP, comb.fixed = T,
                          comb.random = T, sm="PLOGIT", method.ci = "CP",
```

```
                          studlab = dati$Article, data=dati,
                          method = "GLMM", method.incr = "only0")
spec_sum_GLMM

forest.meta(sens_sum_GLMM, xlab = "Sensitività", col.diamond.common
            = "mediumorchid1", col.diamond.random = "mediumorchid1",
            leftcols = "studlab")
forest.meta(spec_sum_GLMM, xlab = "Specificità", col.diamond.common
            = "mediumorchid1", col.diamond.random = "mediumorchid1",
            leftcols = "studlab")

#diagnostic odds ratio (DOR)
#metodo MH (fixed effect)
fit.DOR.MH <- madauni(dati, method = "MH", correction.control="single")
summary(fit.DOR.MH)

#metodo DSL (random effect)
fit.DOR.DSL <- madauni(dati, correction.control="single")
summary(fit.DOR.DSL)


#ANALISI CON APPROCCIO BIVARIATO
#scatterplot
sens <- dati$TP/(dati$TP+dati$FN)
spec <- dati$TN/(dati$TN+dati$FP)
fpr <- 1-spec
par(mfrow=c(1,2))
plot(spec,sens, main="scatterplot (specifictà e sensibilità)",
     xlab="specificità", ylab="sensibilità")
plot(fpr, sens, main="scatterplot (tasso di falsi positivi e
     sensibilità)", xlab="FPR (1-specificità)", ylab="sensibilità")


#modello lineare misto bivariato Reitsma
fit1 <- reitsma(dati, correction.control = "single")
summary(fit1)
```

```
#modello Reitsma escludendo i due outliers
dati_23 <- dati[-c(19,23),]
fit2 <- reitsma(dati_23, correction.control = "single")
summary(fit2)

#SROC
plot(fit1, sroclwd = 2,main = "SROC curve (modello bivariato)")
points(fpr(dati), sens(dati), pch = 2)
legend("bottomright", c("data", "summary estimate"), pch = c(2,1),
       cex=0.8)
legend("bottomleft", c("SROC", "conf. region (0.95)"), lwd = c(2,1),
       cex=0.8)
text(0.6, 0.9, "AUC = 0.896", col = "black", cex = 0.9)


#META-REGRESSIONE BIVARIATA
#esclusione valori NA
dati$Gender <- as.numeric(dati$Gender)
dati$Age <- as.numeric(dati$Age)
dati <- na.omit(dati)

#Reitsma sui 23 studi
fit3 <- reitsma(dati, correction.control = "single")
summary(fit3)

#modello con covariata "AgeAd"
fit4 <- reitsma(dati, formula = cbind(tsens, tfpr) ~ AgeAd,
                correction.control = "single")
summary(fit4)

#test log-rapporto di verosimiglianza
fit_3_ml <- reitsma(dati, formula = cbind(tsens, tfpr) ~ 1,
                    method = "ml", correction.control = "single")
fit_4_ml <- reitsma(dati, formula = cbind(tsens, tfpr) ~ AgeAd,
                    method = "ml", correction.control = "single")
```

```
anova(fit_4_ml, fit_3_ml)



#curve SROC con distinzione adolescenti/adulti
SP <- dati$TN/(dati$TN+dati$FP)
SE <- dati$TP/(dati$TP+dati$FN)
#modello senza covariata
m <- reitsma(dati, correction.control = "single")
plot(m)
legend("bottomright", legend=c("Adulti","Giovani"), pch=19,
       col = c("blue4","mediumorchid1") )
points(1-SP[dati$AgeAd=='Ad'], SE[dati$AgeAd=='Ad'], col="blue4",
       pch=19)
points(1-SP[dati$AgeAd!='Ad'], SE[dati$AgeAd!='Ad'],
       col="mediumorchid1", pch=19)
#modello con covariata
m2 <- reitsma(dati, formula = cbind(tsens, tfpr) ~ AgeAd,
       correction.control="single")
## sovrapponi SE e SP da m2
points(plogis(-1.929+1.164), plogis(1.617+0.054),
       col="mediumorchid1", pch=17)
points(plogis(-1.929), plogis(1.617), col="blue4", pch=17)

values.curve <- function(x){
  mu.eta <- x[1]
  mu.xi <- x[2]
  rho.xieta <- x[5]
  var.xi <- x[4]
  var.eta <- x[3]
  cov.xieta <- rho.xieta * sqrt(var.xi*var.eta)
  a <- mu.eta - mu.xi*cov.xieta/var.xi
  b <- cov.xieta/var.xi
  return(c('intercept'= a, 'slope'= b))
}

ad <- c(1.617 ,-1.929, 1.103^2, 0.837^2, 0.301)
```

```
ado <- c(1.617+0.054,-1.929+1.164 , 1.103^2, 0.837^2, 0.301)
## aggiungere curva ROC per adulti
values.exact <- values.curve(ad)
intercept <- values.exact[1]
slope <- values.exact[2]
curve((exp(intercept)*(x/(1-x))^slope)/(1 + (exp(intercept)*(x/(1-x))^slope)),
      add=TRUE, lwd=2, col="blue4")
## aggiungere curva ROC per adolescenti
values.exact <- values.curve(ado)
intercept <- values.exact[1]
slope <- values.exact[2]
curve((exp(intercept)*(x/(1-x))^slope)/(1 + (exp(intercept)*(x/(1-x))^slope)),
      add=TRUE, lwd=2, col="mediumorchid1")
```

Bibliografia

Arends, L., Hamza, T., Van Houwelingen, J., Heijenbrok-Kal, M., Hunink, M. & Stijnen, T. (2008). Bivariate random effects meta-analysis of roc curves. *Medical Decision Making* **28**, 621–638.

Baker, W. L., Michael White, C., Cappelleri, J. C., Kluger, J., Coleman, C. I., From the Health Outcomes, P. & Group, E. H. C. (2009). Understanding heterogeneity in meta-analysis: the role of meta-regression. *International journal of clinical practice* **63**, 1426–1434.

Bioletto, F., Berton, A. M., Prencipe, N., Parasiliti-Caprino, M., Gasco, V., Maccario, M. & Grottoli, S. (2022). Strumenti per la lettura critica di metanalisi e revisioni sistematiche: il punto di vista del clinico. *L'Endocrinologo* **23**, 620–626.

Borenstein, M., Hedges, L. & Rothstein, H. (2007). Meta-analysis: Fixed effect vs. random effects. *Meta-analysis. com* , 1–162.

Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods* **1**, 97–111.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10**, 101–129.

Deeks, J. J. (2001). Systematic reviews of evaluations of diagnostic and screening tests. *Bmj* **323**, 157–162.

DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials* **7**, 177–188.

Doebler, P., Bürkner, P.-C. & Rücker, G. (2018). Statistical packages for diagnostic meta-analysis and their application. *Diagnostic Meta-Analysis: A Useful Tool for Clinical Decision-Making* , 161–181.

Doebler, P. & Holling, H. (2015). Meta-analysis of diagnostic accuracy with mada. *R Packag* **1**, 15.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J. & Bossuyt, P. M. (2003). The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology* **56**, 1129–1135.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher* **5**, 3–8.

Global Burden of Disease Collaborative Network (2020). Global burden of disease study 2019 (gbd 2019) results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2020. Data available from `https://vizhub.healthdata.org/gbd-results/`.

Gogtay, N. & Thatte, U. (2017). An introduction to meta-analysis. *Journal of the Association of Physicians of India* **65**, 78–85.

Guolo, A. (2017). A double simex approach for bivariate random-effects meta-analysis of diagnostic accuracy studies. *BMC Medical Research Methodology* **17**, 1–12.

Hamza, T. H., Reitsma, J. B. & Stijnen, T. (2008). Meta-analysis of diagnostic studies: a comparison of random intercept, normal-normal, and binomial-normal bivariate summary roc approaches. *Medical Decision Making* **28**, 639–649.

Harbord, R. M., Deeks, J. J., Egger, M., Whiting, P. & Sterne, J. A. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* **8**, 239–251.

Hardy, R. J. & Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in medicine* **15**, 619–629.

Hardy, R. J. & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in medicine* **17**, 841–856.

Harrer, M., Cuijpers, P., Furukawa, T. A. & Ebert, D. D. (2021). *Doing meta-analysis with R: A hands-on guide*. CRC press.

Henderson, M. & Page, L. (2007). Appraising the evidence: what is selection bias? *BMJ Ment Health* **10**, 67–68.

Higgins, J. P. & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine* **21**, 1539–1558.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F. & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or i^2 index? *Psychological methods* **11**, 193.

Jackson, B. R. (2008). The dangers of false-positive and false-negative test results: false-positive results as a function of pretest probability. *Clinics in laboratory medicine* **28**, 305–319.

Kutz, A. M., Marsh, A. G., Gunderson, C. G., Maguen, S. & Masheb, R. M. (2020). Eating disorder screening: a systematic review and meta-analysis of diagnostic test characteristics of the scoff. *Journal of general internal medicine* **35**, 885–893.

Lee, J., Kim, K. W., Choi, S. H., Huh, J. & Park, S. H. (2015). Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers-part ii. statistical methods of meta-analysis. *Korean journal of radiology* **16**, 1188–1196.

Lin, L. & Chu, H. (2020). Meta-analysis of proportions using generalized linear mixed models. *Epidemiology (Cambridge, Mass.)* **31**, 713.

Littenberg, B. & Moses, L. E. (1993). Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Medical Decision Making* **13**, 313–321.

López-López, J. A., Marín-Martínez, F., Sánchez-Meca, J., Van den Noortgate, W. & Viechtbauer, W. (2014). Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *British Journal of Mathematical and Statistical Psychology* **67**, 30–48.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology* **5**, 1315–1316.

McInnes, M. D., Moher, D., Thombs, B. D., McGrath, T. A., Bossuyt, P. M., Clifford, T., Cohen, J. F., Deeks, J. J., Gatsonis, C., Hooft, L. et al. (2018). Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the prisma-dta statement. *Jama* **319**, 388–396.

Morgan, J. F., Reid, F. & Lacey, J. H. (1999). The scoff questionnaire: assessment of a new screening tool for eating disorders. *Bmj* **319**, 1467–1468.

Mulrow, C. D. (1994). Systematic reviews: rationale for systematic reviews. *Bmj* **309**, 597–599.

Poole, C. & Greenland, S. (1999). Random-effects meta-analyses are not always conservative. *American Journal of Epidemiology* **150**, 469–475.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M. & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology* **58**, 982–990.

Riley, R. D., Abrams, K. R., Sutton, A. J., Lambert, P. C. & Thompson, J. R. (2007). Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology* **7**, 1–15.

Rodgers, R. F., Lombardo, C., Cerolini, S., Franko, D. L., Omori, M., Fuller-Tyszkiewicz, M., Linardon, J., Courtet, P. & Guillaume, S. (2020). The impact of the covid-19 pandemic on eating disorder risk and symptoms. *International Journal of Eating Disorders* **53**, 1166–1170.

Rutter, C. M. & Gatsonis, C. A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in medicine* **20**, 2865–2884.

Santomauro, D. F., Melen, S., Mitchison, D., Vos, T., Whiteford, H. & Ferrari, A. J. (2021). The hidden burden of eating disorders: an extension of estimates from the global burden of disease study 2019. *The Lancet Psychiatry* **8**, 320–328.

Shim, S. R. (2022). Meta-analysis of diagnostic test accuracy studies with multiple thresholds for data integration. *Epidemiology and Health* **44**.

Sidik, K. & Jonkman, J. N. (2005). Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society Series C: Applied Statistics* **54**, 367–384.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science* **240**, 1285–1293.

Taquet, M., Geddes, J. R., Luciano, S. & Harrison, P. J. (2022). Incidence and outcomes of eating disorders during the covid-19 pandemic. *The British Journal of Psychiatry* **220**, 262–264.

Thompson, S. G. & Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in medicine* **21**, 1559–1573.

Van Houwelingen, H. C., Arends, L. R. & Stijnen, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in medicine* **21**, 589–624.

Verde, P. E. (2010). Meta-analysis of diagnostic test data: a bivariate bayesian modeling approach. *Statistics in medicine* **29**, 3088–3102.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* **30**, 261–293.

World Health Organization, . (2022). Eating disorders. `https://www.who.int/news-room/fact-sheets/detail/mental-disorders`.