UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia Generale

Corso di Laurea Magistrale in Psicologia Clinica

Tesi di Laurea Magistrale

Nuovi metodi per identificare le risposte simulate in una prova per la valutazione della depressione

New methods for identifying simulated responses in a trial for assessing depression

Relatore

Giuseppe Sartori

Correlatrice

Giulia Melis

Laureanda: **Ilaria Castelli**

Matricola: **1238200**

Anno Accademico **2021/2022**

INDICE

INTRODUZIONE

La depressione maggiore è uno dei disturbi psicologici più antichi, presente nel DSM (Manuale Diagnostico e Statistico dei Disturbi mentali) sin dalla sua nascita negli anni ‘50 con il termine di “reazione depressiva”, diventata poi “nevrosi depressiva” nel secondo manuale. La “reazione depressiva” agli inizi della sua classificazione all'interno del DSM, era considerata, in ottica del tutto dinamica, una risposta ad eventi di vita negativi e a condizioni ambientali stressanti. Oggigiorno lo sviluppo della depressione maggiore invece viene attribuita ad un insieme di fattori di rischio ambientali e genetici che rendono il soggetto maggiormente vulnerabile allo sviluppo del disturbo

La depressione maggiore è anche il disturbo mentale più diffuso nelle diverse fasce d’età, con un forte impatto socioeconomico, sia per i costi diretti che per quelli indiretti.

L’elevata prevalenza e la storicità di tale disturbo, trattate dettagliatamente nel primo capitolo, fa sì che i suoi sintomi siano conosciuti dalla maggior parte delle persone poiché la probabilità di aver esperito almeno una volta nella vita tale sindrome e aver avuto parenti e/o conoscenti con tale sintomatologia è molto alta. Inoltre, i sintomi che caratterizzano questo disturbo psichiatrico non sono difficili da emulare pertanto ottenere un risarcimento economico dalle compagnie assicurative, fingendo di aver sviluppato il disturbo depressivo a seguito di un evento, diventa un “gioco da ragazzi”.

Alla luce di ciò, il rilevamento della simulazione è uno dei problemi più impegnativi nella valutazione dello stato mentale di un soggetto. Questa problematica viene trattata nel II capitolo dell’elaborato, all’interno del quale viene specificato come, essendo i sintomi psichiatrici e cognitivi facili da esagerare e simulare, l'uso di strategie e strumenti appropriati per rilevare i tentativi di simulazione è cruciale sia nel contesto clinico che in quello forense. Riveste una particolare importanza quest'ultimo contesto, in quanto la simulazione di successo è legata a conseguenze economiche e sociali rilevanti, come l'aumento dei premi assicurativi e l'assegnazione di fondi a persone immeritevoli (e non a pazienti idonei a riceverli). Oltre a ciò, ci sono conseguenze sulla giustizia, in quanto ai falsificatori potrebbe essere consentito di evitare il carcere o avere una riduzione della pena. Attualmente sono disponibili diversi strumenti che possono essere utili per individuare la simulazione, come i questionari costruiti *ad hoc* o quelli della pratica clinica dotati di scale di controllo. Sebbene questi strumenti si basino principalmente sull'autodichiarazione (quindi facilmente falsificabili), il loro uso combinato può aumentare la capacità dell'esaminatore di rilevare con precisione i tentativi di simulazione. Inoltre, questi strumenti "tradizionali" possono essere supportati dall'uso di nuovi metodi, basati sulla registrazione automatizzata dell'indice psicometrico,

come i tempi di reazione e gli aspetti cinematici delle traiettorie del mouse, che hanno mostrato capacità promettenti. Tecniche ancora più recenti e oggi molto studiate nella detezione della simulazione arrivano dal campo dell'intelligenza artificiale: modelli di apprendimento automatici (*Machine Learning*), in grado di rilevare la simulazione a livello del singolo item del questionario.

In questo elaborato ci serviremo dell'indice TF-IDF, un indice di frequenza che è stato qui applicato allo strumento di valutazione della depressione PHQ-9 con l'obiettivo di verificare se, con il suo utilizzo, sarebbe stato possibile individuare le specifiche domande a cui i partecipanti allo studio avevano mentito durate la compilazione del questionario in oggetto. Viene poi proposto un nuovo approccio in parte basato su modelli di apprendimento automatico (*Machine Learning*) grazie al quale, una volta rilevate le risposte simulate, è possibile correggerle e ricostruire il profilo di risposte oneste.

All'interno del III capitolo, vengono descritti il metodo sperimentale adottato, lo strumento utilizzato, il campione di partecipanti e la sua composizione, ed infine le ipotesi formulate.

Il capitolo IV invece, tratta l'analisi dei dati, sia grezzi che alla luce delle nuove tecniche, insieme ai risultati ottenuti.

Infine, il quinto ed ultimo capitolo sarà dedicato alla discussione ed interpretazione dei risultati delle analisi eseguite contestualizzandone i limiti e le eventuali prospettive future.

Capitolo I

LA DEPRESSIONE

1.1 Prevalenza

La depressione è considerata il disturbo mentale più diffuso nella popolazione, con circa 350 milioni di persone (5%) che ne soffrono nelle differenti fasce d'età (Lim et al., 2018).
Le indagini sulla sua diffusione sono state condotte sin dalla fine della Seconda guerra mondiale, ma solo agli inizi degli anni '80 sono state effettuate con strumenti diagnostici utili alla sua rilevazione. In queste prime stime, il Disturbo Depressivo Maggiore (DDM) era stato esperito, almeno una volta nella vita, con una prevalenza dal 3,0% al 5,9% e dall'1,7% al 3,4% negli ultimi 12 mesi. Nei dieci anni successivi le stime di prevalenza sono risultate sostanzialmente più alte: 14,9% almeno una volta nella vita e 8,6% negli ultimi 12 mesi fino ad arrivare, nell'indagine più recente dello studio del 2000-2001, ad una prevalenza del 16,2% nell'arco della vita e del 6,6% negli ultimi 12 mesi (Kessler et al., 2003).

La depressione, oltre ad essere considerata il disturbo più comune, negli anni 2000 era la quarta causa principale di *disease burden.* Il *disease burden* è la conseguenza totale e cumulativa di una o più malattie su una determinata popolazione riguardo alla salute, agli aspetti sociali ed economici della società. La depressione rappresenta il 4,4% del DALYs totale (Disability Adjusted Life Years), una misura del *Disease Burden* che mira sia a quantificare l'impatto di una malattia sia a stabilire le priorità nella distribuzione delle risorse a livello mondiale (Ustun et al., 2004). Tale misura è determinata dagli anni di vita persi per mortalità prematura a causa di una patologia (YLLs) e dagli anni di vita trascorsi in condizioni di disabilità (YLDs) (Managen et al., 2013). Pertanto, la depressione si dimostra il maggior problema di salute pubblica (Ustun et al., 2004). Nel 2001, la *World Health Organizzation* (WHO) ha redatto il "*The World Health Report*". Nel documento, uno studio ha previsto che nell'anno 2020 la depressione sarebbe diventata la seconda causa al mondo di *Disease Burden*, con un 5,7% di DALYS totale ovvero la seconda malattia con il più alto numero di giorni persi a causa di morte prematura o di vita trascorsa in condizioni menomate, seconda alla cardiopatia ischemica (The World Health Report, 2001).

In riferimento alla morte prematura, la depressione è considerata il disturbo psichiatrico più comune nelle persone che muoiono per suicidio (Hawton et al.,2013), con un'incidenza più elevata tra i soggetti con disturbo depressivo maggiore (Holma et al., 2010). Il disturbo depressivo è anche associato ad una grave compromissione della qualità della vita, infatti, il 63% dei soggetti con

disturbo depressivo maggiore ottiene un punteggio molto basso (due o più deviazioni standard al di sotto della media) nei questionari che indagano la percezione che il soggetto ha delle relazioni sociali, il funzionamento nel lavoro e nelle attività quotidiane, la situazione economica, la salute fisica e il senso di benessere generale (Rapaport et al., 2005).

Oggi, a 2 anni di distanza dall'inizio della pandemia Covid-19, gli studi più recenti sulla prevalenza della depressione a livello mondiale sono tutti influenzati dalle conseguenze di questo periodo storico. Uno studio di indagine condotto da Ettman et al. (2020) sulla popolazione statunitense ha riscontrato, durante il primo *Lockdown*, un aumento dei sintomi depressivi tre volte superiori alle stime più recenti pre-pandemiche.

Secondo i dati Istat del 2018 la depressione è considerata il disturbo mentale più diffuso, anche in Italia, con 2,8 milioni (5,4%) di persone che ne hanno sofferto nel corso del 2015 e 1,3 milioni (2,5%) nelle due settimane precedenti l'intervista nel 2017 (Istituto nazionale di statistica [ISTAT], 2018), fino a raggiungere picchi del 17% durante i primi 7 mesi di pandemia di Covid-19 (Lakhan et al.,2020).

1.2 Classificazione del disturbo negli anni

Nella prima edizione del Manuale Diagnostico e Statistico dei disturbi mentali (DSM-I), redatto agli inizi degli anni '50, la depressione era definita "Reazione depressiva" poiché era ciò che alleviava l'ansia scaturita da una situazione vissuta dal paziente ed era inserita nel gruppo dei disturbi funzionali, sviluppati in risposta agli eventi di vita e all'ambiente sociale. Questa edizione aveva più che altro una impostazione dinamica e non forniva indicazioni su come misurare o identificare la condizione riferita dal paziente (Horwitz, 2014).

Horwitz evidenzia come sempre sulla stessa scia psicodinamica, si è sviluppato il DSM-II che ha apportato alcune modifiche alla nomenclatura, cambiando il termine "reazione depressiva" in "nevrosi depressiva" che, similmente alla prima edizione, sorgeva in conseguenza ad un "conflitto interno o a un evento identificabile come la perdita di un oggetto d'amore o un possesso caro". La preponderanza di diagnosi aspecifiche e l'assenza di diagnosi specifiche, ha dato vita a trattamenti sia psichiatrici che farmaceutici molto generici che curavano condizioni generali come stress, nervi o ansia. Ripercorrendo la storia del disturbo Horwitz riporta come tale mancanza di specificità, l'assenza di criteri diagnostici oggettivi e di conseguenza la bassa concordanza tra gli esperti del settore, ha comportato una perdita di credibilità della materia che venne svalutata dalle altre specializzazioni mediche. Negli anni '70, lo sviluppo di programmi di assicurazione governativi che coprivano le spese ambulatoriali, la prescrizione di farmaci psicoattivi esclusivamente per il

trattamento di specifici disturbi psichiatrici e la necessità di tali farmaci per la cura dei soggetti deistituzionalizzati, ha comportato la necessità di diagnosi specifiche che permettessero di misurare oggettivamente il disturbo che veniva trattato e i suoi miglioramenti.

Il DSM-III, pubblicato negli anni '80, ha portato un cambiamento radicale nel mondo della psichiatria con una classificazione dei disturbi dettagliata, ateorica e basata sui sintomi piuttosto che sulle loro cause come nelle versioni precedenti (Horwitz, 2014). L'attenzione ai sintomi è quello che viene definito "approccio descrittivo" secondo il quale la depressione, che nel DSM-III è inserita all'interno dei disturbi affettivi, viene diagnosticata solamente se soddisfa un preciso numero di criteri diagnostici che si basa su un insieme di sintomi e su una precisa durata temporale (Richards, 2011). In questa edizione del manuale (American Psychiatric Association, 1980), la classificazione dei disturbi affettivi viene effettuata sulla base di due criteri: la variazione dell'umore e la presenza di una sindrome affettiva (maniacale o depressiva) totale o parziale. All'interno della classe "Disturbi Affettivi Maggiori" troviamo quelli caratterizzati da una sindrome affettiva completa, in "Altri Disturbi Affettivi Specifici" quelli determinati da una sindrome affettiva parziale della durata di due anni e nei "Disturbi Affettivi Atipici", quelli che non rientravano nelle altre due categorie. La Depressione Maggiore si trova all'interno dei "Disturbi Affettivi Maggiori", insieme al disturbo Bipolare e si distingue da questo per l'assenza di episodio maniacale. Inoltre, la Depressione Maggiore è sotto-classificata ad episodio singolo o ricorrente. Per quanto riguarda i criteri diagnostici per la Depressione Maggiore, occorre far riferimento a quelli dell'Episodio depressivo posto all'inizio del paragrafo.

Secondo questi criteri il disturbo veniva diagnosticato riscontrando nel paziente la presenza di "umore disforico o perdita di interesse o piacere in tutte o quasi tutte le attività e i passatempi abituali" (APA, 1980) sintomi propri del criterio A e la presenza nelle ultime due settimane di almeno quattro dei seguenti sintomi del criterio B:

- aumento/perdita di appetito e peso
- insonnia/ipersonnia
- agitazione o rallentamento psicomotorio
- perdita di interesse o piacere nelle attività abituali
- perdita di energia o fatica
- sentimenti di inutilità, auto-rimprovero o colpa eccessiva
- ridotta capacità di pensare e concentrarsi
- pensieri ricorrenti di morte o ideazione suicidaria.

Il Disturbo Distimico invece è incluso in "Altri Disturbi Affettivi" poiché differente dal Depressione Maggiore per gravità e durata; infatti, per essere diagnosticato, il soggetto dovrebbe esperire sintomi

caratteristici della depressione meno gravi, ma per la maggior parte del tempo o per tutto il tempo negli ultimi due anni, con periodi di remissione dei sintomi brevi che durano da pochi giorni a poche settimane, ma comunque non superano qualche mese (APA, 1980).

Nella revisione (DSM-III-R) pubblicata nel 1987, si nota un cambiamento nella nomenclatura: la classe dei Disturbi Affettivi viene ora definita "Disturbi dell'Umore" che si dividono in Disturbi Bipolari e Disturbi Depressivi (American Psychiatric Association, 1987).

In questa edizione del manuale, all'interno dei Disturbi Depressivi troviamo la Depressione Maggiore e la Distimia. La prima è caratterizzata dalla presenza di uno o più Episodi Depressivi Maggiori (da qui deriva la sotto classificazione "Depressione Maggiore ad episodio singolo" e "ricorrente"), la seconda dalla presenza di sintomi depressivi che persistono per un periodo più lungo di tempo, senza però soddisfare i criteri per un Episodio Depressivo Maggiore.

I criteri dell'Episodio Depressivo Maggiore, a cui far riferimento per diagnosticare la Depressione Maggiore, in questa revisione hanno subito alcuni cambiamenti. I primi due criteri (A e B) vengono unificati nel criterio A; pertanto per la diagnosi viene richiesta la presenza di 5 o più sintomi di cui almeno uno è l'umore depresso o la perdita di piacere/interesse nel fare le cose per almeno 2 settimane. Nel criterio A viene sottolineato come tale condizione, per essere considerata disturbo, deve interferire con il funzionamento dell'individuo: *"rappresenta un cambiamento rispetto al funzionamento precedente"* (APA, 1987). Tra i criteri per le sotto classificazioni, oltre a quelli già presenti nella precedente versione (in remissione, con caratteristiche psicotiche, con o senza melancolia e non specificato) sono stati aggiunti quelli per determinare la gravità del disturbo.

Il disturbo viene considerato lieve quando sono presenti *"pochi, se del caso, sintomi in eccesso rispetto a quelli richiesti per la diagnosi, e i sintomi si traducono solo in una minore compromissione delle funzioni lavorative o nelle normali attività sociali o nelle relazioni con gli altri"*; moderato quando sono presenti *"sintomi o compromissione funzionale tra "lieve" e "grave"*; severo, senza però caratteristiche psicotiche, quando sono presenti *"numerosi sintomi in eccesso rispetto a quelli necessari per fare la diagnosi e i sintomi interferiscono notevolmente con il funzionamento lavorativo o con le normali attività sociali o le relazioni con gli altri"*.

La gravità del disturbo, oltre a dipendere dal numero di sintomi presenti, è determinata anche dall'incidenza di esso sul funzionamento del soggetto nei diversi contesti della vita quotidiana: criterio a sé stante nel DSM-IV.

Nella versione del manuale pubblicata nel 1994 (DSM-IV), il criterio C ribadisce: *"i sintomi causano un disagio clinicamente significativo o menomazione in ambito sociale, lavorativo o in altre importanti aree di funzionamento"* (American Psychiatric Association, 1994). In questa versione la "Depressione Maggiore" viene definita "Disturbo Depressivo Maggiore", abbreviato DDM, e viene

diagnosticato quando sono soddisfatti i criteri per l'Episodio Depressivo Maggiore e va specificato se si tratta di un episodio singolo o ricorrente. Tale distinzione può risultare difficile quando si ha un episodio con sintomi crescenti e decrescenti che potrebbero sembrare due episodi separati. In questo caso il DSM-IV specifica che *"un episodio si considera terminato quando i criteri completi per l'Episodio Depressivo Maggiore non sono stati soddisfatti per almeno 2 mesi consecutivi. Durante questo periodo di 2 mesi, c'è una risoluzione completa dei sintomi o la presenza di sintomi depressivi che non soddisfano più i criteri completi per un Episodio Depressivo Maggiore (In Remissione Parziale)".*

Infine, questa edizione divide i vari specificatori in tre classi distinte: una riguardante lo stato clinico e le caratteristiche dell'episodio quando i criteri sono soddisfatti, una per quando i criteri non sono soddisfatti e una per indicare lo schema degli episodi quando questi sono ricorrenti.

Per quanto riguarda lo stato clinico, se i criteri per l'Episodio Depressivo Maggiore sono soddisfatti va specificato se questo è lieve, moderato, grave, con o senza caratteristiche psicotiche, cronico. Se i criteri non sono soddisfatti occorre specificare se questo è dovuto ad una remissione parziale o totale. L'Episodio Depressivo Maggiore inoltre può avere caratteristiche Catatoniche, Melanconiche, Atipiche o con Insorgenza Post-Partum, tali caratteristiche sono uguali sia nel caso in cui i criteri vengano o non vengano soddisfatti. Nel momento in cui i criteri non vengono soddisfatti le caratteristiche faranno riferimento all'episodio più recente. Per gli episodi ricorrenti invece bisogna indicare se questi hanno un andamento stagionale o se tra due episodi si è verificato un recupero completo oppure no (American Psychiatric Association, 1994).

1.3 Il Disturbo Depressivo Maggiore nel DSM-5

Come mostrato nel paragrafo precedente, la depressione è stata sempre presente nel DSM sin dalla sua prima versione (DSM-I) degli anni '50. Nell'evoluzione ha subìto diversi cambiamenti nella nomenclatura, nella classificazione e nei criteri diagnostici. Il termine Disturbo Depressivo Maggiore è comparso per la prima volta nel DSM-IV ed è stato mantenuto nella versione più recente pubblicata nel 2013, ovvero il DSM-5 (American Psychiatric Association, 2013).

Nell'attuale versione, il DDM appartiene alla categoria "Disturbi Depressivi" che, a differenza delle precedenti versioni, non fa più parte dei "Disturbi dell'Umore" in quanto questi sono stati dicotomizzati in due categorie nosografiche distinte: "Disturbi Depressivi" e "Disturbo Bipolare e disturbi correlati".

Per essere diagnosticato, il DSM-5 (American Psychiatric Association, 2013), richiede la soddisfazione dei seguenti criteri:

Criterio A: Cinque o più dei seguenti sintomi sono stati contemporaneamente presenti durante un periodo di 2 settimane e rappresentano un cambiamento rispetto al precedente livello di funzionamento; almeno uno dei sintomi è 1) umore depresso o 2) perdita di interesse o piacere.
Nota: Non comprende sintomi chiaramente attribuibili a un'altra condizione medica.

1) *Umore depresso per la maggior parte del giorno, quasi tutti i giorni, come riportato dall'individuo (per es. si sente triste, vuoto/a, disperato/a) o come osservato da altri (per es., appare lamentoso/a). (Nota: nei bambini e negli adolescenti l'umore può essere irritabile.)*
2) *Marcata diminuzione di interesse o piacere per tutte, o quasi tutte, le attività per la maggior parte del giorno, quasi tutti i giorni (come indicato dal resoconto soggettivo o dall'osservazione).*
3) *significativa perdita di peso, non dovuta a dieta, o aumento di peso (per es., un cambiamento superiore al 5% del peso corporeo in un mese) oppure diminuzione o aumento dell'appetito quasi tutti i giorni. (Nota: Nei bambini, considerare l'incapacità di raggiungere i normali livelli ponderali.)*
4) *Insonnia o ipersonnia quasi tutti i giorni.*
5) *Agitazione o rallentamento psicomotori quasi tutti i giorni (osservabile dagli altri, non semplicemente sentimenti soggettivi di essere irrequieto/a o rallentato/a).*
6) *Faticabilità o mancanza di energia quasi tutti i giorni.*
7) *Sentimenti di autosvalutazione o di colpa eccessivi o inappropriati (che possono essere deliranti), quasi tutti i giorni (non semplicemente autoaccusa o sentimenti di colpa per il fatto di essere ammalato/a).*
8) *Ridotta capacità di pensare o di concentrarsi, o indecisione, quasi tutti i giorni (come impressione soggettiva o osservata da altri).*
9) *Pensieri ricorrenti di morte (non solo paura di morire), ricorrente ideazione suicidaria senza un piano specifico o un tentativo di suicidio o un piano specifico per commettere suicidio.*

Criterio B. I sintomi causano disagio clinicamente significativo o compromissione del funzionamento in ambito sociale, lavorativo o in altre aree importanti.

Criterio C. L'episodio non è attribuibile agli effetti fisiologici di una sostanza o a un'altra condizione medica.

Nota: I Criteri A-C costituiscono un episodio depressivo maggiore.

Nota: Risposte a una perdita significativa (per es. lutto, tracollo finanziario, perdite derivanti da un disastro naturale, una grave patologia medica o disabilità) possono comprendere sentimenti di intensa tristezza, ruminazione sulla perdita, insonnia, scarso appetito e perdita di peso, annotati nel Criterio A, che possono assomigliare a un episodio depressivo. Nonostante tali sintomi possano essere comprensibili oppure considerati appropriati alla perdita, la presenza di un episodio depressivo maggiore in aggiunta alla normale risposta ad una perdita significativa dovrebbe essere considerata attentamente. Questa decisione richiede inevitabilmente una valutazione clinica basata sulla storia dell'individuo e sulle norme culturali per l'espressione del disagio nel contesto della perdita.

Criterio D. Il verificarsi dell'episodio depressivo maggiore non è meglio spiegato dal disturbo schizoaffettivo, dalla schizofrenia, dal disturbo schizofreniforme, dal disturbo delirante o dal disturbo dello spettro della schizofrenia e altri disturbi psicotici con altra specificazione o senza specificazione.

Criterio E. Non vi è mai stato un episodio maniacale o ipomaniacale.

Nota: Tale esclusione non si applica se tutti gli episodi simil-maniacali o simil-ipomaniacali sono indotti da sostanze o sono attribuibili agli effetti fisiologici di un'altra condizione medica.

Procedure di codifica e registrazione
Il codice diagnostico per il disturbo depressivo maggiore è basato sulla presenza di un episodio singolo o ricorrente[1], sulla gravità attuale, sulla presenza di caratteristiche psicotiche e sullo stato della remissione. La gravità attuale e le caratteristiche psicotiche sono indicate solo se attualmente sono soddisfatta tutti i criteri per un episodio depressivo maggiore. Gli specificatori di remissione sono indicati solo se attualmente non sono soddisfatti tutti i criteri per un episodio depressivo maggiore.

Specificatori di gravità/decorso
Lieve
Moderato
Grave
Con Caratteristiche psicotiche[2]
In remissione parziale
In remissione completa
Non specificato
Nel registrare il nome di una diagnosi, i termini dovrebbero essere elencati nel seguente ordine: disturbo depressivo maggiore, episodio singolo o ricorrente, specificatori di gravità/caratteristiche psicotiche/remissione, seguiti da tutti i seguenti specificatori [...].

Specificare:
Con ansia
Con caratteristiche miste
con caratteristiche melancoliche
con caratteristiche atipiche
Con caratteristiche psicotiche congruenti all'umore
Con caratteristiche psicotiche non congruenti all'umore
Con catatonia
Con esordio nel peripartum
Con andamento stagionale

La diagnosi di DDM è caratterizzata da una grave differenza di genere, infatti il doppio delle donne risultano soffrire di depressione rispetto agli uomini (Salk et al., 2017). Nel DSM-5 tale prevalenza del disturbo nei soggetti di sesso femminile viene riscontrata da 1,5 a 3 volte rispetto ai soggetti di sesso maschile, con una discrepanza maggiore nella prima adolescenza (American

[1] *Perchè un episodio sia considerato ricorrente, deve esservi un intervallo di almeno 2 mesi consecutivi tra episodi separati in cui non sono soddisfatti i criteri per un episodio depressivo maggiore.*
[2] *Se sono presenti caratteristiche psicotiche, codificare con caratteristiche psicotiche, indipendentemente dalla gravità dell'episodio.*

Psychiatric Association, 2013). Tale discrepanza di genere sembra variare con lo sviluppo e il decorso del disturbo: nella metanalisi condotta da Salk et al (2017) è stato riscontrato un picco di differenza di genere nelle diagnosi durante l'adolescenza (OR=2,37 a 12 anni), che va poi a restringersi nell'età adulta (OR=1,95).

Questo divario è stato riscontrato soprattutto in nazioni caratterizzate da una parità di genere maggiore e potrebbe essere spiegato dallo strumento di misura utilizzato per riscontrare la presenza della depressione, ovvero un questionario di auto-valutazione (Wood & Eagly, 2012). Rispetto ad una misura oggettiva, come per esempio il rendimento in una materia scolastica, che mostra una minore differenza di genere nelle nazioni con maggior parità (Else Quest et al., 2010), i questionari di auto-valutazione soggettivi usati nella misurazione delle condizioni psicologiche (e.g. la depressione) hanno mostrato un aumento della differenza di genere nelle popolazioni con maggiore parità (Wood & Eagly, 2012). Secondo Wood e Eagly (2012) Una spiegazione potrebbe essere che il soggetto per rispondere alle domande sul sé, fa una stima della sua condizione scegliendo un insieme di persone con cui confrontarsi; pertanto le variazioni nei processi di confronto sociale tra le varie culture possono spiegare il divario di genere (Guimond et al., 2007). Secondo Guimond e coll. (2007) in contesti in cui c'è maggiore parità, come in Occidente, è più probabile che l'individuo metta in atto un confronto con soggetti dell'altro gruppo (inter-gruppo) piuttosto che del proprio (intra-gruppo). Secondo gli autori, le femmine si confronterebbero maggiormente con i maschi piuttosto che con le altre femmine (auto-stereotipizzazione), portando a differenze di genere maggiori nelle autovalutazioni. In contesti con una bassa parità di genere, essendo le relazioni tra i gruppi limitate, si tende a mettere in atto un confronto intra-gruppo con una conseguente minore differenza di genere su una variabile come la depressione.

Questa differenza di genere viene spesso spiegata anche attraverso fattori psicologici e sociali, Nolen-Hoeksema e Hilt (2009) ritengono che ci sia una prevalenza di Depressione nelle femmine in quanto queste sono esposte ad un maggior tasso di eventi di vita avversi:

- il maggior rischio di subire abusi sessuali infantili;
- lo stress cronico dato dallo status sociale della donna che rispetto all'uomo ha una maggiore probabilità di minor guadagno;
- il maggior rischio di subire molestie sul lavoro,
- il farsi carico della cura di familiari malati ecc.

In realtà, la percentuale maggiore di depressione nelle femmine non può essere spiegata solamente dai tassi di eventi di vita stressanti (Kendler et al., 2001), in quanto l'insorgenza del disturbo depressivo è attribuibile a diversi fattori di rischio: genetici, neurobiologici, personologici e

cognitivi che, interagendo tra di loro o con una varietà di fattori ambientali, aumentano la probabilità di esordio del disturbo (Brigitta, 2002).

In riferimento ai fattori genetici, *"i familiari di primo grado di individui con disturbo depressivo maggiore hanno un rischio di sviluppare il disturbo da due a quattro volte maggiore rispetto alla popolazione generale [...]. L'ereditarietà è di circa il 40%"* (American Psychiatrich Association, 2013). Studi recenti di genetica molecolare hanno individuato come la presenza nell'individuo di una o due copie dell'allele corto del polimorfismo del gene 5-HT (5-idrossitriptamina), trasportatore della serotonina, modera le reazioni psicopatologiche alle esperienze stressanti. Un soggetto portatore di tale allele, rispetto ad uno non portatore, ha quindi un maggior rischio di sviluppare il disturbo in presenza di un evento stressante poiché uno stimolo di paura comporta una maggiore attività neuronale dell'amigdala. La genetica da sola non può causare lo sviluppo di un disturbo: in questo caso infatti ci deve essere un'interazione gene-ambiente, in cui la risposta dell'individuo agli eventi stressanti è moderata dalla genetica (Caspi et al., 2003).

Le cause neurobiologiche della depressione consistono in un deficit dei sistemi modulatori diffusi. Un'ipotesi molto accreditata è quella monoaminergica, secondo la quale un difetto funzionale dei neurotrasmettitori monoaminergici cerebrali (noradrenalina, dopamina e serotonina) comporta una carenza di questi e di conseguenza la comparsa dei sintomi depressivi nel soggetto. Questi neurotrasmettitori sono i responsabili di sintomi comportamentali come agitazione o ritardo, vigilanza, affaticamento e psicomotricità, umore e motivazione (Brigitta, 2002). Un'altra ipotesi avvalorata nel tempo riguarda l'asse ipotalamo-ipofisi-surrene (HPA) che riveste un ruolo centrale nella moderazione delle risposte allo stress attraverso la regolazione di un insieme di ormoni tra cui il cortisolo. Nello specifico, l'iper-attivazione dell'amigdala, in risposta a stimoli stressanti, invia segnali all'ipotalamo che secerne un ormone, il quale, a cascata, innesca il rilascio di diversi ormoni lungo l'asse, fino ad arrivare al rilascio di cortisolo da parte delle ghiandole surrenali (Nolen-Hoeksema & Hilt, 2009). Il cortisolo, attraverso l'attivazione di particolari recettori ippocampali per i glucocorticoidi, dovrebbe inibire l'asse HPA, ma questo non avviene nei pazienti depressi che di conseguenza sono caratterizzati da un'iper-attivazione dell'asse. Anche qui l'ambiente gioca un ruolo importante: gli studi sui ratti hanno mostrato come la mancanza di inibizione è spiegata delle precoci esperienze sensoriali negative che hanno comportato lo sviluppo di un numero di recettori minore rispetto a quelli necessari all'animale per far fronte allo stress (Bear et al, 2007, p. 801-802). Inoltre, la presenza elevata di cortisolo, chiamato anche "ormone dello stress", prodotta dall' iper-attivazione dell'asse comporta cambiamenti nel sistema nervoso centrale: in particolare è associata ad una riduzione del volume dell'ippocampo che rende il soggetto vulnerabile allo sviluppo della depressione e alle recidive (Holsboer, 2000).

Alla base dell'eziologia del disturbo possono esserci anche fattori personologici (o temperamentali) e fattori cognitivi. Il Nevroticismo, per quanto riguarda i fattori personologici, è un tratto della personalità che consiste nel tendere a vivere emozioni negative frequenti ed intense aumentando così il rischio di insorgenza del disturbo. Invece, i fattori cognitivi che aumentano la vulnerabilità sono i pensieri e le convinzioni negative che il soggetto ha sviluppato durante l'infanzia. Secondo la teoria cognitiva della depressione di Aaron Beck, gli schemi di pensiero negativi sono convinzioni al di sotto del livello della coscienza appresi a seguito di eventi stressanti nell'infanzia: si attivano quando il soggetto vive una situazione simile e comportano una elaborazione distorta dell'evento (*cognitive biase*s). Il soggetto tende a leggere la situazione negativamente e a porre l'attenzione sui feedback negativi tralasciando quelli positivi, andando così ad alimentare le proprie convinzioni disfunzionali in un circolo vizioso in cui le conclusioni che trae dall'esperienza confermano gli schemi di base mantenendoli attivi. Tale visione negativa non coinvolge solo il sé, ma anche l'idea che il soggetto ha del mondo e del futuro, portandolo a sviluppare la depressione (Kring et al., 2016/2017).

I fattori ambientali, come eventi avversi o contesti ambientali negativi durante l'infanzia (abuso, mancanza di supporto sociale, critiche e rimproveri), sono un importante fattore di rischio per lo sviluppo del disturbo mentre gli eventi stressanti vissuti dal soggetto in prossimità dell'episodio depressivo (come perdite ed umiliazioni), fungono da fattori scatenanti del disturbo soprattutto nei soggetti maggiormente vulnerabili (Kring et al., 2016/2017).

Nel DSM-5 tra i fattori di rischio troviamo anche la presenza di altri disturbi in quanto diverse condizioni mediche o psicologiche possono aumentare il rischio di sviluppare la depressione. Le condizioni mediche più comuni sono il diabete, l'obesità patologica e le malattie cardiovascolari, mentre quelle psicologiche sono i disturbi da uso di sostanze, di personalità borderline e di ansia, infatti si trovano spesso in comorbidità con il disturbo depressivo oltre al disturbo correlato a sostanze, di panico, ossessivo-compulsivo, anoressia nervosa, bulimia nervosa (American Psychiatric Association, 2013).

In conclusione, tutti i fattori di rischio sopra citati costituiscono diatesi, ovvero una vulnerabilità che rende il soggetto maggiormente a rischio di sviluppo del disturbo quando questo vive un evento stressante, come spiegato dal modello diatesi-stress a cui fanno riferimento tutte le teorie più accreditate (Kring et al., 2016/2017).

1.3.2 La valutazione del disturbo

La valutazione del disturbo avviene attraverso un esame psicodiagnostico, che lo psicologo svolge con diverse funzioni a seconda dei contesti. Nella clinica per esempio, l'esame psicodiagnostico viene effettuato con lo scopo di intraprendere un percorso di sostegno psicologico o di psicoterapia o per valutare le condizioni psicologiche di un soggetto in particolari contesti come il cambio di sesso o l'adozione di un bambino (Sanavio & Sica, 1997). Un altro ambito in cui vengono effettuate valutazioni psicodiagnostiche è quello peritale in cui si va a valutare le componenti psicologiche della condotta criminale (Sanavio & Sica, 1997), o comunque in generale i fattori psicologici rilevanti ai fini della valutazione giudiziaria. Le valutazioni in questo ambito infatti possono essere svolte in diversi ambiti come quello del "danno psichico", l'affidamento del minore in caso di separazione, la condizione di infermità di mente in casi di circonvenzione di incapace e molti altri (Sammicheli, 2019).

Il colloquio clinico è l'asse portante della valutazione psicodiagnostica, ma viene integrato e supportato dai test psicodiagnostici: strumenti che aiutano il professionista a confermare o rigettare un'ipotesi diagnostica (Sanavio & Sica, 1997). Uno studio condotto da Regier e coll. (2013) ha indagato la concordanza tra diversi valutatori nel diagnosticare un disturbo su un soggetto con il solo uso dei criteri definiti dal DSM-5. Ciò che è emerso è come il DDM, a differenza di diagnosi come il Disturbo da Stress Post-Traumatico, mostra un'accuratezza discutibile con un accordo tra valutatori pari a 0,20-0,30, a causa dell'eterogeneità dei pazienti che soddisfano questi criteri diagnostici e della loro comorbilità con altri disturbi.

Fare una diagnosi di DDM non è quindi molto semplice, anche perché è raro che il paziente mostri l'intero pattern di sintomi, così come sono definiti nel DSM-5, e lo riferisca allo specialista. Molto più frequenti sono le lamentele somatiche, riportate dal paziente, che nascondono la depressione (depressione mascherata) o la depressione in comorbidità con altri disturbi che rende difficile una corretta diagnosi differenziale. Il Disturbo d'Ansia, per esempio, si trova molto spesso in compresenza del disturbo depressivo e condivide con questo alcuni sintomi, come il sonno disturbato, che rendono difficile la diagnosi differenziale e il conseguente trattamento. Infatti, se viene diagnosticato un disturbo piuttosto che un altro, potrebbero venire assegnati al soggetto trattamenti psicoterapeutici e/o farmacologici non specifici per il suo problema e questo potrebbe comportare un aggravamento della condizione fino ad arrivare ad esiti come il suicidio, una delle gravi conseguenze della depressione non trattata (Balsamo & Saggino, 2007). La diagnosi di disturbo depressivo, così come tutte le diagnosi psichiatriche, è sindromica: ovvero effettuata in base alla presenza di un insieme di segni e sintomi riferiti dal paziente durante l'anamnesi e l'esame oggettivo (Goldman et

al., 1999). Non essendoci quindi test fisiologici o di laboratorio che individuano la presenza del disturbo, come invece è possibile fare per molte malattie mediche, la diagnosi di un disturbo psicologico si basa su un colloquio condotto da un intervistatore addestrato a valutare se il soggetto soddisfa i criteri stabiliti dal manuale di riferimento più aggiornato, in questo caso il DSM-5 (Williams et al., 2002), aiutato da strumenti di screening e di diagnosi oggettivi (Goldman et al., 1999).

Essendo il disturbo depressivo uno tra i disturbi psichiatrici più diffusi al mondo, negli anni sono stati sviluppati un numero considerevole di strumenti per la valutazione del disturbo depressivo (Furukawa, 2010). Le scale di valutazione ad oggi disponibili possono essere raggruppate in due categorie: interviste semi-strutturate e autosomministrate.

Per quanto riguarda le prime un valutatore addestrato, a seguito delle informazioni raccolte attraverso l'intervista semi-strutturata, compila la scala di valutazione corrispondente.

Una scala di valutazione storica, usata per misurare la gravità della depressione nei pazienti affetti dal disturbo è la *Hamilton Rating scale for depression* (HAM-D) (Hamilton, 1960), costituita da 21 *item*. Di questi, solo i primi 17 costituiscono punteggio poiché gli ultimi 4, secondo l'autore, si riscontrano nel soggetto troppo raramente e hanno così una funzione qualitativa più che quantitativa. Il periodo di riferimento sono gli ultimi 7 giorni e l'intervallo di punteggio ottenibile dalla somma degli item va da 0 a 50, con la scala likert con opzioni variabili da 0-4 a 0-2 per item, più difficili da quantificare in maniera affidabile (Furukawa, 2010). Gli intervalli consigliati per definire la gravità della depressione sono 4: 0-7 assenza di depressione, 8-16 depressione lieve, 17-23 moderata e grave se maggiore o uguale a 24 (Zimmerman et al., 2013).

L'intervista semi-strutturata (*Structured Interview Guide for the Hamilton Depression Rating Scale*) da seguire nel colloquio clinico per compilare al meglio la scala di valutazione è stata redatta da Williams (1988). Oggi è disponibile anche una versione aggiornata dell'*Hamilton Rating Scale for Depression*, sviluppata dallo stesso autore nel 2008 (Williams & Kobak, 2008), chiamata GRID-HAMD, contenente la guida per l'intervista strutturata. Sebbene abbia rappresentato per molti anni un *gold-standard* della valutazione clinica, tralascia la valutazione di alcuni criteri diagnostici fondamentali del disturbo depressivo come l'anedonia, mancanza di concentrazione e assenza di reattività dell'umore (Maier et al., 1988).

Un'altra scala di valutazione molto utilizzata per misurare la gravità della depressione nei pazienti con depressione maggiore è la *Montgomery-Asberg Depression Rating Scale* (Montgomery et al., 1969). Montgomery e coll. hanno costruito questa scala a partire da 65 item della *Comprehensive Psychopathology Rating Scale,* fino ad ottenere i 10 item più comunemente riscontrati nella sintomatologia depressiva e più sensibili al cambiamento. Infatti, questa scala è nata con lo scopo di

evidenziare l'efficacia del trattamento rispetto al placebo, oltre a monitorare la gravità dei sintomi. Il periodo di tempo a cui fare riferimento sono gli ultimi 7 giorni e le risposte vengono assegnate in una scala likert che va da 0 a 6, con un punteggio massimo di 70 (Montgomery et al., 1969). Un'interpretazione dei punteggi stabilita da Bandelow e coll (2006) attraverso un'analisi post-hoc di cinque studi, stabilisce un punteggio di 11 come indicativo di malattia borderline, 19 malattia lieve e 29 malattia moderata. Recentemente Williams e Kobak (2008) hanno sviluppato la guida per l'intervista strutturata (SIGMA), usata dai professionisti per raccogliere tutte le informazioni necessarie al fine di valutare i diversi item della scala.

Molto utilizzate nella pratica clinica sono anche le misure di autovalutazione in cui, a differenza delle scale di valutazione viste precedentemente, è il soggetto stesso a compilarle facendo riferimento alla sua condizione.

Il *Beck Depression Inventory-II* (BDI-II) (Beck et al.1996), sin dalla sua prima versione sviluppata nel 1961, è probabilmente lo strumento di autovalutazione più utilizzato per rilevare la depressione nella popolazione generale e per misurarne la gravità nei pazienti con diagnosi, adulti ed adolescenti (Furukawa,2010). Lo strumento consiste in 21 item, costruiti sui criteri diagnostici per il Disturbo Depressivo Maggiore del DSM-IV, a cui il soggetto risponde facendo riferimento al modo in cui si è sentito nelle ultime due settimane; ogni item corrisponde ad un sintomo della depressione e viene valutato dal soggetto su una scala likert da 0 a 3, ad eccezione degli item 16 e 18 in cui la scala di risposta va da 0 a 7, per meglio evidenziare l'aumento o la diminuzione di sonno ed appetito (Beck et al., 1996). Secondo le linee guida fornite dagli autori un punteggio di 14-19 è indicativo di depressione lieve, 20-28 moderata e 29-63 grave. Del BDI-II è disponibile anche il manuale adattato e validato in italiano da Ghisi e coll. nel 2006.

Zung (1965) ha sviluppato la *Zung Self-Rated Deprression Scale* (SDS) con lo scopo di rilevare la presenza e la gravità dei sintomi depressivi e la sua variazione durante il trattamento; la scala è composta da 20 item, 10 formulati positivamente e 10 negativamente, a cui il soggetto risponde attraverso una scala likert da 1 a 4 punti (un po' del tempo - la maggior parte del tempo). Nelle domande negative il punteggio va invertito, il punteggio massimo è 80 e sono stati indicati 4 range di gravità: 50-59 depressione lieve, 60-69 moderata, 70 o maggiore grave (Behera et al., 2017). La versione italiana della scala è stata testata in un piccolo studio da Innamorati e coll. (2006) che non ha ottenuto risultati soddisfacenti, in quanto poco attendibile e con scarsa capacità di discriminare tra ansia e depressione.

Il *Center for Epidemiological Studies-Depression* (CES-D) è un questionario sviluppato da Radloff (1977), una ricercatrice del *Center for Epidemiological Studies* del *National Institute of Mental Health* (NIMH) e proprio per questo motivo nasce prevalentemente per le valutazioni

epidemiologiche e di screening. Il questionario è composto da 20 item, a cui il soggetto risponde attraverso una scala likert da 0 a 3 (raramente o nessuna delle volte, meno di un giorno - la maggior parte o sempre, 5-7 giorni) (Radloff, 1977). Un punteggio, ottenuto dalla somma delle risposte, pari a 24 funge da allarme clinico mentre un punteggio superiore è indicativo di un possibile stato depressivo che necessita di approfondimento. Non misura quindi la gravità né viene usato nelle valutazioni diagnostiche (Sanavio & Sica, 1997). La validazione italiana dello strumento è stata condotta da Pierfederici e coll. (1982), i quali hanno evidenziato che il valore soglia di 24 per l'individuazione del soggetto depresso andrebbe elevato a 29 nei pazienti ospedalizzati con malattie somatiche, mentre può essere mantenuto per la popolazione generale (Sanavio & Sica, 1997). Nel 2004 William e coll. hanno sviluppato una revisione della scala, adattandola ai criteri diagnostici del DSM-IV, in cui la scala likert diventa a 5 punti allungando la presenza dei sintomi a *"quasi ogni giorno per due settimane"*, come stabilito dal manuale.

Il *Patient Health Questionnaire-9* (PHQ-9) (Kroenke et al., 2001) è un nuovo strumento utilizzato con il duplice scopo di diagnosi e rilevazione della gravità della depressione in contesti di cure primarie. Il questionario fa parte di un questionario più completo, il *Patient Health Questionnaire* (PHQ) (Spitzer et al. 1999), totalmente autosomministrabile, che indaga 8 disturbi frequentemente riscontrati nei contesti di cure primarie. A sua volta quest'ultimo deriva dal PRIME-MD, uno strumento valutativo composto da 2 parti: un questionario autosomministrato e la guida alla valutazione del clinico (Spitzer et al., 1999), un connubio quindi delle due categorie di strumenti descritte in questo paragrafo. Il PHQ-9 è così definito perché è composto da 9 elementi, ognuno dei quali corrisponde ad un criterio del Disturbo Depressivo Maggiore del DSM-IV: ogni sintomo può essere valutato dal soggetto con una risposta a scala likert che va da 0 (per niente) a 3 (quasi ogni giorno), facendo riferimento alle ultime due settimane (Kroenke et al. 2001). Come per i criteri diagnostici del DSM-IV, la depressione maggiore viene diagnosticata se il soggetto risponde "più della metà dei giorni" a 5 o più item, mentre per quanto riguarda la gravità sono indicati 4 range: 0-4 assenza di disturbo, 5-9: disturbo lieve, 10-14 moderato, 15-19 moderatamente grave, 20-27 grave (Kroenke et al. 2001). Se si avesse bisogno di un singolo cut-off di screening per rilevare la presenza di disturbo depressivo Kroenke & Spitzer (2002) raccomandano un punteggio uguale o superiore a 10, con una sensibilità e specificità per la depressione maggiore pari a 88%. Partendo dal presupposto che i risultati ottenuti attraverso questo strumento non variano tra sesso, tipologia di pazienti e background linguistico (Teymoory et al., 2020), il PHQ-9 è stato validato nella popolazione italiana da Picardi e coll. (2005; 2006) come strumento di screening della depressione nei soggetti affetti da malattie della pelle. Inoltre, è possibile trovare il questionario da compilare in molte lingue nel sito web di Pfizer (http://www.phqscreeners.com) insieme al manuale di istruzioni, in cui un cut-off di 10

viene considerato come una “bandiera gialla del disturbo” (ovvero indica una possibile condizione clinica significativa).

Inoltre, tutte queste scale, ad eccezione del CES-D che viene usato per le valutazioni epidemiologiche, possono essere utilizzate per misurare il cambiamento del disturbo in corso di trattamento (Furukawa, 2010; Zung, 1965).

1.3.3 Trattamento del disturbo

Il DDM può essere trattato principalmente in due modi: trattamento farmacologico e psicologico. Il trattamento combinato, ovvero la somministrazione contemporanea dei due trattamenti, è risultato essere quello più efficace soprattutto nei soggetti con DDM ricorrente e sintomi più gravi, rispetto ai soggetti con sintomi più lievi, per i quali la terapia combinata non si è dimostrata più efficace della sola terapia psicologica (Thase et al. 1997). La combinazione dei due trattamenti è molto importante in quanto, oltre a prevenire le recidive, migliorare il funzionamento del soggetto e di conseguenza la qualità della vita, facilita anche l’adesione al trattamento farmacologico. Infatti, nonostante i farmaci vengano utilizzati come trattamento principale nella depressione, studi epidemiologici hanno mostrato come l’aderenza sia un problema molto frequente, con un paziente su tre che non porta a termine il trattamento prescritto (Pampallona et al., 2004).

Per quanto riguarda il trattamento farmacologico, i farmaci antidepressivi agiscono sui neurotrasmettitori che precedentemente (vedi paragrafo 1.3) abbiamo visto essere implicati nell’eziologia della sintomatologia, ovvero dopamina, serotonina e noradrenalina. I primi farmaci antidepressivi sono comparsi circa 50 anni fa con la scoperta dell’imipramina, dando vita alla prima classe di farmaci definita antidepressivi triciclici (TCA); più o meno nello stesso periodo è stata scoperta un’altra classe di farmaci chiamati inibitori delle monoaminossidasi (MAOI), che insieme ai TCA fanno parte degli antidepressivi di prima generazione (Julien et al., 2011/2012). Questi, molto efficaci ma al tempo stesso con gravi effetti collaterali sulla memoria e sulle funzioni cognitive, negli anni ‘80-’90 sono stati sostituiti dagli antidepressivi di seconda generazione o atipici, equivalenti nell’efficacia ma meno tossici e meglio tollerati. I farmaci antidepressivi atipici sono chiamati inibitori selettivi della ricaptazione di serotonina (SSRI) e vi fanno parte: fluoxetina, paroxetina, sertralina, fluvoxamina, citalopram ed escitalopram. Sebbene siano meno tossici degli antidepressivi di prima generazione, non mancano gli effetti collaterali come per esempio la sindrome serotoninergica data da alti dosaggi di SSRI, la sindrome d’astinenza che può verificarsi in seguito alla cessazione improvvisa, le disfunzioni sessuali e il suicidio. Un’ altra classe di farmaci è quella

degli antidepressivi a duplice azione (*dual action*), come la venlafaxina e la mirtazepina, che oltre a bloccare la ricaptazione presinaptica della serotonina come gli SSRI, bloccano anche la ricaptazione della noradrenalina; dal punto di vista del meccanismo d'azione sono simili ai TCA ma gli effetti collaterali sono limitati. Un altro farmaco a duplice azione è il bupropione, che si distingue dagli altri poiché è un inibitore della ricaptazione della dopamina e noradrenalina e quindi non ha alcun effetto sui neurotrasmettitori serotoninergici, evitando tutti gli effetti collaterali associati ai farmaci SSRI (Buccellati et al., 2011/2012).

Dato il numero elevato di farmaci antidepressivi, lo studio clinico su scala internazionale STAR*D (*Sequenced Treatment Alternatives to Relieve Depression)* (Rush et al., 2009) è andato ad indagare quale fosse la strategia migliore di trattamento del DDM per migliorare la remissione della sintomatologia in soggetti affetti da tale disturbo. In questo studio i partecipanti sono stati trattati in 4 fasi, di 12 settimane ognuna, in cui ogni volta il paziente poteva scegliere se cambiare farmaco o aggiungere un nuovo farmaco a quello della fase precedente, mentre chi otteneva la remissione del disturbo attraverso il trattamento sperimentato in uno dei 4 livelli, continuava il trattamento. Nel corso delle varie fasi circa due terzi dei partecipanti hanno raggiunto la remissione, mostrando al follow-up una percentuale di recidive molto più bassa: questo ha confermato l'importanza di arrivare alla remissione per avere una prognosi migliore, anche ricorrendo a diversi trattamenti e strategie. Inoltre, tale studio ha mostrato come il trattamento è più efficace da un lato quando il farmaco iniziale viene assunto per un periodo più lungo di quello previsto per ottenere l'effetto e dall'altro quando, in assenza di miglioramento, il farmaco viene combinato con altri piuttosto che aumentato il dosaggio (Rush et al., 2009). I trattamenti combinati infatti, in particolare mirtazapina più venlafaxina (58%) e mirtazapina più fluoxetina (52%) sono clinicamente più efficaci della monoterapia con venlafaxina (25%), ma ben tollerati allo stesso modo (Blier et al., 2010). Inoltre, nella seconda fase dello studio STAR*D, Thase e coll. (2007) hanno paragonato l'efficacia della terapia cognitivo-comportamentale (CBT), da sola o in aggiunta al farmaco, e della farmacoterapia nei pazienti che nella prima fase non hanno risposto al trattamento con citalopram. Ciò che è emerso dallo studio è l'uguale efficacia delle due terapie (CBT e farmacologica) dal punto di vista del tasso di risposta e della remissione, mentre dal punto di vista temporale, i farmaci hanno mostrato una remissione del disturbo significativamente più rapida, ma la CBT è stata meglio tollerata rispetto al passaggio ad un diverso antidepressivo (Thase et al., 2007).

La CBT è uno degli approcci psicoterapeutici più studiati e validati ed è anche molto utilizzato nel trattamento della depressione (Jobst et al., 2016) e insieme alla Terapia interpersonale (IPT) sono considerate le più efficaci nel trattamento del DDM sia nelle fasi acute, che nel mantenimento (Parikh, 2009). La CBT sviluppata sulla teoria cognitiva-comportamentale di Beck, è un trattamento che

combina tecniche cognitive e comportamentali: il suo obiettivo è modificare le modalità di pensiero disadattive e i comportamenti correlati che mantengono la depressione (come il ritiro sociale e l'anedonia) (Parikh, 2009). La Psicoterapia Interpersonale (IPT) invece, è un'altra forma di psicoterapia breve che si focalizza sulle relazioni interpersonali attuali del soggetto e sul contesto sociale in cui è inserito. Insieme al soggetto viene effettuato un esame sui principali problemi interpersonali che possono aver contribuito allo sviluppo della depressione, che possono riguardare eventi di diversa natura come il cambiamento e conflitti di ruolo, deficit interpersonali, lutto (Parikh et al., 2009). Un'altra terapia con significative prove di efficacia è l'attivazione comportamentale (BA), utile nel contrastare l'assenza di rinforzo positivo da parte dell'ambiente causato da una mancata interazione con questo, quadro tipico nel disturbo depressivo. Consiste quindi nell'aumentare le attività piacevoli e le esperienze gratificanti, così da avere un rinforzo positivo e diminuire l'attenzione del soggetto sui pensieri negativi (Parikh et al., 2009). Un trattamento psicologico applicabile anche dai membri dell'assistenza sanitaria di base, quindi che non necessita di un professionista con specifiche competenze specialistiche come per i trattamenti visti fino ad ora, è la terapia del problem solving (PST), in cui si aiuta il soggetto ad usare le proprie capacità e risorse per affrontare i problemi attuali e futuri che causano i sintomi depressivi. Il soggetto quindi identifica i problemi, decide su quale focalizzarsi, sviluppa diverse soluzioni, ne sceglie una, stila le diverse fasi per metterla in atto e ne valuta i progressi (Mynors-Wallis et al. 1995).

Nella prevenzione delle ricadute è stata sviluppata la Terapia Cognitiva Basata sulla Mindfulness (MBCT), che aiuta i soggetti con DDM a prendere consapevolezza dei loro processi di pensiero negativi con un'attenzione limitata al presente e non giudicante, allenandosi così ad osservare i pensieri negativi senza farsi coinvolgere da questi. In tal modo il soggetto impara a riconoscere quando si sente depresso e di conseguenza a mettere in atto una prospettiva decentrata, che consideri i pensieri come "eventi mentali" senza più associarli a stati d'animo tristi e senza più considerarli un riflesso accurato della realtà (Segal et al. 2018).

Una meta-analisi che confronta l'efficacia del trattamento psicologico della depressione ha trovato che la CBT, l'IPT e la PST siano le terapie con maggiori effetti in ampi studi, mentre l'effetto era minore per l'attivazione comportamentale (Barth et al., 2016).

Capitolo II

LA SIMULAZIONE

"Any neuropsychological evaluation that does not include careful consideration of the patient's motivation to give their best effort should be considered incomplete"
(Iversion, 2003[3])

Nel capitolo precedente, all'interno della valutazione della depressione, veniva esposto il problema della difficoltà della diagnosi del DDM essendo essa sindromica, ovvero basata sui riferiti del paziente. Proprio per questo motivo, la depressione, così come i disturbi psichici in generale, può essere facilmente simulata, poiché il soggetto può modificare la descrizione che dà allo specialista della propria condizione (Ferracuti et al, 2007), così come può manipolare il punteggio ad un questionario somministratogli, alterando le risposte agli item (Stracciari et al., 2010). Inoltre, Stracciari e coll. (2020) sostengono come un'ulteriore difficoltà nasca dal fatto che più i sintomi del disturbo sono "intuibili" più è facile la loro simulazione e aggravamento. Essendo quindi difficile far si che il soggetto non alteri la propria condizione, negli ultimi anni sempre più studi si sono concentrati sulla simulazione e la sua detezione: nella decade 1990-2000, su 139 articoli di neuropsicologia forense, 120 erano incentrati sulla simulazione (89%), con il trauma cranico come categoria diagnostica simulata più studiata (45%) (Sweet et al., 2002).

A partire dall'articolo appena citato, possiamo apprezzare come lo studio della simulazione avvenga soprattutto nell'ambito forense piuttosto che in quello clinico, poiché a differenza di quest'ultimo, è caratterizzato dalla presenza di vantaggi e svantaggi. Il soggetto infatti, potendo ottenere o meno un risarcimento per un danno subito, così come ottenere o meno una condanna minore durante un processo penale, è maggiormente spinto a produrre una rappresentazione di sé alterata in conformità all'obiettivo che deve raggiungere. Il professionista psicologo o neuropsicologo che si trova quindi a valutare il soggetto in questa tipologia di contesti (legali, previdenziali, assicurativi) sarà maggiormente esposto a tentativi di simulazione (Stracciari et al., 2010), con una prevalenza di circa il 15,7% rispetto al 7,4% rilevata nei setting clinici. (Rogers et al., 1994). Inoltre, lo studio di questa si è concentrato maggiormente nel setting forense poiché in quello clinico la simulazione non è un problema, anzi può essere un sintomo che va tenuto in considerazione nella valutazione e nell' assessment psicodiagnostico (come nel caso del disturbo fittizio e della sindrome di Munchausen) (Monaro et al., 2018)

[3] Iverson, G. L. (2003). Detecting malingering in civil forensic evaluations. *Handbook of forensic neuropsychology*, 137-177.

2.1 Definizione, faking bad e faking good

Il DSM-IV-TR colloca la simulazione di malattia tra le "Ulteriori condizioni che possono essere oggetto di attenzione clinica" e così la definisce:

"La caratteristica fondamentale della simulazione è la produzione intenzionale di sintomi fisici o psicologici falsi o grossolanamente esagerati, motivata da incentivi esterni, come evitare il servizio militare, il lavoro, ottenere risarcimenti finanziari, evitare procedimenti penali, oppure ottenere farmaci. In alcune circostanze la simulazione può rappresentare un comportamento adattivo - per esempio fingere una malattia quando si è prigionieri del nemico in tempo di guerra" (APA, 2000).

Tale classificazione e definizione della simulazione viene riportata anche nel DSM-5, che, come l'edizione precedente, suggerisce di sospettare la presenza di simulazione quando si riscontra una combinazione dei seguenti quattro fattori:

1. Presentazione dei sintomi all'interno di un contesto medico-legale (e.g. invio del soggetto a valutazione clinica da parte di un magistrato o richiesta di autovalutazione da parte dell'individuo per cause legali che lo riguardano);
2. Eccessiva discrepanza tra la condizione lamentata dal soggetto e i dati oggettivi e le osservazioni;
3. Assenza di collaborazione nella valutazione clinica e rifiuto del regime terapeutico assegnato;
4. Paziente affetto da disturbo antisociale della personalità.

Andando ad analizzare la definizione di simulazione è possibile evidenziarne due caratteristiche fondamentali: produzione intenzionale e consapevole dei sintomi e presenza di vantaggi secondari o incentivi esterni, per ottenere i quali il soggetto ha prodotto i primi, motivato da un incentivo esterno, come per esempio il denaro. Nel suddetto manuale, viene sottolineato come queste due caratteristiche, oltre a permettere il riconoscimento della simulazione, consentono la diagnosi differenziale dai disturbi fittizi che vedremo nel prossimo paragrafo (APA, 2013). I disturbi fittizi, così come la simulazione, sono caratterizzati dalla produzione intenzionale di sintomi, e per questo sono facilmente confondibili, ma il vantaggio è quello di assumere il ruolo di malato (vantaggio intrapsichico) e non esterno o materiale (Stracciari et al., 2010). Secondo Rogers (2008), ciò che distingue il disturbo fittizio dalla simulazione è la motivazione, che nel caso del disturbo fittizio è inconscia mentre nella simulazione è conscia. Tuttavia, accertare la presenza/assenza di

simulazione sulla base della consapevolezza è un compito arduo, in quanto la simulazione può assumere diverse sfaccettature come vediamo di seguito.

Nella definizione data dal DSM-IV-TR e poi dal DSM-5, ed in particolare in: "*produzione intenzionale di sintomi fisici o psicologici falsi o grossolanamente esagerati",* emerge un altro aspetto importante della simulazione: il soggetto può produrre intenzionalmente i sintomi di un disturbo dapprima inesistenti, può esagerarne la gravità, oppure può simulare prolungandone la presenza (Stracciari et al., 2010). Resnick (citato in Rogers, 2008) definisce queste sfaccettature della simulazione come: simulazione completa, parziale e falsa imputazione. La cosiddetta simulazione "completa" consiste nella creazione di una sintomatologia o la grossolana esagerazione di questa per scopi secondari, ed è la più semplice da riconoscere. La simulazione "parziale" si ha quando il soggetto esagera i sintomi che esperisce o quando descrive sintomi passati come attuali, prolungando quindi la presenza di questi. La falsa imputazione consiste invece, nell' attribuzione intenzionale e consapevole dei sintomi realmente esperiti dal soggetto ad una causa diversa da quella reale, come nel caso dell'adolescente che per non seguire i corsi di matematica, attribuisce la sua mancanza di concentrazione ad un disturbo dell'apprendimento, invece di attribuirlo all'abuso di sostanze quotidiano.

Secondo Stracciari e coll. (2010) la simulazione può essere di due tipi: generalizzata e specifica. La prima consiste in un atteggiamento simulatorio generale in cui i sintomi simulati appartengono a diversi disturbi (ansia, depressione, deficit cognitivi) mentre la seconda, ovvero quella specifica, consiste nella simulazione di un preciso disturbo psichico, con sintomi circoscritti ed ascrivibili alla specifica diagnosi. Infine, un particolare tipo di simulazione, molto frequente nei contesti forensi, assicurativi o in generale medico-legali è quello del *coaching* (addestramento). Si ha questo tipo di simulazione quando il periziando, prima di essere sottoposto alla valutazione psicodiagnostica della controparte, viene preparato dall'esperto (che sia esso avvocato, medico, psicologo) a rispondere in modo tale da alterare il quadro clinico per raggiungere l'obiettivo, senza rendere la simulazione evidente.

Nella direzione opposta ma speculare alla simulazione, troviamo la dissimulazione definita *faking good*. Mentre nella simulazione il soggetto aggrava i sintomi o li crea da zero per mostrare un'immagine patologica di sé, nella dissimulazione nasconde i propri tratti negativi o migliora i sintomi presenti per mostrare un'immagine particolarmente positiva di sé (Mazza et al., 2020). In entrambi i casi c'è quindi un'intenzionalità ed un vantaggio secondario a motivare la produzione di una rappresentazione alterata della propria condizione: in una direzione, il vantaggio sta nell'accentuare i sintomi (*faking bad*) mentre nell'altra direzione il vantaggio è insito nel negare i sintomi (*faking good*).

Nello specifico *faking bad* è il costrutto della simulazione maggiormente studiato sia per la frequenza in ambito forense che per la sua rilevanza in termini economici (prevalentemente sociali/assistenziali) essendo spesso messo in atto per ottenere dei risarcimenti (Mazza et al., 2020). L'attuazione della simulazione nei contesti forensi, e la simulazione della depressione come nel caso di questo elaborato, può avvenire per diversi motivi: in un procedimento penale, per esempio, il soggetto può essere spinto a simulare per ottenere vantaggi processuali (es. una riduzione della pena per vizio di mente parziale/totale) o per ottenere benefici sulla modalità di detenzione (valutazione della compatibilità con il regime carcerario); nel contesto civile, invece, il soggetto potrebbe mentire per ottenere un risarcimento (danno alla persona) o per ottenere l'interdizione/inabilitazione (capacità di provvedere ai propri interessi); ancora, in ambito previdenziale, la motivazione potrebbe essere l'ottenimento della pensione di invalidità o di accompagnamento.

Negli stessi contesti è probabile imbattersi in tentativi di *faking good,* ma con motivazioni differenti: in un contesto penale, per esempio, il soggetto dissimula, ovvero tende a mostrarsi sotto una luce positiva, nella valutazione sulla pericolosità sociale, così da ottenere i vantaggi sulla modalità di sconto della pena, propri di chi non è considerato socialmente pericoloso; in corso di un procedimento civile la dissimulazione più frequente risiede nel contesto di capacità genitoriale, per ottenere l'affido dei figli o sulla capacità di provvedere ai propri interessi, ma al contrario della situazione precedente, con lo scopo di evitare l'interdizione o l'inabilitazione; il soggetto può essere spinto a dissimulare anche per ottenere il porto d'armi o riottenere la patente (Stracciari et al., 2010). Un contesto in cui si trova molto frequentemente il *faking good* è quello lavorativo, in cui la persona tende a dare una impressione di sé molto positiva con l'obiettivo di ottenere il lavoro, con una diffusione elevata del comportamento dissimulatorio pari a circa il 30-50% dei casi nella selezione del personale (Mazza et al., 2020).

2.2 Simulazione e diagnosi differenziale

La simulazione, come è illustrato nel paragrafo precedente, si rileva e si differenzia da altre forme psicopatologiche simili ad essa, grazie a due caratteristiche principali:

1. intenzionalità nella produzione e nell'esagerazione di sintomi del disturbo
2. presenza di incentivi o vantaggi secondari.

Nella valutazione del disturbo, soprattutto in ambito forense, è molto importante tenere conto di queste due caratteristiche sia per riconoscere la simulazione e quindi evitare di diagnosticare un disturbo mentale, in assenza del reale disturbo, solo per raggiungere obiettivi personali, sia per

effettuare una diagnosi differenziale che ci permette di distinguere questo comportamento strategico da forme psicopatologiche simili alla simulazione.

In questo ultimo caso, ciò che distingue la simulazione dalle altre forme psicopatologiche che condividono con questa alcune caratteristiche (come i disturbi fittizi) sono la motivazione (intenzionale/inconsapevole) ed i vantaggi (esterni/intrapsichici). Nel lavoro peritale, dove il controllo della simulazione e la sua detezione è uno degli aspetti primari della consulenza tecnico-scientifica, potrebbe essere comodo collocare il periziando in una tabella 2x2 come quella di seguito riportata (tabella 1.1), così da aiutare il professionista a riconoscere un probabile tentativo di simulazione.

Tab.1.1 *- Tabella utile al professionista per collocare il periziando e distinguere la simulazione dalle altre psicopatologie*

	Produzione intenzionale	Produzione inconsapevole
Vantaggio esterno e/o materiale	Simulazione ed esagerazione	Disturbo di conversione ed altre forme di disturbo somatoforme
Vantaggio intrapsichico	Disturbo fittizio	Tutte gli altri disturbi psicopatologici

Alla luce delle caratteristiche già menzionate vediamo di seguito le altre psicopatologie passate in rassegna da Stracciari e coll (2010), che devono essere escluse per poter parlare di simulazione:

- Il disturbo di conversione ed altre forme di disturbo somatoforme come il disturbo algico, disturbo di somatizzazione, disturbo somatoforme indifferenziato sono caratterizzati dalla produzione di sintomi fisici in assenza di condizioni neurologiche o mediche sottostanti (APA, 2013). Tali condizioni psicopatologiche potrebbero essere scambiate facilmente come simulazione poiché possono anche avere un vantaggio secondario (e.g. esonero da responsabilità), ma si differenziano da questa principalmente per la produzione inconsapevole del disturbo (Stracciari et al., 2010).
- I disturbi dissociativi sono caratterizzati da sintomi neuropsicologici come disorientamento, amnesia, disturbi dell'ideazione, deficit di ragionamento e comprensione, pseudodemenza che, come per i disturbi somatoformi, non sono spiegati da una perdita cognitiva o disfunzione cerebrale (APA, 2013). Come per i disturbi somatoformi quindi, questi si distinguono dalla simulazione, e vanno di conseguenza esclusi, per l'assenza di intenzionalità da parte del soggetto.

- Il disturbo fittizio è caratterizzato dalla messa in atto di sintomi fisici o psicologici falsi o autoinduzione di malattia, con un atteggiamento ingannevole palese (APA, 2013). Questa condizione è la più simile alla simulazione in quanto c'è una produzione intenzionale e consapevole di malattia con lo scopo di ottenere un vantaggio secondario. A distinguere quindi le due condizioni che sono tra loro molto simili è il tipo di vantaggio secondario: mentre nella simulazione la produzione intenzionale di sintomi è motivata da incentivi esterni e/o materiali, nel disturbo fittizio il vantaggio è intrapsichico, ovvero il soggetto finge per il piacere e /o bisogno psicologico di essere considerato malato o usufruire delle cure, in assenza di altri incentivi esterni (Stracciari et al. 2010).
- Altre categorie diagnostiche che non sono ben differenziabili dalla simulazione attraverso le due caratteristiche di intenzionalità e vantaggi secondari come le precedenti, ma che spaziano tra queste e vanno quindi valutate quando ci si trova di fronte alla possibilità di simulazione. Come evidenziato da Stracciari e coll (2010) tre in particolare sono da tenere in considerazione:
    1. Sindrome di Münchausen: condizione psichiatrica in cui il soggetto si causa intenzionalmente sintomi e disturbi fisici (anche procurandosi degli infortuni) per ricevere attenzione medica e condurre uno stile di vita incentrato sull'ospedalizzazione. Sono soggetti che passano da un ospedale all'altro e di medico in medico, arrivando anche ad avere spesso contenziosi con queste figure. Viene considerata una forma di grave e intrattabile del disturbo fittizio in quanto il soggetto con questa sindrome, a differenza di quello con disturbo fittizio, non si ferma alla simulazione o alla lamentela, ma si infligge realmente i disturbi di natura fisica
    2. Sindrome di Münchausen per procura (definita anche "disturbo fittizio per procura"): condizione psichiatrica simile alla precedente se non per il fatto che i sintomi e i disturbi fisici vengono perpetrati su un soggetto terzo. Però è importante sottolineare che, nel caso in cui si evidenzia un vantaggio esterno rilevante nell'indurre il disturbo su un'altra persona, come nel caso di un genitore che infligge uno stato di malattia al figlio per creare un danno all'ex compagno, si parla di simulazione per procura.
    3. Sindrome di Ganser: condizione psichiatrica molto simile alla demenza, caratterizzata da linguaggio assurdo ed evasivo, decadenza di ogni competenza semantica, amnesia grave, incapacità di ragionamento logico-deduttivo, ma con coscienza, comprensione e orientamento preservati. Tale condizione può far pensare alla simulazione in quanto alcune capacità che solitamente vengono perse insieme alle altre, vengono preservate, ma va presa in considerazione soprattutto nelle valutazioni in contesto carcerario, in

cui la situazione altamente stressogena, come può essere l'attesa di esecuzione, può aver comportato una perdita delle funzionalità cognitive superiori ma mantenuto le altre.

In ultimo Stracciari e coll. (2010) riportano un tipo di simulazione chiamata nevrosi da compenso o sindrome da indennizzo. Tale condizione è caratterizzata da una produzione o amplificazione strutturata dei sintomi neuropsicologici o psicopatologici con lo scopo di frode assicurativa, molto comune nell'ambito del danno psichico.

2.2.1 La simulazione della depressione nel contesto forense: il danno psichico

Come accennato nei paragrafi precedenti, un'applicazione pratica della simulazione della depressione è quella del danno psichico nel contesto forense. Prima di trattarla nello specifico, si ritiene necessario un breve excursus sulla definizione di danno psichico nel linguaggio giuridico e sulla sua valutazione nel contesto forense.

Il danno psichico è un sottocomponente del danno biologico, che costituisce uno dei tre tipi di danno alla persona, che rientra a sua volta nella categoria di danno non patrimoniale. Oltre al danno biologico, sono riconosciuti come danni alla persona:

- il danno esistenziale, definito come un peggioramento della qualità della vita a seguito della violazione di uno dei diritti riconosciuti dalla costituzione come inviolabili o di un interesse della persona che non costituisce diritto ma è meritevole di tutela;
- il danno morale, caratterizzato da sofferenza psicologica esperita dal soggetto a seguito di un reato, che deve essere presente per definizione (Gulotta, 2011).

Il danno biologico viene definito dalla Società Italiana di Medicina legale e delle Assicurazioni (S.I.M.L.A.) come *"menomazione permanente e/o temporanea dell'integrità psico-fisica della persona, comprensiva degli aspetti personali dinamico-relazionali, passibile di accertamento e di valutazione medico-legale ed indipendente da ogni riferimento alla capacità di produrre reddito [...]. Il danno biologico è nozione unitaria ed univoca, da valere in ogni ambito in cui per norma ne sia richiesta la stima: responsabilità civile, assicurazione sociale contro i rischi del lavoro, [...] assicurazione privata contro gli infortuni e le malattie e in ogni ambito di assistenza e previdenza sociale"* (Camerini, 2011). Come si evince dalla definizione, a costituire il danno biologico sono il danno biologico di natura fisica e il danno biologico di natura psichica (più brevemente chiamato danno psichico), che possono turbare l'esistenza del soggetto in maniera temporanea, per un determinato periodo espresso in giorni di invalidità, o permanente, ovvero per tutta la vita. L'invalidità temporanea o permanente può essere assoluta, quando il soggetto leso è

totalmente incapace di lavorare o di attendere alle sue ordinarie occupazioni, o parziale, ovvero quando il soggetto è impedito allo svolgimento del lavoro e delle ordinarie occupazioni in una percentuale che varia in base alla gravità di menomazione. Questi due tipi di danno possono essere considerati in una relazione di causa oppure indipendenti: *"per danno psichico si intende sia la conseguenza del danno fisico prodotta nella psiche della vittima, sia l'alterazione di tipo mentale verificatasi indipendentemente dalle lesioni fisiche"* (Franzoni, 2004, citato in Camerini, 2011, p. 13).

Sulla base delle caratteristiche appena descritte, il danno psichico viene definito come *"un'ingiusta turbativa dell'equilibrio psichico di un soggetto che gli causa una modificazione menomante della sua salute psichica con alterazione - temporanea o permanente - delle sue funzioni psichiche [...], che impedisce alla vittima di attendere - in modo totale o parziale, temporaneo o permanente - alle sue ordinarie occupazioni"* (Giannini & Pogliani, 1996). Quindi il danno psichico, per essere considerato tale, a differenza degli altri danni non patrimoniali, deve aver causato una psicopatologia nosograficamente inquadrabile che va ad influenzare il funzionamento del soggetto e a provocare dolore e sofferenza. Fornari (2008) indica le cause del danno psichico e le distingue in dirette ed indirette: il danno psichico si può sviluppare come diretta conseguenza di traumi cranio-encefalici, *mobbing*, *stalking*, lesioni personali, sequestri di persona, maltrattamenti, abusi e violenze; mentre può essere causato indirettamente dal lutto a seguito di morte di un familiare o dal carico pisco-fisico dovuto all'assistenza di un familiare non autosufficiente a causa di un evento lesivo altrui.

Nella valutazione del danno psichico, il primo aspetto che deve essere accertato è quindi la sussistenza del disturbo psicopatologico, essendo esso *conditio sine qua non* per la sua esistenza, attraverso l'esame neuropsicologico forense, strumento molto usato ed apprezzato nelle valutazioni post evento traumatico. Suddetto esame comprende la parte clinica di colloquio ed osservazione e la parte psicometrica di somministrazione dei test neuropsicologici. Una volta rilevata la sussistenza del danno, ci sono altri tre aspetti da dimostrare empiricamente nella valutazione e sono:

1) la presenza di un evento che può essere considerato proporzionale a scaturire il danno e che quindi è definibile come colpa (idoneità psicolesiva);
2) la modificazione in termini peggiorativi rispetto alla situazione premorbosa a causa dell'evento;
3) il nesso di causalità tra l'evento e il danno (Gulotta, 2011).

Un esempio di danno psichico che il soggetto può sviluppare come conseguenza diretta o indiretta di un fatto illecito è il disturbo depressivo. Tale disturbo, come accennavamo all'inizio, è facilmente simulabile in quanto è definito da un insieme di segni e sintomi facili da falsificare: mancanza di concentrazione, mancanza di interesse e di piacere nel fare le cose, sensi di colpa, umore

depresso, disturbi del sonno e così via. Inoltre, in Italia le compagnie assicurative spendono somme considerevoli per il risarcimento del danno psichico (Monaro et al., 2018). La facilità di simulazione e il vantaggio economico che la persona può derivare dall'essere clinicamente depresso devono mettere in guardia il professionista che viene chiamato a fare una valutazione, il quale deve tenere sempre in considerazione la probabilità di simulazione o aggravamento del disturbo.

I casi di simulazione di un disturbo psichico infatti, non sono così rari: nello studio condotto da Mittenberg e coll. (2002), in ambito di cause civili per danno alla persona, in cui rientra l'esempio di danno psichico qui riportato, è stata riscontrata una percentuale di casi sospetti di simulazione pari al 30,43%. In particolare, in un campione di soggetti che chiedono risarcimento da danno psichico è stata rilevata una prevalenza di simulazione pari al 64% (Heaton et al., 1978). In merito al gruppo diagnostico, Mittberg e coll. (2002), hanno riscontrato una percentuale di finte diagnosi depressive pari al 16%. Questi dati rendono quindi evidente la necessità di integrare nell' esame psicodiagnostico strumenti di valutazione oggettiva, capaci di rilevare la simulazione e meno influenzati dalle decisioni dei soggetti di fingere un disturbo, rispetto al semplice colloquio (Monaro et al., 2018).

2.3 Strumenti tradizionali per la detezione della simulazione

Come appena accennato, il colloquio clinico o l'osservazione, colonne portanti della valutazione clinica classica, non sono affidabili quando si tratta di diagnosticare la presenza di un disturbo psichico in contesti di valutazione forense, nei quali la simulazione è molto frequente. L'incapacità dei valutatori esperti di distinguere un vero depresso da un simulatore è stata largamente approfondita negli anni: Rosenhan (1973) nel suo famoso studio ha mostrato come 12 "pseudo-pazienti" di diversi ospedali psichiatrici specializzati, che fingevano di avere delle allucinazioni sono stati scambiati per pazienti con un disturbo schizofrenico conclamato e necessità di ricovero. Uno studio più recente ha confermato la difficoltà di distinguere un simulatore da un non-simulatore attraverso il solo colloquio clinico, con un'accuratezza che si attesta intorno al livello del caso (Rosen et al., 2004), fino ad arrivare ad un tasso di errore di classificazione pari all'80% (Sartori et al., 2016).

Alla luce dei risultati emersi sono diventati sempre più necessari strumenti integrativi al colloquio, che permettessero una corretta diagnosi differenziale (Monaro et al., 2018). Negli anni sono stati sviluppati diversi metodi e strumenti testistici per la detezione della simulazione e della dissimulazione che possono essere distinti in:

a) strumenti psicometrici specifici, costruiti *ad hoc* per l'ambito forense;
b) comuni strumenti psicodiagnostici già esistenti e applicati nella pratica clinica e psichiatrica.

Nella prima categoria di strumenti troviamo il SIMS e il SIRS, che sono ad oggi due tra i più utilizzati nel settore.

- Lo *Structured Interview of Malingered Symptomatology* (SIMS) (Smith & Burger, 1997) è un questionario autosomministrato creato appositamente per la rilevazione della simulazione delle patologie psichiatriche e cognitive. Può essere usato sia come strumento di screening, per diminuire il numero di soggetti che necessitano di valutazioni più approfondite come la SIRS, sia in aggiunta ad una batteria di test come conferma di simulazione. Il test è composto da 75 item a risposta dicotomica (vero/falso), 15 per ognuna delle seguenti condizioni psicopatologiche indagate dallo strumento:
  - bassa intelligenza (LI);
  - disturbi affettivi (FA);
  - danno neurologico (N);
  - Psicosi (P);
  - disturbi amnestici (AM).

  Alcuni di questi item sono stati presi da un altro questionario, l'MMPI, perché considerati molto efficaci nel rintracciare la presenza di simulazione mentre gli altri sono stati costruiti con diverse strategie di individuazione di questa, come lamentele improbabili, sintomi bizzarri e risposte approssimative. La logica sottostante è che il simulatore quando deve fingere un disturbo tende a mostrare sintomi bizzarri, atipici, rari o estremi; tale strumento però riesce anche ad individuare il soggetto che è stato istruito sulle risposte da dare, non solo i simulatori ingenui (Van Impelen et al., 2014). Secondo gli autori un punteggio totale maggiore di 14 e un punteggio nella sottoscala FA maggiore di 26 è indicativo di sospetta simulazione ed è quindi necessario un ulteriore approfondimento. Invece, il singolo punteggio di ogni sottoscala può essere usato qualitativamente come indicatore del disturbo che il soggetto sta cercando di simulare, e.g. un alto punteggio alla scala AM indica che la persona ha tentato di fingere sintomi mnestici (Smith & Burger, 1997).

- Lo *Structured Interview of Reported Symptoms* (SIRS) (Rogers, 1986, in Stracciari et al, 2010) è un *gold standard* per la detezione della simulazione di una vasta gamma di patologie psichiatriche e di sintomi che è molto improbabile possano essere veri. Il test è un'intervista strutturata, somministrata da un clinico che ha seguito uno specifico *training* formativo, ed è composta da 172 item suddivisi in 13 scale che corrispondono ai 13 stili di risposta solitamente associati alla simulazione. Le scale si dividono in primarie e supplementari, le prime sono otto e sono:

- *Rare Symptoms* (RS): sintomi apparentemente validi, ma che compaiono raramente anche nelle popolazioni psichiatriche;
- *Symptoms Combination* (SC): combinazioni di sintomi che, come nella scala precedente, sono improbabili anche nelle popolazioni psichiatriche;
- *Improbable or Absurd Symptoms* (IA): sintomi che non si possono trovare nemmeno nella popolazione psichiatrica, proprio perché improbabili o assurdi;
- *Blatant symptoms* (BS): sintomi evidenti di disturbi psichiatrici, ma che vengono riferiti dal simulatore in misura maggiore rispetto ai veri pazienti;
- *Subtle Symptoms* (SU): sintomi caratteristici di disturbi psichiatrici e molto comuni nei pazienti ma che non vengono segnalati dal simulatore in quanto considerati dal contenuto innocuo;
- *Severity of Symptoms* (SEV): riguarda l'eccessivo numero di sintomi riportati dal soggetto rispetto alla popolazione psichiatrica;
- *Selectivity of Symptoms* (SEL): riguarda il range di sintomi che il simulatore riferisce, in questo caso troppo ampio rispetto a quello della sintomatologia vera;
- *Reported vs. Observed Symptoms* (RO): sintomi riportati dal soggetto su base autodescrittiva, ma che il valutatore può verificare la presenza grazie all'osservazione diretta.

Per ognuna di queste scale, attraverso lo *scoring* dei punteggi è possibile interpretare la descrizione come "onesta", "dubbio", "finzione probabile" e "finzione certa". Le cinque scale supplementari aiutano il professionista a valutare la credibilità del soggetto nelle conclusioni generali e sono:

- *Direct Appraisal of Honesty* (DA): costituita da item che chiedono al soggetto di indicare esplicitamente la correttezza con cui si descrive ai clinici (molti simulatori per esempio riferiscono di provare piacere nel lasciare nel dubbio i dottori su quanto gli sta davvero accadendo);
- *Defensive Symptoms* (DS): contiene item che misurano la difensività del soggetto attraverso domande riguardanti sintomi quotidiani esperiti dalla maggior parte delle persone. Se questi vengono negati sono indicativi della messa in atto di un atteggiamento di chiusura.
- *Symptom Onset* (SO): indaga la veridicità della sintomatologia attraverso item che valutano l'esordio del disturbo. Se questo è improvviso o atipico, in mancanza di importanti traumi improvvisi, è indicativo di dubbia veridicità;

- *Overly Specified Symptoms* (OS): contiene item che indagano la simulazione in base a quanto il soggetto è in grado di definire alcuni aspetti dei disturbi, come la loro durata o frequenza. Il simulatore potrebbe riferire un grado di precisione elevato, impossibile da definire anche per il paziente vero;
- *Inconsistency of Symptoms* (INC): costituita da 32 item ripetuti che servono ad indicare il livello di attenzione e di consistenza delle risposte e l'accuratezza del soggetto (Stracciari et al., 2010)

Al di là del metodo di interpretazione suggerito dal manuale che permette di interpretare il punteggio di tutte le scale all'interno delle quattro aree di interpretazione sopra citate, Rogers (1997, in Stracciari et al, 2010) ha pubblicato *scores* compositi e nuovi *cut-off* che permettono di rilevare i simulatori con un'accuratezza pari al 97%.

Tra i comuni strumenti della pratica clinica e diagnostica che possono essere usati anche in ambito forense per la detezione della simulazione troviamo:

- il *Minnesota Multiphasic Personality Inventory* (MMPI-2), definito da Stracciari e coll. (2010) come lo strumento migliore dal punto di vista dell'attendibilità e dell'utilità in ambito forense. L'MMPI-2 per gli adulti e l'MMPI-A per gli adolescenti è un questionario che indaga i disturbi di personalità multiscala, è nato nel 1942 e l'adattamento italiano è stato condotto da Sirigatti e Pancheri (2001, in Stracciari et al., 2010). Il questionario è composto da 567 a risposta dicotomica (vero/falso) per gli adulti e 478 per gli adolescenti. Caratteristica fondamentale che lo rende utile al fine della detezione della simulazione sono le scale e indici di validità o di controllo che, insieme alla logica degli item ovvi e sottili in cui chi simula ottiene differenze maggiori, permettono di verificare la tendenza del soggetto a mettere in atto atteggiamenti simulatori o dissimulatori (difensivi). Nella direzione della simulazione le scale e gli indici utili alla sua rilevazione sono:
  - la scala F (Frequenza), con le scale aggiuntive di validità Fb, Fp ed Fbs utili per una migliore interpretazione, valuta la tendenza del soggetto ad esagerare i propri problemi o falsificare il test rispondendo in maniera eccessiva ad elementi estremi. Gli item di questa scala contengono sintomi bizzarri o rari, per cui punteggi elevati sono tipici di soggetti che simulano oppure che hanno risposto in maniera casuale o si sono confusi/disorientati (Butcher, 2010);
  - l'indice F-K (*Dissimulation Index*) invece si ottiene dalla differenza tra il punteggio grezzo ottenuto alla scala F e alla scala K ed è molto utile in quanto riesce ad individuare il simulatore con un'accuratezza molto elevata (circa 90%) (Stracciarti et

al., 2010). Nella simulazione si rileva un punteggio F elevato, indicativo di sintomi psicofisici insoliti, e un punteggio K basso (Sartori et al., 2000).

Mentre nella direzione della dissimulazione gli indici e le scale e utili sono:

- la scala L (Menzogna) indaga la disponibilità del soggetto ad ammettere difetti o problemi, per cui un punteggio elevato è tipico dell'individuo che cerca di mostrarsi sotto una luce favorevole e che quindi vuole dare un'immagine favorevole di sé, dissimulando (Butcher, 2010);
- la scala K (Correzione difensiva) valuta la tendenza a minimizzare i propri problemi. Inoltre, se il soggetto risulta sulla difensiva, il punteggio che ottengono cinque scale cliniche specifiche viene aggiustato alla luce di questo risultato, così da aumentare il potere discriminante delle scale (Butcher, 2010);
- l'indice F-K (*Dissimulation Index*) che viene calcolato come nella simulazione, ma nella dissimulazione mostra un profilo opposto: K risulta elevato, essendo questo indice di comportamento difensivo, mentre F basso (Sartori et al., 2000).

Inoltre, le due scale di validità VRIN e TRIN valutano l'incoerenza delle risposte date dall'individuo attraverso item che implicano risposte semanticamente incoerenti. Solitamente il simulatore di un disturbo psichico fornisce un *pattern* di risposta caratterizzato dalla presenza di sintomi tipici di diverse diagnosi e l'esagerazione di questi a livelli di gravità tanto elevati, non riscontrati nemmeno nei soggetti psichiatrici. Nonostante l'utilità e l'attendibilità di tale strumento, occorre ricordare che questo permette solo di individuare l'atteggiamento simulatorio generale messo in atto dal soggetto, per cui non è possibile individuare nello specifico in quali sintomi/disturbi abbia simulato (Sartori et al., 2000).

❖ il *Personality Assessment Inventory* (PAI) (Morey, 1996 in Zennaro et al., 2015) è un altro questionario che indaga i disturbi di personalità ed è stato considerato, insieme all'MMPI-2 e al SIRS, uno degli strumenti migliori usabili in ambito forense dagli specialisti del settore statunitensi (Lally, 2003). Come il questionario precedente anche questo fornisce scale ed indici di validità per valutare i vari aspetti che potrebbero modificare i risultati del test e quindi falsarlo. Le scale principali di validità sono 4:

- Incoerenza (ICN): valuta la coerenza con la quale il soggetto ha risposto alle domande di contenuto simile. Tale scala, nonostante correli con gli indicatori di desiderabilità sociale e quindi con la dissimulazione, è più probabile che risulti elevato quando il soggetto è stato disattento o ha fatto confusione/difficoltà nel compilare il questionario essendoci anche molti item in forma negativa;

- Infrequenza (INF): rileva, similmente la scala precedente, le risposte atipiche date dal soggetto a causa di disattenzione, confusione o "tirando a caso". Gli item selezionati hanno la caratteristica comune di essere poco frequenti all'interno del campione normativo perché riguardano affermazioni insolite ma non bizzarre, per cui è raro che un soggetto concordi su più di uno. I soggetti sani e clinici ottengono solitamente punteggi simili che differiscono di molto da quelli derivati dalla simulazione di risposte casuali;
- Impressione Negativa (NIM): costituita da item che rilevano la tendenza del soggetto a dare un'impressione esagerata o distorta di sé stesso e a riferire sintomi estremamente bizzarri ed improbabili, rari anche all'interno della popolazione clinica. Sebbene tale scala sia sensibile a misurare la distorsione in negativo messa in atto dal soggetto, non può essere considerata una scala di simulazione del disturbo, però può fungere da spia di tentativo di simulazione: i soggetti istruiti a dare risposte per fingere un disturbo depressivo grave ottengono punteggi considerevolmente maggiori rispetto al campione clinico;
- Impressione Positiva (PIM): valuta la tendenza del soggetto a rispondere dando un'immagine positiva di sé o negando difetti personali di minore entità. Il soggetto che sta cercando di modificare il risultato del test mostrandosi più favorevolmente di quanto sia nella realtà (*fake good*), ottiene punteggi molto elevati rispetto a quelli ottenuti dai pazienti clinici o dalla popolazione generale.

A queste si aggiungono altri sei indicatori supplementari di validità, tra i quali meritano particolare attenzione il *Malingering Index* (MAL) e la *Rogers Discriminant Function* (RDF).

- Il MAL è molto utile nella rilevazione della simulazione, in quanto è stato costruito sulle otto caratteristiche del profilo PAI che sono state riscontrate molto più frequentemente nei simulatori dei disturbi mentali gravi. Tale indice è quindi un indicatore specifico della tendenza a simulare la malattia e si basa sulle elevazioni delle scale di validità e su alcuni aspetti delle scale e sottoscale cliniche.
- La RDF invece ha funzione discriminante e aiuta il professionista a distinguere il profilo di un paziente da quello di un simulatore di disturbo psichiatrico, attraverso la combinazione ponderata di 20 punteggi ottenuti nell'intero questionario (Zennaro et al., 2015).

❖ il *Millon Clinical Multiaxial Inventory* (MCMI-III) (Millon et al., 1994 in Daubert & Metzler, 2000) è anche questo, come i questionari sopra, dotato di quattro scale di validità per il

rilevamento di stili di risposta *fake-good* e *fake-bad*. La Scala V, fungendo da indice di validità, rileva la risposta casuale, mentre le altre tre scale sono indici di modifica, nello specifico la scala X permette di adeguare l'interpretazione dei punteggi delle scale cliniche, la Scala Y rileva la potenziale tendenza del soggetto a mostrarsi sotto una luce migliore (*fake-good*), mentre la scala Z la potenziale tendenza del soggetto a peggiorare la propria condizione (*fake-bad*) (Daubert & Metzler,2000). Nonostante fosse stato ritenuto appropriato per l'uso in ambito forense (Millon et al., 1994 in Daubert & Metzler, 2000),gli esperti non si sono mostrati concordi nel suo uso in tale ambito (Stracciari et al., 2010).

Benché tali strumenti tradizionali siano molto utili nel rilevare la messa in atto del comportamento simulatorio da parte dell'individuo durante la compilazione del questionario, ha lo stesso limite della valutazione clinica: basandosi sui sintomi autoriferiti dal soggetto, i test neuropsicologici possono essere facilmente alterati e, se ben istruiti, anche facilmente ingannati (Storm & Graham, 2000).

2.4 Strumenti innovativi per la detezione della simulazione

Gli strumenti tradizionali appena descritti sopra, sono accomunati da tre grandi limiti:

1. sono basati totalmente sull'autodichiarazione del soggetto
2. sono facilmente falsificabili dal soggetto in base ai propri scopi (Monaro et al., 2018)
3. sebbene siano in grado di rilevare la tendenza generale del soggetto a simulare, non riconoscono la simulazione del sintomo specifico o dello specifico item simulato (Stracciari et al., 2010).

Alla luce di ciò e considerata l'importanza della detezione della simulazione in un ambito come quello forense, negli ultimi anni si è sentita la necessità di sviluppare strumenti più sofisticati, in grado di rilevare la simulazione attraverso misure implicite, al di fuori del controllo esplicito e consapevole del soggetto (Monaro et al., 2018), che consentono la verifica di specifiche informazioni simulate e non la sola tendenza generale del soggetto a mentire (Sartori et al., 2017).

Le tecniche più moderne, per rispondere a tale necessità, si sono basate sulla misurazione dei tempi di risposta (TR) ovvero il tempo che intercorre tra lo stimolo presentato e la risposta data dal soggetto. Secondo la teoria cognitiva l'atto del mentire richiede maggiori risorse cognitive rispetto al dire la verità per diversi motivi: inibire la risposta vera automatica e formulare una menzogna è innanzitutto lo sforzo cognitivo più importante, a cui si aggiungono poi il monitoraggio del proprio comportamento per cercare di mostrarsi il più onesti possibile, e delle reazioni dell'intervistatore, per capire se sta riuscendo ad ingannarlo, e infine la preoccupazione data dal ricordarsi di interpretare la

parte in modo coerente con la menzogna che si sta riferendo. Tali carichi cognitivi implicano quindi che il soggetto che mente ci impieghi più tempo nel rispondere, così come, se viene sottoposto ad un compito cognitivo aggiuntivo, il mentitore, rispetto al sincero, mostrerà difficoltà maggiori come balbettii, discorso più lento, pause, qualità dei dettagli inferiore, maggiori incongruenze (Vrij et al., 2008).

Un esempio di strumento di misura indiretta è l'aIAT (*autobiographical IAT*) il quale, somministrato via computer, misura la specifica alterazione intenzionale di un'informazione presente in memoria attraverso i tempi di reazione, i quali sono veloci se il soggetto risponde all'item presentato congruentemente con quanto presente in memoria e lenti se deve rispondere in modo incongruo. Tale metodo ha mostrato una precisione di circa il 90% nell'identificare il simulatore del colpo di frusta (Stracciari et al., 2010).

L'uso di queste tecniche implicite nella simulazione di disturbi psichiatrici è ostacolato da due limiti: innanzitutto tali tecniche possono indagare solamente un sintomo alla volta, per cui nella valutazione di una sindrome psichiatrica andrebbe ripetuto più volte (Monaro et al., 2018), in secondo luogo, anche se tali misure sono al di fuori del controllo esplicito del soggetto, nel caso del bugiardo patologico e del soggetto addestrato a mentire, rispondere con una menzogna può risultare più facile e veloce che rispondere sinceramente, ingannando lo strumento (Van Bockstaele et al., 2012).

Negli ultimi anni, con l'obiettivo di superare i due limiti sopra citati, diversi studi (Monaro et al. 2017; Mazza et al. 2020) hanno implementato nel contesto forense di detezione della simulazione una nuova tecnica computerizzata chiamata *Mouse Tracking* (Freeman & Ambady, 2010). Nello specifico, questa tecnica, oltre a misurare il tempo di risposta, si basa sugli indici cinematici di movimento, ovvero sulla traiettoria del mouse e sull'analisi di diversi parametri, tra i quali velocità e accelerazione; questo la rende più resistente alle contromisure che possono essere messe in atto dal simulatore rispetto alla sola registrazione dei tempi di reazione, poiché misurando più parametri è più difficile da controllare. Questo studio pilota si è basato su un'indagine pionieristica di rilevazione dell'inganno mediante analisi cinematica condotta da Duran e coll. (2010). Questi, usando un controller Nintendo Wii, hanno registrato il movimento motorio effettuato dalla mano del soggetto nel rispondere sinceramente o mentendo ad un compito binario. I risultati hanno mostrato una differenza significativa nelle due tipologie di risposte per quanto riguarda: il tempo totale per rispondere, il tempo per mettere in atto il movimento di risposta, la velocità, l'accelerazione e la traiettoria della risposta, la quale è stato visto riflettere la complessità cognitiva della menzogna. Il *mouse tracking* è stato utilizzato in un recente studio finalizzato al rilevamento delle false identità, in cui domande inaspettate su informazioni autobiografiche combinate con l'analisi dei movimenti del mouse durante la risposta, hanno evidenziato la capacità di questa tecnica nel rilevare la veridicità di

quanto dichiarato dal soggetto, con un tasso di accuratezza pari al 95% e quindi applicabile al contesto forense (Monaro et al., 2017).

Tali tecniche innovative fanno parte della categoria di strategie per la detezione della simulazione che usano l'informatica per rilevare segni psicometrici di inganno impliciti. Recentemente, l'informatica e le nuove tecnologie di bioingegneria hanno preso piede anche all'interno della psicologia forense, che, in una progressiva fusione con le neuroscienze, hanno come obiettivo quello di creare un sistema informatico capace di riconoscere i simulatori attraverso elementi oggettivi (Sartori et al., 2017).

Tuttavia, ad oggi non ci sono ancora studi che applichino tali metodi alla rilevazione della simulazione di disturbi psichiatrici, seppur in alcuni studi di Monaro e Sartori (Monaro et al., 2017) è emerso che è possibile usare l'analisi cinematica nella rilevazione della simulazione di disturbi come la depressione, l'ansia e il disturbo da stress-post traumatico in casi di risarcimento del danno. Come sostenuto da Sartori e coll. (2017), la situazione ideale in casi come quello dell'accertamento del danno psichiatrico sarebbe lo sviluppo di un algoritmo in grado di identificare oggettivamente la falsificazione o l'autenticità dei sintomi riferiti.

Sulla scia di queste nuove prospettive, gli studi effettuati da Monaro e coll. (2018) e da Mazza e coll. (2020) al fine di identificare rispettivamente i simulatori nei disturbi depressivi e nei disturbi di personalità simulati, hanno analizzato i risultati ottenuti con la tecnica del *Mouse Tracking* e il punteggio di risposta ai test attraverso il *Machine Learning* (ML). Il ML è un modello di apprendimento automatico, che permette di costruire algoritmi capaci, in questo caso, di distinguere il simulatore dal soggetto sincero, raggiungendo una precisione superiore al 90%, come dimostrato nello studio pilota condotto da Sartori e coll. (2017). L'apprendimento automatico è uno dei campi più promettenti nell'area dell'intelligenza artificiale (Mitchell, 1997): è una disciplina associata alla statistica computazionale che ha come obiettivo la creazione di nuove conoscenze o di previsioni su nuovi dati attraverso uno o più algoritmi che, basandosi su osservazioni reali, categorizza gli elementi in categorie differenti. In questo contesto, l'apprendimento automatico avviene su determinate caratteristiche come gli indici oggettivi di simulazione, che poi vengono analizzati e studiati in relazione con le variabili osservate in modo da definire le regole secondo cui verranno classificati automaticamente i dati (Sartori, 2017).

Nello studio pionieristico condotto da Monaro e coll. (2018) sulla simulazione del disturbo depressivo è stato visto come, attraverso questa tecnica innovativa, è possibile identificare fino al 96% dei simulatori.

Tuttavia, tali studi hanno utilizzato la tecnica del *Mouse Tracking* che per quanto sia promettente, essendo essa resistente alle contromisure e di facile/economica somministrazione ed

utilizzazione, per la detezione accurata della simulazione occorre prendere in considerazione più fonti di dati, tra cui strumenti di misurazione validi e strutturati, specificamente creati per rilevare la menzogna, supportate poi da tecniche automatizzate e non controllabili dall'esaminatore (Sartori et al., 2017).

2.4.1 TF-IDF

Come visto precedentemente, i questionari creati appositamente per la detezione della simulazione sono in grado di rilevare l'attitudine del soggetto a simulare, ma non lo specifico item in cui il soggetto ha simulato.

La necessità di una detezione della simulazione oggettiva ed accurata attraverso l'uso di tecniche automatiche applicate a strumenti di valutazione validi ed esistenti, espressa da Sartori e coll (2017) come riportato alla fine del paragrafo precedente, è stata di ispirazione nel validare una nuova metodologia oggetto del presente elaborato. Nello specifico, con l'obiettivo di indentificare la simulazione a livello del singolo item in un questionario, ci serviremo dell'indice TF-IDF (*Term Frequency - Inverse Document Frequency*). Tale indice, verrà infatti utilizzato per permette di identificare sia la collocazione di una risposta mentita, quindi di un singolo item, rispetto al campione di validazione, sia lo stile di risposta del soggetto (Baeza-Yates & Ribeiro-Neto, 1999). Se applicato alle risposte date dal soggetto ad un questionario già validato quindi, permette di scovare la simulazione a livello del singolo item.

Il TF-IDF è uno strumento matematico nato nel campo dei motori di ricerca del web ed usato da questi per recuperare le pagine web pertinenti alla domanda dell'utente. Nello specifico è un metodo standard di valutazione dei termini, utilizzato nel recupero delle informazioni per calcolare numericamente i dati testuali in base a come sono distribuiti i termini all'interno di un testo (Zhang, 2011). Si basa sull'intuizione che un determinato termine all'interno di un documento *d* è indicativo del suo significato generale in maniera proporzionale alla sua Frequenza di Termine (TF) all'interno di *d* e alla sua frequenza inversa nella raccolta (IDF). In altre parole, quando un termine è molto frequente nel documento *d* e poco negli altri documenti della raccolta vuol dire che tale termine è più rappresentativo del contenuto di *d* rispetto ad altri termini che o non vengono usati frequentemente in *d* o sono molto usati anche negli altri documenti della raccolta.

Nella pratica, data una collezione di documenti *C*, il valore TF-IDF per ogni termine *t* in un documento *d* che appartiene a *C* si calcola nel seguente modo:

$$\text{TF-IDF}(t, d, C) = \text{TF}(d,t) * \text{IDF}(t, C),$$

in cui

- TF (*d,t*) è il numero di volte in cui un termine *t* compare in un documento *d;*
- IDF (t,C) è uguale a log(N/n), dove N è il numero di documenti in tutta la collezione *C* e n è il numero di documenti in cui *t* è usato.

Nonostante il TF-IDF nasca nell'ambito dell'informatica, non è usato esclusivamente in questa area: già in precedenza infatti è stato applicato anche in psicologia, nello specifico in neuropsicologia cognitiva per la modellazione della memoria semantica e dei suoi disturbi (Sartori & Lombardi, 2004).

Quando il TF-IDF viene applicato nella rilevazione degli item simulati all'interno di un questionario *q*, viene calcolato considerando TF come il numero di volte in cui quella determinata risposta dell'item (es. 5) è stata fornita dal soggetto in tutto il questionario e se tale valore è stato assegnato ad un solo item su 9 allora TF(5, *q*)=1. L'IDF invece, che, come visto precedentemente, si calcola con log(N/n), considera N come il numero di partecipanti che hanno risposto al test e n come il numero di volte in cui tutti i partecipanti hanno utilizzato quello stesso valore di risposta (es. 5) in quello specifico item.

Il punteggio ottenuto dal prodotto dei valori TF e IDF per ogni item del questionario avrà valori elevati se la risposta all'item è atipica rispetto alle risposte date dal campione di validazione a quel singolo item nella condizione onesta, ovvero non simulata. In questo modo il TF-IDF funge da indicatore di anomalie a livello di singolo item e non a livello di soggetto.

Nella pratica, occorre innanzitutto calcolare i punteggi TF e IDF di tutti gli item forniti da ogni soggetto, poi per individuare il simulatore, si moltiplica il TF di uno specifico item all' IDF ottenuto dalle risposte date a quell'item dai partecipanti sinceri (le quali sappiamo essere oneste in quanto ogni partecipante ha compilato il questionario prima onestamente e poi simulando un disturbo). Se il prodotto che si ottiene è superiore ad una certa soglia viene contrassegnato come anomalo e la risposta indicata come falsa. Tale soglia di rilevamento della falsificazione non è stimata sul punteggio grezzo TF-IDF ma sul percentile corrispondente ed è diversa per ogni item, in quanto viene stabilita sulle risposte date a quella singola domanda dal gruppo degli onesti. Il valore soglia scelto è quello considerato come maggiormente affidabile nel rilevare i simulatori in quella specifica domanda, capacità che rende tale metodo innovativo nell'ambito della detezione della simulazione.

Un ulteriore vantaggio di questo metodo è la possibilità di identificare il simulatore sulla sola comparazione con un gruppo di partecipanti che precedentemente hanno risposto in maniera onesta allo stesso questionario, quindi senza l'uso di alcuna scala di controllo e indipendentemente dal contesto. Quest'ultima caratteristica ha anche un importante risvolto pratico essendo la strategia di

falsificazione sempre diversa in base agli obiettivi del soggetto (la dissimulazione che un soggetto mette in atto per ottenere la custodia del figlio è diversa da quella per ottenere un lavoro come venditore, seppure entrambi cercano di mostrarsi sotto una luce favorevole).

Capitolo III

LA RICERCA SPERIMENTALE

3.1 Descrizione dello studio

La ricerca che verrà presentata ora è stata svolta dal Dipartimento di Psicologia Generale (D.P.G.) dell'Università degli Studi di Padova. Lo scopo di tale studio è indagare l'efficacia dei due metodi innovativi per la detezione della simulazione quando applicati ad uno strumento *self-report*, esposti nei due paragrafi precedenti (2.4.1 e 2.4.2).

Lo studio si propone dapprima di indagare l'efficacia dell'indice TF-IDF nel rilevare la probabile simulazione del soggetto in ogni singolo item del questionario e nel suo individuale stile di risposta. Una volta rilevate le risposte falsate, è stata applicata una metodologia, che verrà più avanti descritta, per ricostruire il profilo onesto del singolo individuo partendo dalle risposte simulate. L'obiettivo è quello di "ripulire" le risposte dei partecipanti dalla menzogna e cercare di sovrapporle a quelle oneste, nel tentativo di costruire una valida metodologia che permetta di rispondere al quesito: "Come avrebbero risposto i partecipanti mentitori se avessero dato risposte sincere?"

Il questionario *self-report* su cui sono state applicate queste due nuove metodologie è il PHQ-9 (Kroenke et al., 2001), per indagare la detezione della simulazione del disturbo depressivo maggiore.

3.2 Metodologia sperimentale

La scelta dello strumento da utilizzare per raccogliere le risposte dei partecipanti è ricaduta sul questionario PHQ-9 per motivi che verranno più avanti dettagliatamente riportati quali brevità e validità di costrutto e criterio, che lo rendono un ottimo strumento sia per diagnosticare il disturbo depressivo maggiore che per valutarne la gravità (Kroenke et al., 2001). Al termine del questionario (riportato in appendice in originale) sono state aggiunte alcune domande utili al raccoglimento di informazioni demografiche come l'età, il genere, la scolarità. La scelta di inserire queste domande alla fine invece che all'inizio, come di consueto, nasce dalla necessità di evitare l'accentuazione dell'auto-stereotipizzazione, come descritto al paragrafo 1.3, che il soggetto potrebbe mettere in atto ponendo attenzione al suo genere di appartenenza nel rispondere alla domanda. All'inizio del questionario invece sono state poste al soggetto domande relative alla presenza di una diagnosi e al

trattamento farmacologico e psicoterapeutico di disturbo depressivo, al fine di individuare i partecipanti con diagnosi o con disturbo depressivo non diagnosticato ma trattato con farmaci antidepressivi ed escluderli dal presente campione di studio.

Il questionario, creato attraverso la piattaforma Qualtrics, è stato somministrato completamente online e la durata per il suo completamento è stata di circa 5 minuti.

All'inizio del questionario ogni soggetto ha preso visione del consenso informato pre-esperimento ed ha accettato di partecipare, in caso contrario il questionario si chiudeva automaticamente, così da avere la garanzia della protezione dei dati personali e l'assoluto anonimato; la partecipazione alla ricerca è stata su base volontaria, senza essere previsto alcun compenso.

Dopo il consenso il formato e la risposta alle domande riguardanti la diagnosi e il possibile trattamento in atto, il soggetto ha ricevuto le seguenti istruzioni:

"Immagini di essere coinvolto/a in un processo in cui richiede un risarcimento danni a causa di un evento molto stressante (es. lutto improvviso di un familiare a causa di un incidente, stalking da parte del partner, abuso sessuale) che le ha causato lo sviluppo di un disturbo depressivo.
Le chiediamo di leggere attentamente e di rispondere alle domande scegliendo una risposta su una scala da 0 a 3, dove: 0 = mai, 1 = alcuni giorni, 2 = per più della metà dei giorni, 3 = quasi ogni giorno.
Attenzione, ogni domanda le verrà posta due volte:
- alla domanda in rosso deve rispondere immaginando di essere sottoposto/a a valutazione psichiatrica: il suo obiettivo è compilare il questionario simulando un disturbo depressivo grave, ai fini di ottenere il risarcimento dei danni. Deve cercare, quindi, di dare un'immagine patologica di sé, ma allo stesso tempo di essere il più credibile possibile.
- all'altra, ovvero quella in nero, le chiediamo di rispondere facendo riferimento alla sua vera condizione attuale, quindi scegliendo il punteggio in maniera onesta.
La invitiamo a utilizzare tutto il tempo necessario e a leggere ciascuna affermazione, nel pieno tentativo di immedesimarsi nella simulazione e di rispondere onestamente".

A tali istruzioni seguivano due domande di controllo per verificare la comprensione di queste e se risposte in maniera sbagliata, venivano fornite le seguenti istruzioni semplificate ed a seguire un'altra domanda di controllo:

"OPS! Ha sbagliato una delle due risposte.
Le chiediamo di leggere più attentamente la consegna prima di proseguire:
Risponda alle domande in rosso immaginando di essere sottoposto/a a valutazione psichiatrica: il suo obiettivo è compilare il questionario simulando un disturbo depressivo grave, ovvero dando un'immagine patologica di sé, ai fini di ottenere il risarcimento dei danni, cercando allo stesso tempo di essere il più credibile possibile.
Alle domande in nero, le chiediamo di rispondere facendo riferimento alla sua vera condizione attuale, quindi scegliendo il punteggio in maniera onesta".

Il partecipante è stato randomizzato casualmente dalla piattaforma Qualtrics in uno dei due gruppi che differivano l'uno dall'altro per l'ordine di presentazione della risposta onesta o simulata.

Un gruppo di soggetti (231) soggetti ha risposto ad ogni domanda del PHQ-9 prima in maniera onesta (condizione onesta - *honest*), poi simulando un disturbo depressivo (*dishonest*) (gruppo HD), mentre un altro gruppo (298) è stato assegnato alla condizione opposta in cui dovevano rispondere ad ogni domanda prima in maniera simulata (*dishonest*) e poi onesta (*honest*) (gruppo DH). Tale randomizzazione è stata effettuata per escludere la presenza di qualche possibile effetto dovuto all'ordine di presentazione sulle risposte dei soggetti.

Nella pratica, ai partecipanti sono state presentate le 9 domande del PHQ-9 una alla volta e veniva loro chiesto di rispondere sia sinceramente che simulando (l'assegnazione dell'ordine della modalità di risposta, se sincera primo e simulata dopo o viceversa, dipendeva dalla condizione a cui era assegnato mediante randomizzazione). Ad esempio, chi è stato assegnato al gruppo HD (honest-dishonest), per ogni domanda presentata doveva rispondere prima in maniera onesta e poi, sempre alla stessa, simulando.

l'obiettivo di individuare la simulazione a livello del singolo item: con il metodo precedente pur chiedendo ai partecipanti di mentire, non tutti hanno deciso di mentire a tutti gli item, seguendo quello che è un loro pattern personale di risposta. Per questo studio si è deciso di richiedere ai partecipanti di rispondere a ciascun item prima onestamente e poi mentendo, di modo da ottenere due set di dati: un set di risposte oneste e un set di risposte tutte mentite.

Le risposte erano forzate ed a scelta multipla, con una sola opzione di risposta in una scala Likert da 0 a 3, in cui:

- 0 = "mai"
- 1 = "alcuni giorni"
- 2 = "per più della metà dei giorni"
- 3 = "quasi ogni giorno".

Gli obiettivi della ricerca sono quindi due: individuare le singole risposte agli item simulate dal soggetto attraverso l'indice TF-IDF e ricostruirle attraverso una nuova metodologia che verrà mostrata più avanti, così da ottenere un profilo del test onesto usando solamente il test falsato del singolo individuo. Senza dimenticare l'obiettivo generale di verificare l'efficacia di questo nuovo metodo di detezione della simulazione rispetto ai metodi tradizionali e ai nuovi metodi esistenti.

3.2.1 Lo strumento: il Patient Health Questionnaire-9 (PHQ-9)

Per la raccolta dati ci siamo serviti del questionario Patient Health Questionnaire-9 (PHQ-9, Kroenke et al., 2001) uno dei questionari citati nel paragrafo 1.3.2 come tra i più usati nella

valutazione del disturbo depressivo maggiore. La scelta di questo questionario tra i diversi *gold standard* della valutazione è dettata da diverse motivazioni. A differenza degli altri strumenti, questo è un questionario autosomministrato molto breve, tradotto e validato in lingua italiana (Picardi e coll. 2005; 2006), con una scala di risposta di tipo Likert (0-3) uguale per tutti gli item e utilizzato in diversi contesti, senza tralasciare le caratteristiche psicometriche come la validità di costrutto e di criterio che fanno di esso uno strumento breve, ma valido ed affidabile nella rilevazione del disturbo (Kroenke et al., 2001).

Come accennato nel paragrafo di riferimento, il PHQ-9 è molto breve, infatti è composto da soli 9 item, i quali fanno riferimento ai 9 criteri diagnostici del DSM-IV e quindi ai criteri diagnostici attuali, essendo rimasti gli stessi anche nell'ultima edizione del DSM-5. Il tempo di somministrazione è di pochi minuti e il soggetto deve rispondere ai 9 item facendo riferimento alla frequenza con cui nelle due ultime settimane ha esperito i problemi presentati nelle domande, come per esempio sentirsi stanco, avere scarso interesse nel fare le cose, sentirsi uno/a fallito/a, così via.

Per rispondere il soggetto deve indicare il livello di frequenza in una scala Likert che va da 0 (mai) a 3 (quasi ogni giorno), che è uguale per ogni item del questionario. Questo è uno dei motivi per cui nello studio qui riportato è stato adottato come questionario il PHQ-9, a differenza di altri strumenti che hanno diverse scale di risposta dal punto di vista del numero e/o delle etichette, come per esempio il BDI-II in cui gli item 16 e 18 a differenza degli altri hanno una scala Likert di risposta a 7 punti invece che 4.

Oltre a questi 9 item che costituiscono la parte diagnostica del questionario, nella parte finale è stato inserito un item che chiede al soggetto che ha dichiarato di riscontrare problemi nelle precedenti domande, quanto è stato difficile occuparsi del resto delle attività della vita quotidiana. Tale item ha lo scopo di ottenere una valutazione globale della compromissione funzionale, utile a valutare la qualità della vita e lo stato funzionale del soggetto che, come evidenziato nel paragrafo 1.1, sono correlate negativamente con il disturbo depressivo. Essendo quindi un item qualitativo non viene considerato nel calcolo del punteggio per rilevare la presenza del disturbo e la gravità di questo (Kroenke & Spitzer, 2002).

Per quanto riguarda lo *scoring*, questo può essere calcolato in due modi:

1. il primo consiste in un algoritmo diagnostico, in cui per fare diagnosi di depressione maggiore il soggetto deve aver risposto con un punteggio positivo, ovvero deve aver risposto selezionando il punteggio "2" (per più della metà dei giorni) o "3" (quasi ogni giorno) della scala Likert, ad almeno 5 dei 9 item (tranne che nell'item 9 sul suicidio e i pensieri suicidari perchè in questo caso è considerato punteggio positivo anche "1"), ad almeno uno dei primi

due item e deve essere presente una menomazione nel funzionamento dell'individuo nell'item 10;

2. il secondo invece consiste nel semplice calcolo matematico dei punteggi dati dal soggetto ai singoli item (escluso il decimo), la cui somma è compresa tra 0 e 27 ed un punteggio ≥10 è considerato indicativo della diagnosi di depressione maggiore (Iglesias-González & Diez-Quevedo, 2021).

Per questa ricerca, la modalità di *scoring* utilizzata è la seconda in quanto comprensiva di punteggio *cut-off*, che anche gli autori raccomandano debba essere di 10 o superiore in quanto ha sensibilità e specificità pari all'88% nella rilevazione del disturbo depressivo maggiore (Kroenke & Spitzer, 2002).

In meno di un decennio, l'uso di questo questionario è stato largamente diffuso in ambito clinico, di ricerca e di screening.

Innanzitutto, la validità diagnostica di questo strumento è stata appurata su due numerosi campioni derivanti da 8 cliniche di cure primarie (Spitzer et al., 1999 in Kroenke et al, 2001) e 7 cliniche di ostetricia e ginecologia (Spitzer et al., 2000, in Kroenke et al., 2001). Successivamente a questi studi iniziali, il PHQ-9 è stato validato ed utilizzato sia in ampi sondaggi sulla popolazione generale, sia sulla popolazione clinica, in particolare viene utilizzato per fare screening nei soggetti con una varietà di condizioni mediche, spesso in comorbidità con la depressione, come disturbi neurologici, malattie cardiovascolari, disturbi dermatologici, malattie infettive come l'HIV e molte altre, così come si è rivelato essere uno strumento altamente specifico per la rilevazione della depressione post-partum (un'ampia rassegna è consultabile nell'articolo di Kroenke et al., 2010). Per quanto riguarda i sondaggi, anche l'Istat (Istituto Nazionale di Statistica) utilizza il PHQ-9 nelle indagini sulla salute mentale della popolazione generale (Istat, 2018).

Il regno Unito, nello specifico il sistema sanitario nazionale, ha adottato il PHQ come strumento *gold standard* per rilevare la depressione nelle cure primarie (Kroenke et al., 2010). Infatti, sia per il MMG (Medico di Medicina Generale) che per qualsiasi altro professionista del sistema sanitario, avere un semplice strumento autosomministrato da far completare al soggetto in clinica o in chiamata telefonica è molto utile, poiché permetterebbe di risparmiare il tempo necessario per informarsi sulla presenza dei nove sintomi che definiscono il disturbo e fare un rapido screening della presenza o meno di depressione e della sua gravità (Kroenke et al., 2001).

Inoltre, il PHQ-9 è molto utilizzato in ambito clinico anche per la sua sensibilità al cambiamento, ovvero alla rilevazione di un miglioramento della sintomatologia durante il trattamento o a seguito di questo (Kroenke & Spitzer, 2002)

Dal punto di vista della validità psicometrica, ovvero il grado in cui il questionario misura effettivamente il costrutto che dovrebbe misurare, il PHQ-9 ha una validità di costrutto (interna) e di criterio eccellenti, come mostrato dallo studio pionieristico condotto da Kroenke e coll. (2001) su un campione di 6000 pazienti provenienti da cliniche di cure primarie e di ostetricia-ginecologia. Per quanto riguarda la validità di costrutto, la coerenza interna, ovvero il grado di accordo tra i diversi item che compongono lo strumento, il PHQ-9 ha un valore dell'Alpha di Cronbach pari a 0.89 nel contesto cure primarie e 0.86 in quello ostetrico-ginecologico, così come anche l'affidabilità test-retest, ovvero la stabilità della capacità di misurazione dello strumento, è stata eccellente. Per quanto concerne la validità di criterio, il PHQ-9 è stato valutato confrontandolo con quanto ottenuto da una intervista strutturata somministrata da un professionista della salute mentale (MHP), così da valutare la corrispondenza tra questo e un'altra misura di valutazione del disturbo esistente e riconosciuta come valida, ed è stata riscontrata una correlazione di 0.84.

3.3 Partecipanti

Hanno partecipato alla ricerca un totale di 1.099 soggetti, reclutati via e-mail, Facebook, Instagram o Whatsapp. Abbiamo informato i partecipanti circa lo scopo dello studio, garantendo l'anonimato e che i dati sarebbero stati utilizzati solo per scopi di ricerca. Tutti i partecipanti hanno accettato di unirsi alla ricerca volontariamente dopo aver letto e accettato il consenso informato.

Di questi, 346 sono stati esclusi dalle analisi in quanto non hanno portato a termine il questionario o non hanno risposto correttamente alle domande di comprensione delle istruzioni.

Dei restanti 753, sono stati esclusi dal campione i soggetti che hanno dichiarato di aver ricevuto una diagnosi di disturbo depressivo, che assumono farmaci antidepressivi, anche senza avere una diagnosi, e che hanno raggiunto il *cut-off* prestabilito dalla letteratura, ovvero un punteggio maggiore o uguale a 10 (Kroenke & Spitzer, 2001), per un totale di 224 partecipanti che verranno utilizzati negli sviluppi futuri di tale ricerca come campione della popolazione clinica e clinica a livello psicometrico.

Escludendo i soggetti eliminati (346 + 224), abbiamo ottenuto un campione finale composto da 529 soggetti, di cui 403 femmine (76%), 122 maschi (23%) e 4 che si sono riconosciuti sotto il termine "altro" (0,7%). Dei 529 soggetti che formano il campione finale, 231 sono stati assegnati casualmente al gruppo HD (*honest-dishonest*) in cui dovevano, per ogni domanda, rispondere prima onestamente e poi simulando, mentre i restanti 298 nel gruppo DH (*dishonest-honest*) in cui dovevano prima simulare e poi rispondere onestamente.

L'età dei soggetti del campione è compresa tra i 18 e 74 anni (M = 33,5; DS = 12,3) e la scolarità va dagli 8 anni di istruzione con la licenza media a più di 16 con il dottorato/master ecc (M = 14,8; DS = 2,6).

3.4 Ipotesi

H1. Nella condizione *honest* e *dishonest* i punteggi grezzi presentano una differenza statisticamente significativa.

H2. Sia i punteggi grezzi che i valori TF-IDF risultano inferiori nella condizione *honest* rispetto alla condizione *dishonest*.

H3. I valori TF-IDF permettono una detezione più facile e accurata dei soggetti che simulano il disturbo ansioso-depressivo rispetto ai soli dati grezzi.

H4. I valori TF-IDF *dishonest* si collocano prevalentemente al di sopra di un determinato valore soglia.

Inoltre, fra gli obiettivi i questo studio c'è la ricostruzione delle risposte oneste tramite una nuova metodologia che verrà più avanti descritta (vedi paragrafo 4.3).

Capitolo IV

ANALISI DEI DATI E RISULTATI

4. Analisi dei dati grezzi

Come descritto al paragrafo 3.2.3, i soggetti sono stati randomizzati nelle due condizioni *honest-dishonest* (HD)/*dishonest-honest* (DH), quindi prima di mostrare le osservazioni grafiche e condurre le analisi statistiche tradizionali sui 529 soggetti che compongono il campione finale della ricerca, occorre verificare se l'ordine di presentazione delle domande possa aver influito sulla modalità di risposta.

Per escludere un possibile effetto dell'ordine nel dare la risposta, è stata fatta innanzitutto un'analisi della distribuzione del campione attraverso il test di normalità Shapiro-Wilk per campioni indipendenti: in tabella 4.1 è riportato il valore di significatività. Questo test assume come ipotesi nulla che i dati provengano da una distribuzione normale. Se emerge un *p-value* > 0,05, allora l'ipotesi nulla viene confermata. Al contrario, nel caso in cui il *p-value* sia inferiore a 0,05 è molto probabile che la distribuzione dei dati non segua un andamento normale. Nel presente caso, i valori del *p-value* sono inferiori a 0,001 per tutti gli *item*, per questo si può rifiutare l'ipotesi nulla e ritenere che i dati in analisi non provengano da una distribuzione normale. La non normalità della distribuzione ci permette di scegliere i test da utilizzare nelle prossime analisi

Tabella 4.1 *- Test di normalità Shapiro-Wilk eseguito tra i due gruppi del campione HD e DH*

		W	p
Piacere_Q1_H	DH	0.684	< .001
	HD	0.735	< .001
Piacere_Q1_D	DH	0.688	< .001
	HD	0.632	< .001
Triste_Q2_H	DH	0.664	< .001
	HD	0.721	< .001
Triste_Q2_D	DH	0.631	< .001
	HD	0.666	< .001
Sonno_Q3_H	DH	0.746	< .001
	HD	0.762	< .001
Sonno_Q3_D	DH	0.743	< .001
	HD	0.746	< .001
Stanco_Q4_H	DH	0.720	< .001
	HD	0.691	< .001

Appetito_Q5_H	DH	0.724	< .001
	HD	0.631	< .001
Appetito_Q5_D	DH	0.823	< .001
	HD	0.837	< .001
Fallito_Q6_H	DH	0.672	< .001
	HD	0.668	< .001
Fallito_Q6_D	DH	0.668	< .001
	HD	0.679	< .001
Concentrazione_Q7_H	DH	0.680	< .001
	HD	0.626	< .001
Concentrazione_Q7_D	DH	0.821	< .001
	HD	0.814	< .001
Psicomotorio_Q8_H	DH	0.347	< .001
	HD	0.347	< .001
Psicomotorio_Q8_D	DH	0.871	< .001
	HD	0.853	< .001
Suicidio_Q9_H	DH	0.215	< .001
	HD	0.275	< .001
Suicidio_Q9_D	DH	0.846	< .001
	HD	0.851	< .001

Note: *In tabella sono indicati i valori psicometrici ottenuti attraverso il test di Shapiro-Wilk per ogni item del questionario, nella condizione honest (H) e dishonest (D) e in entrambe le distribuzioni HD e DH. I p-value sono risultati essere tutti significativi in quanto < di 0,001.*

Il *p-value* è risultato essere significativo in ogni item, indicativo di una distribuzione non simmetrica del *dataset*, quindi di violazione dell'assunzione di normalità e di conseguenza l'applicazione di test non parametrici. Inoltre, essendo il campione formato da due gruppi indipendenti con variabili di risposta di tipo "qualitativa ordinale", il test da utilizzare in alternativa al t-test di *Student* per dati non parametrici è il Mann-Whitney (tabella 4.3), il quale ci permette di stabilire se c'è una differenza statistica tra le medie dei due campioni. In questo caso, un *p-value* > 0,05 confermerebbe l'ipotesi nulla, secondo cui le due distribuzioni non presentano differenze statisticamente significative, mentre al contrario un *p-value* < 0,05 comporterebbe un rifiuto dell'ipotesi nulla, per cui le due distribuzioni presentano differenze significative.

Tabella 4.3 *- Test non parametrico Mann-Whitney eseguito trai due gruppi HD e DH.*

	W	df	p	Hodges-Lehmann Estimate	Rank-Biserial Correlation
Piacere_Q1_H	32248.000		0.138	−1.243e−5	−0.063
Piacere_Q1_D	33026.500		0.348	−5.003e−5	−0.040
Triste_Q2_H	32418.000		0.174	−3.148e−5	−0.058
Triste_Q2_D	36288.500		0.195	1.904e−5	0.054
Sonno_Q3_H	33759.500		0.675	−2.614e−5	−0.019
Sonno_Q3_D	34755.000		0.830	3.492e−5	0.010
Stanco_Q4_H	36317.500		0.171	2.601e−5	0.055
Stanco_Q4_D	34723.500		0.836	4.768e−5	0.009
Appetito_Q5_H	39038.000		0.002	9.523e−6	0.134
Appetito_Q5_D	36302.500		0.247	5.190e−5	0.055
Fallito_Q6_H	35603.000		0.432	4.940e−5	0.034
Fallito_Q6_D	34742.000		0.827	4.669e−5	0.009
Concentrazione_Q7_H	36525.500		0.153	5.889e−5	0.061
Concentrazione_Q7_D	33722.000		0.668	−3.895e−5	−0.020
Psicomotorio_Q8_H	34857.500		0.638	4.093e−5	0.013
Psicomotorio_Q8_D	30788.500		0.029	−6.583e−5	−0.105
Suicidio_Q9_H	33652.000		0.273	−4.562e−6	−0.022
Suicidio_Q9_D	34856.500		0.792	6.885e−5	0.013

Note: *In tabella si presenta il confronto tra le medie delle due distribuzioni honest e dishonest in ogni item. Nella colonna "p" è rappresentato il p-value che risulta significativo solamente negli item "Appetito_Q5_H" e "Psicomotorio_Q8_D". Infine, nell'ultima colonna "Rank-Biserial" (rB) è indicato il valore di effect-size.*

I valori della tabella 2.2 hanno quasi tutti un *p-value* > 0.05, ad eccezione di due *item* in cui il *p-value* risulta significativo (*item* "Appetito_Q5_H" e "Psicomotorio_Q8_D"). Quindi l'ipotesi nulla, secondo cui le due distribuzioni sono uguali, viene accettata, ad indicare che non c'è un effetto dell'ordine di presentazione tale da dover considerare i due gruppi in modo separato. Per quanto riguarda la dimensione dell'effetto, il *range dell'effect-size* è compreso tra 0 e 0,1. In accordo con la correlazione di Cohen, si tratta di un effetto con magnitudo *small*, a conferma di una differenza tra i valori medi delle distribuzioni molto piccola. Anche dal punto di vista delle analisi descrittive, nella tabella 4.4 si evince come le medie delle risposte dei soggetti nelle due condizioni HD e DH a livello del singolo item, nelle due condizioni *honest* e *dishonest*, sono molto simili, confermando l'assenza di differenza tra i due campioni.

Tabella 4.4 *- Tabella dei valori, dei punteggi medi e della mediana di ciascun item honest e dihsonet nelle due condizioni di risposta HD e DH*

	Piacere_Q1_H		Piacere_Q1_D		Triste_Q2_H		Triste_Q2_D	
	DH	HD	DH	HD	DH	HD	DH	HD
Valid	298	231	298	231	298	231	298	231
Missing	0	0	0	0	0	0	0	0
Median	1.000	1.000	3.000	3.000	1.000	1.000	3.000	3.000
Mean	0.728	0.814	2.520	2.528	0.668	0.740	2.628	2.524
Std. Deviation	0.553	0.615	0.678	0.779	0.506	0.553	0.608	0.739
Minimum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	3.000	3.000	3.000	3.000	2.000	2.000	3.000	3.000

	Sonno_Q3_H		Sonno_Q3_D		Stanco_Q4_H		Stanco_Q4_D	
	DH	HD	DH	HD	DH	HD	DH	HD
Valid	298	231	298	231	298	231	298	231
Missing	0	0	0	0	0	0	0	0
Median	0.000	1.000	3.000	3.000	1.000	1.000	3.000	3.000
Mean	0.607	0.636	2.386	2.377	0.966	0.892	2.570	2.545
Std. Deviation	0.708	0.727	0.780	0.775	0.613	0.568	0.644	0.696
Minimum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	3.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000

	Stanco_Q4_H		Stanco_Q4_D		Appetito_Q5_H		Appetito_Q5_D	
	DH	HD	DH	HD	DH	HD	DH	HD
Valid	298	231	298	231	298	231	298	231
Missing	0	0	0	0	0	0	0	0
Median	1.000	1.000	3.000	3.000	0.000	0.000	2.000	2.000
Mean	0.966	0.892	2.570	2.545	0.557	0.403	2.131	2.056
Std. Deviation	0.613	0.568	0.644	0.696	0.700	0.671	0.820	0.814
Minimum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	3.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000

	Fallito_Q6_H		Fallito_Q6_D		Concentrazione_Q7_H		Concentrazione_Q7_D	
	DH	HD	DH	HD	DH	HD	DH	HD
Valid	298	231	298	231	298	231	298	231
Missing	0	0	0	0	0	0	0	0
Median	0.000	0.000	3.000	3.000	0.000	0.000	2.000	2.000
Mean	0.463	0.450	2.550	2.545	0.436	0.355	2.148	2.165
Std. Deviation	0.526	0.601	0.686	0.670	0.572	0.497	0.799	0.833
Minimum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	2.000	3.000	3.000	3.000	3.000	2.000	3.000	3.000

	psicomotorio_Q8_H		Psicomotoro_Q8_D		Sucidio_Q9_H		Suicidio_Q9_D	
	DH	HD	DH	HD	DH	HD	DH	HD
Valid	298	231	298	231	298	231	298	231
Missing	0	0	0	0	0	0	0	0
Median	0.000	0.000	2.000	2.000	0.000	0.000	2.000	2.000
Mean	0.141	0.104	1.748	1.935	0.047	0.069	1.919	1.900
Std. Deviation	0.450	0.320	0.961	0.904	0.212	0.254	0.884	0.862
Minimum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	3.000	2.000	3.000	3.000	1.000	1.000	3.000	3.000

Note: *In tabella è presentata la statistica descrittiva delle risposte honest (H) e dishonest (H) date dai partecipanti nelle due condizioni HD e HD. Nelle colonne sono presenti gli item H e D nelle due condizioni HD e DH, mentre nelle righe sono presenti i valori psicometrici, tra cui media e mediana*

che ci permettono di osservare come queste siano simili tra oro, ad indicare l'assenza di differenza tra i due campioni.

Per i motivi elencati fin qui, d'ora in avanti il campione verrà considerato come proveniente dalla stessa popolazione e quindi i soggetti trattati come campione unico.

4.1.1 Rappresentazione grafica dei dati grezzi

Considerando il campione unito, si propone una presentazione grafica dei punteggi grezzi medi nelle due condizioni *honest* e *dishonest* (fig. 4.1) e di seguito i valori del punteggio medio e della deviazione standard per ogni item (tabella 4.5).

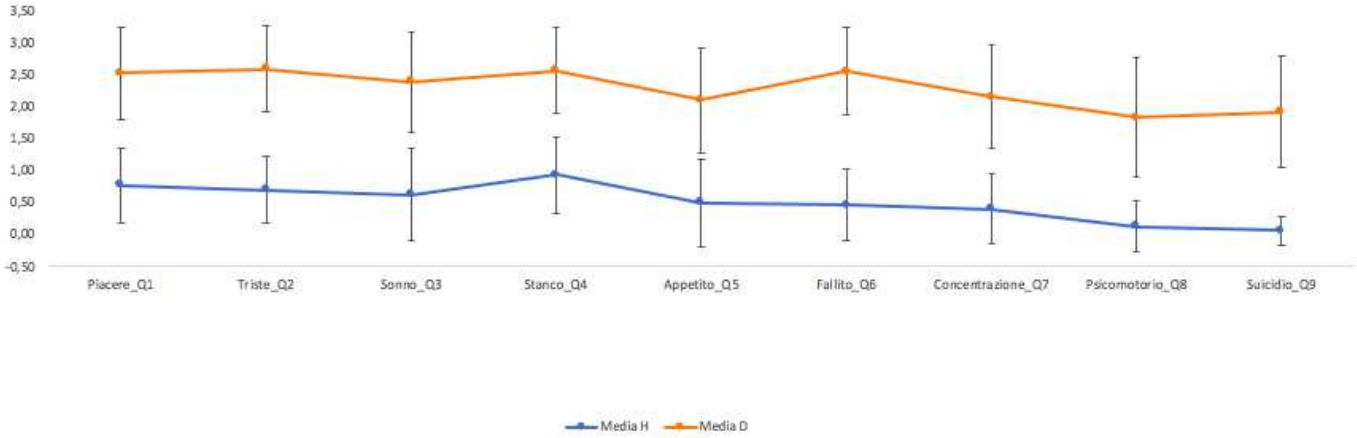


Figura 4.1 *- PHQ-9: grafico delle distribuzioni dei punteggi medi nelle condizioni honest e dishonest per ciascun item.*
Note: *Grafico delle distribuzioni dei punteggi grezzi medi e rispettive deviazioni standard nelle condizioni honest (blu) e dishonest (arancione). Sull'asse delle ascisse è raffigurato l'item, mentre sull'asse delle ordinate il punteggio medio ottenuto per il relativo item, all'interno di ciascuna condizione.*

Osservando la figura 4.1 si può notare l'andamento generale dei dati grezzi delle due condizioni a livello del singolo item, dal quale si evince un'evidente differenza tra le risposte che i soggetti danno quando rispondono in maniera sincera o simulando la presenza di un disturbo grave. Questo è dato dalla differenza a livello di statistica descrittiva, infatti i punteggi medi nella condizione *dishonest* risultano essere maggiori rispetto a quelli nella condizione *honest*, come mostrato nella tabella di seguito.

Tabella 4.5 *- Media e deviazione standard dei valori di ciascun item (Qi) nelle condizioni honest (H) e dishonest (D)*

Misure	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
Media (H)	0,77	0,70	0,62	0,93	0,49	0,46	0,40	0,12	0,06
Media (D)	2,52	2,58	2,38	2,56	2,10	2,55	2,16	1,83	1,91
SD (H)	0,58	0,53	0,72	0,59	0,69	0,56	0,54	0,40	0,23
SD (D)	0,72	0,67	0,78	0,67	0,82	0,68	0,81	0,94	0,87

Note: *In tabella sono rappresentati i valori del punteggio medio e della rispettiva deviazione standard per ogni item. Nelle colonne sono indicati gli item, mentre nelle righe sono indicate le misure indicate in ciascuna riga. La condizione onesta viene abbreviata con H, la condizione disonesta con D.*

Questa differenza tra le due condizioni indica che quando viene chiesto ai soggetti di rispondere simulando il disturbo, questi tendono a scegliere un punteggio più elevato rispetto alla condizione onesta, con una M (D) = 2,29 e DS (D) = 0,10, a differenza di una M (H) = 0,51, SD (H) = 0,15, considerando una scala likert di risposta con un punteggio massimo di 3 (“quasi ogni giorno”). Quindi in generale, quando i soggetti devono rispondere alle domande simulando un disturbo depressivo grave, questi scelgono punteggi elevati nella maggior parte degli item, indipendentemente dallo specifico contenuto semantico.

La media delle risposte delle due condizioni, si trovano quindi nei poli opposti: le medie delle risposte nella condizione *honest* nei valori più bassi della scala Likert mentre quelle delle risposte nella condizione *dishonest* nei valori più alti. Nella condizione *honest* gli item che hanno ottenuto i valori più bassi sono, in ordine crescente: Q9 (M= 0,06), Q8 (M=0,12), Q7 (M=0,40), Q6 (M=0,48), Q5 (M=0,49), Q3 (M=0,62); mentre nella condizione *dishonest* gli item con i valori più alti sono, in ordine decrescente: Q2 (M= 2,58), Q4 (M=2,56), Q6 (M=2,55), Q1 (M=2,52), Q3 (M=2,38), Q7 (2,16). Come si può notare, gli item Q3 (“Problemi ad addormentarsi o a dormire tutta la notte senza svegliarsi, o a dormire troppo”), Q6 (“Avere una scarsa opinione di sé, o sentirsi un/una fallito/a o aver deluso se stesso/a o i propri familiari”) e Q7 (“ Difficoltà a concentrarsi su qualcosa, per esempio leggere il giornale o guardare la televisione”), sono tra i primi sei punteggi medi più bassi nelle risposte oneste e i primi 6 punteggi medi più alti delle risposte disoneste, e quindi potremmo aspettarci che tali elementi siano i più utili per discriminare i soggetti onesti dai disonesti. Possiamo anche confermare questa ipotesi osservando i grafici di densità sottostanti (fig. 4.2) che mostrano che i dati per questi tre elementi sono quasi linearmente separabili. Infatti, le osservazioni *honest* (rappresentate

dal colore verde) sono concentrate in basso a sinistra, ovvero sui valori bassi sia nell'*item* in ascissa che in ordinata; al contrario le osservazioni *dishonest* (rappresentate dal colore rosa) sono concentrate soprattutto sui valori alti, in alto a destra in entrambi gli *item*. Questi tre *item* quindi si configurano in maniera similare, in cui la frequenza delle risposte è divisa nettamente: gli *honest* rispondono con maggior frequenza con punteggi come 0 e 1, mentre i *dishonest* con punteggi 2 e 3.

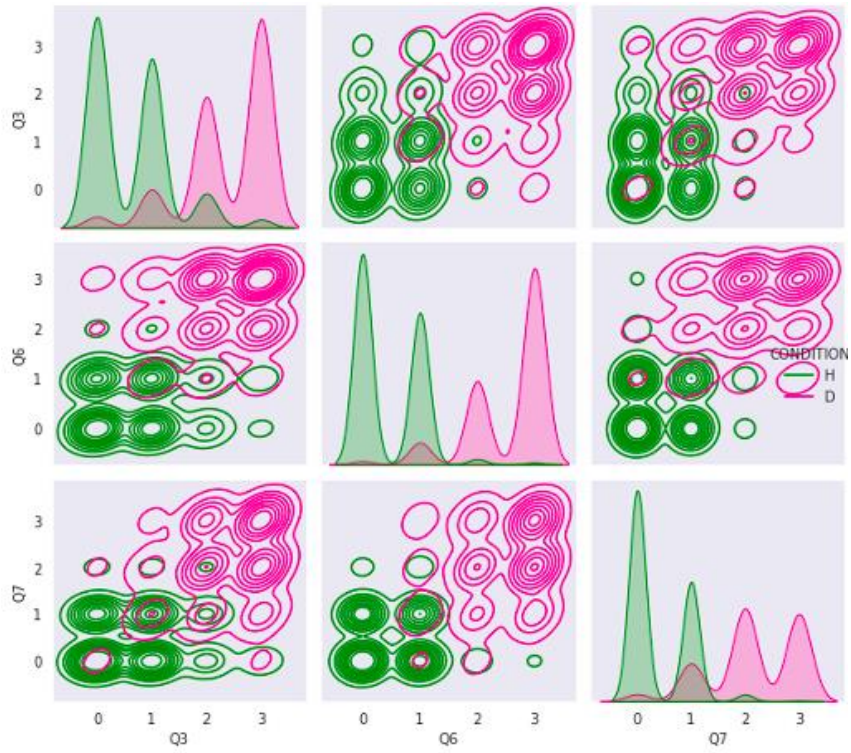


Figura 4.2 *- Grafici di densità delle risposte agli item Q3, Q6 e Q7 nelle due condizioni honest (color verde) e dishonest (color rosa). Le osservazioni honest sono concentrate in basso a sinistra, ovvero sui valori bassi sia nell'item in ascissa che in ordinata; al contrario le osservazioni dishonest sono frequenti soprattutto sui valori alti in alto a destra in entrambi gli item.*

Successivamente, è stata calcolata la differenza tra le risposte grezze dei partecipanti nelle due condizioni *honest* e *dishonest*, al fine di individuare le eventuali modifiche apportate dal soggetto nel rispondere alle domande in maniera sincera e simulando il disturbo depressivo. Nella tabella 4.6 sono riportate le percentuali di risposte che dalla condizione *honest* alla *dishonest* hanno subito un incremento (per esempio da 0 a 2) o un decremento (per esempio da 2 a 0), o che non hanno subito nessun cambiamento, rimanendo quindi invariate (il soggetto ha risposto 3 in entrambe le condizioni). Il PHQ-9 è un questionario costruito in modo tale che ad un punteggio elevato corrisponda una maggiore gravità del disturbo, infatti, come si può notare nella tabella, quasi tutte le risposte (96%) hanno subito un incremento del punteggio dalla condizione *honest* alla *dishonest*, comportamento coerente con il concetto di *faking bad* (paragrafo 2.1) secondo cui il soggetto per simulare un disturbo risponde alle domande scegliendo un punteggio più alto. Il 6% delle risposte non ha subìto

cambiamenti nelle due diverse condizioni, mentre una percentuale molto bassa (2%), ha subìto un decremento della risposta dalla condizione sincera a quella simulata. Tali modifiche, in direzione opposta rispetto a quella attesa, potrebbero essere spiegate dalle differenti strategie di simulazione nella condizione *dishonest*: i soggetti, dovendo fingere un disturbo grave in maniera credibile, potrebbero aver deciso di attuare la tecnica di non mentire in alcune domande specifiche.

Tabella 4.6 *- Numero di risposte che nella condizione dishonest subiscono modifiche rispetto alla condizione honest*

Totale risposte	Incremento risposte	Decremento risposte	Risposte invariate
4761	4388 (92%)	86 (2%)	287 (6%)

Note: *nella prima colonna è indicato il numero totale di risposta fornite nel complesso al questionario. nelle colonne successive è indicato il numero di risposte che subiscono un incremento, un decremento oppure non subiscono alcuna modifica dalla condizione honest alla condizione dishonest. Infine, tra parentesi è indicato il rispettivo valore in percentuale.*

4.1.2 Analisi statistiche sui dati grezzi

In primis è stata valutata la normalità della distribuzione del campione attraverso il test Shapiro-Wilk. Come spiegato precedentemente nelle analisi sulla normalità della distribuzione dei due gruppi HD e DH, questo test assume come ipotesi nulla una distribuzione normale del campione e quindi di forma gaussiana, la quale viene accettata quando il *p-value* è maggiore di 0,05. Nel nostro caso i *p-value* ottenuti per ogni item sono risultati significativi, ovvero inferiori a < 0,05, nello specifico sono inferiori a 0,001, a significare che l'ipotesi nulla può essere rifiutata e quindi che la distribuzione dei dati non è normale.

Secondariamente è stata condotta un'analisi statistica più approfondita sulla differenza dei dati nelle due condizioni *honest* e dishonest (tabella 4.7). Seppure tale differenza sia già emersa nelle analisi precedenti effettuate attraverso i punteggi medi (paragrafo 4.1.1), tale passaggio occorre per evidenziare la significatività a livello statistico. Poiché il campione non è distribuito normalmente, come emerso dal test *Shapiro-Wilk*, le analisi sono state condotte attraverso il test di Wilcoxon per i dati non parametrici in campioni non indipendenti, equivalente al *t-test* di *Student* per dati parametrici. Qui, un *p-value* > 0,05 confermerebbe l'ipotesi nulla, vale a dire che le due distribuzioni sono uguali e quindi non differiscono statisticamente, mentre un valore < 0,05 comporterebbe un rigetto dell'ipotesi nulla a favore dell'ipotesi alternativa, secondo la quale c'è una differenza significativa tra la condizione *honest* e *dishonest*.

Tabella 4.7 *– P-value ed effect-size calcolati per ogni item nelle due condizioni H e D attraverso il test di Wilcoxon*

Measure 1		Measure 2	W	df	p	Rank-Biserial Correlation
Piacere_Q1_H	-	Piacere_Q1_D	3909.000		< .001	−0.941
Triste_Q2_H	-	Triste_Q2_D	860.000		< .001	−0.987
Sonno_Q3_H	-	Sonno_Q3_D	1442.000		< .001	−0.976
Stanco_Q4_H	-	Stanco_Q4_D	1223.000		< .001	−0.980
Appetito_Q5_H	-	Appetito_Q5_D	1364.500		< .001	−0.976
Fallito_Q6_H	-	Fallito_Q6_D	158.000		< .001	−0.998
Concentrazione_Q7_H	-	Concentrazione_Q7_D	501.000		< .001	−0.992
Psicomotorio_Q8_H	-	Psicomotorio_Q8_D	1072.500		< .001	−0.981
Suicidio_Q9_H	-	Suicidio_Q9_D	0.000		< .001	−1.000

Note: *In tabella si presenta il confronto tra lo stesso item nelle due diverse condizioni (H, honest e D, dishonest). Nella colonna W è indicato il valore statistico di Wilcoxon, mentre nella colonna p è rappresentato il p-value. Poiché questo risulta essere sempre <,001, si può rifiutare l'ipotesi nulla e affermare che vi è la presenza di una differenza significativa tra le due condizioni i tuti gli item. Infine, nell'ultima colonna è indicato il valore di effect size.*

Nella tabella 4.7, è possibile osservare come i *p-value* ottenuti in questa analisi siano tutti inferiori a 0,001 e sono quindi indicativi di una differenza statisticamente significativa.

La dimensione dell'effetto, indicata in tabella come "*Rank-Biserial Correlation"*, è compresa tra 0,94 e 1 e, in accordo con la classificazione di Cohen, è indicativa di un *effect-size* con magnitudo *large*, ovvero di una differenza tra le due condizioni molto elevata. Infatti, una *d* di Cohen pari a 1 significa che i punteggi medi delle due condizioni H e D differiscono di una deviazione standard.

Matrici di correlazione

Di seguito vengono proposte le matrici di correlazione dei punteggi grezzi, in cui viene analizzata la correlazione tra le condizioni *honest-honest, dishonest-dishonest* e *honest-dishonest* (fig. 4.3). In questo caso, quando i valori si avvicinano ad 1, questo indica che le due variabili H e D sono tra loro correlate, al contrario più questi sono vicini allo 0, più le variabili non sono correlate e quindi mancano di una correlazione lineare secondo cui al variare dell'una varia anche l'altra, rendendo meno prevedibile il loro comportamento.

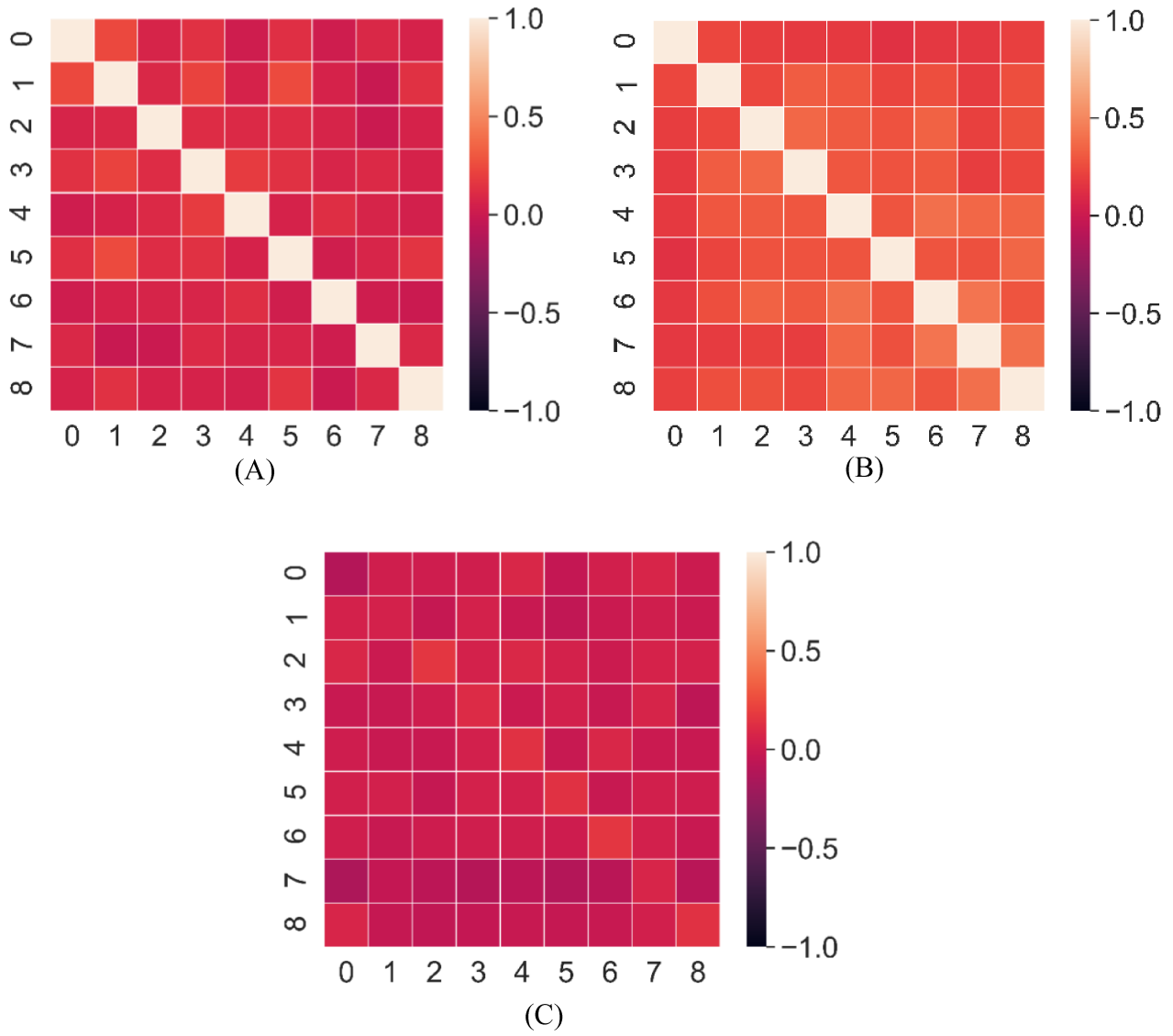


Figura 4.3 *- Matrici di correlazione dei paragoni tra le condizioni honest e dishonest. Nello specifico A) paragone honest/honest; B) paragone dishonest/dishonest ed infine C) matrice di correlazione honest/dishonest. Si può notare come nella condizione honest/dishonest, a differenza delle altre due matrici, la correlazione si avvicini allo 0.*

Nella figura 1.2 si può osservare come le correlazioni *honest-honest* (A) e *dishonest-dishonest* (B) sono lineari, con una correlazione più forte all'interno di quest'ultimo *dataset*, pari a 0,35. La figura 1.2 C, invece, raffigurante la condizione di paragone *honest-dishonest*, differisce dalle precedenti, in quanto, come si evince dai suoi colori, si avvicina allo 0 ed è quindi indicativa di un'assenza di correlazione e di conseguenza di una correlazione non lineare. Questo implica che non c'è una procedura immediata per predire le risposte oneste, pertanto l'identificazione della risposta simulata all'interno di un questionario, partendo dai soli dati grezzi, non è un'operazione banale.

Test Kruskal-Wallis

Sono state condotte altre analisi sulla condizione sperimentale *honest-dishonest*, al fine di verificare l'interazione di questa con le altre varabili indagate nel questionario, come genere, età e anni di scolarizzazione. Per fare ciò, essendo il nostro campione caratterizzato da un andamento della distribuzione non normale, è stato usato il test della varianza Kruskal-Wallis, l'alternativa non parametrica dell'ANOVA, considerando come significativo un *p-value* < 0,05.

GENERE. Per quanto riguarda la variabile "genere", attraverso *il Kruskal-Wallis* è emersa una differenza tra le due condizioni: nel rispondere onestamente non è stata rilevata nessuna differenza tra "maschi", "femmine" e "altro"; invece nella condizione di simulazione il valore p < 0,01 indica che c'è una differenza statisticamente significativa tra le risposte date dal genere "femmina" e "maschio".

È stato poi condotto il test di Dunn, un test Post-hoc non parametrico, sulle condizioni sopracitate al fine di indagare quale dei gruppi differissero tra di loro nel rispondere alle domande in maniera simulata. Come si può osservare nella tabella 4.8, il genere "femmina" quando mente tende a rispondere con punteggi più elevati rispetto al genere "maschio", mentre non è stata riscontrata alcuna differenza con chi si è identificato in "altro".

Tabella 4.8 *– Test di Dunn per la variabile "genere" nella condizione dishonest*

Comparison	z	W_i	W_j	p	p_{bonf}	p_{holm}
F – M	4.455	281.923	211.779	< .001	< .001	< .001
F – Altro	1.151	281.923	203.000	0.125	0.374	0.250
M – Altro	0.126	211.779	203.000	0.450	1.000	0.450

Note: *In tabella è riportato il confronto post-hoc tra i tre gruppi riguardanti il "genere" attraverso il test di Dunn. Nella colonna p è indicato il p-value, che se < 0,05 comporta un rifiuto dell'ipotesi nulla e quindi l'accettazione dell'ipotesi alternativa, secondo la quale i due gruppi hanno risposto alle domande simulate con una differenza statisticamente significativa.*

ETÀ. Per valutare l'interazione di questa variabile in relazione alla condizione sperimentale *honest-dishonest*, i soggetti sono stati raggruppati in classi di età, nello specifico: 18-30, 31-40, 41-50, 51-60, 60-74. Anche in questo caso, nella condizione onesta non è stata rileva una differenza nella modalità di risposta alle domande del questionario, mentre nella condizione disonesta l'appartenere a classi di età differenti è risultato essere significativo.

Nello specifico, come mostra la tabella 4.9, attraverso il test *Post-hoc* di Dunn è stato visto come i soggetti più giovani, appartenenti alla fascia d'età 18-30 e 31-40, quando mentono danno

punteggi più bassi rispetto ai soggetti con un'età compresa tra i 41 e i 60 anni, appartenenti alle rispettive fasce d'età 41-50 e 51-60.

Tabella 4.9 *– Test di Dunn per la variabile "età" nella condizione dishonest*

Comparison	z	W_i	W_j	p	p_{bonf}	p_{holm}
18-30 - 31-40	0.060	246.910	245.759	0.476	1.000	1.000
18-30 - 41-50	−3.311	246.910	312.297	< .001	0.005	0.004
18-30 - 51-60	−3.577	246.910	325.553	< .001	0.002	0.002
18-30 - 61-74	−0.609	246.910	273.192	0.271	1.000	1.000
31-40 - 41-50	−2.716	245.759	312.297	0.003	0.033	0.023
31-40 - 51-60	−3.030	245.759	325.553	0.001	0.012	0.010
31-40 - 61-74	−0.603	245.759	273.192	0.273	1.000	1.000
41-50 - 51-60	−0.494	312.297	325.553	0.311	1.000	1.000
41-50 - 61-74	0.854	312.297	273.192	0.197	1.000	0.983
51-60 - 61-74	1.118	325.553	273.192	0.132	1.000	0.790

Note: *In tabella è riportato il confronto post-hoc tra i cinque gruppi "classi di età" attraverso il test di Dunn. Nella colonna p è indicato il p-value, che se < 0,05 05 comporta un rifiuto dell'ipotesi nulla e quindi l'accettazione dell'ipotesi alternativa, secondo la quale i due gruppi hanno risposto alle domande simulate con una differenza statisticamente significativa.*

ANNI DI SCOLARIZZAZIONE. Infine, per quanto riguarda gli anni di scolarità, attraverso il *Kruskal-Wallis* è possibile notare come questa variabile non influisca sulla modalità di risposta alle domande sia nella condizione *honest* che *dishonest.*

4.2 TF-IDF

Dopo aver condotto le analisi sui dati grezzi, questi sono stati trasformati in valori TF-IDF per andare a verificare la capacità discriminatoria di questo nuovo metodo di detezione della simulazione. L'ipotesi H3 infatti, implica che l'indice TF-IDF permetta una discriminazione più accurata tra onesti e disonesti rispetto ai dati grezzi. Per svolgere questa analisi sono state utilizzate le formule riportate nel paragrafo 2.4.1; inoltre, si specifica che per calcolare l'indice TF-IDF delle risposte *dishonest* è stato utilizzato il valore IDF delle risposte *honest*, così da valutare come i *dishonest* si collocavano all'interno della distribuzione onesta.

KL-divergence (KLD)

Dovendo verificare la correttezza dell'ipotesi numero tre, ovvero che l'indice TF-IDF riesce a discriminare meglio gli onesti dai non, per prima cosa è stato indagato il KLD. Su questo ci occorre solamente sapere che consiste in una misura che si basa sull'entropia e che permette di effettuare un

paragone tra due diverse distribuzioni di probabilità. Per interpretare i suoi valori, basta sapere che il suo *range* varia da 0 a infinito, in cui lo 0 indica una sovrapposizione completa delle due distribuzioni (in questo caso *honest* e *dishonest*), mentre un valore alto è indicativo di un distanziamento tra di queste. In tabella 4.8 viene messo a confronto il KLD dei dati grezzi e dell'indice TF-IDF.

Tabella 4.10 *– Valori KLD delle distribuzioni honest e dishonest per quanto riguarda i punteggi grezzi e i valori TF-IDF*

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
Dati grezzi	2,28	9,37	1,82	2,20	1,72	2,87	2,04	1,93	2,89
TF-IDF	6,53	7,53	6,21	6,11	5,64	7,20	6,64	9,93	9,98

Note: *In tabella sono indicati i valori KLD di ciascun item, sia per quanta riguarda i dati grezzi che il TF-IDF. Maggiore è il valore KLD, tanto più le distribuzioni sono discriminabili tra loro. Si noti come appunto i valori KLD corrispondenti al TF-IDF, ad eccezione di uno, siano più elevati rispetto a quelli dei dati grezzi, ad indicare che le distribuzioni sono più discriminabili nella prima condizione.*

Dalla tabella si evince come l'indice KLD, quando applicato ai valori TF-IDF è maggiore rispetto a quando applicato ai dati grezzi, a dimostrazione che la discriminazione tra le risposte oneste e simulate è più accurata se effettuata con l'indice TF-IDF, avvalorando così l'ipotesi numero 3. Sebbene nell'item Q2 il valore KLD, seppur di poco, sia maggiore per i dati grezzi, tale risultato può essere trascurato in quanto il TF-IDF riesce a discriminare gli *honest* dai *dishonest* in maniera accurata in tutti gli item, a differenza dei dati grezzi.

Valutazione della performance

Al fine di valutare, ancora una volta, la capacità del TF-IDF di discriminare i soggetti *honest* dai *dishonest,* è stata effettuata una ricerca del valore soglia (*cut-off*) più accurato, al di sopra del quale ricadono le risposte simulate e al di sotto le risposte oneste.
Per fare questo è stata utilizzata la *K-fold cross validation*[4] (*K*=10) una procedura di *Machine* Learning in cui il campione viene diviso in 10 sezioni (gruppi): 9 usate come *training* su cui l'algoritmo si allena con lo scopo di trovare il valore soglia che possa essere valido per l'individuazione delle risposte anomale in tutti gli *item* e una in cui il *cut-off* emerso viene testato. Tale procedura viene iterata più volte ed ogni volta viene usato un gruppo di *test* diverso e, di

[4] La *k-fold cross validation*, o cosiddetta anche convalida incrociata, consiste nella suddivisione dell'intero dataset in k parti di uguale numerosità e, ad ogni passo, una di queste k parti viene a essere quella di convalida, mentre le restanti costituiscono l'insieme di addestramento. In altre parole, si suddivide il campione in gruppi di numerosità uguale, si esclude poi iterativamente un gruppo alla volta e, con i gruppi non esclusi si cerca di predirlo, in maniera tale da verificare la bontà del modello di predizione utilizzato. (https://it.wikipedia.org/wiki/Convalida_incrociata)

conseguenza, anche gruppi di *training* diversi. Alla fine del procedimento si ottengono 10 valori percentili, uno per ogni interazione, ovvero quello che ha dimostrato una precisione maggiore. Dei 10 valori finali, si sceglie come valore soglia quello più frequente, in quanto risulta essere quello che in tutti gli *item* riesce meglio a categorizzare le risposte anomale dalle sincere e che quindi massimizza la precisione nella discriminazione delle due distribuzioni. Questa procedura è stata applicata anche ai dati grezzi (*Distribution Based Model*), così da effettuare un paragone tra questi e il TF-IDF nella capacità di discriminare gli onesti dai simulatori.

Dalla *k-fold cross validation* è emerso che il *cut-off* con la migliore capacità discriminativa è risultato essere il 95° percentile sia per i dati grezzi per il TF-IDF, comparendo dieci interazioni su dieci in entrambe le tecniche.

Successivamente, è stata effettuata un'analisi della prestazione in termini di *precision*, *recall*, *F1-score* e *accuracy*, sia per i dati grezzi che per i valori TF-IDF, attraverso l'implementazione di un ulteriore modello (*Multi-Label Classification task*). Tali indici sono stati calcolati nei seguenti modi:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 * \frac{precision * recall}{precision + recall} \qquad Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

In cui:

- TP = *true positive* (veri positivi) cioè numero di risposte categorizzate correttamente dal modello come simulate;
- TN = *true negative* (veri negativi) cioè numero di risposte categorizzate correttamente come non simulate;
- FP = *false positive* (falsi positivi) cioè numero di risposte categorizzate come simulate ma che in realtà sono sincere;
- FN = *false negative* (falsi negativi) cioè numero di risposte categorizzate come sincere ma che in realtà sono simulate.

Gli indici di *performance* appena calcolati sono indicati nella tabella 4.11, permettendo un confronto immediato.

Tabella 4.11 *– Valori degli indici di performance indagati, per i dati grezzi e per il TF-IDF*

	Precision	*Recall*	*F1-score*	*Accuracy*
Dati grezzi	0,96	0,95	0,95	0,93
TF-IDF	0,97	0,78	0,85	0,80

Note: *In tabella sono indicati i valori degli indici di prestazione per ciascuna metodologia. Tali indici sono indicati nelle colonne, mentre le tecniche (dati grezzi, TF-IDF) nelle righe.*

Tra i vari indici di *performance* riportati nella tabella 4.11, quello che più ci interessa è la *Precision*, poiché risulta essere l'indice di prestazione più informativo e massimizzato all'interno del modello creato, in quanto è quello che ha meno probabilità di categorizzare un soggetto onesto come disonesto, riducendo il rischio di falsi positivi. Il TF-IDF, come si può vedere in tabella risulta leggermente massimizzato in termini di *Precision* rispetto ai dati grezzi nel discriminare i soggetti *honest* dai *dishonest* al 95° percentile. Mentre, per quanto riguarda gli altri tre indici (*Recall, F1-score* e *accuracy*), questi presentano un valore più elevato quando applicati ai dati grezzi, ad indicare che per questo tipo di analisi, l'utilizzo della distribuzione dei dati grezzi si dimostra utile ed informativa metodologia nella discriminazione delle due distribuzioni. Nei grafici riportati in figura 4.4, A e B, si può osservare il confronto tra la *Precision* dei dati grezzi e del TF-IDF: esaminando la funzione di densità della probabilità (linea rossa osservabile nei grafici della figura 4.4), si può notare come questa cresca all'aumentare del numero di domande simulate, raggiungendo il picco quando tutti gli *item* (9) sono simulati. Questo indica che c'è una relazione proporzionale al numero di domande simulate in entrambe le tecniche: più domande vengono simulate, più è alta la *Precision*. Inoltre, tutti i partecipanti alterano la propria risposta almeno in almeno due domande.

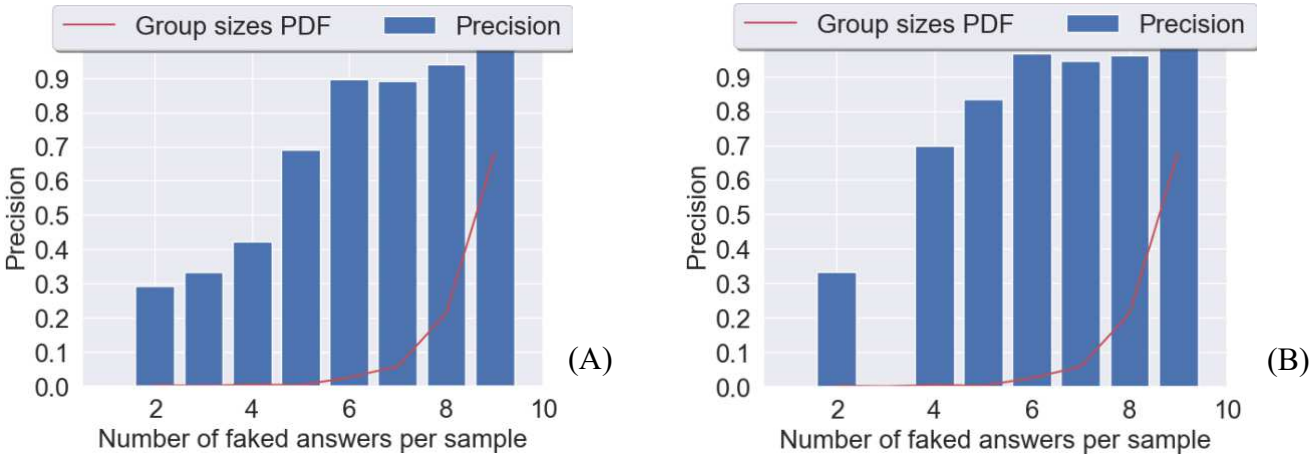


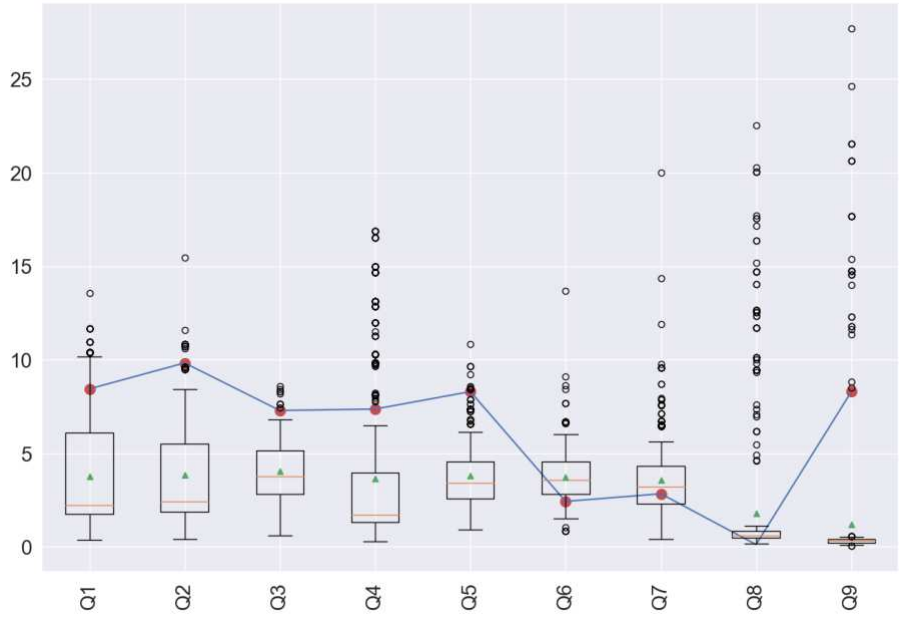
Figura 4.4 *– Grafici degli indici di Precisione per i dati grezzi (A) e per il TF-IDF (B). Nell'asse delle ascisse è rappresentato il numero di domande simulate, mentre sull'asse delle ordinate i valori da 0 a 1 dell'indice di Precision. In ciascuna barra degli istogrammi riportiamo la Precision media*

ottenuta dal modello sui diversi sottoinsiemi di questionari. Ogni sottoinsieme contiene tutti i questionari in cui i soggetti hanno falsificato un numero di risposte indicate sull'asse x.

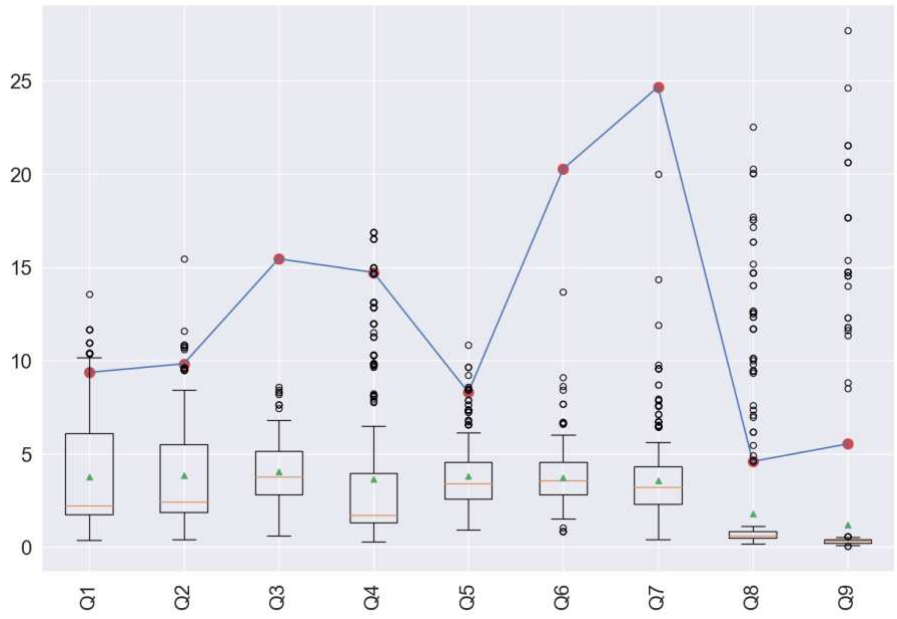
Identificazione della simulazione nel singolo item

Come spiegato precedentemente, il nuovo metodo di detezione della simulazione attraverso l'indice TF-IDF, a differenza dei metodi tradizionali che indagano la tendenza generale dell'individuo a simulare, riesce ad individuare dove il soggetto ha mentito a livello di ogni singolo item. Al fine di comprendere meglio il suo funzionamento sono presentati in figura 4.4 due grafici *boxplot* di due soggetti del campione, in cui viene comparata la distribuzione *dishonest* di un soggetto con la distribuzione dei punteggi *honest* dell'intero campione. Per una comprensione dei grafici occorre sapere che:

- sull'asse delle ascisse sono rappresentati i 9 item del questionario;
- sull'asse delle ordinate sono raffigurati i valori TF-IDF;
- i *box* nella parte bassa del grafico rappresentano la distribuzione dei punteggi *honest*, in cui:
    - la linea arancione all'interno dei box sta ad indicare il punteggio medio dell'*item* di riferimento,
    - le estremità del *box* rappresentano il primo ed il terzo quartile,
    - gli *whiskers* inferiori e superiori rappresentano rispettivamente il punteggio minimo e il punteggio massimo ottenuto in quello specifico *item;*
- la linea blu raffigura la distribuzione dei punteggi *dishonest* ottenuti dal singolo soggetto;
- i puntini rossi segnati sulla linea blu indicano l'alterazione della risposta rispetto alla condizione onesta in quello specifico *item*, evidenziano quindi le domande nelle quali è stata rilevata un'anomalia, ovvero è stata simulata.



(A)



(B)

Figura 4.5 *– Grafici boxplot rappresentanti l'andamento dei valori TF-IDF dishonest (linea blu) dei soggetti n. 13 (A) e 14 (B) comparati con la distribuzione honest (box nella parte inferiore del grafico) del TF-IDF dell'intero campione. Sull'asse delle ascisse sono presentati i 9 item del questionario, mentre sull'asse delle ordinate i valori TF-IDF. Il pallino rosso indica che la risposta alla domanda cui fa riferimento è stata alterata rispetto alla condizione honest.*

Nella figura 4.4 A, possiamo vedere la distribuzione *dishonest* dei punteggi del partecipante n. 13 Si può vedere come i pallini rossi sono presenti in otto *item* su nove, ad indicare che il soggetto ha simulato in tutto il questionario tranne che nell'*item* Q8, in cui la risposta è stata identificata come sincera ed infatti ricade all'interno dei baffi del *boxplot honest*. Inoltre, ciò che si evince dal grafico è che l'accuratezza di rilevazione è pari a 6/8, poichè la simulazione è stata rilevata correttamente dal TF-IDF in sei risposte simulate su otto. Infatti, nell'*item* Q1, Q6 e Q7 il pallino rosso ricade nel *box* delle risposte *honest,* per cui tre domande oneste sono state identificate dal TF-IDF come simulate. Nella figura 4.4 B invece, è riportato l'esempio di un altro soggetto, il numero 14 del campione. Qui, come si può notare nel grafico, l'alterazione riguarda tutte le domande, però la risposta data all'item Q1 è stata riconosciuta dal TF-IDF come alterata, ma non correttamente individuata poiché ricade nel *box* della distribuzione onesta. Perciò l'accuratezza qui risulta essere di 8/9.

4.3 Ricostruzione delle risposte oneste

In seguito alle analisi implementate e sopra descritte, che permettono di identificare a livello del singolo item la risposta simulata, un secondo obiettivo del presente studio è stato quello di ricostruire le risposte oneste partendo da quelle simulate dai partecipanti. L'obiettivo in questa seconda analisi è quello di "pulire" i dati simulati dalla menzogna e riuscire, attraverso un modello, a sovrapporli a

quelli del gruppo onesto, in modo da rispondere alla domanda: "come avrebbero risposto i partecipanti che hanno simulato se fossero stati onesti?". Di seguito verrà descritta la metodologia implementata.

4.3.1 Classificazione

Il primo passaggio di questo processo di ricostruzione delle risposte è consistito in un compito di classificazione di tipo binaria, poiché la variabile, essendo di tipo categorico, può assumere solo questi due valori. La classificazione è stata implementata utilizzando l'algoritmo *Logistic Regression*, un efficiente modello di classificazione binaria che ha come obiettivo quello classificare un'osservazione - in questo caso un partecipante del *dataset* - nell'una o nell'altra categoria della variabile dipendente, vale a dire riconoscerla come onesta o disonesta, in base alle sue caratteristiche[5];

Questo algoritmo di *Machine Learning* è stato addestrato su una parte di dati (*training set*), poi una volta allenato, l'algoritmo è stato testato sui restanti dati (*test set*). In tabella 4.12 sono mostrati i risultati della *performance* ottenuti dall'applicazione dell'algoritmo.

Tabella 4.12 *- Valori degli indici di performance indagati per l'algoritmo Logistic Regression*

Modello	*Accuracy*	*AUC*	*Recall*	*Precision*	*F1*	*Kappa*	*MCC*	*TT (Sec)*
Logistic Regression	0,99	1	0,99	0,98	0,99	0,98	0,98	0,52

Note: *In tabella sono indicati i valori degli indici di prestazione per l'algoritmo Logistic Regression. Tali indici sono indicati nelle colonne, mentre l'algoritmo su cui sono stati applicati, nella riga.*

Secondariamente alla classificazione nei due gruppi *honest* e *dishonest* attraverso l'algoritmo *Logistic Regression, è* stata estratta la probabilità di ogni partecipante di appartenere al gruppo dei mentitori. Ad esempio, la probabilità che il soggetto 500 ha di essere un mentitore è pari a 0.9066 (ovvero 91%), quindi questo viene classificato come mentitore.

Dopo aver calcolato la probabilità di ogni soggetto, è stata parallelamente calcolata una seconda misura, definita come *Delta*. Il *Delta* corrisponde alla sottrazione della media delle risposte oneste, dalla media delle risposte mentite. Tale operazione è stata effettuata per ogni item:

Delta 1 = Media Q1 D – Media Q1 H = 1,76

Delta 2 = Media Q2 D – Media Q2 H = 1,88

[5] https://it.wikipedia.org/wiki/Modello_logit

Delta 3 = Media Q3 D – Media Q3 H = 1,76

Delta 4 = Media Q4 D – Media Q4 H = 1,63

Delta 5 = Media Q5 D – Media Q5 H = 1,61

Delta 6 = Media Q6 D – Media Q6 H = 2,10

Delta 7 = Media Q7 D – Media Q7 H = 1,75

Delta 8 = Media Q8 D – Media Q8 H = 1,71

Delta 9 = Media Q9 D – Media Q9 H = 1,85

Quindi, mentre il *Delta* è diverso per ogni *item,* ma uguale fra i partecipanti, la probabilità è uguale per tutti gli *item* ma diversa tra un soggetto e l'altro

Una volta calcolata la probabilità e il *Delta*, come si può osservare nella tabella 4.13, per ogni *item* di ogni partecipante la probabilità del soggetto di essere un mentitore è stata moltiplicata per il *Delta* corrispondente a quell' *item*.

Tabella 4.13 *– Moltiplicazione effettuata per ogni item di ciascun partecipante*

Partecipante	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
Sogg 1	0.91* Delta1 (1,76)	0.91* Delta2 (1,88)	0.91* Delta3 (1,76)	0.91* Delta4 (1,63)	0.91* Delta5 (1,61)	0.91* Delta6 (2,10)	0.91* Delta7 (1,75)	0.91* Delta8 (1,71)	0.91* Delta9 (1,85)
Sogg 2	0.85* Delta1 (1,76)	0.85* Delta2 (1,88)	0.85* Delta3 (1,76)	0.85* Delta4 (1,63)	0.85* Delta5 (1,61)	0.85* Delta6 (2,10)	0.85* Delta7 (1,75)	0.85* Delta8 (1,71)	0.85* Delta9 (1,85)
Sogg 3	0.60* Delta1 (1,76)	0.60* Delta2 (1,88)	0.60* Delta3 (1,76)	0.60* Delta4 (1,63)	0.60* Delta5 (1,61)	0.60* Delta6 (2,10)	0.60* Delta7 (1,75)	0.60* Delta8 (1,71)	0.60* Delta9 (1,85)

Note: *In tabella sono indicate le moltiplicazioni effettuate per ogni item (colonne) di ciascun partecipante (righe) tra la probabilità del soggetto di essere un mentitore e il Delta, ovvero la differenza tra la media delle risposte mentite e la media delle risposte oneste.*

In questo ultimo passaggio della ricostruzione della risposta, il valore ottenuto dalla moltiplicazione tra la probabilità e il *Delta*, che è diverso per ogni domanda di ciascun partecipante, è stato sottratto al punteggio grezzo dato dal mentitore a quel determinato *item*. Questa operazione viene effettuata in tutti i punteggi grezzi dei disonesti, ottenendo così per ogni domanda un nuovo punteggio, ovvero quello ricostruito.

Il *dataset* finale è quindi costituito dai dati grezzi originali onesti e dai dati onesti ricostruiti partendo dai grezzi simulati.

Per comprendere meglio la ricostruzione delle risposte oneste partendo da quelle simulate dai partecipanti, effettuata attraverso i passaggi sopra elencati, occorre osservare la rappresentazione grafica dei dati qui di seguito (figura 4.6). La distribuzione in colore blu rappresenta le medie dei punteggi grezzi originali degli *honest*, quella in rosso le medie dei punteggi grezzi originali dei *dishonest*, mentre quella in verde rappresenta i punteggi medi disonesti ricostruiti attraverso il processo mostrato fin qui.

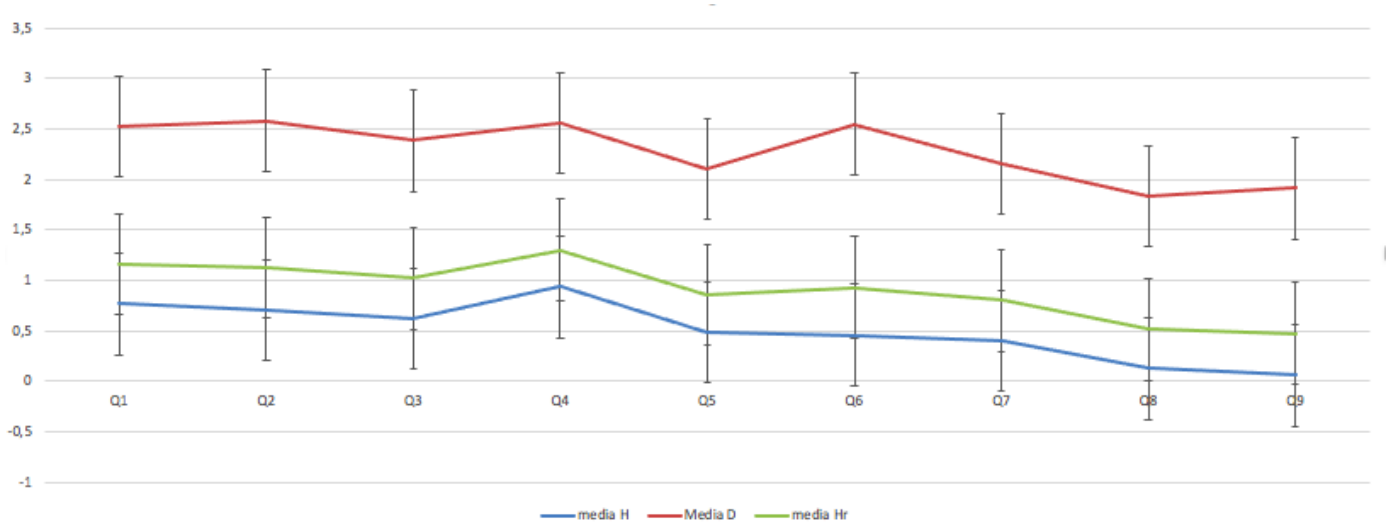


Figura 4.6 *- grafico delle distribuzioni delle medie dei punteggi onesti grezzi originali (colore blu), dei disonesti grezzi originali (colore rosso) e dei punteggi disonesti ricostruiti (colore verde).*
Note: *Sull'asse delle ascisse è raffigurato l'item, mentre sull'asse delle ordinate il punteggio medio ottenuto per il relativo item, all'interno di ciascuna condizione.*

Come si può osservare nel grafico (figura 4.6), la distribuzione dei dati *honest* ricostruiti (colore verde), si avvicina molto alla distribuzione *honest,* infatti attraverso la moltiplicazione della probabilità per il *Delta* e la sottrazione di questo valore dal punteggio grezzo, si riesce ad eliminare la "quota" di menzogna ed avvicinarsi alla risposta che il soggetto avrebbe dato se avesse risposto onestamente. Questo abbassamento dei valori medi, si riscontra anche nei dati grezzi riportati in tabella 4.14.

Tabella 4.14 *– Mediana, Media, deviazione standard, minimo e massimo dei valori di ciascun item (Qi) nelle condizioni honest (H) e honest ricostruita (Hr)*

	Q1		Q2		Q3		Q4		Q5		Q6		Q7		Q8		Q9	
	H	Hr	H	Hr	H	Hr	H	Hr	H	Hr	H	Hr	H	Hr	H	Hr	H	Hr
M	0,77	1,16	0,70	1,13	0,62	1,02	0,93	1,30	0,50	0,85	0,46	0,93	0,40	0,80	0,13	0,51	0,06	0,48
Med	1,00	1,45	1,00	1,35	1,00	1,40	1,00	1,57	0,00	0,76	0,00	1,16	0,00	0,69	0,00	0,60	0,00	0,48
DS	0,58	0,66	0,53	0,58	0,72	0,68	0,60	0,58	0,69	0,72	0,56	0,57	0,54	0,70	0,40	0,84	0,23	0,76
Min	0,00	-1,52	0,00	-1,51	0,00	-1,42	0,00	-1,30	0,00	-1,37	0,00	-1,51	0,00	-1,30	0,00	-1,44	0,00	-1,44
Max	3,00	2,31	2,00	2,26	3,00	2,05	3,00	2,36	3,00	2,11	3,00	2,17	3,00	2,00	3,00	2,18	1,00	2,41

Note: *In tabella è riportata la statistica descrittiva (media, mediana, deviazione standard, minimo e massimo) dei valori di ogni item nelle due condizioni honest (H) e honest ricostruiti (Hr). Nelle colonne sono indicati gli item nelle due condizioni, mentre nelle righe sono riportate le misure indicate in ciascuna riga.*

Infatti, come si può osservare nella tabella 4.15 di seguito riportata, confrontando le differenze tra le medie *honest* (H) e *dishonest* (D) originali e le differenze tra le medie delle risposte *honest* (H) e *honest* ricostruite (Hr) si può apprezzare come i valori sono molto più piccoli nella seconda condizione rispetto alla prima, ad indicare una riduzione della differenza e quindi un avvicinamento delle risposte ricostruite a quelle oneste.

Tabella 4.15 *– Differenza delle medie honest (H) e dishonest (D) originali e differenza delle medie delle risposte honest (H) e honest ricostruite (Hr) per ogni item*

item	Differenza delle medie H e D	Differenza delle medie H e Hr
Q1	1,76	0,40
Q2	1,88	0,43
Q3	1,76	0,40
Q4	1,63	0,37
Q5	1,61	0,36
Q6	2,10	0,47
Q7	1,75	0,40
Q8	1,71	0,39
Q9	1,85	0,42

Note: *in tabella sono riportate le differenze delle medie H e D e delle medie H e Hr per ogni item. Nelle righe sono indicati gli item mentre nelle colonne le due differenze confrontate.*

Osservando la tabella è possibile vedere un netto avvicinamento della riposta ricostruita alla distribuzione onesta, con una distanza massima di 0,47 tra le due distribuzioni, a differenza di un 2,10 nella condizione onesta e simulata.

In ultimo, è stato applicato nuovamente l'algoritmo *Logistic Regression* a questo nuovo campione di dati ed è stata valutata nuovamente la *performance*, in tabella 4.16 sono mostrati i risultati ottenuti.

Tabella 4.16 *- Valori degli indici di performance indagati per l'algoritmo Logistic Regression nel nuovo campione di dati con le domande ricostruite*

Modello	***Accuracy***	***AUC***	***Recall***	***Precision***	***F1***	***Kappa***	***MCC***	***TT (Sec)***
Logistic Regression	0,79	0,87	0,81	0,79	0,80	0,57	0,58	0,32

Note: *In tabella sono indicati i valori degli indici di prestazione per l'algoritmo Logistic Regression. Tali indici sono indicati nelle colonne, mentre l'algoritmo su cui sono stati applicati, nella riga.*

Se confrontati con i valori della *performance* della campione iniziale H e D presenti (tabella 4.12) si può notare come i valori ottenuti qui, nel nuovo campione ricostruito, sono diminuiti, dimostrando una difficoltà dell'algoritmo (sebbene comunque risulti una buona performance) nel classificare le distribuzioni *honest* e *honest* ricostruita.

Capitolo V

DISCUSSIONE DEI RISULTATI

5.1 Struttura e scopo dell'esperimento

Lo scopo di questo elaborato è duplice: innanzitutto è stata testa una nuova tecnica di *lie detection*, ovvero il TF-IDF, poi, sulla base delle risposte simulate individuate da tale tecnica, è stata testata una procedura con l'obiettivo di correggere queste risposte ed ottenere il profilo che, ripulito, si avvicinasse il più possibile ai dati onesti. Mentre quest' ultimo, in maniera del tutto innovativa, di ricostruire le risposte oneste partendo dai dati disonesti dei soggetti item per item, il TF-IDF è una nuova tecnica di detezione della simulazione che si inserisce in un ambito molto sviluppato della neuropsicologia forense. Infatti, sebbene siano già presenti molti strumenti che hanno l'obiettivo di indagare il comportamento simulatorio del soggetto in diversi disturbi mentali, a differenza di questi, il TF-IDF, permette la rilevazione della menzogna a livello del singolo item del questionario. Inoltre, mentre i metodi convenzionali sono creati per cogliere la simulazione di disturbi specifici, il TF-IDF, così come la procedura di ricostruzione delle risposte, potrebbero essere utilizzati in un qualsiasi test che rileva la presenza di un disturbo nell'individuo, in quanto opera sulle singole rispose date al questionario.

Nello specifico, il TF-IDF consiste nel trasformare i punteggi grezzi in valori ottenuti dal confronto della singola risposta del soggetto con le risposte fornite da questo in tutto il questionario e quelle fornite da tutti gli altri soggetti in quello specifico *item* di riferimento. Questo permette che due soggetti che rispondono con lo stesso punteggio in una determinata domanda, mentre nei metodi tradizionali venivano considerati allo stesso modo, qui ottengano un TF-IDF differente, poiché questo è messo in relazione al suo modo di rispondere al questionario in generale e a quello degli altri in quella specifica domanda, ottenendo così un'analisi più informativa. Il TF-IDF viene poi interpretato come indice di anomalia: più il valore ottenuto è elevato, maggiore è la probabilità che il soggetto abbia mentito all'*item* di riferimento. Per fare ciò viene calcolato il valore soglia (*cut-off*), al di sopra del quale la risposta è da considerarsi come simulata, confrontando le risposte oneste date dagli altri soggetti in quello stesso *item.*

L'algoritmo di *Machine Learning* permette di ricostruire le risposte oneste partendo da quelle simulate dai partecipanti. Tale procedura avviene sottraendo ai punteggi grezzi dei disonesti la "parte" di menzogna calcolata attraverso la moltiplicazione del valore in percentuale della probabilità di appartenere alla categoria dei mentitori per il *Delta*, ovvero la differenza tra le medie delle risposte

simulate e delle risposte oneste di uno specifico *item* (probabilità x *Delta*). Ogni domanda, di ogni soggetto, avrà quindi un proprio valore da sottrarre al punteggio grezzo per ricostruire la risposta. Infatti, così come per il TF-IDF, due soggetti che rispondo allo stesso modo in una determinata domanda simulata, nella ricostruzione di questa otterranno due punteggi diversi, poiché differiscono sia nella probabilità di appartenere alla categoria mentitori, sia nel *Delta*.

Queste tecniche sono state applicate alla simulazione del Disturbo Depressivo Maggiore, nello specifico alle risposte date al questionario PHQ-9 da un campione di 529 soggetti che, nel rispondere onestamente alle domande del questionario, non hanno raggiunto il *cut-off* di 10.

5.2 Discussione dei risultati

In questo capitolo verranno verificate le ipotesi alla base della ricerca, attraverso l'interpretazione dei risultati ottenuti dalle analisi applicate.

5.2.1 Dati grezzi

Prima di analizzare i dati al fine di verificare le ipotesi proposte nel paragrafo 3.4, si è proceduto a verificare se i due gruppi HD e DH, nati dalla randomizzazione del campione in due gruppi che si differenziavano per l'ordine di somministrazione delle domande (HD prima in maniera onesta e poi disonesta, mentre DH prima risponde in maniera disonesta e poi onesta), fossero stati influenzati dal differente ordine o se potevano essere considerati come gruppo unico. Per verificare questo è stato condotto innanzitutto il test *Shapiro-Wilk*, per analizzare la distribuzione del campione, e visto che quest'ultimo risulta non seguire un andamento normale è stato effettuato il test non parametrico Mann-Whitney da cui non è emersa una differenza statisticamente significativa tra i due gruppi. Alla luce di queste evidenze, il campione è stato trattato come campione unico.

Detto ciò, partendo dall'analisi dei dati grezzi, nel grafico rappresentato in figura 4.1 (grafico delle distribuzioni dei punteggi medi nelle condizioni *honest* e *dishonest* per ciascun item) si può osservare la netta differenza tra la media dei punteggi ottenuti quando il soggetto deve rispondere mentendo e quando deve rispondere in maniera onesta. Infatti, le medie delle risposte simulate sono maggiori rispetto a quelle delle risposte oneste, indicando che i partecipanti quando devono mentire, cercando di simulare un disturbo depressivo grave, tendono a farlo dando un punteggio più elevato. Quanto affermato quindi conferma, anche se solo parzialmente, la seconda ipotesi della ricerca (H2), secondo la quale sia i punteggi grezzi che i valori TF-IDF risultano inferiori nella condizione *honest,*

rispetto alla condizione *dishonest*. Tale evidenza si riscontra anche dallo studio della percentuale di variazione delle risposte dalla condizione *honest* a *dishonest*: su un totale di 4761 risposte, il 92% dei partecipanti hanno aumentato il punteggio quando dovevano simulare il disturbo. Inoltre, gli item che sono risultati essere più utili a discriminare i soggetti onesti dai simulatori sono quelli che indagano la qualità del sonno, l'autostima e la concentrazione.

Questa differenza nei punteggi di risposta tra le due condizioni è stata avvalorata anche a livello statistico, confermando quindi la prima ipotesi di ricerca (H1), vale a dire che nella condizione *honest* e *dishonest* i punteggi grezzi presentano una differenza statisticamente significativa. Questa ipotesi è stata confermata andando innanzitutto ad analizzare la distribuzione del campione attraverso il test *Shapiro-Wilk* e, approvato che il campione non ha una distribuzione normale, è stato applicato il test non parametrico di Wilcoxon, che è risultato significativo (tabella 4.7) con valori inferiori allo 0,001 in ogni item, confermando l'ipotesi alternativa secondo cui le due distribuzioni (*honest* e *dishonest*) sono tra di loro differenti statisticamente.

Successivamente sono state proposte le matrici di correlazione, attraverso le quali si va ad indagare la correlazione delle due variabili, ovvero come l'una varia al variare dell'altra. Nella figura 4.3 C, corrispondente alla matrice *honest/dishonest*, è possibile osservare una correlazione molto vicina allo zero, a dimostrazione di quanto sia difficile, con il solo utilizzo di dati grezzi, poter dedurre le risposte simulate partendo da quelle oneste. È per questo motivo che non esistono infatti delle tecniche per la detezione dei singoli *item* simulati tramite l'analisi dei dati grezzi.

Infine, è stata studiata l'influenza delle variabili indagate nel questionario (come genere, età e scolarizzazione) sulla modalità di risposta nelle due condizioni *honest* e *dishonest*. Per fare ciò è stato condotto il test Kruskal-Wallis, un'equivalente dell'ANOVA per dati non parametrici e, quando questo risultava significativo, veniva effettuato un test *Post-hoc*, nello specifico il test di Dunn, per comprendere quale gruppo ha risposto in modo statisticamente differente. Per quanto riguarda il genere, è emerso che le partecipanti "femmine", che rappresentano gran parte del campione, quando devono simulare un disturbo depressivo grave tendono a rispondere con punteggi più elevati rispetto ad i partecipanti "maschi", mentre questa differenza non è stata riscontrata nelle risposte oneste. In merito alla variabile età, il Kruskal-Wallis è risultato essere significato solamente nella condizione *dishonest*, in particolare, attraverso il test di Dunn, è emerso come i soggetti più giovani appartenenti alle fasce d'età 18-30 e 31-40, quando si trovano nella condizione di mentire tendono a rispondere con punteggi più bassi rispetto ai partecipanti con un'età compresa tra i 41 e i 60. Invece, per quanto riguarda gli anni di scolarità, non è stata evidenziata alcuna differenza significativa nel rispondere alle domande, ad indicare che i differenti anni di scolarizzazione non influiscono sul modo di rispondere.

5.2.2 TF-IDF

I punteggi grezzi sono stati poi trasformati in valori TF-IDF, con l'obiettivo di verificare l'ipotesi numero tre (H3), secondo cui i valori TF-IDF permettono una detezione più facile e accurata dei soggetti che simulano il disturbo ansioso-depressivo rispetto ai soli dati grezzi.

Per verificare questa ipotesi sono stati calcolati i valori KLD (misura che permette di paragonare due diverse condizioni) di ogni item, sia per i dati grezzi che per i valori TF-IDF. Occorre ricordare che il *range* di punteggi calcolati dal KLD possono essere compresi tra zero ed infinito: più il numero è alto, maggiore è la differenza tra le due distribuzioni. Come si può notare nella tabella 4.10, i valori ottenuti dal KLD applicato ai punteggi TF-IDF sono risultati più alti rispetto ai dati grezzi, ad indicare una maggiore accuratezza del TF-IDF nel discriminare gli onesti dai non onesti.

La capacità del TF-IDF e dei dati grezzi di discriminare le due distribuzioni è stata valutata anche attraverso l'analisi della *performance* con *k-fold cross validation*, la quale ha indicato il novantacinquesimo come miglior percentile per la discriminazione tra soggetti *honest* e *dishonest*. Tale risultato ha soddisfatto anche l'ultima ipotesi (H4), secondo la quale i valori TF-IDF *dishonest* si collocano prevalentemente al di sopra di un determinato valore soglia.

Su questo poi è stato applicato un *Mutli-label Classification Task* per valutare i quattro indici di prestazione: *precision, accuracy, F1-score* e *recall.* Nella tabella 4.11 (Valori degli indici di *performance* indagati, per i dati grezzi e per il TF-IDF) è possibile osservare come i quattro indici siano relativamente alti sia per i dati grezzi che per il TF-IDF, ma che la *precision*, l'indice massimizzato dal modello e più informativo, risulta essere leggermente più elevato nel TF-IDF. Inoltre, nelle figure 4.4 A e B è possibile osservare come la funzione di densità di probabilità degli indici cresce all'aumentare del numero di *item* mentiti: maggiore è il numero, maggiore è l'accuratezza sia per i dati grezzi che per il TF-IDF.

Al fine di rilevare la simulazione a livello del singolo item, per ogni partecipante sono stati costruiti i *box-plot* del TF-IDF. In questo modo, le risposte *dishonest* del singolo soggetto vengono comparate con le risposte *honest* dell'intero campione ed è possibile osservare come la distribuzione delle risposte simulate si colloca superiormente rispetto a quella onesta data dalle risposte di tutti i soggetti del campione. Per quanto riguarda la detezione della simulazione, le risposte riconosciute come alterate vengono contrassegnate, in maniera molto chiara, da un pallino rosso. La rilevazione della risposta alterata può essere considerata errata solamente quando il pallino rosso ricade tra il baffo superiore ed inferiore del *box-plot*, in questo caso vuol dire che una domanda a cui il soggetto ha risposto onestamente è stata riconosciuta dal modello come disonesta e quindi mal identificata.

Con il TF-IDF quindi, a differenza dei dati grezzi, è possibile identificare la risposta simulata con un'accuratezza elevata, infatti nei due esempi riportati nella figura 4.4 è possibile osservare come le risposte simulate sono state ben identificate 6 volte su 8 (figura A) e 8 volte su 9 (figura B).

5.2.3 Ricostruzione delle risposte oneste

Per raggiungere il secondo obiettivo del presente studio, ovvero quello di ricostruire le risposte oneste partendo da quelle simulate dai partecipanti è stata utilizzata una procedura di correzione delle risposte che permettesse di ottenere un profilo ripulito che si avvicinasse il più possibile ai dati onesti.

Tale procedura comprende diverse fasi, la prima dei quali consiste nell'applicazione di un algoritmo di *Machine Learning* di classificazione binaria, chiamato *Logistic Regression*, che permette di classificare un partecipante del *dataset* nell'una o nell'altra categoria della variabile dipendente, vale a dire riconoscerlo come onesto o disonesto. Una volta addestrato l'algoritmo e testato, è stata valutata la *performance* di questo, analizzando i diversi indici tra cui *Accuracy, Recall, Precision* e F1. I valori ottenuti sono tutti molto elevati (0,98 – 1) ad indicare un'ottima capacità dell'algoritmo di discriminare tra soggetti onesti e simulatori.

Dopo aver riconosciuto il soggetto come mentitore, l'algoritmo restituisce un valore in percentuale della probabilità di appartenere alla categoria dei mentitori. Tale valore, moltiplicato per il *Delta*, ovvero la differenza tra le medie delle risposte simulate e delle risposte oneste di uno specifico *item*, restituisce il valore che va poi sottratto al punteggio grezzo dato dal disonesto ad una specifica domanda. Il valore da sottrarre è quindi diverso per ogni domanda di tutti i partecipanti, questo perché la probabilità di essere un mentitore è uguale in ogni domanda ma diverso fra i partecipanti, mentre il *Delta* è diverso fra le domande ma lo stesso per tutti i partecipanti, ovvero ogni domanda ha un proprio *Delta*, ma ogni partecipante ha gli stessi nove. Eseguite le operazioni fino a qui elencate, come si può osservare nella figura 4.6, si è ottenuto un notevole avvicinamento della distribuzione *honest* ricostruita alla distribuzione *honest* originale. Questo vuol dire che tolta la "quota" di menzogna, il risultato che otteniamo è un valore molto vicino a ciò che il soggetto avrebbe risposto se fosse stato onesto. Sebbene il risultato ottenuto sia promettente, in quanto la distribuzione delle risposte ricostruite si avvicina molto a quella delle risposte oneste, non è del tutto soddisfacente. Infatti, le due distribuzioni, come è possibile osservare nella colonna destra della tabella 4.15, non si sovrappongono del tutto, ma differiscono in tutti gli item di circa 0,40. Una spiegazione potrebbe essere data dal fatto che la probabilità di appartenere al gruppo dei simulatori, che viene poi

moltiplicata al *Delta*, non assume mai il valore 1, ma decimali come per esempio 0,60, 0,91 e 0,85 riportati nella tabella 4.13. Il risultato ottenuto quindi, non è sovrapponibile a quello originale come se moltiplicassimo per il *Delta* il valore 1, ovvero la probabilità massima che il partecipante sia un mentitore. Inoltre, una seconda spiegazione di questa differenza tra le due distribuzioni, essendo essa costante ed uguale in tutti gli *item,* potrebbe risiedere in un errore sistematico che, se eliminato, comporta lo scorrimento dei dati di un valore costante, in questo caso 0,40.

Su questo nuovo campione è stato applicato valutata nuovamente l'algoritmo *Logistic Recgression* e valutata la sua *performance*, ma a differenza dei risultati ottenuti sopra, qui gli indici hanno ottenuti valori molto più bassi (0,57 – 0,87) ad indicare che l'algoritmo discrimina i soggetti onesti dai simulatori con più difficoltà, in quanto le due distribuzioni *honest* e *honest* ricostruita sono molto simili tra di loro.

5.3 Limiti e prospettive future

Nonostante questo elaborato abbia ottenuto dei risultati molto buoni, le ipotesi formulate siano state tutte confermate e il secondo obiettivo della ricerca abbia portato promettenti conclusioni, non è esente da alcuni limiti ed è quindi opportuno esaminarli nel dettaglio.

Un primo limite da tenere in considerazione riguarda la composizione del campione: la forte prevalenza del genere femminile (76%), impedisce di trarre conclusioni generalizzate. Seppur è stato dimostrato che il disturbo depressivo colpisce maggiormente il genere femminile, sarebbe opportuno avere un campione più bilanciato.

Va inoltre considerato come altro limite il fatto che i partecipanti sono stati forzati a dare delle risposte simulate, cercando di immedesimarsi in un soggetto che vuole ottenere un risarcimento del danno a seguito di un evento negativo. Seppure sia stato specificato ai partecipanti di simulare in maniera credibile, è probabile che questi abbiano spesso utilizzato punteggi elevati e che la simulazione non rispecchi il reale contesto forense in cu il soggetto non è istruito a simulare o è particolarmente bravo a farlo, mettendo in atto strategie sottili di simulazione. Alla luce di questo limite, potrebbe essere interessante in futuro applicare queste tecniche in reali contesti forensi di simulazione, come per esempio in una richiesta di risarcimento di danno biologico, come nell'esempio riportato in questo elaborato, e confrontare questi risultati con quelli ottenuti dalle tecniche tradizionali di detezione della simulazione.

Per quanto riguarda il modello di ricostruzione delle risposte, le ricerche future dovrebbero innanzitutto colmare il *gap* presente tra le due distribuzioni (*honest* e *honest* ricostruita) a partire dalle due possibili spiegazioni raggiunte alla conclusione di questo elaborato, in modo tale da correggere tale differenza sistematica e permettere la ricostruzione precisa della risposta. Un altro limite senza

dubbio rilevante consiste nella sua futura applicazione in un contesto reale. Infatti, per riuscire a ricostruire le risposte oneste, occorre avere a disposizione un esiguo *dataset* di dati onesti e disonesti per allenare l'algoritmo e di conseguenza testare la sua efficacia. In un contesto reale però è molto difficile avere a disposizione questi due campioni, l'unica soluzione per sviare questo problema sarebbe quella di validare l'algoritmo per i principali questionari utilizzati, così, nel caso venissero rilevate risposte simulate, poter direttamente applicare questo modello al singolo soggetto e ricostruire le risposte oneste.

In futuro, sarebbe interessante analizzare anche la popolazione del campione composta dai soggetti che hanno superato il *cut-off* e quindi considerati soggetti clinici a livello psicometrico e i soggetti che hanno una diagnosi di depressione e/o assumono farmaci antidepressivi, per testare la capacità del TF-IDF di discriminare il simulatore dal soggetto realmente patologico e provare a ricostruire le risposte. In questo elaborato non è stata presa in considerazione questa parte del campione perché si è preferito testare l'efficacia di queste tecniche innovative in primis nei due campioni *honest* e i *dishonest* della popolazione generale.

Infine, un'altra prospettiva futura di ricerca dovrebbe valutare la capacità del TF-IDF di individuare le risposte simulate quando sono mentite solo alcune, variando di volta in volta il numero di risposte simulate, così da ricreare un contesto più ecologico possibile.

CONCLUSIONI

Nella presente ricerca è stata indagata innanzitutto la capacità di una innovativa metodologia di *lie detection*, ovvero il *Term Frequency–Inverse Document Frequency* (TF-IDF), in una prova per la valutazione del disturbo depressivo maggiore e, in un secondo momento, è stata testata una nuova tecnica, per correggere le risposte simulate, in modo tale da ricostruire il profilo onesto del soggetto. Sebbene il TF-IDF sia una tecnica innovativa di detezione della simulazione, si inserisce in un ambito molto sviluppato e studiato della neuropsicologia forense. Però, a differenza delle tecniche tradizionali che hanno l'obiettivo di indagare il comportamento simulatorio del soggetto in diversi disturbi mentali, il TF-IDF permette la rilevazione della menzogna a livello del singolo *item* del questionario. Del tutto innovativa invece è l'applicazione di una nuova semplice procedura che permette di ricostruire le risposte mentite dai soggetti ai singoli item. Quindi lo scopo di questo elaborato è sia di valutare se il TF-IDF riesce a discriminare in maniera migliore i simulatori dagli onesti, sia di ricostruire le risposte simulate, a livello del singolo *item*.

Come è stato già accennato, si è andati ad indagare la simulazione del disturbo depressivo maggiore, essendo esso uno dei disturbi mentali più diffusi nel mondo e, allo stesso tempo, con un *cluster* di sintomi molto facilmente simulabile. Lo strumento utilizzato è il *Patient Health Questionnaire-9* (PHQ-9), un questionario molto breve, composto da 9 item che vanno ad indagare i nove criteri che definiscono il disturbo depressivo maggiore.

Il questionario è stato compilato da 1.099 soggetti, ma il campione su cui sono state condotte le analisi è composto da 529 soggetti. Oltre ad aver eliminato i soggetti che non avevano portato a termine la compilazione del questionario o non aveva risposto correttamente alle domande di comprensione delle istruzioni (346), sono stati esclusi dal campione (a) i soggetti che hanno dichiarato di aver ricevuto una diagnosi di disturbo depressivo, (b) che assumono farmaci antidepressivi, anche senza avere una diagnosi, (c) che hanno raggiunto il *cut-off* prestabilito dalla letteratura, ovvero un punteggio maggiore o uguale a 10.

I partecipanti sono stati randomizzati in due gruppi che si distinguevano per l'ordine di presentazione delle domande: un gruppo rispondeva alla domanda 1, e così per tutte le domande, prima in maniera onesta e poi disonesta (gruppo HD) mentre l'altro, al contrario, rispondeva prima simulando e poi onestamente (gruppo DH). Tale assegnazione casuale è stata effettuata per verificare la possibile presenza di un effetto dell'ordine di presentazione della risposta.

Per quanto riguarda le analisi, in primo luogo si è proceduto ad indagare la possibile presenza di un effetto dell'ordine, così da decidere se considerare il campione come unico o separato nei due gruppi HD e DH. Per fare questo è stata verificata innanzitutto la distribuzione del campione ed è

stato visto che questa non segue un andamento normale, per cui è stato effettuato un Mann-Withney, che non ha evidenzialo una differenza significativa tra i due gruppi, quindi il campione è stato considerato come proveniente dalla stessa popolazione.

A questo punto sono state condotte le analisi considerando il campione non indipendente *honest/dishonest*. Verificata la distribuzione non normale del campione, si è proceduto ad analizzare la presenza di una differenza significativa tra i due gruppi sia graficamente attraverso le medie, che statisticamente con il test di Wilcoxon. Le medie delle risposte *dishonest* sono maggiori rispetto a quelle delle risposte *honest*, indicando che i partecipanti quando devono mentire tendono a farlo dando un punteggio più elevato. Successivamente, attraverso le matrici di correlazione è stato visto come, a partire dai soli dati grezzi, sia molto difficile individuare quante e quali riposte siano state simulate. I punteggi grezzi sono stati poi trasformati in valori TF-IDF e attraverso l'indice KLD è stato dimostrato come questi permettono una discriminazione più accurata e informativa tra risposte oneste e simulate, rispetto ai dati grezzi. È stata poi effettuata un'analisi della *performance* attraverso la *k-fold cross validation*, la quale ha indicato il novantacinquesimo come miglior percentile per la discriminazione tra soggetti *honest* e *dishonest* sia per i dati grezzi che per i valori TF-IDF. Individuato tale valore soglia, al di sopra del quale ricadono i simulatori e al di sotto gli onesti, è stato applicato un *Mutli-label Classification Task* per valutare i quattro indici di prestazione: *precision, accuracy, F1-score* e *recall* in entrambe le tecniche. L'indice *precision*, ovvero quello massimizzato dal modello e più informativo, è risultato essere leggermente maggiore per i valori TF-IDF, con una precisione del riconoscimento delle risposte simulate pari al 97%. Sebbene tale risultato sia molto soddisfacente, è di poco superiore a quanto ottenuti nei dati grezzi. Per tale ragione si può concludere che il TF-IDF non è una metodologia sostitutiva all'analisi dei dati grezzi, ma la loro utilità aumenta di molto quando usato in affiancamento ai dati grezzi, perché ha la peculiarità di individuare la simulazione a livello del singolo *item*. Al fine di indagare quest'ultimo aspetto, sono stati costruiti i rispettivi *box-plot* dei valori TF-IDF, in cui le risposte *dishonest* del singolo soggetto vengono comparate con le risposte *honest* dell'intero campione. In figura 4.5 è possibile osservare molto facilmente le risposte riconosciute come simulate perché evidenziate da un pallino rosso. La rilevazione della risposta alterata può essere considerata errata solamente quando il pallino rosso ricade tra il baffo superiore ed inferiore del *box-plot*, in questo caso vuol dire che una domanda a cui il soggetto ha risposto onestamente è stata riconosciuta dal modello come disonesta e quindi mal identificata.

Infine, per quanto riguarda il secondo obiettivo di questo studio, ovvero la ricostruzione delle risposte oneste, questo è stato raggiunto con l'applicazione di un algoritmo di *Machine Learning* chiamato *Logistic Regression*, attraverso il quale tutti i partecipanti del *dataset* sono stati classificati

in onesti o simulati. L'algoritmo, per fare ciò si è basato sulla percentuale della probabilità di quel valore di appartenere alla categoria dei mentitori, valore che è stato poi moltiplicato per il *Delta,* ovvero la differenza ottenuta dalla sottrazione tra le medie delle risposte oneste e quelle simulate. Quanto ottenuto dalla moltiplicazione, che è diverso per ogni domanda di ogni soggetto che ha partecipato alla ricerca, è stato poi sottratto al punteggio grezzo dato dal partecipante nel rispondere alle domande, a quello specifico *item*. Da questa sottrazione si è ottenuto un nuovo valore, ovvero il dato *honest* ricostruito, che, come si può osservare sia graficamente che statisticamente nella figura 4.6 e nella tabella 4.14, si avvicina di molto alla distribuzione *honest* originale, ad indicare che le risposte ricostruire sono molto simili alle risposte date dai partecipanti nella condizione onesta.

In conclusione, seppur si è appena agli inizi della ricerca in questa nuova direzione, si può essere soddisfatti degli importanti risultati che sono stati ottenuti in questo elaborato. Infatti, le tecniche innovative qui esposte possono portare importanti risvolti nell'ambito della simulazione, in particolare nel contesto forense. Poter individuare gli *item* in cui il soggetto ha simulato e ricostruire le risposte oneste sulla sola base di quelle mentite, permette di evitare l'ottenimento di vantaggi processuali, invalidità o risarcimenti economici a soggetti a cui non spettano. La scelta di indagare la simulazione del disturbo depressivo maggiore è stata spinta dalla diffusione di questo disturbo e dalla facilità di simulazione, che consente a qualsiasi persona di fingersi depresso o di aggravare la propria condizione per beneficiare di tali vantaggi.

Bibliografia

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders (3rd ed.).* Washington DC.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., revised). Washington, DC.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of disorders (4th ed.).* Washington, DC.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders, DSM-5.* Arlington, VA.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). New York: ACM press.

Balsamo, M., & Saggino, A. (2007). Test per l'assessment della depressione nel contesto italiano: un'analisi critica. *Psicoterapia Cognitiva e Comportamentale*, *13*(2), 167.

Bandelow, B., Baldwin, D. S., Dolberg, O. T., Andersen, H. F., & Stein, D. J. (2006). What is the threshold for symptomatic response and remission for major depressive disorder, panic disorder, social anxiety disorder, and generalized anxiety disorder?. *Journal of Clinical Psychiatry*, *67*(9), 1428-1434.

Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S., Znoj, H., ... & Cuijpers, P. (2016). Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. *Focus*, *14*(2), 229-243.

Bear Mark, F., Connors Barry, W., & Paradiso Michael, A. (2007). Neuroscienze. Esplorando il cervello.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck depression inventory (BDI-II)* (Vol. 10, p. s15327752jpa6703_13). Pearson.

Behera, P., Gupta, S. K., Nongkynrih, B., Kant, S., Mishra, A. K., & Sharan, P. (2017). Screening instruments for assessment of depression. *Indian Journal of Medical Specialities*, *8*(1), 31-37.

Blier, P., Ward, H. E., Tremblay, P., Laberge, L., Hébert, C., & Bergeron, R. (2010). Combination of antidepressant medications from treatment initiation for major depressive disorder: a double-blind randomized study. *American Journal of Psychiatry*, *167*(3), 281-288.

Brigitta, B. (2002). Pathophysiology of depression and mechanisms of treatment. *Dialogues in clinical neuroscience*, *4*(1), 7.

Butcher, J. N. (2010). Minnesota multiphasic personality inventory. *The Corsini Encyclopedia of Psychology*, 1-3.

Camerini, G. B. (2011). *La valutazione del danno psichico nell'infanzia e nell'adolescenza: danno, pregiudizio e disabilità: aspetti clinici, medico-legali e giuridici*. Giuffrè.

Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H., ... & Poulton, R. (2003). Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science*, *301*(5631), 386-389.

Daubert, S. D., & Metzler, A. E. (2000). The detection of fake-bad and fake-good responding on the Millon Clinical Multiaxial Inventory III. *Psychological Assessment*, *12*(4), 418.

Duran, N. D., Dale, R., & McNamara, D. S. (2010). The action dynamics of overcoming the truth. *Psychonomic bulletin & review*, *17*(4), 486-491.

Eaton, W. W., Smith, C., Ybarra, M., Muntaner, C., & Tien, A. (2004). Center for Epidemiologic Studies Depression Scale: review and revision (CESD and CESD-R).

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychological bulletin*, *136*(1), 103.

Ettman, CK, Abdalla, SM, Cohen, GH, Sampson, L., Vivier, PM e Galea, S. (2020). Prevalenza dei sintomi della depressione negli adulti statunitensi prima e durante la pandemia di COVID-19. *Rete JAMA aperta* , *3* (9), e2019686-e2019686.

Ferracuti, S., Parisi, L., & Coppotelli, A. (2007). *Simulare la malattia mentale*. Centro scientifico.

Fornari, U. (2008). Trattato di psichiatria forense, quarta edizione. *Torino, Ed. UTET*.

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior research methods*, *42*(1), 226-241.

Furukawa, T. A. (2010). Assessment of mood: guides for clinicians. *Journal of psychosomatic research*, *68*(6), 581-589.

Ghisi, M., Flebus, G. B., Montano, A., Sanavio, E., & Sica, C. (2006). Beck depression inventory-Adattamento italiano: manuale. *Firenze: Organizzazioni Speciali*.

Giannini, G., & Pogliani, M. (1996). *La responsabilità da illecito civile: assicuratore, magistrato, produttore, professionista*. Giuffrè.

Goldman, L. S., Nielsen, N. H., Champion, H. C., & Council on Scientific Affairs, American Medical Association. (1999). Awareness, diagnosis, and treatment of depression. *Journal of general internal medicine*, *14*(9), 569-580.

Guimond, S., Branscombe, N. R., Brunot, S., Buunk, A. P., Chatard, A., Désert, M., ... & Yzerbyt, V. (2007). Culture, gender, and the self: variations and impact of social comparison processes. *Journal of personality and social psychology*, *92*(6), 1118.

Gulotta, G. (2011). *Compendio di psicologia giuridico-forense, criminale e investigativa* (Vol. 53). Giuffrè Editore.

HAMILTON, M. (1960). A rating scale for depression.

Hawton, K., i Comabella, C. C., Haw, C., & Saunders, K. (2013). Risk factors for suicide in individuals with depression: a systematic review. *Journal of affective disorders*, *147*(1-3), 17-28.

Holma, K. M., Melartin, T. K., Haukka, J., Holma, I. A., Sokero, T. P., & Isometsä, E. T. (2010). Incidence and predictors of suicide attempts in DSM–IV major depressive disorder: a five-year prospective study. *American Journal of Psychiatry*, *167*(7), 801-808.

Holsboer, F. (2000). The corticosteroid receptor hypothesis of depression. *Neuropsychopharmacology*, *23*(5), 477-501.

Horwitz, A. V. (2014). DSM-I and DSM-II. *The encyclopedia of clinical psychology*, 1-6.

Innamorati, M., Lelli, M., Aiello, S., Di Lorenzo del Casale, F. L., Russo, S., & Ferrari, V. (2006). Validazione convergente e discriminante della versione italiana della Zunf Self-Rating Depression Scale. *Psicoterapia Cognitiva e Comportamentale*, *12*(3), 343-353.

Istituto nazionale di statistica. (2018). *La salute mentale nelle varie fasi della vita. https://www.istat.it/it/files//2018/07/Report_Salute_mentale.pdf*

Jobst, A., Brakemeier, E. L., Buchheim, A., Caspar, F., Cuijpers, P., Ebmeier, K. P., ... & Padberg, F. (2016). European Psychiatric Association Guidance on psychotherapy in chronic depression across Europe. *European Psychiatry*, *33*(1), 18-36.

Julien, R. M., Advokat, C. D., Comaty, J. E. (2012). *Droghe e farmaci psicoattivi* (C. Buccellati, Trad.; 2. ed.). Zanichelli. (Originariamente pubblicato nel 2011) 132-288.

Kendler, KS, Thornton, LM, & Prescott, CA (2001). Differenze di genere nei tassi di esposizione a eventi di vita stressanti e sensibilità ai loro effetti depressogeni. *Giornale americano di psichiatria* , *158* (4), 587-593.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., ... & Wang, P. S. (2003). The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *Jama*, *289*(23), 3095-3105.

Kring, A. M., Johnson, S. L., Davidson, G. C., Neale, J.M. (2017). *Psicologia clinica* (D Conti, Trad.; 5. ed.). Zanichelli. (Originariamente pubblicato nel 2016) 129-165.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: a new depression diagnostic and severity measure.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, *16*(9), 606-613.

Lakhan, R., Agrawal, A. e Sharma, M. (2020). Prevalenza di depressione, ansia e stress durante la pandemia di COVID-19. *Giornale di neuroscienze nella pratica rurale* .

Lally, S. J. (2003). What tests are acceptable for use in forensic evaluations? A survey of experts. *Professional Psychology: Research and Practice*, *34*(5), 491.

Lim, G. Y., Tam, W. W., Lu, Y., Ho, C. S., Zhang, M. W., & Ho, R. C. (2018). Prevalence of depression in the community from 30 countries between 1994 and 2014. *Scientific reports*, *8*(1), 1-10.

Maier, W., Heuser, I., Philipp, M., Frommberger, U., & Demuth, W. (1988). Improving depression severity assessment—II. Content, concurrent and external validity of three observer depression scales. *Journal of psychiatric research*, *22*(1), 13-19.

Mangen, M. J. J., Plass, D., Havelaar, A. H., Gibbons, C. L., Cassini, A., Mühlberger, N., ... & BCoDE Consortium. (2013). The pathogen-and incidence-based DALY approach: an appropriated methodology for estimating the burden of infectious diseases. *PloS one*, *8*(11), e79740.

Mazza, C., Monaro, M., Burla, F., Colasanti, M., Orrù, G., Ferracuti, S., & Roma, P. (2020). Use of mouse-tracking software to detect faking-good behavior on personality questionnaires: an explorative study. *Scientific reports*, *10*(1), 1-13.

Mazza, C., Monaro, M., Burla, F., Colasanti, M., Orrù, G., Ferracuti, S., & Roma, P. (2020). Use of mouse-tracking software to detect faking-good behavior on personality questionnaires: an explorative study. *Scientific reports*, *10*(1), 1-13.

Mitchell, T. M. (1997). Machine learning (WBC/McGraw-Hill, Boston). *MA*.

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exeggeration. *Journal of clinical and experimental neuropsychology*, *24*(8), 1094-1102.

Monaro, M., Gamberini, L., & Sartori, G. (2017). The detection of faked identity using unexpected questions and mouse dynamics. *PloS one*, *12*(5), e0177851.

Monaro, M., Toncini, A., Ferracuti, S., Tessari, G., Vaccaro, M. G., De Fazio, P., ... & Sartori, G. (2018). The detection of malingering: a new tool to identify made-up depression. *Frontiers in psychiatry*, *9*, 249.

Monaro, M., Toncini, A., Ferracuti, S., Tessari, G., Vaccaro, M. G., De Fazio, P., ... & Sartori, G. (2018). The detection of malingering: a new tool to identify made-up depression. *Frontiers in psychiatry*, *9*, 249.

Montgomery, S. A., & Åsberg, M. A. R. I. E. (1979). A new depression scale designed to be sensitive to change. *The British journal of psychiatry*, *134*(4), 382-389.

Mynors-Wallis, L. M., Gath, D. H., Lloyd-Thomas, A. R., & Tomlinson, D. B. M. J. (1995). Randomised controlled trial comparing problem solving treatment with amitriptyline and placebo for major depression in primary care. *Bmj*, *310*(6977), 441-445.

Nolen-Hoeksema, S., & Hilt, L. M. (2009). Gender differences in depression.

Pampallona, S., Bollini, P., Tibaldi, G., Kupelnick, B., & Munizza, C. (2004). Combined pharmacotherapy and psychological treatment for depression: a systematic review. *Archives of general psychiatry*, *61*(7), 714-719.

Parikh, S. V., Segal, Z. V., Grigoriadis, S., Ravindran, A. V., Kennedy, S. H., Lam, R. W., & Patten, S. B. (2009). Canadian Network for Mood and Anxiety Treatments (CANMAT) clinical guidelines for the management of major depressive disorder in adults. II. Psychotherapy alone or in combination with antidepressant medication. *Journal of affective disorders*, *117*, S15-S25.

Picardi, A., Adler, D. A., Abeni, D., Chang, H., Pasquini, P., Rogers, W. H., & Bungay, K. M. (2005). Screening for depressive disorders in patients with skin diseases: a comparison of three screeners. *Acta dermato-venereologica*, *85*(5).

Picardi, A., Mazzotti, E., & Pasquini, P. (2006).

Prevalence and correlates of suicidal ideation among patients with skin disease. *Journal of the American Academy of Dermatology*, *54*(3), 420-426.

Pierfederici, A., Fava, G. A., Munari, F., Rossi, N., Baldaro, B., Pasquali Evangelisti, L., ... & Zecchino, F. (1982). Validazione italiana del CES-D per la misurazione della depressione. *R Canestrari (a cura di), Nuovi metodi in psicometria. Firenze: Organizzazioni Speciali*.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied psychological measurement*, *1*(3), 385-401.

Rapaport, M. H., Clary, C., Fayyad, R., & Endicott, J. (2005). Quality-of-life impairment in depressive and anxiety disorders. *American Journal of Psychiatry*, *162*(6), 1171-1178.

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *American journal of psychiatry*, *170*(1), 59-70.

Richards, D. (2011). Prevalence and clinical course of depression: a review. *Clinical psychology review*, *31*(7), 1117-1125.

Rogers, R. E. (2008). *Clinical assessment of malingering and deception*. Guilford Press.

Rogers, R., Sewell, K. W., & Goldstein, A. M. (1994). Explanatory models of malingering. *Law and human behavior*, *18*(5), 543-552.

Rosen, J., Mulsant, BH, Bruce, ML, Mittal, V. e Fox, D. (2004). Rappresentazioni della depressione da parte degli attori per testare l'affidabilità degli interterritori negli studi clinici. *Giornale americano di psichiatria* , *161* (10), 1909-1911.
Rosenhan, DL (1973). Sull'essere sani in posti folli. *Scienza* , *179* (4070), 250-258.

Rush, A. J., Warden, D., Wisniewski, S. R., Fava, M., Trivedi, M. H., Gaynes, B. N., & Nierenberg, A. A. (2009). STAR* D. *CNS drugs*, *23*(8), 627-647.

Salk, R. H., Hyde, J. S., & Abramson, L. Y. (2017). Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. *Psychological bulletin*, *143*(8), 783.

Sammicheli, Luca. La perizia psicologica : prospettive e metodi in psicologia e psicopatologia forense. Il mulino, 2019.
Sanavio, E., & Sica, C. (1997). *I test di personalità: Inventari e questionari.* Il Mulino.

Sartori G., Tenconi E., Lo Priore C. (2000). La simulazione della depressione: un nuovo strumento diagnostico per applicazioni medico-legali. RIVISTA ITALIANA DI MEDICINA LEGALE. vol. XII, pp. 1063-1077

Sartori, G., & Lombardi, L. (2004). Semantic relevance and semantic disorders. *Journal of Cognitive Neuroscience*, *16*(3), 439-452.

Sartori, G., Orrù, G., & Zangrossi, A. (2016). Detection of malingering in personal injury and damage ascertainment. In *Personal injury and damage ascertainment under civil law* (pp. 547-558). Springer, Cham.

Sartori, G., Zangrossi, A., Orrù, G., & Monaro, M. (2017). Detection of malingering in psychic damage ascertainment. In *P5 medicine and justice* (pp. 330-341). Springer, Cham.

Segal, Z. V., Williams, M., & Teasdale, J. (2018). *Mindfulness-based cognitive therapy for depression*. Guilford Publications.

Smith, G. P., & Burger, G. K. (1997). Detection of malingering: validation of the Structured Inventory of Malingered Symptomatology (SIMS). *Journal of the American Academy of Psychiatry and the Law Online*, *25*(2), 183-189.

Spitzer, R. L., Kroenke, K., Williams, J. B., Patient Health Questionnaire Primary Care Study Group, & Patient Health Questionnaire Primary Care Study Group. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *Jama*, *282*(18), 1737-1744.

Storm, J., & Graham, J. R. (2000). Detection of coached general malingering on the MMPI—2. *Psychological Assessment*, *12*(2), 158.

Stracciari, A., Bianchi, A., & Sartori, G. (2010). *Neuropsicologia forense*. Il mulino.

Sweet, J. J., King, J. H., Malina, A. C., Bergman, M. A., & Simmons, A. (2002). Documenting the prominence of forensic neuropsychology at national meetings and in relevant professional journals from 1990 to 2000. *The Clinical Neuropsychologist*, *16*(4), 481-494.

Teymoori, A., Real, R., Gorbunova, A., Haghish, E. F., Andelic, N., Wilson, L., ... & von Steinbüchel, N. (2020). Measurement invariance of assessments of depression (PHQ-9) and anxiety (GAD-7) across sex, strata and linguistic backgrounds in a European-wide sample of patients after Traumatic Brain Injury. *Journal of affective disorders*, *262*, 278-285.

Thase, M. E., Friedman, E. S., Biggs, M. M., Wisniewski, S. R., Trivedi, M. H., Luther, J. F., ... & Rush, A. J. (2007). Cognitive therapy versus medication in augmentation and switch strategies as second-step treatments: a STAR* D report. *American Journal of Psychiatry*, *164*(5), 739-752.

Thase, M. E., Greenhouse, J. B., Frank, E., Reynolds, C. F., Pilkonis, P. A., Hurley, K., ... & Kupfer, D. J. (1997). Treatment of major depression with psychotherapy or psychotherapy-pharmacotherapy combinations. *Archives of general psychiatry*, *54*(11), 1009-1015.

Üstün, T. B., Ayuso-Mateos, J. L., Chatterji, S., Mathers, C., & Murray, C. J. (2004). Global burden of depressive disorders in the year 2000. *The British journal of psychiatry*, *184*(5), 386-392.

Van Bockstaele, B., Verschuere, B., Moens, T., Suchotzki, K., Debey, E., & Spruyt, A. (2012). Learning to lie: effects of practice on the cognitive cost of lying. *Frontiers in psychology*, *3*, 526.

Van Impelen, A., Merckelbach, H., Jelicic, M., & Merten, T. (2014). The Structured Inventory of Malingered Symptomatology (SIMS): A systematic review and meta-analysis. *The Clinical Neuropsychologist*, *28*(8), 1336-1365

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, *5*(1-2), 39-43.

Williams Jr, J. W., Noël, P. H., Cordes, J. A., Ramirez, G., & Pignone, M. (2002). Is this patient clinically depressed?. *Jama*, *287*(9), 1160-1170.

Williams, J. B. (1988). A structured interview guide for the Hamilton Depression Rating Scale. *Archives of general psychiatry*, *45*(8), 742-747.

Williams, J. B., & Kobak, K. A. (2008). Development and reliability of a structured interview guide for the Montgomery-Åsberg Depression Rating Scale (SIGMA). *The British Journal of Psychiatry*, *192*(1), 52-58.

Williams, J. B., Kobak, K. A., Bech, P., Engelhardt, N., Evans, K., Lipsitz, J., ... & Kalali, A. (2008). The GRID-HAMD: standardization of the Hamilton depression rating scale. *International clinical psychopharmacology*, *23*(3), 120-129.

Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. In *Advances in experimental social psychology* (Vol. 46, pp. 55-123). Academic Press.

World Health Organization. (2001). The World Health Report 2001: Mental health: new understanding, new hope, 29-30.

Zennaro A., Santo Di Nuovo A.L., Fulcheri M., Mazzeschi C. (2015) PAI-Personality Assessment Inventory-Adattamento italiano: manuale. *Hogrefe*

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758-2765.

Zimmerman, M., Martinez, J. H., Young, D., Chelminski, I., & Dalrymple, K. (2013). Severity classification on the Hamilton depression rating scale. *Journal of affective disorders*, *150*(2), 384-388.

Zung, W. W. (1965). A self-rating depression scale. *Archives of general psychiatry*, *12*(1), 63-70.

Appendice - A

Questionario[6] utilizzato per la ricerca

QUESTIONARIO SULLA SALUTE DEL/DELLA PAZIENTE-9 (PHQ-9)

Nelle ultime 2 settimane, con quale frequenza le ha dato fastidio ciascuno dei seguenti problemi? *(Segni la sua risposta con una "X")*	Mai	Alcuni giorni	Per più della metà dei giorni	Quasi ogni giorno
1. Scarso interesse o piacere nel fare le cose	0	1	2	3
2. Sentirsi giù, triste o disperato/a	0	1	2	3
3. Problemi ad addormentarsi o a dormire tutta la notte senza svegliarsi, o a dormire troppo	0	1	2	3
4. Sentirsi stanco/a o avere poca energia	0	1	2	3
5. Scarso appetito o mangiare troppo	0	1	2	3
6. Avere una scarsa opinione di sé, o sentirsi un/una fallito/a o aver deluso se stesso/a o i propri familiari	0	1	2	3
7. Difficoltà a concentrarsi su qualcosa, per esempio leggere il giornale o guardare la televisione	0	1	2	3
8. Muoversi o parlare così lentamente da poter essere notato/a da altre persone. O, al contrario, essere così irrequieto/a da muoversi molto più del solito	0	1	2	3
9. Pensare che sarebbe meglio morire o farsi del male in un modo o nell'altro	0	1	2	3

FOR OFFICE CODING 0 + _____ + _____ + _____

=Total Score: _____

Se ha fatto una crocetta su uno qualsiasi di questi problemi, quanto questi problemi le hanno reso difficile fare il suo lavoro, occuparsi delle sue cose a casa o avere buoni rapporti con gli altri?

Per niente difficile	Abbastanza difficile	Molto difficile	Estremamente difficile
☐	☐	☐	☐

[6] Il questionario è disponibile al sito web Pfizer (http://www.phqscreeners.com) ed è scaricabile in più di 90 lingue.