Ministero della Salute – Direzione Generale della Ricerca e dell'Innovazione in Sanità

Fondi 5 per mille ANNO 2019
Abstract ed elenco pubblicazioni scientifiche

Ente della Ricerca Sanitaria

Denominazione Ente: IRCCS E. Medea dell'Associazione La Nostra Famiglia

Codice fiscale: 00307430132
Sede legale: via don Luigi Monza, 1 22037 Ponte Lambro (Co)
Indirizzo di posta elettronica dell'ente: segreteria.scientifica@pec.emedea.it
Dati del rappresentante legale: Luisa Minoli, nata il 14.01.1968 a Busto Arsizio (VA)
CF: MNLLSU68A54B300V

Titolo del progetto: Multimodalità, intelligenza artificiale e big data: rianalisi, razionalizzazione e integrazione di basi dati raccolte in studi esistenti

L'obbiettivo di questo progetto è quello di valutare le opportunità offerte dal riutilizzo di dati raccolti retrospettivamente tra quelli utilizzati in progetti esistenti. Ciò comporta la necessità di avere un sistema informativo in grado di identificare quali dati e in quale numerosità sono disponibili per un nuovo studio. Inizialmente sono stati individuati tre casi di interesse sia per quanto riguarda la possibilità di analisi multimodale che per le possibili applicazioni di metodiche di tipo machine learning.
1. ASD: studio su un gruppo di soggetti con ASD (Autism Spectrum Disorder)
2. EDI: uno studio sull'impatto di depressione e ansia materna in gravidanza (studio EDI): in questo caso sono disponibili dati relativi a marcatori biologici oltre a registrazioni video, già codificate nel tempo.
3. MIMOSA: Studio su gruppo di soggetti con ADHD per i quali sono disponibili registrazioni EEG e NIRS contemporaneamente eseguite durante l'esecuzione di un task cognitivo.

In questi tre case-studies, applicando metodiche di fattorizzazione a tutti raw data l'idea del progetto è quella estrarre pattern significativi che verranno in seguito impiegati sia in funzione predittiva sulle variabili outcome (discrete o continue) con adeguati strumenti di machine learning, sia nel tentativo di identificare gruppi (cluster) di individui che presentino omogeneità nei suddetti patterns per poi studiarne le caratteristiche clinico-comportamentali.

Il progetto si proponeva altresì di valutare in modo più generale la situazione esistente con l'obbiettivo di individuare altre raccolte di dati che appaiano promettenti per una possibile rianalisi. Verranno inizialmente privilegiati studi relativamente recenti per i quali siano stati raccolti dati di tipo genetico/genomico/epigenetico. Nei casi più interessanti e articolati verranno proposti specifici sotto-progetti relativi alle raccolte individuate in cui

verranno specificati l'obbiettivo e i metodi per la rianalisi.
Al tempo stesso, sfruttando la dotazione tecnologica esistente dedicata alla ricerca, verranno individuate le migliori strategie per favorire il passaggio ad una modalità di storage e organizzazione dei dati più sicura e fruibile. L'obbiettivo di lungo termine è quello di costruire un ambiente tecnologico che favorisca la collaborazione tra diversi gruppi di ricerca e, l'integrazione con la cartella clinica in progettazione all'interno dell'istituto e l'utilizzo di tecniche di analisi moderne fin dalla progettazione degli studi.

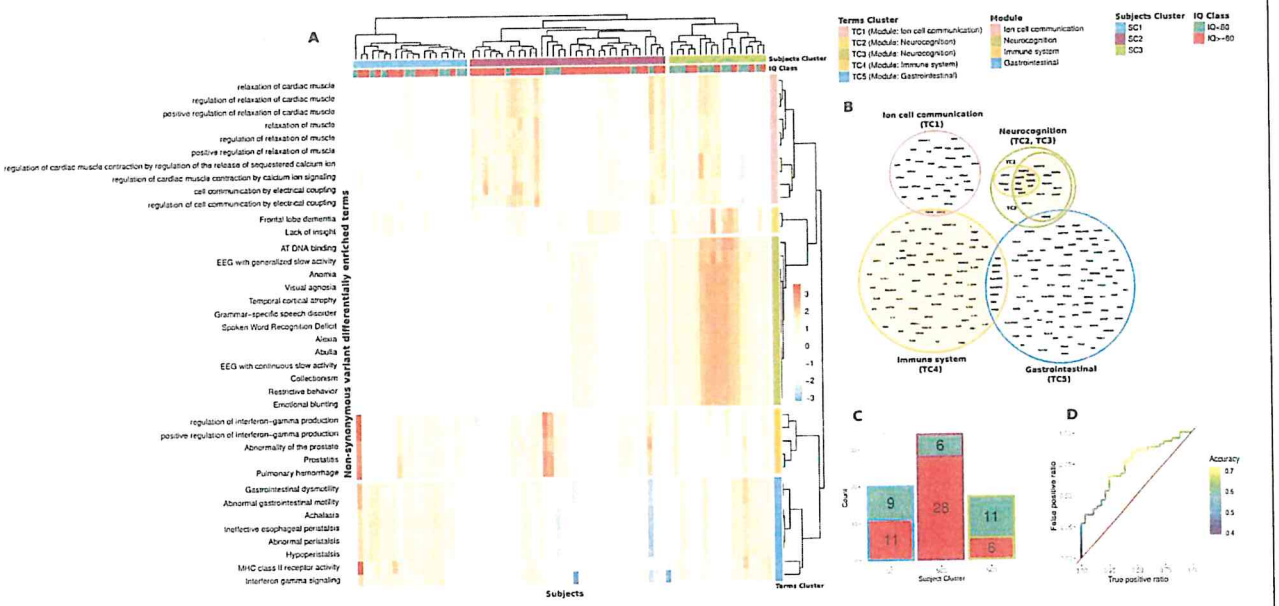
Quest'ultima parte dello studio ha assunto nel corso degli ultimi 2 anni un'urgenza particolare nata dall'esigenza a livello dell'intero IRCCS di razionalizzare la raccolta e la fruizione dei dati di ricerca, anche in relazione a specifiche richieste da parte del ministero. Abbiamo dovuto dedicare molte più risorse a questa parte che doveva rappresentare una semplice ricognizione e si è invece trasformata nella messa a punto di un sistema di database. Per questa ragione la rianalisi dello studio MIMOSA è ancora in corso e al momento non siamo in grado di riportare risultati definitivi.

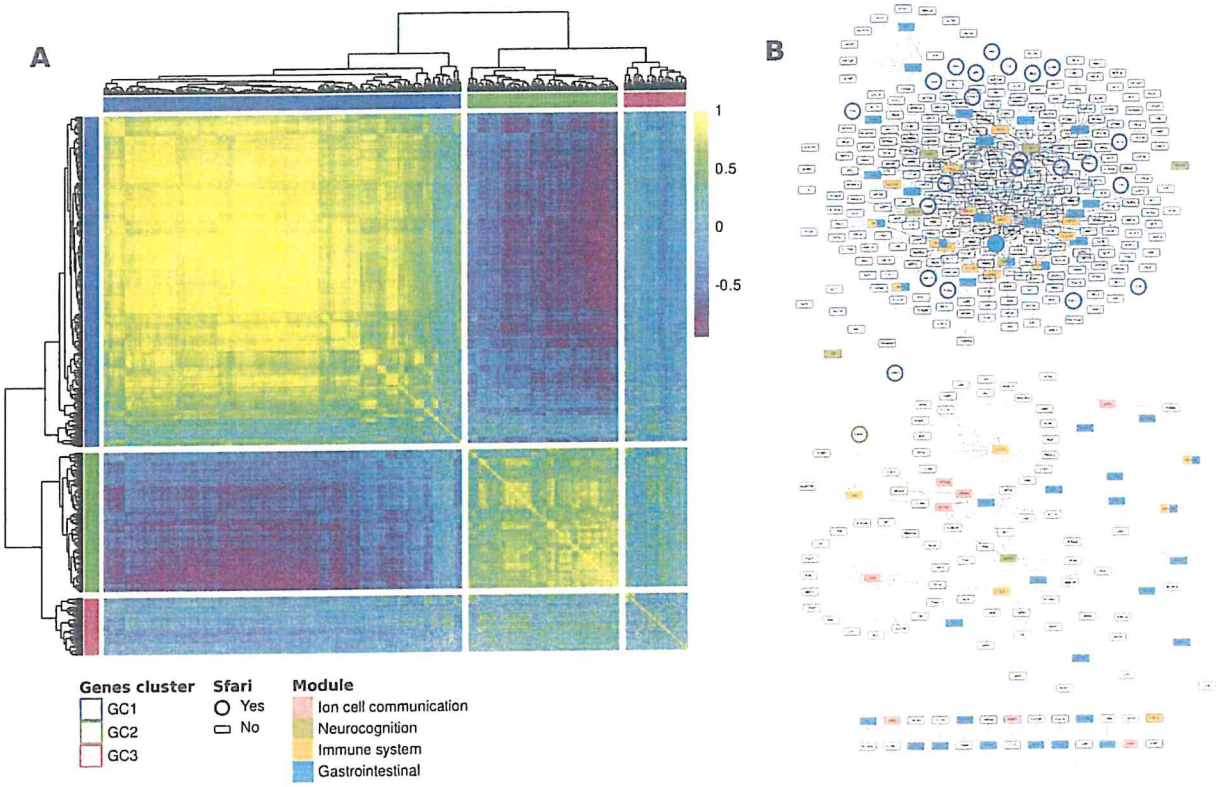
1. STUDIO ASD EXOMES
Data l'eterogeneità della condizione, l'identificazione delle implicazioni funzionali delle varianti dei geni correlate all'autismo richiede comunemente studi su larga scala che coinvolgano migliaia di partecipanti. Una prospettiva alternativa è quella di spostare l'attenzione dai singoli geni a gruppi di geni associati a processi biologici funzionalmente rilevanti. Questo approccio consente di esaminare l'effetto combinato di più varianti potenzialmente dannose per il DNA (PDV) che hanno un impatto cumulativo su percorsi biologici, funzioni e processi potenzialmente rilevanti per l'autismo. Abbiamo mirato a esplorare se un'analisi unbiased potesse identificare set di geni caratterizzati funzionalmente con rilevanza per la sottotipizzazione dell'autismo in termini di carico di varianti tra due sottogruppi di bambini autistici.
Dopo aver suddiviso il nostro campione di 71 bambini autistici (3-12 anni) in due sottogruppi con quoziente intellettivo (QI) più alto (>80; n=43) e più basso ($\leqslant$80; n=28), è stata applicata un'analisi di arricchimento delle varianti per identificare set di geni con incidenza significativamente diversa di varianti potenzialmente dannose tra i due sottogruppi.
Utilizzando un approccio basato sui dati e non supervisionato, i set di geni sono stati quindi raggruppati in moduli con punteggi di arricchimento più omogenei. Successivamente, per caratterizzare funzionalmente i geni nei moduli, abbiamo studiato i loro profili di espressione secondo il BrainSpan Atlas of the Developing Human Brain. Quindi, abbiamo esteso ogni modulo selezionando quei geni che interagiscono fisicamente e mostrano un profilo di espressione cerebrale spazio-temporale altamente correlato con quelli nei moduli. La co-espressione e le interazioni spazio-temporali sono state inoltre testate tra l'intero set di geni nei quattro moduli estesi. Infine, abbiamo esplorato l'incidenza di geni correlati all'autismo ad alta affidabilità, come riportato dal database Simons Foundation Autism Research Initiative (SFARI), tra i geni nei moduli originali ed estesi.

La nostra analisi ha identificato 38 set di geni significativi (tasso di falsa scoperta, $q<0,05$) con diverso carico di varianti nei due sottogruppi di bambini autistici. Questi set di geni, che includevano 219 geni in totale, si sono raggruppati in quattro moduli che rappresentano specifici processi biologici: comunicazione delle cellule ioniche, neurocognizione, funzione gastrointestinale e sistema immunitario. Quei moduli genici erano altamente espressi in specifiche strutture cerebrali in diverse fasi dello sviluppo, ad eccezione del modulo immunitario. La coespressione cerebrale spazio-temporale attraverso lo sviluppo e le interazioni fisiche delle proteine sono state trovate tra cluster significativi di geni da diversi moduli estesi. Inoltre, abbiamo trovato una

sovrarappresentazione di geni SFARI ad alta affidabilità (OR=2,59; p=0,001) quando si considerano i quattro moduli estesi.

La nostra analisi ubiased e multi-step ha identificato moduli di geni funzionalmente significativi per l'autismo anche in un insieme relativamente piccolo di partecipanti, fornendo prove della loro potenziale implicazione nelle differenze fenotipiche dei sottogruppi di autismo con diversi livelli di QI. I geni associati ai moduli sono espressi precocemente nel cervello e significativamente co-espressi spazio-temporalmente. Presentano anche tra loro un elevato livello di interazione proteina-proteina, confermando così l'interazione funzionale tra quei percorsi biologici. La rilevanza dei moduli genici identificati nella manifestazione dell'autismo è stata ulteriormente supportata dalle loro interazioni con molti geni noti per essere correlati all'autismo. Nel complesso, i risultati suggeriscono che la diversità nell'autismo probabilmente ha origine da più percorsi interagenti. Sebbene queste osservazioni debbano essere considerate con cautela, la ricerca futura potrebbe sfruttare l'attuale approccio per identificare percorsi genetici rilevanti per la sottotipizzazione dell'autismo, verso l'identificazione di biotipi distinti di pazienti. Nelle figure seguenti sono riportati i principali risultati di questa paste dello studio.

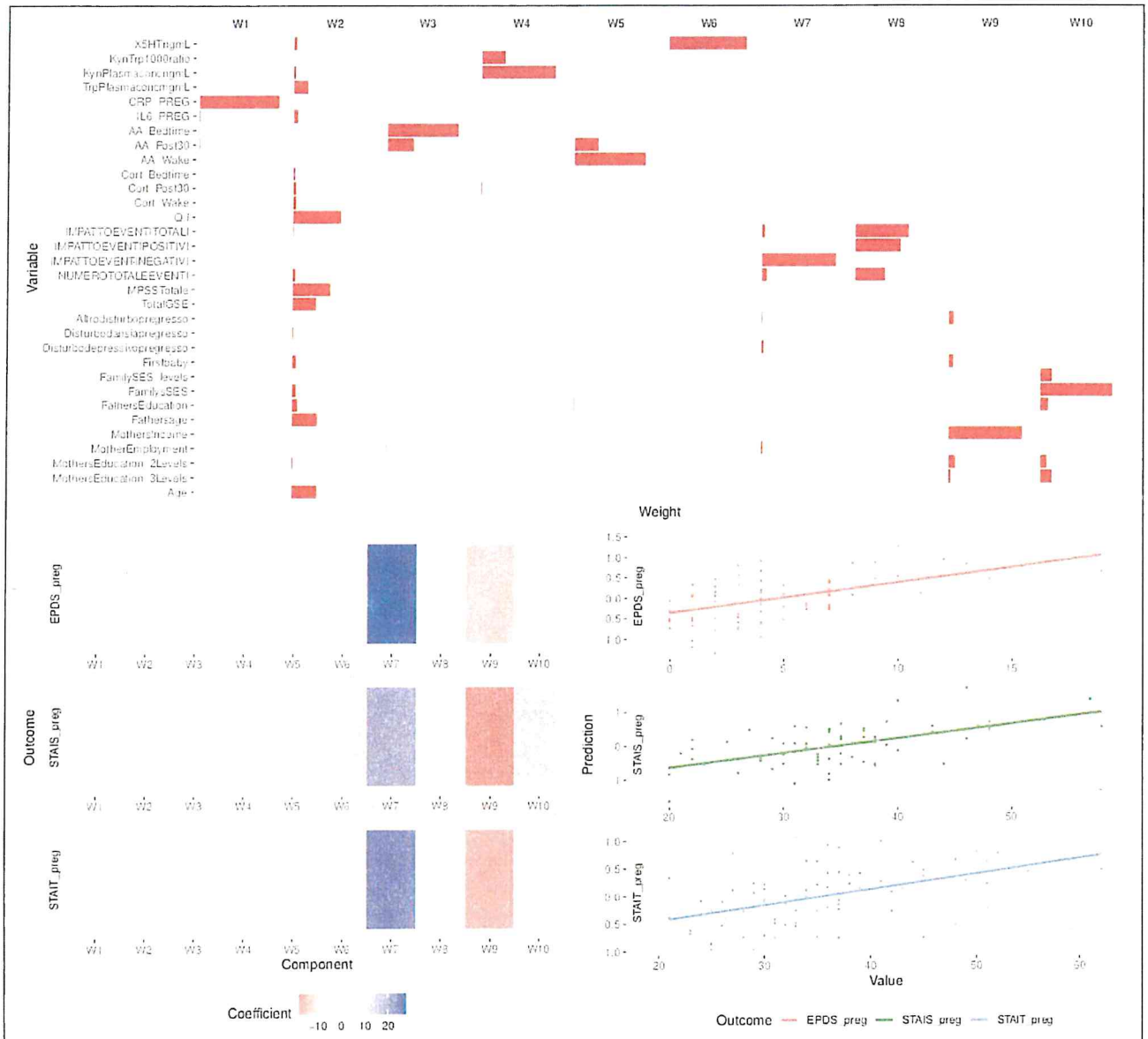




2. STUDIO EDI

In questo studio sono state raccolte una serie di variabili di natura biologica e psico-sociale per identificare un possibile legame con l'insorgenza di problematiche di tipo depressivo nelle madri dopo una gravidanza. L'analisi che abbiamo svolto in questo caso mirava ad identificare dei sottogruppi di variabili che mostrassero un effetto dello stesso segno sugli outcome misurati. Lo scopo di questa analisi è quello di rendere maggiormente interpretabile l'impatto delle diverse variabili sull'outcome analizzandole per gruppi identificati n modo automatico direttamente dai dati.

Con un approccio esclusivamente numerico e non supervisionato vengono inizialmente identificati una serie di "Componenti" ottenute dalla matrice delle variabili nei soggetti in studio. Tale tecnica (Projective non-negative factorization) scompone la matrice dei dati in due matrici il cui prodotto "ricostruisce" la matrice originale. Mentre prima matrice definisce le Component osservate in termini di combinazioni lineari delle variabili originali. La seconda fornisce il "peso" di ciascuna componente nel ricostruire la matrice dei dati originale. Ciò che rende particolarmente interessante questa metodica è il fatto che tutti i coefficienti sono vincolati ad essere non negativi. Ciò rende il metodo adatto a produrre scomposizioni in componenti molto più interpretabili rispetto ad altri metodi di riduzione delle variabili (PCA,SVD etc). I pesi delle componenti ottenute sono poi stati utilizzati in modelli regressivi per la predizione degli outcome. I coefficienti di tali modelli possono essere interpretati come in una normale analisi di regressione per interpretare il significato di ciascuna componente in relazione all'outcome osservato. Nella seguente figura è riportata la definizione delle 10 componenti ottenute e il loro coefficiente nei modelli predittivi dell'outcome. I risultati sono attualmente in fase di valutazione da parte del gruppo di psicologi che hanno portato avanti lo studio originario, il metodo di fattorizzazione è stato implementato in un pacchetto di R che è in grado di sfruttare la presenza di un GPU per accelerare l'analisi (piuttosto intensiva dal punto di vista computazionale)



3. MATCHA

Matcha è un sistema costituito da un database e da un'interfaccia web e che ha lo scopo di gestire nel modo semplice e flessibile la pseudonimizzazione dei soggetti partecipanti a progetti di ricerca garantendo comunque (ove consentito) la rintracciabilità delle informazioni relative allo stesso individuo in diversi progetti di ricerca e, in prospettiva, nei sistemi informativi istituzionali (Cartella clinica, Dossier etc).

Le principali caratteristiche sono:

1. Pseudonimizzazione del soggetto attraverso la generazione di un id (record_id) specifico per progetto che identifica le informazioni relative al soggetto in un progetto senza che sia possibile da questo identificare l'individuo (al di fuori del database Matcha).
2. Generazione di un codice (hash_code) a partire dal codice fiscale del soggetto che

sarà sempre lo stesso in tutti i progetti e per lo stesso individuo. Anche in questo caso, al di fuori del database Matcha il codice non consente di risalire all'identità dell'individuo.

Il sistema si compone di diversi componenti:

- Server che ospita il database MATCHA: è presente un motore database MariaDB che gestisce il database MATCHA. E' accessibile esclusivamente dalla rete interna dell'Associazione (Inf.it)
- Server che ospita il software e l'interfaccia web: ospita il software necessario all'interrogazione del database Matcha e l'interfaccia web. Sia il software che l'interfaccia sono basati su R. Il server è interno all'associazione ma è accessibile, esclusivamente per quanto riguarda l'interfaccia web attraverso l'indirizzo sicuro https://matcha.emedea.it
- Server dedicato a REDCap per ospitare i database di progetto basati su REDCap: Il server ospita il software REDCap e il motore database che gestisce il database di REDCap. E' ospitato dal GARR ed accessibile tramite l'indirizzo sicuro https://redcap.emedea.it
- Eventuali server esterni che ospitano database di progetto non sviluppati all'interno dell'associazione (es studi multicentrici)

L'accesso è consentito esclusivamente ad un utente di sistema oltre che agli amministratori del sistema. Sia l'utente di sistema che gli amministratori hanno accesso illimitato alle informazioni contenute qualora si connettano al server. Informazioni selezionate possono tuttavia essere estratte in base ai permessi e alle regole illustrate in seguito. Ciò è esclusivamente possibile dall'interfaccia web di Matcha previa autenticazione. L'autenticazione avviene tramite le credenziali istituzionali per gli utenti interni all'associazione (la scadenza e il rinnovo delle password è gestito quindi come per tutti gli altri servizi informatici dell'associazione). Per gli utenti esterni (altri istituti, nel caso di studi multicentrici o altri operatori che non hanno credenziali Inf) la gestione dell'autenticazione è gestita internamente al database Matcha e prevede la scadenza delle password ogni 90 giorni

L'accesso al database Matcha attraverso l'interfaccia è controllato a livello di singolo utente con diritti specifici relativi ai progetti e in base alle regole stabilite. Le informazioni contenute nel database matcha sono separate (anche fisicamente) da quelle nei diversi sistemi di archiviazione dei dati di progetto. I rispettivi database sono fisicamente separati e i loro unici punti di contatto sono il record_id e l'hash_code.

Il record_id funziona a livello di singolo progetto e consente a chi ne ha l'autorizzazione di inserire le informazioni relative a un individuo necessarie per il progetto

L'hash_code unifica i diversi database garantendo, a chi ne abbia il diritto, la possibilità di ottenere informazioni relative a un soggetto già presenti in un altro progetto.

Definiamo un record come l'insieme delle informazioni (cliniche, diagnostiche, specifiche del progetto, etc) presenti all'interno del database di progetto. Tali informazioni NON devono includere dati che possano ricondurre direttamente o indirettamente all'identità del soggetto stesso. Per indirettamente intendiamo che consentano di risalire all'identità con uno sforzo "ragionevole", anche commisurato all'eventuale interesse di questa operazione. Non è infatti possibile escludere in modo assoluto che da certe tipologie di dati si possa risalire all'identità.

Al momento il database Matcha può gestire due tipologie di progetto:

1. Progetti basati su REDCap sviluppati e gestiti da ricercatori dell'IRCCS Medea:
   in questo caso il sistema Matcha è in grado registrare le informazioni anagrafiche e contemporaneamente aggiungere il record relativo a un soggetto al database di progetto. Il record verrà poi completato ad opera dei ricercatori autorizzati direttamente in REDCap. Le uniche informazioni anagrafiche che vengono trasferite al database di progetto sono il sesso e la data di nascita del soggetto.
2. Progetti basati su REDCap non sviluppati e gestiti direttamente da ricercatori dell'IRCCS Medea o progetti basati su altri sistemi di archiviazione: in questo caso il sistema Matcha deve essere usato per registrare unicamente le informazioni anagrafiche e l'esistenza di un record relativo al soggetto. La creazione o la rimozione del record all'interno del database di progetto deve essere eseguita direttamente nel database di progetto stesso. Questa modalità può essere utilizzata per qualsiasi sistema di archiviazione dei dati di progetto.

In ogni caso vengono registrate la data di inserimento e l'operatore che ha compiuto l'operazione. Il sistema Matcha è tecnicamente in grado di gestire progetti multicentrici tenendo traccia di quale centro partecipante ha reclutato un soggetto e rendendo visibili le informazioni anagrafiche dello stesso esclusivamente agli operatori autorizzati afferenti a quel centro. Tuttavia, dato che tali informazioni verrebbero comunque registrate nel database Matcha che è ospitato su un server dell'IRCCS Medea, riteniamo più ragionevole che i dati anagrafici relativi a soggetti reclutati presso altri centri siano gestiti autonomamente dal centro reclutatore stesso. In ogni caso, per i progetti sviluppati e gestiti dal Medea (plausibilmente quindi nei casi in cui il Medea è capofila del progetto) sarà comunque necessario, anche per gli operatori di altri centri, utilizzare matcha per creare il record nel database di progetto, senza che le informazioni anagrafiche vengano registrate. Il ricercatore dovrà inserire unicamente il codice fiscale del soggetto per cui intende creare il record. Il codice fiscale verrà utilizzato per generare l'hash_code ma non verrà registrato nel sistema. L'operatore riceverà invece il record_id che utilizzerà poi per accedere al record creato per quel soggetto nel database di progetto. Sarà sua responsabilità mantenere la corrispondenza tra questo record_id e le informazioni anagrafiche. Il sistema così configurato consente di implementare qualsiasi regola relativa alla conservazione, rintracciabilità, rimozione e riutilizzo di tutte le informazioni pertinenti a un individuo, regole che devono essere esplicitate in base alle normative vigenti e che possono variare nel tempo. Al momento il sistema Matcha NON gestisce i consensi al trattamento dei dati nei consensi informati alla partecipazione al progetto di ricerca. L'interfaccia web del sistema consente di inserire e visualizzare le associazioni tra record e dati anagrafici fino alla conclusione del progetto. Da quel momento in avanti non sarà più possibile consultare attraverso l'interfaccia tali informazioni che saranno però conservate finchè non verrà deciso di eliminarle. Fino a quel momento sarà possibile attraverso una richiesta non automatica ricollegare un record all'identità del soggetto qualora si renda necessario. Nonostante il sistema configuri la possibilità di riutilizzare informazioni già presenti nei record relativi a un soggetto (fino alla loro eliminazione) è ancora poco chiaro con quali modalità e con quali regole ciò sia consentito. Per questa ragione al momento non è stato implementato nessun automatismo per facilitare questa operazione.

Dal punto di vista dell'implementazione l'intero sistema è basato sul motore database

open source mariaDB, mentre per quanto riguarda l'interfaccia è stato usato l'ambiente Shiny basato sul software R. Tutto il software è stato costruito come un R-package che può essere facilmente installato e distribuito.
Nella figura seguente è riportato, a titolo esemplificativo, uno screenshot dell'interfaccia.

Prodotti della Ricerca (correlati al progetto):

Software:
R-package: matchaR
R-package: Rpnmf

Pubblicazioni:
Gaia Scaccabarozzi, Luca Fumagalli, Maddalena Mambretti, Roberto Giorda, Marco Villa, Silvia Busti Ceccarelli, Laura Villa, Elisa Mani, Maria Nobile, Massimo Molteni, Uberto Pozzoli, Alessandro Crippa, "Potentially damaging variants' analysis in autism subgroups uncovers early brain-expressed gene modules relevant to autism pathophysiology" under review Autism research.

Data 10/12/2024
Il Responsabile del Progetto
Ing. Pozzoli Uberto

Il Legale Rappresentante
D.ssa Luisa Minoli

Si autorizza al trattamento dei dati ai sensi del d.lgs. 196/2003

Il Legale Rappresentante
D.ssa Luisa Minoli