Big Data Integration & Data-Centric AI for eHealth

Domenico Beneventano, Sonia Bergamaschi, Luca Gagliardelli, Giovanni Simonini[#], Luca Zecchini

Dipartimento di Ingegneria "Enzo Ferrari", Università degli Studi di Modena e Reggio Emilia
[#]Dipartimento di Economia "Marco Biagi", Università degli Studi di Modena e Reggio Emilia
{nome.cognome}@unimore.it

Abstract

La *big data integration*, ovvero l'integrazione di grandi quantità di dati provenienti da molteplici sorgenti, rappresenta una delle principali sfide per l'impiego di tecniche e strumenti basati sull'intelligenza artificiale in ambito medico (*eHealth*). In questo contesto risulta inoltre di primaria importanza garantire la qualità dei dati su cui operano tali strumenti e tecniche (*Data-Centric AI*), che rivestono un ruolo ormai centrale nel settore. Le attività di ricerca del Database Group (DBGroup) del Dipartimento di Ingegneria "Enzo Ferrari" dell'Università degli Studi di Modena e Reggio Emilia si muovono in questa direzione. Presentiamo quindi i principali progetti di ricerca del DBGroup nel campo dell'eHealth, che si inseriscono nell'ambito di collaborazioni in diversi settori applicativi.

1 Introduzione

L'integrazione di grandi quantità di dati provenienti da molteplici sorgenti (*big data integration)* è di fondamentale importanza per l'impiego dell'intelligenza artificiale in ambito medico (*eHealth*). Tale esigenza è diventata particolarmente significativa con lo sviluppo della telemedicina e il crescente utilizzo di dispositivi IoT (si parla in particolare di IoMT, ovvero Internet of Medical Things), che consentono ad esempio di raccogliere e monitorare da remoto i parametri dei pazienti. In questo scenario, caratterizzato dall'acquisizione di una grande quantità di dati eterogenei da molteplici sorgenti, diventa quindi necessario saper integrare tali dati, al fine di poter disporre delle informazioni di interesse in un quadro completo e coerente.

Con l'impiego sempre crescente dell'intelligenza artificiale in ambito medico e la rilevanza che l'analisi dei dati riveste in tale settore, risulta estremamente importante garantire la qualità dei dati di cui si dispone. Infatti, se i dati che si considerano in input sono sporchi (ad esempio se sono presenti duplicati, informazioni mancanti o errate, ecc.), anche il risultato che si otterrà in output sarà di scarsa qualità e perciò inaffidabile (si pensi ad esempio al training di un modello di machine learning). L'attenzione alla qualità dei dati in tutte le fasi di vita del progetto è il principio cardine alla base delle pratiche di *MLOps*[1], che mirano a rendere la cosiddetta *Data-Centric AI*[2] un processo efficiente e sistematico.

Infine, trattandosi di dati sensibili, è sempre necessario operare nel rispetto della privacy, in conformità al GDPR. Diventa così fondamentale lo studio e l'utilizzo di tecniche di pseudonimizzazione, che garantiscano la privacy e allo stesso tempo consentano di poter effettuare le operazioni di data integration.

2 DBGroup

La data integration rappresenta la principale area di ricerca del Database Group (DBGroup[3]) del Dipartimento di Ingegneria "Enzo Ferrari" dell'Università degli Studi di Modena e Reggio Emilia, guidato dalla Prof. Sonia Bergamaschi. Il DBGroup opera infatti da oltre vent'anni su questo tema [Bergamaschi *et al.*, 2018], avendo incentrato per lungo tempo le proprie attività sul sistema di data integration `MOMIS` (Mediator envirOnment for Multiple Information Sources) [Bergamaschi *et al.*, 1999; Magnotta *et al.*, 2018], che nel corso degli anni ha trovato numerose applicazioni nell'ambito dell'eHealth. Una versione open-source è attualmente gestita da DataRiver[4], fondata come spin-off del DBGroup nel 2009.

Con l'avvento dei big data e l'importanza sempre crescente acquisita dall'intelligenza artificiale (e in particolare dal machine learning) nel campo della big data integration, il DBGroup ha saputo sviluppare diversi progetti innovativi, con collaborazioni internazionali di primo piano e un impatto significativo sulle attività di ricerca della comunità scientifica del settore [Simonini *et al.*, 2019a; Simonini *et al.*, 2019b; Papadakis *et al.*, 2020], prestando sempre attenzione anche alle possibili applicazioni concrete di tali innovazioni [Gagliardelli *et al.*, 2018].

3 Scoperta di nuovi farmaci

La scoperta di nuovi farmaci tradizionalmente si basa sull'identificazione di composti che colpiscono selettivamente un singolo target di interesse limitando reazioni avverse. Tuttavia, è correntemente accertato che i farmaci molto spesso

[1] https://www.youtube.com/watch?v=06-AZXmwHjo
[2] https://www.datacentricai.cc
[3] https://dbgroup.unimore.it
[4] https://www.datariver.it

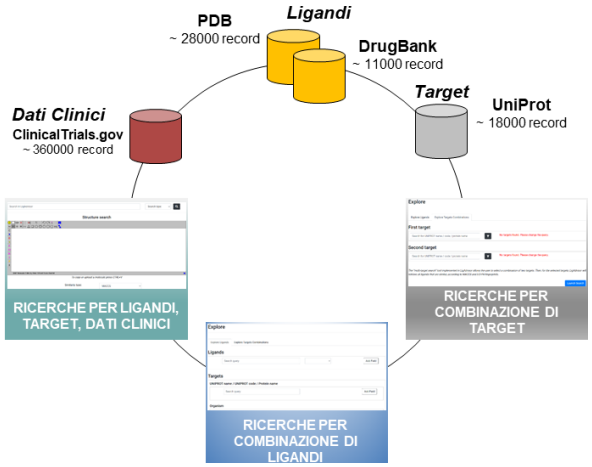

Figura 1: Architettura della piattaforma `LigAdvisor`

esercitano i loro effetti colpendo più target in modo simultaneo. La possibilità di modulare gli effetti di un farmaco su più target ha aperto la strada alla polifarmacologia, che viene principalmente sviluppata tramite approcci *in silico* (ossia tramite simulazioni al computer). Similmente, il riposizionamento di farmaci noti e già approvati verso nuove esigenze terapeutiche (*drug repurposing*) ha suscitato un crescente interesse nel campo della scoperta di nuovi farmaci. In questo contesto, recentemente sono stati sviluppati diversi tool e piattaforme web che basandosi su approcci data-driven e big data facilitano le operazioni per la scoperta di nuovi farmaci. Su queste basi il DBGroup in collaborazione con MMDLab[5] ha sviluppato il progetto multidisciplinare `LigAdvisor` [Pinzi *et al.*, 2021], finanziato tramite FAR - Fondo di Ateneo per la Ricerca 2019.

`LigAdvisor` è uno strumento web liberamente accessibile[6] che integra informazioni raccolte da DrugBank, Protein Data Bank, UniProt, Clinical Trials e Therapeutic Target Database, creando una piattaforma intuitiva per facilitare le operazioni di scoperta di nuovi farmaci (drug repurposing, polifarmacologia, target fishing e profilazione), la cui architettura è rappresentata in Figura 1.

Lo scopo principale di `LigAdvisor` è quello di assistere i ricercatori nell'identificazione di nuovi target per farmaci già conosciuti consentendo di eseguire varie tipologie di ricerche sui dati integrati. Per questo è stato implementato un database di composti integrando dati prelevati da DrugBank e PDB che per ogni possibile coppia di composti contiene la similarità tra le loro rappresentazioni cristallografiche in due dimensioni. Questa informazione dovrebbe aiutare nell'identificazione di una potenziale attività biologica di un set di ligandi di interesse su un ampio set di target. Ad esempio, un utente può disegnare una nuova molecola ed eseguire una ricerca in grado di trovare molecole simili. Utilizzando poi le informazioni pre-calcolate sulla base delle rappresentazioni cristallografiche è possibile ottenere i target su cui le molecole risultate simili sono attive.

[5] http://www.mmddlab.unimore.it

[6] http://ligadvisor.unimore.it

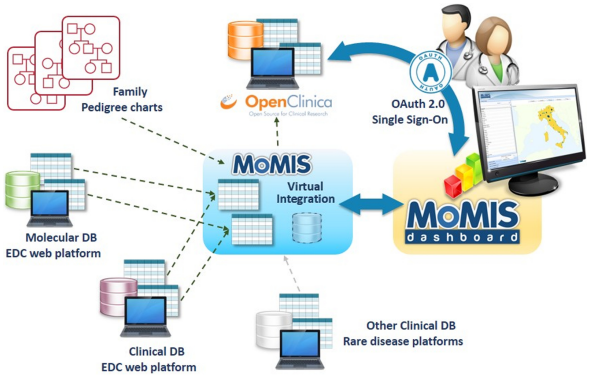

Figura 2: Architettura del registro FSHD

4 Gestione di dati medici relativi a pazienti affetti da FSHD e SLA

Il laboratorio di ricerca Miogen Lab[7] dell'Università degli Studi di Modena e Reggio Emilia si occupa della ricerca sulle malattie neuromuscolari. In particolare, fornisce diagnosi molecolare, analisi e consulenza genetica per pazienti affetti da distrofia muscolare facio-scapolo-omerale (FSHD). Il laboratorio Miogen Lab ha istituito il Registro Italiano per la FSHD [Bettio *et al.*, 2021], uno dei più grandi registri europei che raccoglie dal 1993 i dati clinici e genetici dei pazienti affetti da FSHD e dei loro parenti. Il registro comprende oltre 6600 soggetti e più di 2600 famiglie.

In collaborazione con Miogen Lab e DataRiver, il DBGroup si è occupato della progettazione e dello sviluppo di una piattaforma web per facilitare le ricerche cliniche basate sul registro FSHD. L'architettura di questa piattaforma è rappresentata in Figura 2. In particolare, le attività svolte sono state l'integrazione dei dati molecolari e delle analisi radiologiche presenti nel registro FSHD all'interno della piattaforma OpenClinica e lo sviluppo di funzionalità di big data analytics utilizzando la piattaforma `MOMIS Dashboard` [Magnotta *et al.*, 2018] per facilitare l'analisi dei dati clinici e molecolari al fine di comprendere meglio l'evoluzione e l'ereditarietà della malattia. La piattaforma sviluppata offre ai ricercatori un'interfaccia avanzata che nel rispetto delle norme del GDPR consente di: visualizzare ed analizzare i dati relativi ai fenotipi e agli alleli dei pazienti; eseguire indagini sulle modalità di trasmissione per via ereditaria della FSHD in relazione ad altri fattori esterni; geolocalizzare i pazienti e i legami genealogici delle patologie e relativi fenotipi; individuare nuovi fattori genetici e ambientali utili in chiave diagnostica.

In collaborazione con Miogen Lab, il DBGroup svilupperà anche il nuovo progetto "Artificial Intelligence for the Management and Analysis of Clinical and Molecular Data of the Italian National Registry of Facio-Scapulo-Humeral Muscular Dystrophy and the Emilia-Romagna Registry of Amyotrophic Lateral Sclerosis", che sarà finanziato dal bando FAR MISSION ORIENTED 2021 e si proporrà di applicare tecniche di analisi dei dati basate sull'intelligenza artificiale, quali *cluster analysis* e *deep neural networks*, con l'obiettivo di isolare elementi clinici e molecolari per la stratificazione dei

[7] https://www.dsv.unimore.it/site/home/ricerca/laboratorio-di-ricerca.html?id=247

pazienti e di identificare i determinanti ambientali che possono influenzare l'insorgenza e la progressione della malattia, per migliorare quindi la diagnosi personalizzata, la valutazione del rischio e il trattamento. Nel nuovo progetto verranno considerati ed integrati anche i dati relativi ad un'altra malattia neuromuscolare rara con eziologia sconosciuta, la Sclerosi Laterale Amiotrofica (SLA); tali dati provengono dal registro ERRALS [Mandrioli *et al.*, 2014], un registro prospettico attivo in Emilia-Romagna dal 2009 che raccoglie tutti i casi di SLA tra la popolazione residente nella regione. Inoltre, per aumentare l'efficacia delle analisi, i dati saranno continuamente integrati con informazioni relative alle abitudini di vita dei pazienti (attività fisiche, alimentazione, frequenza cardiaca, frequenza respiratoria, saturazione di ossigeno, idratazione, sonno) ottenute attraverso dispositivi indossabili. Infine, è prevista una fase di arricchimento e standardizzazione di tali dati integrati mediante annotazione rispetto alla *Human Phenotype Ontology*.

5 Monitoraggio della salute e della qualità di vita del paziente

Il DBGroup ha recentemente partecipato in qualità di relatore a workshop su *Big Data Integration & Data-Centric AI for eHealth*[8], tema su cui collabora in modo costante con DataRiver, attiva su diversi progetti di rilevanza internazionale in tale ambito. Uno dei risultati più significativi di tale collaborazione è rappresentato da `MyHealth`[9], una piattaforma web e mobile il cui scopo è quello di migliorare la qualità della vita degli individui.

La piattaforma `MyHealth`, validata secondo gli standard internazionali per la sperimentazione clinica, è stata progettata per consentire il monitoraggio continuo della salute e della qualità di vita dei pazienti, garantendo inoltre l'efficacia dell'interazione tra medico e paziente. `MyHealth` consente infatti di raccogliere automaticamente i parametri fisiologici del paziente e i dati sull'attività fisica svolta, grazie all'integrazione con dispositivi IoMT (medical devices e wearable devices). La piattaforma permette inoltre di stabilire un canale di comunicazione diretto tra medico e paziente, accessibile tramite app su smartphone e tablet e basato sull'uso di assistenti vocali e chatbot (oltre all'eventuale interazione da remoto tramite video), grazie al quale è possibile contattare i pazienti e offrire loro supporto (ad esempio, riguardo le modalità di somministrazione della terapia). La raccolta di informazioni sullo stato di salute dei pazienti, che può essere integrata dall'eventuale somministrazione di questionari accessibili dall'applicazione, consente quindi di gestirne la terapia, sfruttando anche la possibilità di impostare funzionalità di alert e notifiche automatiche (pill reminder, alert di farmacovigilanza, ecc.). Il fine ultimo è quello di rendere realmente personalizzabile l'interazione con il singolo paziente, impiegando tecniche di intelligenza artificiale per sfruttare le informazioni raccolte e i feedback forniti dal paziente stesso per personalizzare la gestione di contenuti e notifiche e stabilire i migliori canali di interazione, incentivando così un utilizzo effettivo ed efficace della piattaforma da parte del paziente.

[8] https://dbgroup.unimore.it/site/home/research/articolo1250035049.html
[9] https://www.datariver.health/piattaforma-myhealth

Figura 3: Architettura della piattaforma `MyHealth-COV`

La piattaforma `MyHealth` è impiegata attivamente per diversi scopi, come il supporto di pazienti con patologie specifiche, il monitoraggio di persone anziane (in contesti di assistenza residenziale e case di cura) o la promozione di stili di vita attivi, in ottica inclusiva e rivolta a tutte le fasce d'età (si veda in particolare la Sezione 5.2). `MyHealth` ricopre inoltre un ruolo centrale in diversi progetti internazionali. È il caso ad esempio di PARTNER (Paediatric Rare Tumours Network)[10], progetto che prevede la creazione di un registro europeo dedicato a bambini e adolescenti affetti da tumori rari, attraverso il collegamento dei registri nazionali esistenti, al fine di migliorarne l'assistenza. Il processo di data integration sfrutta `MOMIS`, che consente di effettuare un'integrazione virtuale dei dati, garantendo l'autonomia e la sicurezza delle sorgenti.

Le funzionalità di `MyHealth` sono inoltre particolarmente adatte al contesto dell'assistenza di pazienti positivi al SARS-CoV-2 e saranno pertanto sviluppate a tale scopo nella piattaforma dedicata `MyHealth-COV`[11], progetto selezionato per il finanziamento europeo nell'ambito dell'Open Call INNO4COV-19. Questa piattaforma, la cui architettura è illustrata in Figura 3, consente il monitoraggio da remoto dei pazienti, implementando anche un sistema di televisita, e offre la possibilità di scambiare dati con le cartelle cliniche ospedaliere.

5.1 Tutela della privacy e pseudonimizzazione

Un aspetto cruciale per l'utilizzo dei dati medici è la tutela della privacy. I dati personali utilizzati in ambito medico includono infatti tutte le informazioni relative allo stato di salute di un individuo (dati clinici e biometrici derivanti da esami o visite, storia clinica, terapie in corso, ecc.), che sono da considerarsi *dati sensibili* ai sensi del GDPR. Questo impone vincoli molto stringenti all'utilizzo e all'elaborazione di tali dati, legati in particolare al consenso esplicito del paziente, che possono rendere la realizzazione della data integration estremamente difficile o inficiarne l'efficacia.

Le tecniche di *anonimizzazione* rendono impossibile ricostruire l'identità dell'individuo al quale il dato anonimizzato si riferisce. Se questo permette di superare le limitazioni definite dal GDPR, non consente però di effettuare operazioni

[10] https://www.raretumors-children.eu/partner-project
[11] https://www.datariver.health/myhealth-cov

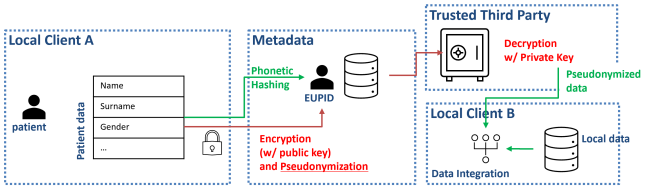

Figura 4: Soluzione per la pseudonimizzazione proposta da EUPID

di data integration su tali dati; è quindi impossibile associare ed integrare dati clinici relativi allo stesso individuo da sorgenti dati diverse (come mostrato ad esempio in Figura 2). La soluzione è quella della *pseudonimizzazione*: in questo caso, l'attribuzione del dato ad un individuo specifico è ancora possibile, ma solo con l'utilizzo di informazioni aggiuntive conservate separatamente, affidate ad esempio ad una terza parte fidata, come avviene nel caso di EUPID (European Patient Identity Management)[12], soluzione progettata appositamente per l'ambito medico e rappresentata in Figura 4.

5.2 Promozione di stili di vita attivi

Il progetto PLEINAIR (Parchi Liberi E Inclusivi in Network per Attività Intergenerazionale Ricreativa e fisica)[13], finanziato dal Bando per progetti di ricerca industriale strategica rivolti agli ambiti prioritari della Strategia di Specializzazione Intelligente (DGR n. 986/2018) e coordinato da DataRiver, si propone di realizzare contesti inclusivi per promuovere l'adozione di stili di vita attivi e la buona salute per tutti, rivolgendosi a tutte le fasce di età con strategie motivazionali personalizzate. PLEINAIR intende realizzare un parco attrezzato "smart" che implementi nuove tipologie di arredo urbano dotate di elementi di intelligenza distribuita. Gli elementi chiave del progetto PLEINAIR sono gli OSO (Outdoor Smart Objects), ovvero arredi ed attrezzi ludici dotati di sensori ed attuatori e resi interoperabili grazie ad una infrastruttura IoMT. Gli OSO saranno in grado di riconoscere l'utente e di adattare dinamicamente le loro prestazioni sia morfologiche che funzionali. Le attività svolte dagli utenti nel parco PLEINAIR verranno monitorate e sarà effettuata una valutazione delle prestazioni e dello stile di vita al fine di produrre feedback personalizzati in grado di mantenere l'utente motivato e consapevole. Sempre nel rispetto delle normative relative all'etica, alla sicurezza e alla privacy dei dati, verrà sviluppato un servizio cloud orientato alla comunità.

In questo contesto, l'attività di ricerca del DBGroup sarà incentrata su metodologie e tecniche di intelligenza artificiale e machine learning per l'analisi dei big data raccolti dalla piattaforma IoMT. I dati raccolti, riguardanti l'attività fisica svolta e lo stato di salute, verranno sfruttati per definire modelli di analisi comportamentale adatti a fornire trend e messaggi motivazionali personalizzati con lo scopo di promuovere l'attività fisica in base al profilo dell'utente.

Riferimenti bibliografici

[Bergamaschi *et al.*, 1999] Sonia Bergamaschi, Silvana Castano, e Maurizio Vincini. Semantic Integration of Semi-structured and Structured Data Sources. *SIGMOD Record*, 28(1):54–59, 1999.

[Bergamaschi *et al.*, 2018] Sonia Bergamaschi, Domenico Beneventano, Federica Mandreoli, Riccardo Martoglia, Francesco Guerra, Mirko Orsini, Laura Po, Maurizio Vincini, Giovanni Simonini, Song Zhu, Luca Gagliardelli, e Luca Magnotta. From Data Integration to Big Data Integration. In *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*, pages 43–59. Springer, 2018.

[Bettio *et al.*, 2021] Cinzia Bettio, Valentina Salsi, Mirko Orsini, Enrico Calanchi, Luca Magnotta, Luca Gagliardelli, June Kinoshita, Sonia Bergamaschi, e Rossella Tupler. The Italian National Registry for FSHD: an enhanced data integration and an analytics framework towards Smart Health Care and Precision Medicine for a rare disease. *Orphanet Journal of Rare Diseases*, 16(1):1–13, 2021.

[Gagliardelli *et al.*, 2018] Luca Gagliardelli, Song Zhu, Giovanni Simonini, e Sonia Bergamaschi. BigDedup: A Big Data Integration Toolkit for Duplicate Detection in Industrial Scenarios. In *ISTE 25th International Conference on Transdisciplinary Engineering (TE 2018)*, pages 1015–1023, 2018.

[Magnotta *et al.*, 2018] Luca Magnotta, Luca Gagliardelli, Giovanni Simonini, Mirko Orsini, e Sonia Bergamaschi. MOMIS Dashboard: A Powerful Data Analytics Tool for Industry 4.0. In *ISTE 25th International Conference on Transdisciplinary Engineering (TE 2018)*, pages 1074–1081, 2018.

[Mandrioli *et al.*, 2014] Jessica Mandrioli, Sara Biguzzi, Carlo Guidi, Elisabetta Venturini, Elisabetta Sette, et al. Epidemiology of amyotrophic lateral sclerosis in Emilia Romagna Region (Italy): A population based study. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 15:262–268, 2014.

[Papadakis *et al.*, 2020] George Papadakis, Georgios M. Mandilaras, Luca Gagliardelli, Giovanni Simonini, Emmanouil Thanos, George Giannakopoulos, Sonia Bergamaschi, Themis Palpanas, e Manolis Koubarakis. Three-dimensional Entity Resolution with JedAI. *Information Systems*, 93:101565, 2020.

[Pinzi *et al.*, 2021] Luca Pinzi, Annachiara Tinivella, Luca Gagliardelli, Domenico Beneventano, e Giulio Rastelli. LigAdvisor: a versatile and user-friendly web-platform for drug design. *Nucleic Acids Research*, 49(W1):W326–W335, 2021.

[Simonini *et al.*, 2019a] Giovanni Simonini, Luca Gagliardelli, Sonia Bergamaschi, e H. V. Jagadish. Scaling entity resolution: A loosely schema-aware approach. *Information Systems*, 83:145–165, 2019.

[Simonini *et al.*, 2019b] Giovanni Simonini, George Papadakis, Themis Palpanas, e Sonia Bergamaschi. Schema-Agnostic Progressive Entity Resolution. *IEEE Transactions on Knowledge and Data Engineering*, 31(6):1208–1221, 2019.

[12] https://eupid.eu

[13] https://www.pleinairpark.it